ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Design and Implementation of Property Valuation Model using Artificial Neural Network

---

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF THE ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN COMPUTER SCIENCE

Tedros Fekadu

June 2020
Addis Ababa, Ethiopia

Addis Ababa University

College of Natural Sciences

Tedros Fekadu Gebregziabher

Advisor: Fekade Getahun (PhD)

This is to certify that the thesis prepared by *Tedros Fekadu Gebregziabher*, titled: *Design and Implementation of Property Valuation Model using Artificial Neural Network* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	Fekade Getahun (PhD)	_____	_____
Examiner:	Mulugeta Libsie (PhD)	_____	_____
Examiner:	Dagmawi Lemma (PhD)	_____	_____

ABSTRACT

Land and buildings owned by a person or legal body are termed as real property. Property valuation is a process of determining a cash estimate of a property's value for given market conditions and the value of a property changes with market conditions, Consequently, a property's value is often updated to reflect changes in market conditions, including for example, recent real estate transactions. Property price changes over time result from specific and general effects of economic and social forces. Currently, there is no automated technique that can valuate a given property. Rather, property valuation is done manually. Naturally, this manual process of property valuation is prone to bias, inconsistencies, and corruption (bribery).

The goal of this research work is to design and implement a property valuation model capable of predicting a given residential property price at current time using artificial neural network based on the known and currently used approaches of property valuation and Ethiopian law on residential property. A novel identification technique is proposed to identify the significant attributes which have highest impact on property price. A total of 43 attribute have been used and 24 significant attributes are identified to model property valuation.

The model is designed by employing unsupervised learning approach using Artificial Neural Network based on sales comparison, replacement cost and income capitalization property valuation approaches. In order to strength the dataset and to fill the gap of missing values in the dataset geo-database is integrated in the model. The network is trained and its performance is tested with different accuracy and loss functions and the performance is extended using optimization metrics. For the purpose of training and testing the model, a total of 20113 data have been collected from FDRE documents authentication and registration agency, Abay bank, Dashen bank, Ayat real estate, Noah Real Estate Plc. The training data is randomly allocated into training (70%) and testing (30%). The model achieved an overall success rate of property price prediction using significant attributes is 97.47% and 89% using all attributes.

Keywords: Property, Artificial neural network, Property valuation, Sales comparison, Replacement cost, Income capitalization, Geo-database, valuation model, Nearest Neighbor, Attribute selection

Dedicated To

My Dad (Fekadu Gebregziabher)
and
My mom (Meselech Ejigu)

ACKNOWLEDGEMENT

First and foremost, praises and thanks to Almighty God, for his countless blessings throughout my life and this research.

I would like to express my deep and sincere gratitude to my research advisor, Dr. Fekade Getahun for his valuable time, guidance, constructive ideas, comments which enabled me to gain good research experience.

I would also like to thank individuals who are working in different organization and gave me historical property sales data for my research. This whole research would not be successful without their support.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. They have been my source of strength next to God. Last but not least, I would like to thank all my classmates.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

ANN	Artificial Neural Network
CSA	Comparable Sales Approach
DARO	Documents Authentication and Registration Office
GIS	Geographic Information System
HPM	Hedonic Price Model
ICA	Income Capitalization Approach
ML	Machine Learning
MRA	multiple regression analysis
MSE	Mean Squared Error
NN	nearest neighbors
OLS	ordinary least squares
PCC	Pearson correlation coefficient
PV	Property Valuation
RCA	Replacement Cost Approach
RE	Real Estate
ReLU	Rectified Linear Unit
RMSprop	Root Mean Square Propagation
RQ	range queries
SAF	Sigmoid Activation function
SGD	Stochastic Gradient Descent
SQL	Structured Query Language
UL	Unsupervised Learning

Chapter One: Introduction

1.1 Background

In Ethiopian Civil Code [14], land and buildings owned by a person or legal body are termed as real property. The concern of this research is on valuation of one category of this real property, which is the residential property. Property valuation is a process of determining a cash estimate of a property's value for given market conditions and the value of a property changes with market conditions, Consequently, a property's value is often updated to reflect changes in market conditions, including for example, recent real estate transactions [1]. Recent advances in the fields of valuation for different purposes have led to a growing interest in design and implementation of property valuation models for different countries [5, 6, 7]. Indeed implementing property valuation could open the door for other novel related applications such as those that include processing tax of a property or in banking where loans are taken against the security of the property and so on [2].

The value of real estate property rights is the function of the property's physical, locational, and legal characteristics [8]. The physical characteristics include the age, size, design and construction quality of the structure, as well as the size, shape, and other natural features of the land. For residential property, the locational characteristics include convenience and access to places of employment, schools, shopping, health center, and other places important to households. The locational characteristics of commercial properties may involve visibility, access to customers, suppliers, and employees, or other availability of reliable data and communications infrastructure [8].

Several valuation techniques have been proposed recently for estimating property values for different countries [3, 4, 5, 6, 7]. Ethiopia lacks effective mechanisms for capturing the value of urban property in a way that is sustainable, redistributive and developmental [13]. Because of the undeveloped real property market in the country, the valuation method followed is the replacement cost approach, which is partially attributable because the market price of houses and buildings is greatly based on the price of construction materials, instead of the value and location of the land, where the building is situated. So property valuation in the Ethiopian market is an

active area of research and has many other challenges also, for example commercial banks [9] survey property market once and use that data for more than five years which makes property valuation complex since the market value of properties are changing dynamically due to several reasons. During expropriation, in urban areas, land has no value and Expropriated people are not compensated for their losses associated with the location, which by itself affects their businesses, living conditions and standards, means of transportation, and access to facilities.

Instead of only using replacement cost approach valuation method in order to value a property, we can develop and implement a generic model which comprises other valuation approaches in order to estimate property's value using property's physical, locational, and legal characteristics.

1.2 Motivation

Developing property valuation model lets people to keep track of the value of their assets from time to time. Most people intend to sell and want to know what to expect in terms of a cash outcome. A buyer needs to know what a specific property is worth, but typically leaves the chore in the hands of a fee appraiser. Commercial Banks, Insurance Companies, Courts, and other institutional organizations and lenders have all been clients of property valuations.

Buyer and seller have a particular bias with an expectation; the buyer would normally want to pay a price at or below market value while the seller expects to receive at or above market value. Two inheritors dividing a real property asset may also have different ideas of value when one inheritor intends to keep the real estate and the other inheritor wants to be cashed out. Collectively, these expectations occupy a range of value. Situations wherein any two value biases diverge, splitting the difference is mutually equitable.

In the banking and insurance sector If the valuation comes in too low, you may have trouble borrowing amount you've applied for or have trouble to be insured for the damage happened. A bank and insurance valuation protects the bank or the insurance company from risks.

1.3 Problem Statement

Currently, Ethiopia is facing with rapid growing urbanization and modernization of infrastructures [8]. Most towns and cities in regions have been expanding twice their size within the past ten years. Thus, presently large tracts of land are being taken by way of expropriation for public use or benefit. In banking sector loans are taken against the security of the property [2, 9, 14]. Because of the above and other so many reasons like court order, property valuation is needed.

Generally, there are three primary valuation methodologies for arriving at the fair market value of real property: the comparable sales method; the income capitalization method; and, the replacement cost method. Since all are methods designed to reach fair market value, different countries use them alternatively.

1.3.1 Comparable Sales Approach

The comparable sales approach simply requires searching for similar properties that have been sold in the marketplace within a reasonable time period preceding the taking date, and then adjusting the sales price of those comparable properties to reflect differences between the comparable and the subject property.

1.3.2 Income Capitalization Approach

It gives value to the property in relation to the income it produces. It usually consists of arriving at an independent value of the underlying property involved, and adding to it the value of improvements, by converting reasonable or actual income at a reasonable rate of return (capitalization rate) into an indication of value. The capitalization of income is not used to project future profits or to compensate the owner for lost profits, but rather, to calculate the fair market value of the property at the time of the taking. The income capitalization approach is an accepted method for determining market value when there are no available comparable sales data, and the income is directly attributable to the land.

1.3.3 Replacement Cost Approach

The replacement cost method values the property by determining the replacement or reproduction cost of improvements, less depreciation, plus the market value of the property.

Hence, this predominantly serves to value buildings as well as utilities, but not the land itself. This approach can be used in countries where the market value of real property is not developed. The method develops the value in terms of current labor and materials required in assembling a similar asset of comparable utility.

There are many researches [5, 6 ,7] done on property valuation for different countries, countries where the property market is stable and where the market is easily managed using automated systems. Ethiopia has no any developed real property market [8] and the valuation method followed is the replacement cost approach.

All Property valuation professionals in Ethiopia use a survey data for valuing a property which is collected before and has not considered the current market price. Using such kind of valuation method has a limitation of estimating current value of a given property and also a time taking task in order to update the current market value. In addition, the existing valuation method followed in all sectors [8, 9, 15] doesn't take so many factors such as convenience and access to places of employment, schools, shopping, health center, and other places important to households, availability of reliable data and communications infrastructure in to consideration which has a direct impact. More specifically, the following research questions need to be addressed:

- How to use other valuation approaches along with cost replacement approach?
- How to use locational valuation function with the physical valuation function?
- How to project future profits or to compensate the owner for lost profits in case of expropriation?
- Which attributes affect the property valuation in Ethiopia?

1.4 Objective of the Research work

1.4.1 General Objective

The general objective of this research work is to design and implement a generic property valuation model for residential property in urban areas.

1.4.2 Specific Objective

In order to achieve the general objective stated above the following specific objectives should be addressed.

- Perform an extensive review on previous research works focusing on property valuation for different countries.
- Perform an extensive review on previous research works focusing on property valuation and related areas for Ethiopia.
- Collect data on current property valuation methods and techniques in use from different companies like bank, insurance company, Real Estate Company.
- Identify efficient techniques, methods, tools for modeling property valuation.
- Design algorithms to develop and test the model to compare against the current market price
- Test performance and accuracy of the designed model.
- Make conclusion on what is achieved and try to show further works to be done.

1.5 Methodology

Design and implementation of property valuation model is estimating value of a property, thus it follows the following fundamental phases [1]: attribute data collection, method selection, design the model, implement the model, estimating the current price. To accomplish each step efficiently different state of the art techniques and methods should be selected and implemented. In order to attempt this research work, the following methodologies are going to be used to select and implement state of the art methods and techniques. Market approach will be used mostly to accomplish the goal of the research because the market approach provides an indication of value by comparing the asset with identical or comparable (that is similar) assets for which price information is available.

1.5.1 Literature Review

Different literatures like books, journals, proceedings, thesis etc. will be intensively reviewed to study different valuation methods, techniques, factors that have impact on property valuation and Ethiopia's property market uniqueness. After that, appropriate tools and techniques for modeling property valuation will be selected. Moreover, a review is made in order to understand the domain knowledge: how Ethiopia's property Market operates, how property valuation rules are applied, which attributes are significant and so on.

1.5.2 Deep understanding of different property Valuation approaches

In order to design and implement property valuation that estimates current market price, different valuation approaches has to be learned. To identify unique features from each valuation approaches, valuation approaches have to be studied first [3].

1.5.3 Data Collection

In order to generate accurate estimation value for a given property, current valuation standards, property attributes, valuation parameters, current market value have to be collected from selected companies like FDRE documents authentication and registration agency, bank, insurance company and real estate company.

1.5.4 Tools and Techniques

Different tools and techniques will be used to design and implement a property valuation technique. Unsupervised learning paradigm of Artificial Neural Network is mainly used for this research using clustering method.

1.5.5 Prototype Development

The performance and accuracy of the designed model will be tested using the prototype implemented and the collected data. The result will be analyzed and evaluated from which a conclusion and further works will be recommended.

1.6 Scope and Limitations

This study focuses on designing and implementing property valuation model and estimating current market price of a residential property existing in urban area and not includes nonresidential property as well as rural areas.

1.7 Application of the study

This research work will have several applications. These are:

- It is used to guide buyers and Sellers with making purchasing decisions.
- Property valuation is also usually necessary to grant most new mortgages against the security of the property in banks, as well as to evaluate mortgage packages that may be available for purchased.
- Insurance companies are other beneficiaries of this research work because they give insurance for properties based on the value of that property.
- Customs and revenue organization is user of property valuation for income tax and other fee collection.
- Private property can be subject to expropriation for public purposes. In order to compensate the owners for lost property the concerned agency will use such property valuation work.

- The objective of property valuation is to avoid the subjectivity, bias, inconsistency and tiredness associated with human nature.
- It is used as a core system for other applications to be built on top of it. Example: automating taxation system and so on.
- It facilitates access of property values online by integrating with different online applications.

1.8 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter two deals with the literatures reviewed so far in the area of property valuation. It includes Concepts Related to valuation, Approaches to Property valuation, artificial neural network, unsupervised learning paradigm, and Property valuation Methods. Chapter 3 presents related works in the development of property valuation models using different approaches and for different scenarios. Chapter four describes the Design and Implementation of the proposed property valuation model. The Experiment and Evaluation of the system developed is discussed in chapter five. Finally, conclusions drawn from the thesis result, the contributions of this research work and recommendations on possible future works related to this research are given in chapter six.

Chapter Two: Literature Review

In this chapter, extensive reviews of general concepts in valuation, primary property valuation approaches and methodologies, characteristics of property valuation, Unsupervised learning paradigm, Artificial Neural Network, and finally review of clustering methods are presented. The extensive literatures have been reviewed to understand the problem associated to the realm of this thesis and also to identify appropriate solution.

2.1 Property Valuation Concepts

A Merriam Webster dictionary definition of the noun value is: 'amount of commodity, money, etc. considered equivalent for something else; material or monetary worth of thing; worth, desirability, utility, quantities on which these depend'. A property valuer's definition of value could be the present price for the rights to receive income and/or capital in the future [16]. Value, by itself, can be a subjective concept in that a property will have different values at any point in time according to the purpose for which it is being valued and the circumstances of the party for whom it is being valued. For example, let's say we have one wool coat and the weather is extremely cold outside; you will want that coat to wear and keep you from freezing. In a case like this, the wool coat might be worth more to you than a diamond necklace. If, on the other hand, the temperature is warm, you will not want to use the coat, so your desire for – and amount you value – the coat decreases. In effect, the value of the coat is based on your desire and need for it, and so it is the value you placed on it, not any inherent value of the coat. However, normally when value is assessed subjectively, from a specific person or organisation's viewpoint, it will be referred to as a calculation of worth [16]. Property is a legal concept encompassing all the interests, rights and benefits related to ownership. As the Constitution of the Federal democratic republic of Ethiopia, article 40(8), land is a common property of the Nations, Nationalities and Peoples of Ethiopia and shall not be subject to sale or other means of exchange. The FDRE constitution defines private property as a tangible or intangible product which has value and is produced by the labour, creativity, enterprise or capital of a person. The Constitution recognizes private property whose contents include the right to acquire, to use and to dispose of such property by sale or inheritance or other means of transfer subject to public interest and the

rights of other persons. Private property can be subject to expropriation for public purposes subject to payment in advance of compensation commensurate with the value of the property. Property valuers, are those who deal with a special discipline of economics associated with preparing and reporting valuations. Property valuation is a process of determining a dollar estimate of a property's value for given market conditions and the value of a property changes with market conditions, Consequently, a property's value is often updated to reflect changes in market conditions, including for example, recent real estate transactions [1].

Property price changes over time result from specific and general effects of economic and social forces [17]. General forces may cause changes in price levels and in the relative purchasing power of money. Specific forces may generate shifts in supply and demand and can create significant price changes. Present price is what it is worth today and it is a vital aspect of property valuation. All values are calculated at a specific date (the 'valuation date') and are only valid for a limited period after that day. How long this validity lasts will depend on the state of the market. In a strongly inflationary market where prices are rising by large percentages over short time periods, the value calculated today could have changed in as short a period as one or two months and no longer be valid. It is essential therefore to establish when a valuation was or is to be carried out, as it is a statement of value at that date only [16]. Present price should be assessed objectively, that is without bias or favour to a particular person's viewpoint. This type of price is usually expressed as present value, which is a term used frequently by valuers and forms an essential ingredient of all valuation theory and formulae. Most values calculated by property valuers are present values as at a stated valuation date.

Property valuations are usually necessary to grant most new mortgages, as well as to evaluate mortgage packages that may be available for purchased. By way of further example, property valuations are also used to guide buyers and Sellers with making purchasing decisions, and are needed for a variety of insurance purposes and also to obtain a compensation payment.

2.2 Characteristics of Property Valuation

Building a property today requires both planning permission (which covers size, shape, design and siting aspects of the property) and building regulations (which covers the technical aspects relating to the construction process) approval. Without these, a property may have to be demolished [8]. Owning the freehold or leasehold is an important factor in value.

The valuer must know how the many and varied characteristics of real property can affect value and how changes in social, economic and political factors, in the local, national and international contexts, are likely to influence it [16].

Analysing the physical and environmental factors in a property market analysis benefits the financial analysis in a number of ways. A thorough analysis of the location's physical and environmental factors provides accurate estimates of development costs. Property owners and developers can better determine the financial feasibility of any project. The analysis is also crucial in determining the listing price of a given property because it helps to understand the market conditions of the property and constantly updating owners and valuers with such conditions allows them to adjust investment strategy promptly. Paying attention to the physical and environmental factors of a property can help owners and valuers to make the best valuation possible.

2.3 Property Valuation Approaches and Methodologies

The term valuation approach refers to generally accepted analytical methodologies that are in common use and valuation of any type whether undertaken to estimate market value or a defined non-market value require that the valuer apply one or more valuation approaches [17]. The approach to the estimation of market value in one case may well be inappropriate to another and consequently, over the years, several distinct approaches which constitute separate methods of valuation have evolved.

Generally, there are three primary valuation methodologies for arriving at the fair market value of real property [8]: the comparable sales method; the income capitalization method; and, the replacement cost method.

2.3.1 Comparable Sales Approach

The comparable sales approach considers the sales of similar or substitute properties and related market data, and establishes a value estimate by processes involving comparison [17]. In general, a property being valued (a subject property) is compared with sales of similar properties that have been transacted in the open market. The comparison approach is the simplest and most reliable method of valuation [16]. Many people in everyday life employ the basic premise of the comparable method. When looking to buy or sell a property they will make comparisons between the property in question and others for sale or that have been sold recently on the market. The property to be valued is compared to others on the market now or that have recently been sold or let on the market. Adjustments are then made to allow for the advantages and disadvantages of the subject property in relation to each comparable to arrive at a figure that can be considered the current market value of the subject [16].

In order to use the comparable sales approach we have to focus on the following three main requirements. Finding similar property type to the subject, similar location to the subject and finally identifying the evidence obtained whether it is recent or not. How recent will depend on the state of the property market. In unstable market where property values are increasing rapidly, sales comparable evidence will need to be obtained from the past few weeks. In a slow or static property market, sales comparable evidence within the last twelve months could still be applicable. It is thus difficult to be too precise over the question of what constitutes ‘recent’, but as a general rule evidence taken from up to six months prior to the valuation date will normally be required and considered.

The major problem that faces valuers in the gathering of sales comparable data is how to get all the essential information. It is critical to get as much relevant data as possible on each comparable property [16]. The property market in Ethiopia is not completely transparent [8, 11]. Not all deals are fully inserted in to a system and some are kept private and confidential. Determining the exact and complete details of a given property and the transaction that has taken place can require extensive time and work. The results of this work should be carefully and accurately recorded before the valuation of the subject property is carried out and all such records kept and filed for future reference. It is always better to have too much than too little information, both in terms of the number of sales comparable and the details on each

comparable, when carrying out a property valuation. The valuer has to be sure that all the available relevant sales comparable evidence has been found and has to be confident that no relevant data has been missed.

The principal factors for sales comparison are time of sale, location and physical characteristics. The physical characteristics include size, square feet of liveable area, number and type of rooms, layout, shape, condition, construction type and quality, special amenities, design, age and whether it is located within the same location or not [16].

2.3.2 Income Capitalization Approach

This approach considers income and expense data relating to the property going to be valued and estimates the value of the property using capitalization process [17]. Capitalization relates income and a defined value type by converting an income amount into a value estimate or in other words by conversion of expected benefits like cash flows and reversion into value of property. The capitalization of income is not used to calculate future profits or to compensate a property owner for lost profits, but rather, to calculate the fair market value of the property at the time of the taking. The income capitalization approach is an accepted method for determining market value when there are no available comparable sales data, and the income is directly attributable to the property [8]. Especially Commercial properties are usually income-producing properties acquired for their ability to generate income and because of this characteristic estimating value via the income capitalization approach is a viable option [19]. It is applicable for those properties that generate income like the rental properties which includes non-owner occupied building, houses and duplex, apartment building, etc. The income from rent that an owner expects from a property is also a part of the value of that property. This approach is not suitable for purely residential properties that do not generate any income. The value of any income producing property like office building, cell tower rental and storage facility can be determined by the income capitalization approach.

Income-producing real property is usually purchased for the right to receive future income. The appraiser evaluates this income for quantity, quality, direction, and duration and then converts it by means of an appropriate capitalization rate into an expression of present worth [19]. Income capitalization approach calculates the present or prospective capital value through direct capitalization or Yield capitalization [19]. The Direct capitalization can be applied with Net

income and Capitalization rate or Gross income and Gross rent multiplier. In order to calculate the present value of the property, accurate data have to be chosen from income statements. The formula for capitalization rate is equal to Net Operating Income (NOI) divided by the current market value of the asset.

Capitalization rate (R) = Net operating income (NOI) ÷ Current market value of the asset

Net Operating Income = Effective Gross Income – Operating Expenses

The net operating income refers to gross potential income (GPI) from which collection loss is deducted and then from it operating expenses is deducted. In operating expenses, income tax, depreciation charges made by accountants and debt service is excluded. In order to estimate net operating income the appraiser needs to have access to income and expense statements for the property going to be valued and for similar properties. A capitalization rate is similar to a rate of return; that is, the percentage that the investors hope to get out of the building in income.

The correctness of market value calculations depend on the correct estimation of rental rates, capitalization rates, discount rates, and other parameters reflecting the market segment represented by a valued property. Capitalization rate on the property market reflects the ratio between income generated by a property (net or gross operating income) and transaction price of a property, which is similar in nature, localization, housing standard, technical and functional condition, neighbourhood, accessibility and communications convenience, other features affecting price.

2.3.3 Replacement Cost Approach

The replacement cost method values the property by determining the replacement or reproduction cost of improvements, less depreciation, plus the market value of the property [8]. Traditionally it was considered that the replacement cost approach should only be used when value cannot be arrived at by any of the other valuation methods and it was considered the least reliable and accurate [16]. Now it was an acceptable method to be employed in the valuation of larger industrial field, for example, oil refineries and steelworks, and also public buildings, universities, etc. [16]. Such properties are often located in particular geographical areas for special reasons or are an unusual size, design or arrangement.

Replacement cost approach is used on a property for which there is no market sales data, or for which there is insufficient direct comparable market evidence or that will produce nothing or insufficient profits for a prospective occupier. This approach is most appropriate for the appraisal of new property, not old. This is because the older the building is, the more variables there are for estimating its value.

Costs of construction can be assessed on either the replacement or renewal approach [16]. The replacement cost approach assesses the costs, including fees of reconstructing the existing building in exactly the same style and materials, as it currently exists; whereas the renewal approach is to construct a new building of the same size and to perform the same function as the present structure, but in modern materials and style. The replacement cost is the approach most often used because it uses the most modern materials and features, eliminating functional obsolescence, such as rooms of an undesirable size or high maintenance construction materials. The replacement cost is also usually lower than the reproduction cost.

There are 3 major methods of estimating the reproduction or replacement cost:

1. The square-foot method: it takes the cost per square foot of a recently developed comparable property and multiplies it by the square footage, using the external dimensions of the structures of the subject property. Sources of building cost information are contractors, analysis of recent contracts for similar buildings, professional cost estimators, building consultants and reference to cost manuals.
2. The unit-in-place method: This estimates the cost of the subject property by summing the costs of the individual components of the structures, such as materials, labor, overhead, and profit. In this method, the whole building is subdivided into parts which are measured and costed on a per-unit basis.
3. The quantity-survey method: This method is used by professional quantity surveyors and is far more detailed than the unit-in-place method. This estimates the separate costs of construction materials (wood, plaster, etc.), labor, and other factors and adds them together. This method is the most accurate and the most expensive method, and is mainly used for historical buildings.

2.4 Property Valuation Techniques

Various researchers have implemented techniques such as multiple regression analysis (MRA), hedonic model, ANN, expert system, case based and rule based reasoning, fuzzy logic, genetic algorithm for house price prediction [21]. This topic makes an attempt to understand the various property value prediction models and techniques developed.

2.4.1 Hedonic method of property valuation

Methods for valuation of assets are divided into direct ones as well as indirect ones. The hedonic price method, also called the hedonic regression method, is a kind of indirect valuation method [20]. It is based on replacement cost approach by valuating properties specific features and attributes with regard to the whole market. Hedonic price theory assumes that a property can be viewed as an aggregation of individual components or attributes of the property. The hedonic price method is used to describe impact of property features on the property value. Based on the regression method, a function determining the relation of its features to the price is created [20].

Hedonic model or hedonic regression is the relationship between property price and set of properties feature. The valuation process consists of decomposing the price of the property into combination of specific characteristics. In fact, only the specific features of the asset are valuated, and not the asset itself [20]. The hedonic price model has also been used to estimate individual external effects (e.g. environmental attribute like noise, air pollution) on house prices. While the hedonic technique is an acceptable method for accommodating attribute differences in a house price determination model, it is generally unrealistic to deal with the housing market in any geographical area as a single unit [22]. When data patterns show non- linearity hedonic price model performance can be seriously affected. The hedonic price model relies on regression technology, which is criticized by some authors for a series of econometric problems that can lead to the bias of estimation, such as function specification, spatial heterogeneity, spatial autocorrelation, housing quality change.

The Hedonic Price Model (HPM) was first introduced by Lancaster in 1964 G.C and formalised by Rosen in 1974 G.C [26]. However, Rosen did not provide any guidance as to the exact functional form of the equation. Traditional HPM regresses the house prices on a range of its constituent attributes (e.g. number of bedrooms, type of house, or distance to the city centre and so on) and is usually calibrated using the ordinary least squares (OLS) method. The resultant

regression coefficients are estimates of the contribution of each element to the price, known as the 'hedonic' prices for each attribute [26].

There are numerous studies using HPM to model house price (e.g. Kong et al. 2007; Anselin & Le Gallo 2006; Adair et al. 1996; Li & Brown 1980 amongst others) and a number of reviews have also been conducted including Freeman(2011), Chau et al (2004), Sirmans et al. (2005), and Malpezzi (2002) according to author in [26]. However, this approach does not recognize that property and neighbourhood characteristics are in fact measured at different scales. Moreover, all the unexplained variation is assumed to be between individual houses. The number of houses is not distinguished from the number of neighbourhoods (there are usually considerably fewer of the latter than the former), the standard errors of the measured neighbourhood variables will have miss-estimated precision and will be too low [26].

Even though the hedonic price model has been widely recognized, issues such as model specification procedures, one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy, independent variable interactions, a collection of random variables, non-linearity and outlier data points can seriously hinder the performance of hedonic price model in property valuations [22]. The artificial neural network model has been offered as a possible solution to many of these problems, especially when the data patterns show non-linearity.

2.4.2 Artificial Neural Network Model of property valuation

Artificial Neural networks are machine learning system that simulates structure and function like a biological neuron. Neural network is an artificial intelligence model originally designed to replicate the human brain's learning process [22] and it consists of three main layers: input data layer (example the property attributes), hidden layer(s) (commonly referred as "black box"), and output layer (estimated house price) as seen in Figure 2.1.

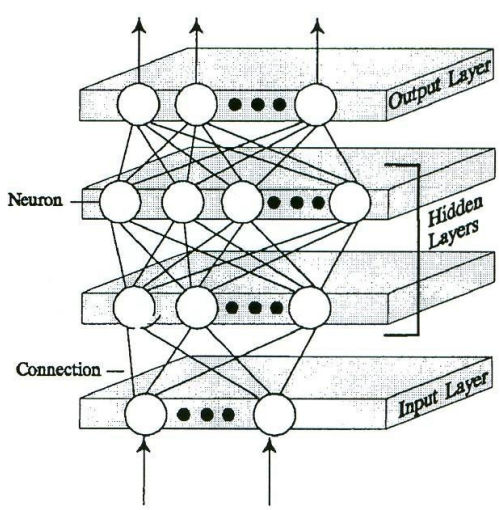

Figure 2.1: Neural Network structure with two hidden layers.

Input layer brings the initial data into the system for further processing by subsequent layers of artificial neurons, hidden layer is a layer in between input layers and output layers, where artificial neurons take in a set of weighted inputs and produce an output through an activation function and output layer is the last layer of neurons that produces given outputs for the program. ANN has self-learning capabilities that enable them to produce better results of recognizing patterns, classifying data, and forecasting future events as more data become available and also ANN is disrupting the traditional way of doing things. The use of the neural network model is similar to the process utilized in building the hedonic price model. However, the neural network must first be trained from a set of data. Using ANNs, there is no need to assume explicit functions between input and output of the studies because an ANN learns directly from observed data [6]. An artificial neural network is a nonlinear approach that provides a new alternative to linear methods, especially in the situations where the dataset possesses complex relationships between the independence of the nonlinear variables. The performance of the neural network can be influenced by the number of hidden layers and the number of nodes that are included in each hidden layer.

An ANN initially goes through a training phase where it learns to recognize patterns in data, whether visually, aurally or textually. During this supervised phase, the network compares its actual output produced with what it was meant to produce, i.e. the desired output. The difference between both outcomes is adjusted using back propagation. This means that the network works

backwards going from the output unit to the input units in order to adjust the weight of its connections between the units until the difference between the actual and desired outcome produces the lowest possible error. During the training and supervisory stage, the ANN is taught what to look for and what its output should be, using Yes/No question types with binary numbers. ANN behaviour is defined by the way its individual elements are connected and by the strength, or weights, of those connections [6]. These weights are automatically adjusted during training according to a specified learning rule until the neural network performs the desired task correctly.

To date, there is no generally accepted approach to design ANN architecture to solve all real life prediction problems and the architecture could be determined through pruning the algorithm, network information criterion, a grid search to optimise the parameters [31]. To develop an ANN model, the data available is to be divided into three parts i.e. for the training of the model, for the validation of the model and for the testing of the model. There is no generally acceptable division ratio; however, a splitting of the data into 50% (for model training), 30% (for model validation) and 20% (for model testing) is common in the literatures [4]. This justifies the fact that property valuation (either traditional or advanced methods) is an art which requires the valuers skills and experience.

Chiarazzoa, Caggiania, Marinellia and Ottomanellia applied a model based on Artificial Neural Network (ANN) to real estate appraisal [6]. Moreover, an evaluation of ANN performances in estimating the sale price of residential properties has been carried out. The study also analysed the impact of such key environmental conditions that represent a problem related to many industrial cities where pollution and landscaping consequences affect the real estate market and residential location choices. They have considered a set of asking prices of houses collected in the urban area of Taranto (Italy) where the biggest European steel factory and the 2nd industrial harbour are located. Furthermore, a sensitivity analysis has been carried out in order to identify the most significant input variables. The traditional multiple regression models and the estimated neural network models were useful in highlighting how different transport characteristics as well as the environmental quality affect the prices of real estate properties. The proposed model has a good fit in both training and test results. The results provided by the neural network can support investments in transport system. The paper concluded that ANN model can help appraisers in

making valuations and also the paper suggested that clustering methods will be applied to improve the statistical performance of the ANN in order to capture specific characteristic of groups of properties.

Artificial neural network based property valuation works has been extensively reviewed and written in the next chapter.

Common machine learning techniques for designing neural network applications include supervised and unsupervised learning, classification, regression, pattern recognition, and clustering [22].

Supervised Learning

Supervised neural networks are trained to produce desired outputs in response to sample inputs, making them particularly well suited for modelling and controlling dynamic systems, classifying noisy data, and predicting future events. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. It is called supervised learning because the process of algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers; the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. All data is labelled and the algorithms learn to predict the output from the input data.

Supervised learning can be further grouped into regression and classification. Classification is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease” and Regression is when the output variable is a real value, such as “dollars” or “weight”. Regression describes the relationship between a response (output) variable and one or more predictor (input) variables.

Unsupervised Learning

Unsupervised neural networks are trained by letting the neural network continually adjust itself to new inputs. They are used to draw inferences from data sets consisting of input data without labelled responses. You can use them to discover natural distributions, categories, and category relationships within data. Unsupervised learning is where you only have input data (X) and no

corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. These are called unsupervised learning because unlike supervised learning above there are no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning can be further grouped into clustering and association. Clustering is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour. Using clustering neural networks can be used for exploratory data analysis to find hidden patterns or groupings in data. Association is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y. All data is unlabelled and the algorithms learn to inherent structure from the input data.

Based on the above techniques clustering which is one of unsupervised learning approach is suitable for predicting the price of the property based on previous sales data.

Chapter 3: Related Work

Researchers [24, 26, 27, 28, 29, 30] studied different property valuation techniques in order to estimate property values at a given time and they have reported different results and there research findings are summarized in this section. There experimental results show that different valuation techniques results different performance.

Melichar and his colleagues developed hedonic price model to quantify the influence of structural, accessibility and environmental attributes on the price of housing in the city of Prague [23]. The methodology is applied to estimate the implicit value of housing proximity to city center, tube station and urban forest in the city of Prague. Relevant information used in the regression analysis comes from a sample of 1,708 residential dwellings [23]. They have tested four types of regression models to infer the significant determinants of housing market in Prague. Compared to linear, linear-log and log-linear, log-log model accounts for 75 % of the price variance, thus provides the best fit from all these models.

Ansah and Anthony discussed and identified advantages and problems of three broad approaches of hedonic pricing model which are parametric, nonparametric and semi-parametric [24]. The parametric models considered in this paper include the log-linear OLS, Box-Cox OLS and Weighted Least Square models. The parametric approach assumes the regression curve has a pre-specified functional form that is fully described by a finite set of parameters. It has been shown that the WLS reduces the problem which occurs because of a collection of random variables, which is a major problem in parametric models. The approach has the advantage of accurately estimating and interpreting the model. Its major demerit is that when the assumptions underlying the model are violated, the model may give a misleading picture of the regression relationship.

The nonparametric approach avoids specifying the regression function in advance and so prevents the strong assumptions underlying the parametric approach. Some of the nonparametric models discussed include the kernel method, the nearest neighbour estimate and the local polynomial regression. The approach is, however, criticised for its "curse of dimensionality" problem that makes the estimates inaccurate if the number of independent variable is large.

Again, it is very technical and complex and in computing the estimates and this makes it uncomfortable for most valuators in using the method [24].

The semi-parametric approach combines elements of the parametric and the nonparametric approaches. This approach has the advantage of reducing the problems associated with the parametric and the nonparametric approaches but at the same time maintain the advantages they possess. The hybrid approach has the advantage of making the estimate easy to interpret. It also reduces the "curse of dimensionality" problem because it reduces the number of independent variable which enter into nonparametric part. The technical computational problem and some of the underlying assumptions of the parametric approach are, however, still present though they are reduced.

Fletcher M., Gallimore P. and J. Mangan examined whether it is more appropriate to use aggregate or disaggregate data in forecasting house price using the hedonic analysis. It is found that the hedonic price amounts of some attributes are not stable between locations, property types and age. However, they have argued that this can be effectively modelled with an aggregate method. They also used hedonic price model to estimate individual external effects (e.g. environmental attribute) on house prices. The models are developed using a dataset of nearly 18,000 transactions in the UK Midlands region in 1994. The comparative performance of these models is then considered using two approaches. The error differences between actual price and the price predicted by the models suggest that the submarket models lead to statistically significant, though small, improvements. A second approach, using comparison of the root mean square errors, is conducted on the models' forecasts for a 10 per cent sample of nearly 2,000 transactions excluded from the modelling process. This shows little practical difference in the forecasting ability between the two approaches. Great care needs to be taken over sample size if a disaggregate model is used.

Mok et al. (1995) apply a hedonic price model to transaction data in Hong Kong and find that a good view of the harbour, large estate size, high floor level, and accessibility to recreational facilities are more desirable housing attributes. This is in line with the analyses done by Choy et al. (2007) and Tse and Love (2000). The former identify the positive effects of property size and floor level on property value; the latter conclude that property values are higher for estate-type

housing properties and lower for “dwelling units with a cemetery view”. Nevertheless, the hedonic price models still have some issues open for discussion such as the choice of function form, identification of hedonic price functions and data access to the dwelling-level information.

Yacim and Boshoff extended the use of artificial neural networks (ANNs) training algorithms in mass appraisal [27]. The goal was to verify the comparative performance of ANNs with linear, semi-log, and log-log models. The methods were applied to a dataset of 3,232 single-family dwellings sold in Cape Town, South Africa. The results reveal that the semi-log model and the Levenberg-Marquardt trained artificial neural networks (LMANNs) performed best in their respective categories. The best performing models were tested in terms of prediction accuracy within the 10% and 20% of the assessed values, performance, and reliability ranking and explicit explains ability ranking order. The LMANNs outperform the semi-log model in the first two tests, but fail the explain ability ranking order test.

Sayali, B. Chaphalkar Studied identification of variables affecting property value, data collection, data standardization, and application of neural network for predicting value of property for Pune City, India [28]. More than 3000 sale instances have been recorded and treated as input data. They applied neural network for value prediction and they found a robust network by testing various combinations of network parameters and dataset. Results of prediction have shown varying percentage of rate of prediction. Results obtained by change in training sample, training sample size, network characteristics like number of runs of training, and the combinations of cases given for training, testing, and cross-validation are compared. The process ensured better generalization resulting in better predictability [28].

Birkeland, D'Silva and Daniel [29] developed an automated valuation model (AVM) for the residential real estate market, and evaluated it by analysing its accuracy on all dwellings sold in the first quarter of 2018 in Oslo, Norway using ANN. They designed the AVM to utilise data on dwellings' transactions and characteristics by leveraging the concepts of stacked generalisation and comparable market analysis. In particular, they combined four novel ensemble learning methods with a repeat sales method and tailored the data selection for each value estimate. They calibrated and evaluated the model for the residential real estate market in Oslo, by producing out-of-sample value estimates for the 1 979 dwellings sold in first quarter of 2018. The research

yields highly promising results; the AVM achieves a median absolute percentage error (MdAPE) of 5.4%, and more than 96% of the dwellings are estimated within 20% of their actual selling price. This is comparable to the accuracy of local estate agents in Oslo, and they observed similar performances by the industry leader for AVMs in the U.S. These results demonstrate that their AVM is a viable tool for both industrial and private applications. Furthermore, they found their novel approach of applying stacked generalisation to residential real estate valuation to be successful and using machine learning in the field of real estate price prediction is a promising.

Rahman, Abd, et al [30] examined the capability of ANN in forecasting house prices in Johor Bahru. They acquired a total of 2,325 double storey sale transactions spanning from year 2000 to 2016 in Mukim Pulai, Johor Bahru from the Valuation and Property Services Department Johor Bahru (VPSDJB). They only extracted from the dataset house attributes theorised to affect the property prices and cleansed the dataset in order to analysis and remove outliers. Samples were discarded based on sales transaction over 233,800.00 Malaysian dollars, land area over 146.03 square metres, main floor area over 137.64 square metres, transaction years between 2013 and 2016 and incomplete information. The cleansing process reduced the sample to a total of 640 observations for training and prediction. Transaction price, measured in Malaysian dollar per unit, was used as the dependent variable. Meanwhile, land area and main floor area measured in square metres were used as independent variables.

A feed-forward structure with one hidden layer was applied in there study. Then they trained the neural network using a back propagation algorithm to adjust the weight and thresholds of the network to minimise forecasting errors in the training set. Datasets were split into three sets: training, testing and validation dataset. Out of 215 datasets, 193 datasets were used for training (years 2000 to 2010), 22 datasets were used for prediction (years 2011 to 2013) and validation separately.

In their paper, the learning and momentum rates were determined through five phases of trial-and-error. A series of trial and error process was performed by identifying the number of hidden neurons randomly, starting with the smallest (one) to the largest number (five). Training and testing were executed by increasing hidden neurons after each training and testing process. The network minimised the difference between the given output and the prediction output monitored

by the minimum average error while the training process was conducted. A decrease in value will minimise the error. Their process continued until 30,000 cycles of test sets were achieved. The result of this process suggests that the best neural network to forecast Johor Bahru house prices is 2-1-1 (2 indicates the number of neurons in input layer, 1 number of neurons in hidden layer and 1 number of neuron in output layer) with 0.0.1 learning rate and 0.1 momentum rate.

Their validation test was implemented after the network output under prediction sets are transferred into validation sets. The validation test examined the performance of ANN in forecasting local house prices. The validation process was performed by comparing the actual and forecasted values for years 2009 to 2011. The performance of the ANN is evaluated through statistical tests, namely Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). Low MAD, MAPE and RMSE values produced by the ANN indicate good predictive performance.

Finally they assessed the performance of ANN by observing the values of R^2, MAD, RMSE and MAPE for two sets of selected housing schemes. All statistical tests indicated good fit. Both sets in these two housing schemes produced high R^2 with low values for MAD, RMSE and MAPE. Superior goodness of fit was observed for Sets 1 and having a higher value of R^2 at 0.99 and 0.93 respectively. Meanwhile, the MAPE showed a percentage error of 4.41% and 4.55% for Sets 1 and 2 respectively, both with less than the 10% error threshold. This implies that the ANN is able to predict house prices with low errors. However, the second datasets showed slightly higher MAPE with 14.32% for Set 1 and 16.31% for Set 2. The results suggest that models with large sample sizes (Sets 1 of first and second dataset) have superior performance compared to models with small sample sizes (Sets 2 of first and second dataset).

Chapter Four: Design and Implementation

4.1 Overview

In this chapter the design and development of the proposed property valuation model which uses previous sales data and current material cost is presented in detail. As described in different research works, property valuation using previous sales data or using current material cost or using the property income passes through different phases: historical valuation data acquisition, current material cost acquisition, data preprocessing, model generation, choosing property indicators for the type of valuation methods used, applying valuation methods and estimation. The sections in this chapter present details of components of the proposed architecture. The different algorithms which are used in each step are also presented in more detail. Section 4.2 presents the general overview of the proposed system architecture. Section 4.3 presents Historical valuation data acquisition, Current material cost data acquisition, Geospatial data acquisition. Section 4.4 explains Data pr-processing. Section 4.5 explains Attribute selection (property indicators) for each valuation methods. Section 4.6 briefly explains and discusses ANN model generation and the tools and techniques used to create the model. Finally, a summary for the chapter is briefly presented in the last section.

4.2 The Proposed Property Valuation Model Architecture

The system architecture shown in Figure 4.1 is the proposed architecture for estimating a property value. As depicted in the figure the proposed system architecture is composed of five main modules: Historical property valuation data acquisition, Current material cost data acquisition and Geospatial data acquisition, Data Preprocessing, Attribute selection (property indicators) for each valuation methods, valuation method integration, ANN Model Generation and Estimation.

Attribute selection (property indicators) component identifies the variables used for property valuation depending on valuation method used. As discussed in previous chapters' property

valuation is done for different purposes: compensation, mortgages, insurances and so on. Based on the valuation purpose the valuation methods applied and variables used or the critical variables may vary.

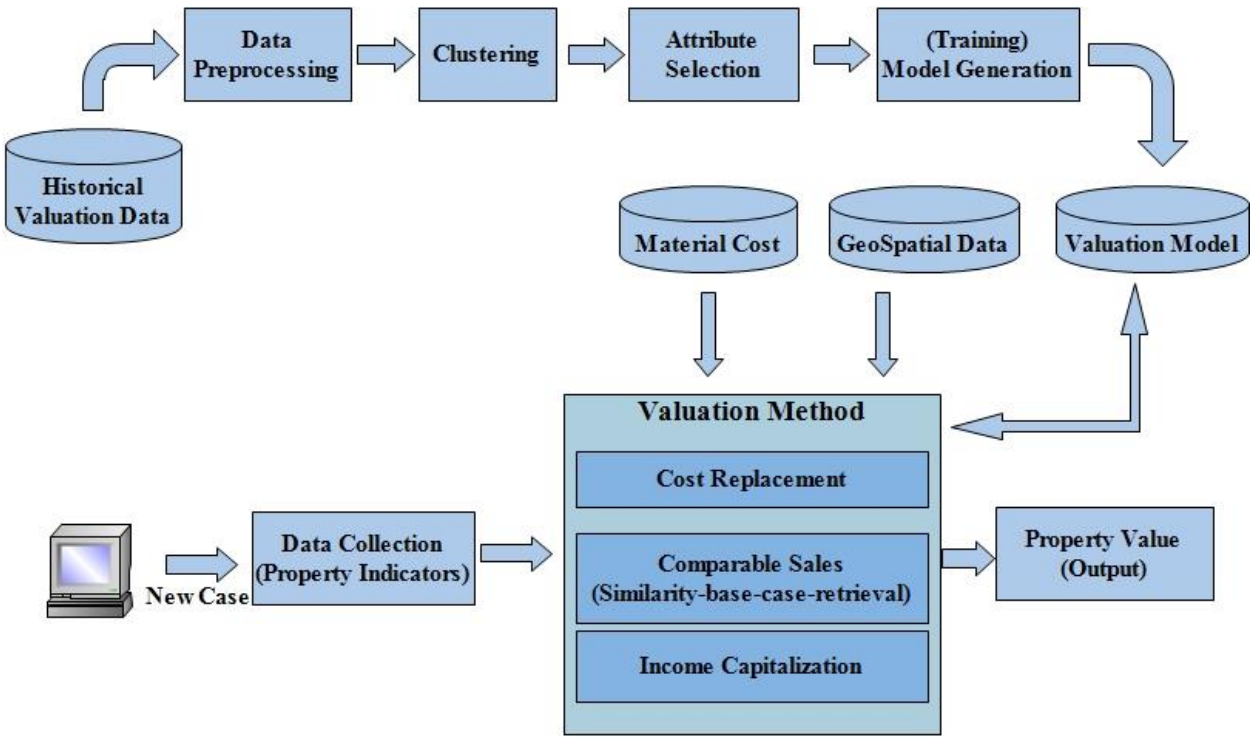

Figure 4.1: Architecture of the Proposed ANN based Property Valuation Model

Historical property valuation data acquisition, Current material cost data acquisition and Geospatial data acquisition module acquires previous property sales data, current material cost and geospatial data from different sources in different format for processing. The dataset is then Pre-processed using different preprocessing tools and techniques and also splits the dataset for training, validation and testing. The third module selects property attributes (property indicators) depending on the property valuation method used. The fourth part of the proposed architecture which is model implementation is the main part of the architecture because the accuracy and performance of the developed model relay on it. This part tells what kind of neural network

architecture is good for valuation and also the number of hidden layers we have to use, type of activation, optimizer, metrics and other subcomponents. The model is developed using keras because it uses the idea of adding layers to the network instead of defining the entire network at once. The final module tests the proposed architecture according to the developed model using the dataset and shows the accuracy and performance of the model. The details of each component and how they are working are briefly explained in the following sections.

4.3 Data Acquisition

This is the first step in property value estimation. For this system, previous property sales data which contains many none spatial and spatial attribute values, current material cost are collected from different sources.

4.3.1 Historical Valuation Data

Various studies have examined factors affecting property values and have identified that: age, location, size, neighbourhood characteristics, economic activity, population, transport etc. [34] are the factors affecting property value and they grouped the variables into; environmental variables, neighbourhood variables, accessibility (location) variables and property variables [34] and further classified factors affecting property values as show in Table 4.1.

Table 4.1: Factors affecting property values

Attribute type	Attribute
Structural attributes	type of house, number of floor, number of rooms, numbers of bedrooms, bathrooms, kitchen, Salon, fireplaces, garages, square footage of house, area per room, lot size, age of the house, condition of the house, parking area size, basement area size, existence of pool, construction phase, veranda width, made of (material type used), is there elevator, roof

	terrace, roofing and ceiling is made of
Locational attributes	Distance from main road, proximity and accessibility to various (dis)amenities including waste sites, power lines, highways, shopping centres, churches, schools, cultural opportunities, airport, public transportation, hospital, and restaurant. Distance to the city center.
Neighbourhood attributes	Socio-economic characteristics of neighbouring residents (population density, single family/small or large multifamily, median income, etc.), quality of neighbouring structures, ownership/rental, ethnic composition, violent crime rate, park accessibility.
Environmental attributes	view from property, noise levels, pollution levels, storm water
Community attributes	school and tax districts
Time-related attributes	month and year of sale, number of days on market

Based on this previous property sales data were collected from Ethiopian document authentication and registration office, Real estate Companies mainly from Ayat real estate, Noah Real Estate Plc. and finally from Abay Bank S.C. Property sales data obtained from the Ethiopian document authentication and registration office (DARO) contains all type of properties data and are located in different locations since every property sales agreement must be legalized by this government office according to Proclamation No. 334/2003 (Authentication and

Registration of Document Proclamation). The data collected from DARO also contains properties which are sold 10 years ago.

Real states companies are also our source of data. As it is known real state companies build a new house (villa, apartment, G+) and sale them, acquiring such data will add additional dimensions to our dataset since all the properties are sold after construction is completed. The other advantage of collecting pervious sales data from Real Estate Company is that, we can get detail information about the material used to construct different part of the property and also to know each component the property has. Such data is very helpful in case of valuing a property using cost approach method. Current costs of material used to construct different part of the property are calculated using current material cost data.

As discussed in the first chapter Banks valuate property by their own means of valuation for the purpose of giving a loan against the security of the property. We collected property data from Abay Bank S.C which is appraised by the bank's valuators. Such data lets us to identify the main attributes used by the valuators for valuating a property for the purpose of giving a loan against the security of the property. Identifying the main variables used in case of giving a loan against the security of the property will simplify the development of property valuation model for the purpose of giving mortgage against the security of the property

Collecting data from different sources helped us to consider different dimensions while developing property valuation model and also helped us to have a bigger dataset. As discussed in previous chapters different organizations use property valuation for different purposes, therefore in order to develop a generic property valuation model our dataset contains all necessary data.

4.3.2 Material cost

Current material cost data help valuators in determining the cost required to reconstruct (replace) the property at current time. Every raw material needed to construct a property using past time property construction style to the latest has to be collected with their current market price.

Based on this current material cost has to be collected from different construction material shops to get the average price of each material. The collected costs of materials are the one which are basic to construct a house or building.

4.3.3 Geospatial Data

Geospatial or geographic data describes information tied to some locations on earth's surface. It is data about objects, events, or phenomena that have a location on the surface of the earth. The location may be static in the short-term (e.g., the location of a road, an earthquake event, children living in poverty), or dynamic (e.g., a moving vehicle or pedestrian, the spread of an infectious disease). Geospatial data combines location information (usually coordinates on the earth), attribute information (the characteristics of the object, event, or phenomena concerned), and often also temporal information (the time or life span at which the location and attributes exist) [40].

As described in the previous sub section locational attributes are one of the factors affecting the property valuation, a satisfactory estimation can be done by analysing a certain amount of property characteristics in an objective way, with the assistance of Geographical data. The spatial nature of real estate data allows the development of specialized models that increase the likelihood for better predictions. The geospatial database is used to map with properties location data in order to easily analyse and show property values in different location. Based on this geographical database is used specifically targeting Addis Ababa city. The geographical data is obtained in two formats, raster and vector geographical data formats. Vector data structures represent specific features on the Earth's surface, and assign attributes to those features. Vectors are composed of discrete geometric locations (x, y values) known as vertices that define the shape of the spatial object. The organization of the vertices determines the type of vector that we are working with: point, line or polygon [41] whereas raster data model is a representation of the world as a surface divided into a regular grid of cells. Raster models are useful for storing data that varies continuously, as in an aerial photograph, a satellite image, a surface of chemical concentrations, or an elevation surface [41].

4.4 Data Preprocessing

4.4.1 Historical valuation data

After acquiring previous sales data, the data should be pre-processed in order to put the data in a format that our algorithm wants. The preprocessing step is vital in formatting our data set in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set and best out of them is chosen. Preprocessing helps in converting the raw data into a clean data set like managing anomalies ,null values from the original raw data set, rescaling the dataset [32] since our data is comprised of attributes with varying scales in magnitudes, units and range.

To start the processing phase we should create databases, one for storing historical valuation data, second for storing current material costs and finally for storing geospatial data. All of the historical valuation data collected from different sources (which are mentioned in the previous section) were in Excel file format. Microsoft SQL server is used to create the database and the Excel data file was imported to the created database. After importing our historical valuation data to our Microsoft SQL server database, we preprocessed it to handle outliers, fill missing values and make dimensionally reduction and clustering.

Outliers are extreme values that deviate from other observations on data; they may indicate variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample [41, 42]. Outliers can be of two kinds: univariate and multivariate. Univariate outliers can be found when looking at a distribution of values in a single feature space. Multivariate outliers can be found in a n-dimensional space (of n-features). Most common causes of outliers on a data set are: Data entry errors (human errors), Measurement errors (instrument errors), Experimental errors (data extraction or experiment planning/executing errors), Intentional (dummy outliers made to test detection methods), Data processing errors (data manipulation or data set unintended mutations), Sampling errors (extracting or mixing data from wrong or various sources), Natural (not an error, novelties in data) [41].

In machine learning and in any quantitative discipline the quality of data is as important as the quality of a prediction or classification model. ML algorithms are very sensitive to the range and

distribution of attribute values. Data outliers can spoil and mislead the training process resulting in longer training times, less accurate models and ultimately poorer results. Outlier detection was done on all of our attributes and for successful outlier detection we applied transformation to the data i.e.: scaling it. Our algorithms can also benefit from rescaling the attributes to have the same scale. Rescaling our data is achieved by Scikit-Learn as shown in Appendix A, which is Python library using the MinMaxScaler class.

MinMaxScaler transforms features by scaling each feature individually to a given range, e.g. between zero and one by preserving the shape of the original distribution and also doesn't meaningfully change the information embedded in the original data [32]. The other advantage of MinMaxScaler is that, it doesn't reduce the importance of outliers. MinMaxScaler follows the following formula for each feature:

$$\frac{x_{\mathrm{a}} - \min(x)}{\max(x) - \min(x)}$$

Where x_a is the original cell value, min(x) is the minimum value of the column and max(x) is the maximum value of the column. Using this formula, we will see that the values of each column will now be shrank and between zero and one. The MinMaxScaler works well for cases when the distribution is not Gaussian or when the standard deviation is very small. However, it is sensitive to outliers [32].

MinMaxScaler Algorithm:
Input: Array: previous property sales data containing detail information about the property and the sale transaction **Variables:** List of attributes with different data types **Output:**

```
BEGIN
def dataset_minmax(dataset):
/ / assign the dataset to the  function which finds the min and max values for each
    column
 minmax = list()
 / / declaring a new empty list variable
 for i in range(len(dataset[column])):
 / / iterate throw dataset column
      col_values = [row[i] for row in dataset]
      / / get all values of the column
      value_min = min(col_values)
      value_max = max(col_values)
      / / get the minimum and maximum value of the column
      minmax.append([value_min, value_max])
      / / assign the values to the list variable created at line 7
      def normalize_dataset(dataset[column], minmax)
      / / call the normalize function to rescale dataset columns to the range 0-1 using
      the above min max values the column in the dataset
     for row in dataset[column]
         for i in range(len(row))
              row[i] = (row[i] - minmax[i][0]) / (minmax[i][1] - minmax[i][0])
                       / / according to minmax scalar math function calculate the value
            of each cell in the column
        Next
    Next
 Next
Return O
END
```

Algorithm 4.1: MinMaxScaler Algorithm

After making the appropriate transformations to all feature space of our historical valuation data using Algorithm 4.1, outliers has to be detected by using outlier detection techniques and methods. Outliers are observations or measures that are suspicious because they are much smaller or much larger than the vast majority of the observations [44]. Anomaly detection techniques can be highly influenced by the way we define anomalies, type of input data and expected output and these differences lead to wide variations in problem formulations, which need to be addressed through different analytical techniques [44].

The problems of anomaly detection in high-dimensional data are threefold, which involves detection of: global anomalies, local anomalies and micro clusters or clusters of anomalies [44]. Global anomalies are very different from the dense area with respect to their attributes, local anomaly is only an anomaly when it is distinct from, and compared with, its local neighbourhood and Micro clusters or clusters of anomalies may cause masking problems.

We used 'Search and Trace Anomaly' algorithm [41], which address the limitation of previous anomaly detection algorithms. The algorithm focuses specifically on fast, accurate anomalous score calculation using simple but effective techniques for improved performance and also it can be applied to both one dimensional and high-dimensional data and it can detect outliers as well as inliers. The algorithm is based on the distance definition of an anomaly; nearest neighbour distances between data instances in the high-dimensional data space are the key information for the algorithm to detect anomalies. To make the variables of equivalent weight, the columns of the data are first normalised such that the data are bounded by the unit hypercube. This normalisation is commonly referred as min-max normalisation, which involves a linear transformation of the original data, with the result data ranging from 0 to 1. We already transformed our data accordingly as described previously.

Search and Trace Anomaly Algorithm:
Input: Array: min-max normalised previous property historical sales data containing detail information about the property and the sale transaction **Output:**

```
BEGIN
def knn(data, query, k, distance_fn, choice_fn):
/ / assign the dataset to the  function which finds the k nearest neighbour values for
    each column value
 neighbor_distances_and_indices = [] / / declaring a new empty array variable
 for index, row in enumerate(data):
 / / iterate throw dataset column
      distance = distance_fn(example[:-1], query)
      / / Calculate the distance between the query and the current
      neighbor_distances_and_indices.append((distance, index))
      / / Add the distance and the index of the data to an ordered collection
      sorted_neighbor_distances_and_indices=
      sorted(neighbor_distances_and_indices)
      / / Sort the ordered collection of distances and indices in ascending order
      k_nearest_distances_and_indices = sorted_neighbor_distances_and_indices[:k]
      / / Pick the first K entries from the sorted collection
      k_nearest_labels= [data[i][1] for distance, i in k_nearest_distances_and_indices]
      // iterate the sorted distances and indices and compare with the first distance to
      remove the bigger gaps.
 Next
Return O
END
```

Algorithm 4.2: Anomaly Detection Algorithm

Running the Algorithm 4.2 outliers in each attribute which are values that are far away from normal value are detected and removed from the dataset. In addition to the outliers missing values are also treated in every attributes. Attributes which have 20 % of missing values were removed from the dataset.

After the dataset is rescaled, outliers are detected, missing values are treated the next task handled in our preprocessing step was clustering the dataset into groups of similar data points. Having a good separation of data points into clusters is fundamental for our artificial neural network based prediction. The purpose of clustering is to identify groups of objects, or clusters, which are more similar to each other than to other clusters. It automatically finds patterns in the data without the need for labels, unlike supervised machine learning. Such an approach to data analysis is closely related to the task of creating a model of the data, which is, defining a simplified set of properties that can provide intuitive explanation about relevant aspects of a dataset and clustering methods are generally more demanding, but provide more insights about complex data [45].

Clustering algorithms involve several parameters, often operate in high dimensional spaces, and have to cope with noisy, incomplete and sampled data and several different approaches to clustering have been proposed in the literature[45]. Each clustering algorithm relies on a set of parameters that needs to be adjusted in order to achieve viable performance, which corresponds to an important point to be addressed. Several clustering techniques have been introduced for prediction such as K-Means, Hierarchical, Probabilistic and Density – based Clustering.

In the all kinds of clustering algorithm, K-means clustering method is the most interesting, the use is also more common. K-means algorithm is an iterative algorithm that tries to partition the dataset into K-pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. K-means clustering method is the most commonly used to calculate the similarity between two samples object, is the Euclidean distance, defined as follows:

$$d(x_{\mathrm{l}}, \mathrm{x_m}) = \sqrt{(x_{\mathrm{l1}} - x_{\mathrm{m1}})^2 + (x_{\mathrm{l2}} - x_{\mathrm{m2}})^2 \ \ \ldots + (x_{\mathrm{lp}} - x_{\mathrm{mp}})^2}$$

Where X_{lp} and X_{mp} show properties of two p data objects and it is implemented using the Algorithm 4.3.

Clustering Algorithm:

Input:

X: Dataset

K: number of clusters

M: number of iterations

Output:

BEGIN

C← randomly choose cluster centroids from X

P← K-means (X, C, K):

$F_{ratioprev}$← calculate f-ratio of P

For j = 1 to m

W←solve fisher discriminant based on class label P

X_w←project all data X into discriminant based on class label p:

P_w←optimally solve clustering problems on X_w by dynamic programming

C, P ←k-means (X, $P_{w, K}$):

F_{ratio}←calculate f-ratio of P

If $F_{ratio} < F_{ratioprev\ then}$

P_{cpt} ← P

$F_{ratioprev}$ ← F_{ratio}

End if

End for

END

Algorithm 4.3: Clustering Algorithm

By implementing all the steps for historical valuation database we cleaned the data since it have many irrelevant and missing parts by filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data because Noisy data is a meaningless data that can't be

interpreted by machines. Those noisy data can be generated due to faulty data collection, data entry errors etc. We also transformed the data in to fixed range in order to make the data suitable for the learning phase by normalizing the data and data with the same values grouped together. After finalizing the above phases the next and last phase of data preprocessing for historical valuation dataset is feature selection.

Attribute Selection

In real-estate valuation it is not possible to talk about exact methods and exact attributes. Therefore, determining the number of most appropriate variables affects the operating speed of the model and increases the model's performance [33]. Thence, before deciding the real-estate valuation model, the attributes which affects valuation, must be detected [33]. Property value is an essential aspect of property markets worldwide and determined by a variety of factors and the determination of those factors is a significant part of property valuation [34].

The objective of variable selection is three-fold [35]: improving the estimation performance of the estimating algorithms, providing faster and more cost-effective estimators, and providing a better understanding of the underlying process that generated the data. There are many potential benefits of variable and feature selection: eliminates irrelevant and redundant data, facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance. The benefits of feature selection for learning can include a reduction in the amount of data needed to achieve learning, improved predictive accuracy, learned knowledge that is more compact and easily understood, and reduced execution time and recent research has shown common machine learning algorithms to be adversely affected by irrelevant and redundant training information [36].

In this thesis feature selection for unsupervised machine learning tasks was accomplished on the basis of correlation based approach. Correlation based feature selection procedure can be beneficial to common machine learning algorithms. Experiments on artificial datasets showed that correlation based feature selection quickly identifies and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features [36].

Variables within a dataset can be related for lots of reasons, for example: One variable could cause or depend on the values of another variable, one variable could be lightly associated with another variable and two variables could depend on a third unknown variable. The performance of some algorithms can deteriorate if two or more variables are tightly related. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features. The correlation between input variables with the output variable is identified in order to provide insight into which variables may or may not be relevant as input for developing our model. The features in our dataset were explored as follows:

Correlation between features

Pearson correlation is applied to our dataset in order to know the correlation between attributes by measuring the strength of a linear association between attributes and is denoted by r. Basically, a Pearson correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit). The PCC, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable and vice versa. The Pearson correlation coefficient, r is calculated by the following formula:

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Where N is the number of attribute values, $\sum xy$ is the sum of the paired attributes, $\sum x$ is the sum of one of the paired attribute, $\sum y$ is the sum of the other attribute, $\sum x^2$ is the sum of squared one of the paired attribute and $\sum x^2$ is the sum of squared of the other attribute. Based on the Pearson correlation formula the correlation between features and identifying the dependent, unnecessary variables was done using Algorithm 4.4. After implementing the algorithm the correlation between features can easily be plotted using the matplotlib library as shown in Figure 4.2.

Correlation coefficient Algorithm:

Input:

Column1: one of the paired attribute

Column2: the other paired attribute

Output:

BEGIN

```
Length1 =  Column1.count
Length2 =  Column2.count
/ / length of the paired attributes
 If Length1 == Length2 == 0
 Return 0
 / / no values to compare
 sum1 = sum2 = sum1_sq = sum2_sq = psum = 0
 for row in Column1
      sum1 = sum1 + Column1[row]
      sum1_sq = sum1_sq + (Column1[row] * Column1[row])
Next
for row in Column2
      sum2 = sum2 + Column2[row]
      sum2_sq = sum2_sq + (Column2[row] * Column2[row])
 Next
 Psum = sum1 + sum2
 num = psum - (sum1 * sum2 / len)
 den = ((sum1_sq - (sum1 ** 2)/len) * (sum2_sq - (sum2 ** 2) / len)) ** 0.5
 den == 0 ? 0 : num / den
```

END

Algorithm 4.4: Correlation coefficient Algorithm

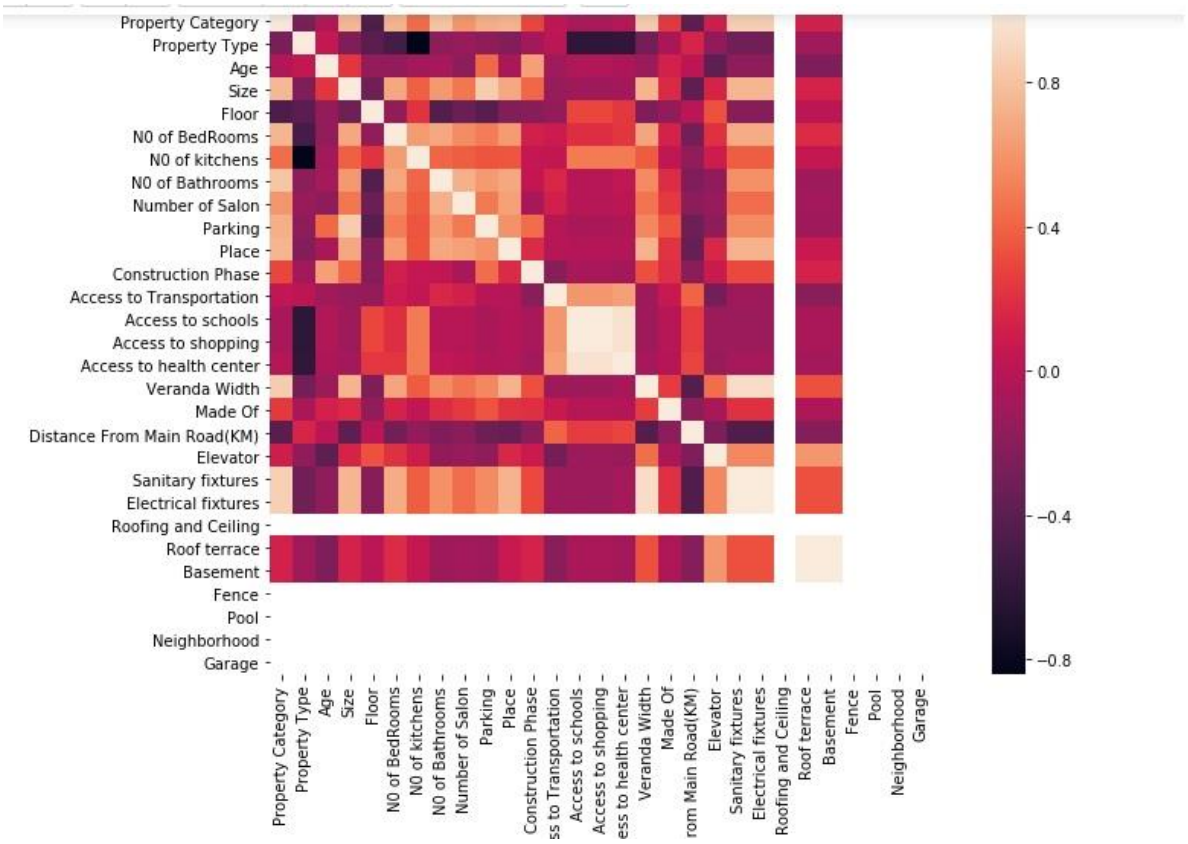

Figure 4.2: Correlation between features

As seen in Figure 4.2 there exist strong correlations between some of the features. For example, Fence, Pool, Neighbourhood, Garage, roofing and ceiling are strongly correlated. They actually express more or less the same thing. There will be no negative impact on the accuracy of our new model if we exclude such attributes, rather it will increase the performance of the model.

Correlation between Sale Price and other features

As correlation was done between attributes in the above section, the correlation between our target variable which is sale price and other features is done the same way. Correlating features against our target variable helps us to know the features which have a positive impact on our estimation model. Features which are more correlated with the sale price attribute are the most important features which determine the price of the value. According to the Algorithm 4.4 the correlation between sale price and other features are shown in Figure 4.3.

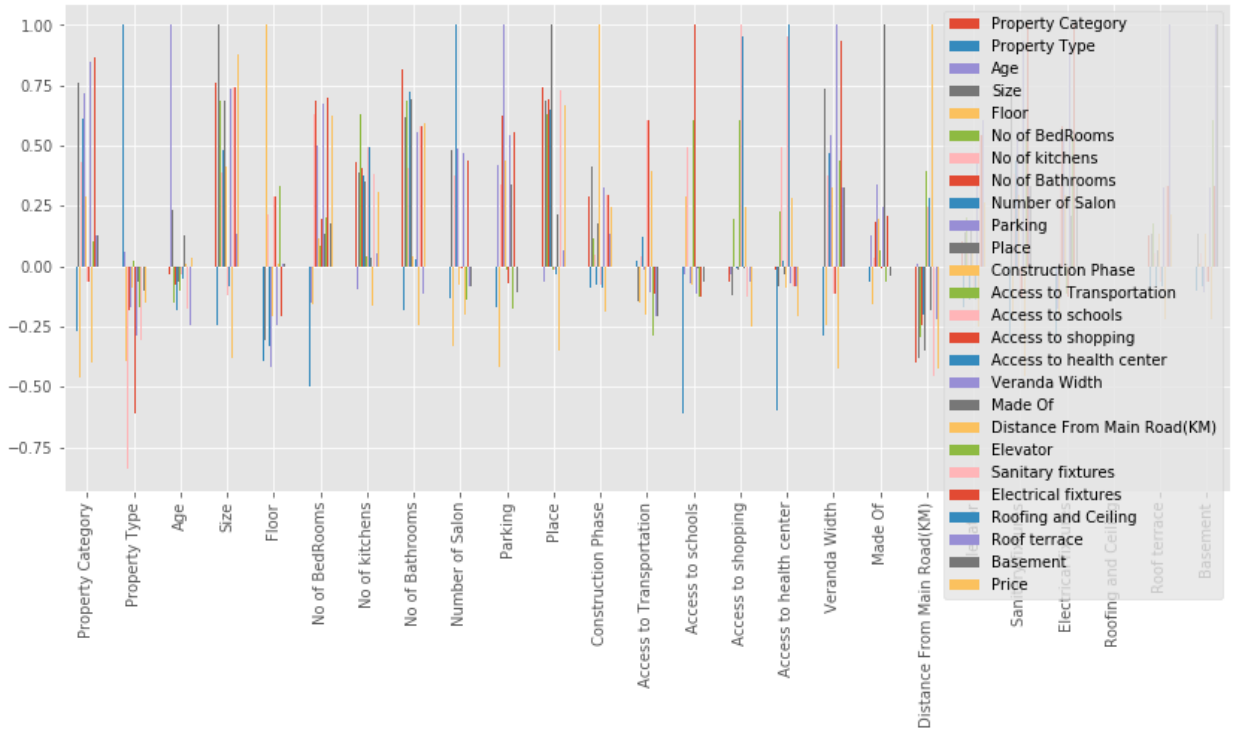

Figure 4.3: Correlation between Sale price and other features

As seen in the above picture the correlation of sale price with property type, size, number of bedrooms, number of kitchens, number of salon, number of bathrooms, place are the greatest (beyond 0.5). These columns have the highest role in determining the sale price of a property.

Based on the above two correlations (correlation between features and correlation of sale price against other features) features which are dependent on other features (redundant features), features which have less impact on the estimation of price, features which have the highest impact of determining property estimate have been identified. In general we determined the degree to which a relationship exists between features. The correlations done are useful because if we found out what relationship variables have, we can make predictions.

After we selected the features which are most important for our property value estimation and we also determined the features which are dependent on other variables and features which have less impact on the estimation process, the next phase was splitting our data for training, validation and testing purpose and finally developing a model. The remaining phases are described in detail in the next sections.

4.4.2 Geospatial Data

The GIS database is similar to the conventional database. However, the ability of the GIS database to store, manipulate, retrieve and display spatial data (i.e. maps) in addition to the normal functions of the conventional database is the main characteristics that set it apart from other databases (Healey, 1991). We use geographic databases because it's a way to put all data in a single container. Within this container, we can build networks, create mosaics and do versioning. The management of Geospatial data and metadata is somewhat different.

The most important and expensive component of geographical information system is data which is generally known as fuel for GIS. Geospatial data is combination of graphic and tabular data. Geospatial data can come from many sources. Geospatial data has been digitized by a wide variety of agencies and commercial enterprises at an increasing pace over the past ten years. A wide variety of data sources exist for both spatial and attribute data. For our research purpose we used a digitized map of addis ababa city which is prepared by one of Ethiopian government agency which is information network security agency (INSA) as shown in Figure 4.4.

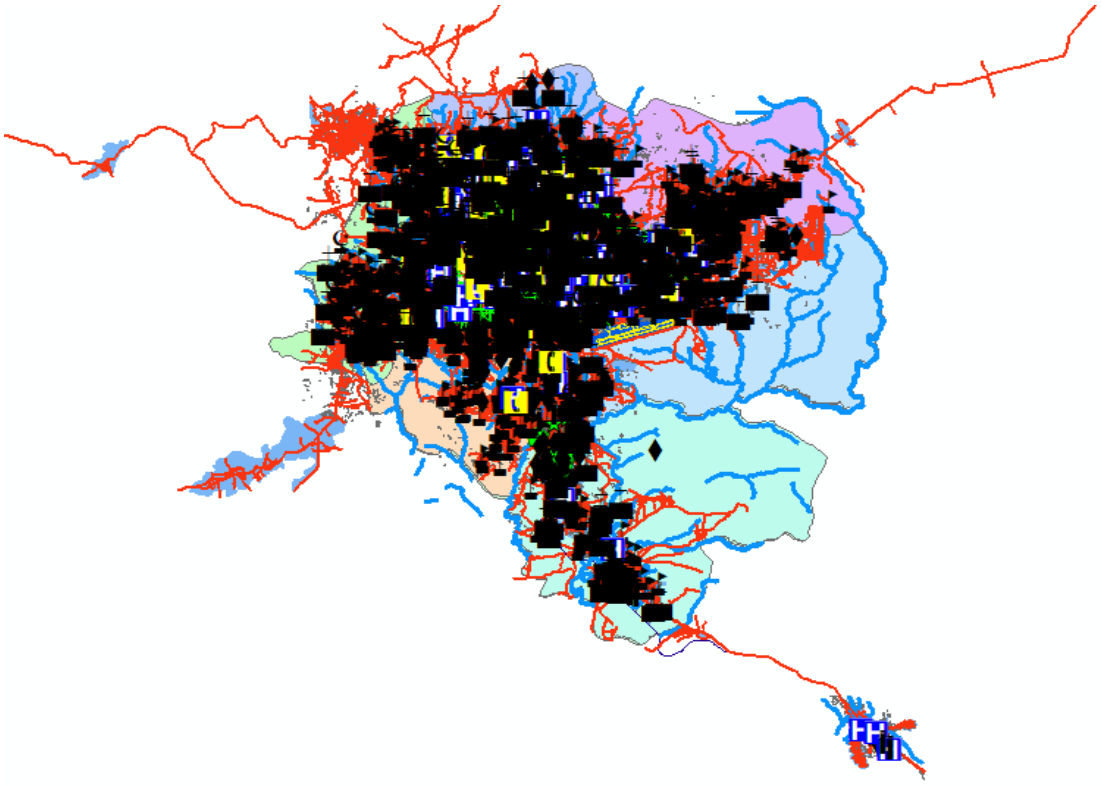

Figure 4.4: Addis Ababa city digitized map

The digitized map of Addis Ababa city as seen in the above image contains detail information of the city by categorizing the spatial data in to different layers. The layers included in the above digitized map include religious and tourist sites as one layer, telecommunication infrastructures and their offices as another layer, cinema and theatre halls as one layer, land mark buildings as another layer, health and medical places, governmental institutions, financial banking institutions, embassies, cultural and educational institutions, companies and enterprises, transport operators, airport runway, road, river, railway layers.

Using a geospatial data containing such a huge collection of layers will help in illustrating the geography of property values over space and time in a way that most property professionals find intuitive, namely in map form and also used for calculation of additional real estate parameters, essential for valuation process. In particular, the buffering functions, belonging to the principle of spatial analyses, which can determine the property distance to other objects, and functions from the domain of geographical networks analyses, making possible calculation of shortest paths, deserve the attention.

It is necessary to determine what kind of objects and in what radius should be looked for, both increasing and decreasing real estate value as shown in Figure 4.5 like access to the road network, information about the course of the mass transport (buses, trams, trains) lines.

Figure 4.5: Parcels with their neighbourhood

The information about parcel neighbourhood, existing or planned, can be obtained from the database. If the parcel is located inside any zone, it is evident that neighbouring parcels have the same destination. In the opposite case topological geometrical data models, available in the database, allow the qualification of neighbourhood and mutual crossing of geographical objects. Distance from the town centre, quality of access road reaching the parcel and accessibility of mass transport facilities, and other essential parameters used for valuation as we seen in the above section can be calculated using the geospatial database.

The use of network analysis functions allow the calculation of route to nearest higher category road on a given area (of course basing on road network data) and then the determination of worst surface type. Regarding the accessibility of mass transport, hospitals, schools, shops and other facilities, several nearly situated stops can be selected, and then the nearest can be found, considering the movement along roads or paths. Generally our latest geospatial data for the city of Addis Ababa fills the missing locational data of each property in the valuation database. Since we already defined the critical factors affecting the valuation, we can know the locational attribute values of the property using different spatial analysis algorithms.

In order to use network analysis, get the accessibility of various facilities the shape file obtained has to be converted in to database file. SQL Server supports two spatial data types: the **geometry** data type and the **geography** data type. The geometry type represents data in a Euclidean (flat) coordinate system and the geography type represents data in a round-earth coordinate system. Using Shape2SQL, which is a popular, simple application specifically designed to load shapefile data into SQL server, the spatial data have been imported into our SQL server database.

After converting every shapefile into database file, every shapefile has now become a table having attribute having the geometry (shape) of the single point or line or polygon and other attributes defining the shape. For example road is represented using one table with fields containing id, road type, description, shape length and the shape. We can know easily retrieve spatial data we want, calculate distance to some point, get the area size, and so on using spatial query language. For example to simply select the main roads from the database we can write simple select statement.

SELECT geom from dbo.road where road_type = 'main road'; and get the result as shown in Figure 4.6.

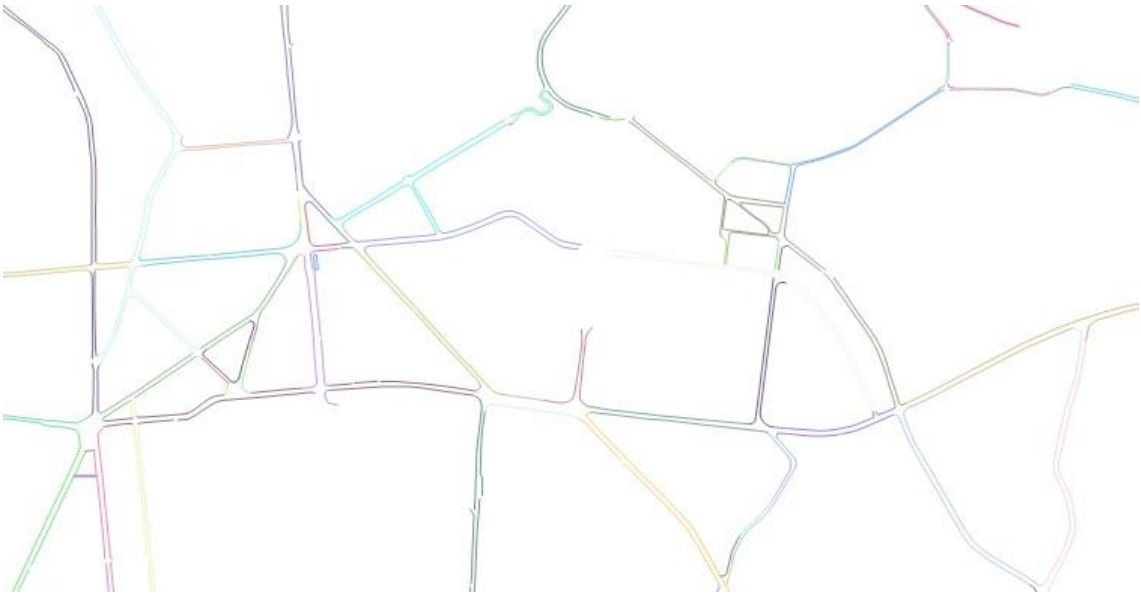

Figure 4.6: Main road sample

As seen in the previous sections, it is a well-known fact that the value of a property is highly affected by the location. So as to regulate the data of neighbourhood and spatial features and to generate location indexes benefited from Geographic information system. Spatial indices are a family of algorithms that arrange geometric data for efficient search. For example, doing queries like "return all buildings in this area", "find 1000 closest gas stations to this point", and returning results within milliseconds even when searching millions of objects. Therefore introducing of spatial data into statistical property assessment models has been considered to be highly efficient.

Spatial data has two fundamental query types: nearest neighbors and range queries. Both serve as a building block for many geometric and GIS problems. For example, retrieving the closest point to a given query point from thousands of points, such as city locations was achieved by calculating the distances from the query point to every other point, sort those points by distance and return the first item. For our valuation purpose we can find the nearest distance to different services like hospitals, schools, public transport by calculating a network accessibility index and also get neighbourhood quality and topology and further helps to show information in the form of map to improve the decision making capabilities.

The historical data of property value was supported by the spatial database for accurate property value estimation.

Nearest neighbors

Nearest neighbor (NN) is a very important problem in the field of information science and data science. NN query, to find the closest point among a set of points to a given query point, is widely required in several applications such as density estimation, pattern classification, information retrieval and spatial analysis [46]. Theoretically, it is not difficult to find the nearest point to a test point q among a point set P. The most straightforward way for the purpose is to compute the distance between q and each point in P and then find the point with minimum distance. Since the query time of the above full search method is proportional to the size of the point set, it is often called linear time algorithm. In practice, however, this approach is not commonly used because of its intolerable inefficiency when faced with large-scale problems. In consequence, how to solve it with a sub-linear complexity is the key to this problem [46].

The majority of the existing spatial search methods are tree-structured. Their inventors hope to realize spatial query in logarithmic computational complexity by taking advantage of the multi-level feature of tree structure. Later, a composite index structure, VoR-tree, was proposed. It integrates a Voronoi diagram into R-tree to improve the kNN query of R-tree [46]. The multi-layer Voronoi Diagrams efficiency is significantly higher than the tree structure indexing such as kd-tree, R-tree and VoR-tree, in terms of the real physical spatial nearest neighbor search.

The Voronoi diagram of a given set P = {p1, . . . , pn} of n points in R^d partitions the space of R^d into n regions. Each region includes all points in R^d with a common closest point in the given set P [47]. A k-NN Voronoi region Vk(H,S) is a polytope in Rd . The common face between two neighboring k-NN Voronoi regions, $V_k(H_1,S)$ and $V_k(H_2,S)$, is a portion of the bisector $B(H_1,H_2)$. In R^2, the boundary between two neighboring k-NN Voronoi regions is a k-NN Voronoi edge, and the intersection point among more than two neighboring k-NN Voronoi regions is a k-NN Voronoi vertex [46]. Therefore, from the first nearest neighbor, by adding its Voronoi neighbor, we update the candidate set. Following that, we can find the second nearest neighbor and include the second nearest neighbor to update the candidate set. Then we find the third nearest neighbor

from the candidate set. We repeat this process until we obtain the k-th nearest neighbour as shown in Algorithm 4.5. From this whole process, we can realize the incremental kNN algorithm based on Voronoi. If the process to update the candidate set and search the nearest neighbor from the candidate set can be seen as an atomic operation, the time complexity of the this algorithm is O(k).

Multi-layer Voronoi Diagrams Algorithm:

Input:

M_p: Multi-layer Voronoi Diagrams

q, k : query points

Output:

BEGIN

nn_q := MVD-NN(M_P , q);

K := [nn_q];

Visited := {nn_q};

V_P := M_P [0];

for i := 1 to k − 1 do

for all n ∈ V_P [K[i]] do

if n ∉ Visited then

Visited := Visited ∪ {n};

for j := i + 1 to Size(K) do

if k q − n k<k q − K[j] k then

Insert(K, j, n);

if Size(K) > k then

Pop(K);

end if

break;

end if

```
            end for
            if Size(K) < k and n ∉ K then
                Append(K, n);
            end if
        end if
    end for
end for
return K
```

Algorithm 4.5: Multi-layer Voronoi Diagrams Algorithm

Range Queries

Spatial range queries are queries that inquire about certain spatial objects related to other spatial objects within a certain distance. There are classical variants of this problem such as computing the number of objects intersecting a query range, checking whether an object intersects a query range, and finding the closest pair of objects intersecting a query range [48]. There are a large number of objects intersecting a query range in many real-world applications, and thus it takes long time to report all of them. Thus one might want to obtain a property of the set of such objects instead of obtaining all such objects with minimal execution time.

To process spatial range queries a standard depth-first search algorithm was applied on the R-tree to retrieve the objects that quantify the spatial predicate of the query as shown in Algorithm 4.6. Then for every object retrieved x, $P_x = E_x$ where p is the probability range and E is entry point. If the query Q is the thresholding query, the threshold t is used to filter out objects with $P_x < t$. if Q is a ranking query, a priority queue maintains the m results with the highest Px, during search, and outputs them at the end of query processing.

The space/time complexity depends on the size of L. as the minimum distance of all entries in L (length) from the query point is at most the distance of the last object seen, the maximum size of L can be estimated by using the value of K. in practice the average size of L is quite small (10m-100m).

Multi-layer Voronoi Diagrams Algorithm:

Input:

P^{first} := 1 probability of number of obhect

P^{m} := 0; P^{min} of m^{th} object

Output:

BEGIN

While P^{first} > P^{m} and more objects in R do

x := next NN of q in R (use BF [10]); /* during BF-search, each non-leaf entry with Pfirst ¢ emaxE < t is removed from Best-First heap and inserted into L*/

compute P^{min}

x and P^{max}

x by using P^{first}, L and E^{x};

while P^{min}

$x < P^{m} * P^{max}$

x do

pick an entry e in L;

remove the entry e from L, read node ne pointed by e;

for each entry e^{0} up to n^{e}

if mindist(q; e^{0}) > d(q; x) then

enheap e^{0} on Best-First heap;

else if n^{e} is a non-leaf node then

insert e0 into L;

else /*e0 is an object*/

P^{first} := P^{first} ¢ (1 ¡ Ee0);

for each entry o up to H such that d(q; e0) * d(q; o)

end for

P^{first}

o := Pfirst

o ¢ (1 ¡ Ee0);

compute P^{min}

x and P^{max}

x by using P^{first}, L and E^{x};

if P^{min}

x > P^{m} then enheap(H,(x,P^{first}x := P^{first},P^{min}x ,P^{max}x));

if H is changed then

recompute, for each o up to H, P^{min}o and P^{max}o by using P^{first}o , L and Eo;

P^{m} := m-th P^{min} in H;

remove entries o from H with P^{max}o < P^{m};

P^{first} := P^{first} ¢ (1 ¡ E^{x});

while jHj > m and jLj > 0 do

repeat Lines 9–19;

repeat Lines 22–25;

return P^{m}

Algorithm 4.6: Spatial range query Algorithm

4.5 Data Splitting

Data splitting is the act of partitioning available data into two or three portions, usually for cross-validator purposes. One portion of the data which is the training set is used to optimize model parameters (train the model), the second portion which is the test set is used for cross validation during training to avoid over-fitting and the other one which is the validation set is used to assess the performance of the trained model. Data splitting is an important consideration during artificial neural network (ANN) development, even for a moderate sample size, the sampling methodology used for data splitting can have a significant effect on the quality of the subsets used for training, testing and validating an ANN [38]. Poor data splitting can result in inaccurate and highly variable model performance. Increased confidence in the sampling is of paramount importance, since the hold-out sampling is generally performed only once during ANN development.

In machine learning, one of the main requirements is to build computational models with a high ability to generalize well the extracted knowledge. When training e.g. artificial neural networks, poor generalization is often characterized by over-training and a common method to avoid over-training is the hold-out cross validation and the basic problem of this method represents, however, appropriate data splitting [37]. If the model over-trains, it just memorizes the training examples and it is not able to give correct outputs also for patterns that were not in the training dataset. Improper split of the dataset can lead especially to an excessively high Variance of the model performance. However, various sophisticated sampling methods can be used to deal with this problem.

The problem of appropriate data splitting can be handled as a statistical sampling problem and various classical statistical sampling techniques can be employed to split the data [37]. Generally, data are divided using a variety of methods, including ad-hoc methods, random methods, stratified methods and optimization based methods, etc. Convenience sampling is an efficient deterministic method widely used when dealing with time series dataset [37].

Based on this we used the code from scikit-learn called ‘train_test_split’, which as the name suggests, split our dataset into a training set and a test set. We used 70 % of the overall dataset for training and the rest 30 % of the dataset for testing and validation. Our testing and validation

dataset have equal amount of data by splitting the 30% of data in to two. In summery we have a total of six (6) variables for our model development: X_train (70% of full dataset), X_val (15% of full dataset), X_test (15% of full dataset), Y_train (1 label, 70% of full dataset), Y_val (1 label, 15% of full dataset), Y_test (1 label, 15% of full dataset).

4.6 ANN Model Implementation

In this section Artificial Neural Network models are created. Anaconda and Jupyter Notebook was installed in order to create a development environment and installed different packages like tensorflow, keras, pandas, scikit-learn and matplotlib for different purposes. Panda is an open source library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. To implement the models Keras is used as shown in Appendix C, because keras uses the idea of adding layers to the network instead of defining the entire network at once. This opens us to quick alter the number of layers and type of layers, which is handy when optimizing the network. ANN model creation consists of two steps. The first step is to specify a template (architecture) and the second step is to find the best numbers from the data to fill in that template. The details of each step and ANN model creation processes are explained briefly as follows.

4.6.1 Setting up the architecture

In ANN model creation the first thing that has to be done is to set up the ANN model architecture. The architecture implemented has input layer, hidden layer and output layer. As described in the previous chapter 4.4. (Feature selection) features which have significant impact on predicting property price are taken as input layer and the output layer is the final estimated value.

The input layer is responsible for receiving the inputs. These inputs can be loaded from an external source such as a web service or a csv file. There must always be one input layer in a neural network. The input layer takes in the inputs, performs the calculations via its neurons and

then the output is transmitted onto the subsequent layers. The number of neurons in an input layer is dependent on the size of our training data.

The output layer takes in the inputs which are passed in from the layers before it, performs the calculations via its neurons and then the output is computed. The output layer is responsible for producing the final result. There must always be one output layer in a neural network. Our prediction model has a single node.

Hidden layers reside in-between input and output layers and this is the primary reason why they are referred to as hidden. The word "hidden" implies that they are not visible to the external systems and are "private" to the neural network. There could be zero or more hidden layers in a neural network. Usually, each hidden layer contains the same number of neurons. The larger the number of hidden layers in a neural network, the longer it will take for the neural network to produce the output and the more complex problems the neural network can solve. The neurons simply calculate the weighted sum of inputs and weights, add the bias and execute an activation function. The main work done here is to find the total number of hidden layers used and also to determine the size (number of nodes) of each hidden layer in order to get best prediction result.

In general, you cannot analytically calculate the number of layers or the number of nodes to use per hidden layer in an artificial neural network to address a specific real-world predictive modelling problem [39]. The number of layers and the number of nodes in each hidden layer are model hyper parameters that we must specify. We are likely to be the first person to attempt to address our specific problem with a neural network [39] but experiments have shown us that the optimum number of neurons in a hidden layer can be determined by:

$$Number\ of\ Neurons = \frac{\text{Number of training data samples}}{\text{Factor} * (\text{Input Neurons} + \text{Output Neurons})}$$

The factor is used to prevent over-fitting and it is a number between 1 and 10.

Based on the above premises initial neural network architecture has been built. The architecture contains one input layer for accepting property features and one output layer for displaying estimated value. As shown in figure 4.4 below, for beginning the architecture contains two hidden layers and the number of neurons of the hidden layers has been computed based on the

above formula. Then the architecture has been described to our keras framework for further development.

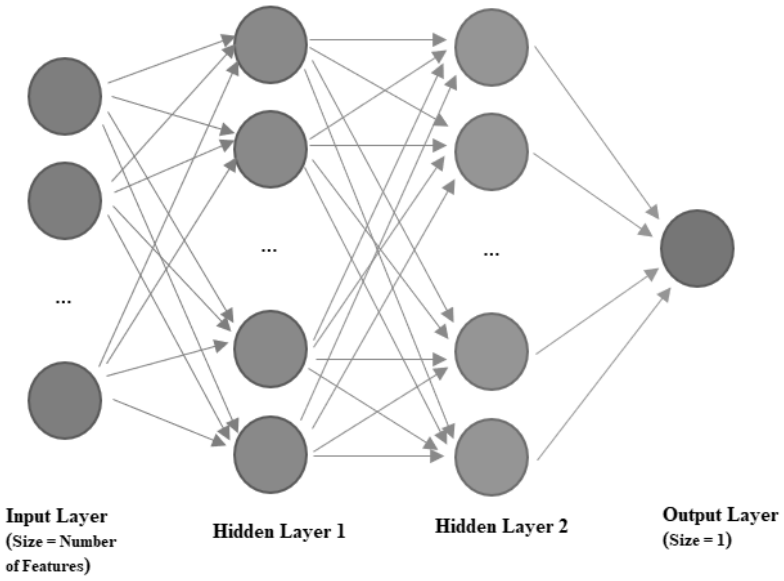

Figure 4.7: Neural Network Architecture

Sequential model is used since it allowed us to build deep neural networks by stacking layers one on top of another and is more customizable. Sequential models have garnered a lot of attention because most of the data in the current world is in the form of sequences – it can be a number sequence, date sequence, image pixel sequence, a video frame sequence or an audio sequence.

Algorithm 4.7 is used to implement sequential model builds high-level features through their successive layers and each layer is associated with a set of candidate mappings. When an input is processed, at each layer, one mapping among these candidates is selected according to a sequential decision process, so that a path from the input to output layer defines a sequence of transformations. Instead of considering global transformations, like in classical multilayer networks, this model allows us for learning a set of local transformations. It is thus able to process data with different characteristics through specific sequences of such local transformations, increasing the expression power of this model with respect to a classical multi-layered network.

The algorithm stores our model in the variable 'model'. It defines the input layer neurons based on the input features. The hidden layers neurons are calculated based on the formula described above. Finally the algorithm defines an output layer.

Algorithm for creating a model:

Input:

input_columns, // total number of columns (features) imported

input_rows // total number of rows (data) in each column imported for training

hidden_layers // the total number of hidden layers to use

factor // number from 1-10 to prevent over-fitting

activation // activation function for input and hidden layers

output_activation // activation function for output layer

Output:

BEGIN

model

neurons = (input_columns * input_rows) / factor * (input_columns + 1)

model.add(Dense(neurons_hidden_layers, activation=activation,

input_shape=(input_columns,)))

for layer in hidden_layers

model.add(Dense(neurons, activation=activation))

Next

model.add(Dense(1, activation= output_activation))

END

Algorithm 4.7: Algorithm for creating a model

As seen in algorithm 4.7 activation functions are used in order to get the output of node by mapping resulting values in between 0 to 1 or -1 to 1 etc. (depending upon the function).. They are also known as Transfer Functions. For output layer Sigmoid Activation function is used because it has a form of f(x) = 1 / 1 + exp(-x) and its Range is between 0 and 1. For input layer and the hiddens layer RELU activation function is used because it has a simple form of R(x) = max(0,x) i.e if x < 0 , R(x) = 0 and if x >= 0 , R(x) = x.

4.6.2 Filling in the best numbers

Now that we've got our architecture specified, we need to find the best numbers for it. Before we start our training, we have to configure the model by adding another algorithm for optimization which also defines what loss function the model uses and the other metrics we want to track apart from the loss function.

The optimization algorithm is used for optimizing our model during deployment and execution time. The optimization starts with defining some kind of loss function/cost function and ends with minimizing it using one or the other optimization routine. The choice of optimization algorithm can make a difference between getting a good accuracy in minutes, hours or days.

Root Mean Square Propagation Algorithm (RMSprop) is used to boost the speed and accuracy of our model. RMSprop is using the same concept of the exponentially weighted average of the gradients like gradient descent with momentum but the difference is the update of parameters. Exponentially weighted average is used to find the moving average. It involves holding the past values in a memory buffer and constantly updating the buffer whenever a new observation is read. RMSprop learns properly and tunes the internal parameters so as to minimize the Cost function.

A loss function is a method of evaluating how well our algorithm models our dataset. It is a measure of how accurately our model is able to predict the expected outcome i.e. the ground truth. If our predictions are totally off, our loss function will output a higher number. If they're pretty good, it'll output a lower number. Loss functions provide more than just a static representation of how our model is performing rather they show whether our algorithms fit our data in the first place. Selection of the proper loss function is critical for training an accurate model. Certain loss functions will have certain properties and help our model learn in a specific way. Some may put more weight on outliers, others on the majority.

Mean Squared Error (MSE) is used as a loss function in compiling our model. To calculate MSE, we take the difference between our predictions and the ground truth, square it, and average it out across the whole dataset. The MSE will never be negative, since we are always squaring the errors. The MSE ensured that our trained model has no outlier predictions with huge errors, since

the MSE puts larger weight on these errors due to the squaring part of the function. The MSE is formally defined by the following equation:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Where N is the number of samples we are testing against.

The other function tracked apart from loss function is metric function. Accuracy is tracked in order to show the performance of our model. A metric function is similar to a loss function, except that the results from evaluating a metric are not used when training the model.

Training the data

Once our network has been structured and filled up with appropriate numbers, it is ready to be trained. Algorithm 4.8 is used to train the dataset iteratively by giving both input value and output value according to data split in section 4.5 above to the network model specified one at a time. The weights associated with the input values are adjusted each time. After all cases are presented, the process often starts over again. During this learning phase, the network learns by adjusting the weights so as to be able to predict the correct output.

The network processes the records in the training data one at a time, using the weights and functions in the hidden layers, and then compares the resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights for application to the next record to be processed. This process occurs over and over as the weights are continually tweaked. During the training of a network the same set of data is processed many times as the connection weights are continually refined. The algorithm will specify for how long the dataset has to train by getting the loss and accuracy values. If the reduction in loss value and the increase in accuracy value becomes' constant after some iteration the algorithm will stop the training.

Algorithm for training a data:

Input:

input_array, // input columns data which are split for training purpose

output_array // ouput column data which are split for training purpose

training_iterations // default number of iteration

Output:

BEGIN

accuracy // accuracy value in each training iteration

loss // loss value in each training iteration

for iteration in training_iterations

output = self.learn(input_array) // Pass the training set through the network.

error = output_array – output // Calculate the loss

factor = dot(input_array.T, loss * self.__ Root Mean Square Propagation _derivative(output))

self.synaptic_weights += factor

// Adjust the weights by a factor

If (loss == error)

Count = count + 1

If (count == 3)

break

Next

return self.__sigmoid(dot(input_array, self.synaptic_weights))

END

Algorithm 4.8: Algorithm for training a data

Evaluating a Model

Once we are happy with our final model, we evaluated on the test dataset. The accuracy and performance of the network on a separate dataset, unseen during training was evaluated. This will provide an estimate of the performance of the network at making predictions for unseen data in the future. The purpose of the our ANN model validation is to ensure its ability to generalize within limits set by the test data in a robust fashion, rather than simply memorized the input-output relationships that are contained in the training data. This evaluation would be the equivalent to conducting an infinite number of network simulations and to calculating the average error performed when learning every possible output value.

Keras evaluate function is used to know the accuracy and performance of our model. The function returns the loss value & metrics values for the model in test mode. The function takes input testing data and output testing data as array.

4.7 Summary

The chapter systematically went through the designing and implementation of ANN based property valuation model. The designed model has components working together. In the first component, previous property sales data is acquired and pre-processed to manage null values from the original raw data set, rescaling the dataset since our data is comprised of attributes with varying scales in magnitudes, units and range. Our algorithms can benefit from rescaling the attributes to have the same scale. Rescaling our data is achieved by min-max scalar algorithm. In the second component, feature selection is done in order to determine the number of most appropriate variables which affect the operating speed of the model and which increase the model's performance. In the third component, data split is implemented to split the dataset in to training, testing and validation set for cross-validator purposes.

Finally, the model is created by calculating the number of hidden layers and number of neurons to be used. The algorithm created to implement sequential model builds high-level features through their successive layers and each layer is associated with a set of candidate mappings. Root mean squared propagation algorithm is used for optimizing our model during deployment

and execution time and Mean Squared Error (MSE) is used as a loss function in compiling our model. To calculate MSE, we take the difference between our predictions and the ground truth, square it, and average it out across the whole dataset. Once our network has been structured and filled up with appropriate numbers, it is trained using the training dataset and evaluated using the validation dataset.

Chapter Five: Experiment and Evaluation

5.1 Overview

As mentioned in the previous chapter, artificial neural network based property valuation model has been developed with the intention of using previous sales data, so that its performance can be tested. Thus, in this chapter, the set of experiments conducted to validate the proposed model are presented.

We have followed some set of procedures to conduct the experiment. The test environment, the set of activities defined under the procedures, and the findings from the experiment are described in detail in the following sub sections. Section 5.2 presents the datasets used in testing the system. Section 5.3 describes the tools used and the overview of the system. Section 5.4 presents the test results found. Finally, the proposed model is evaluated in the last section of the chapter.

5.2 Datasets

In order to train and test the proposed model for property valuation using artificial neural network, we used previous property sales data with the following properties. These documents are collected from Ethiopian document authentication and registration office, Real estate Companies mainly from Ayat real estate, Noah Real Estate Plc. and finally from Abay Bank S.C and Dashen Bank S.C.

The collected data are then pre-processed and divided in to three parts for training, validation and testing purpose according to Table 5.1.

Table 5.1: Experimental Data

Dataset property	Value
Total number of attributes	43
Input attributes	42
Output attribute	1
Total number of rows (data)	20,113
Training data	70 % of the total data = 14,079
Validation data	15 % of the total data = 3,017
Test Data	15 % of the total data = 3,017

The attributes are shown in Appendix D.

5.3 Implementation

The prototype is implemented using keras package using Python as programming language. Python is the most common programming language used by Deep Learning systems. Anaconda software is used as a platform in order to use Python and Keras packages. Anaconda is a free open source Development environment for data science, machine learning applications, etc. The prototype takes previous sales data as a csv file format as an input using the graphical user interface shown in Figure 5.1 and stores the data in to our SQL Server database.

Keras is an open source, high level machine learning library which is written in Python, capable of running on top of TensorFlow and wraps around the functionalities of other Machine Learning libraries. Keras supports convolutional networks and recurrent networks, as well as combinations of the two.

Moreover, the specification of the computer on which the system is implemented is Intel Core i5 laptop computer with 6GB RAM and 1.8 GHz processor.

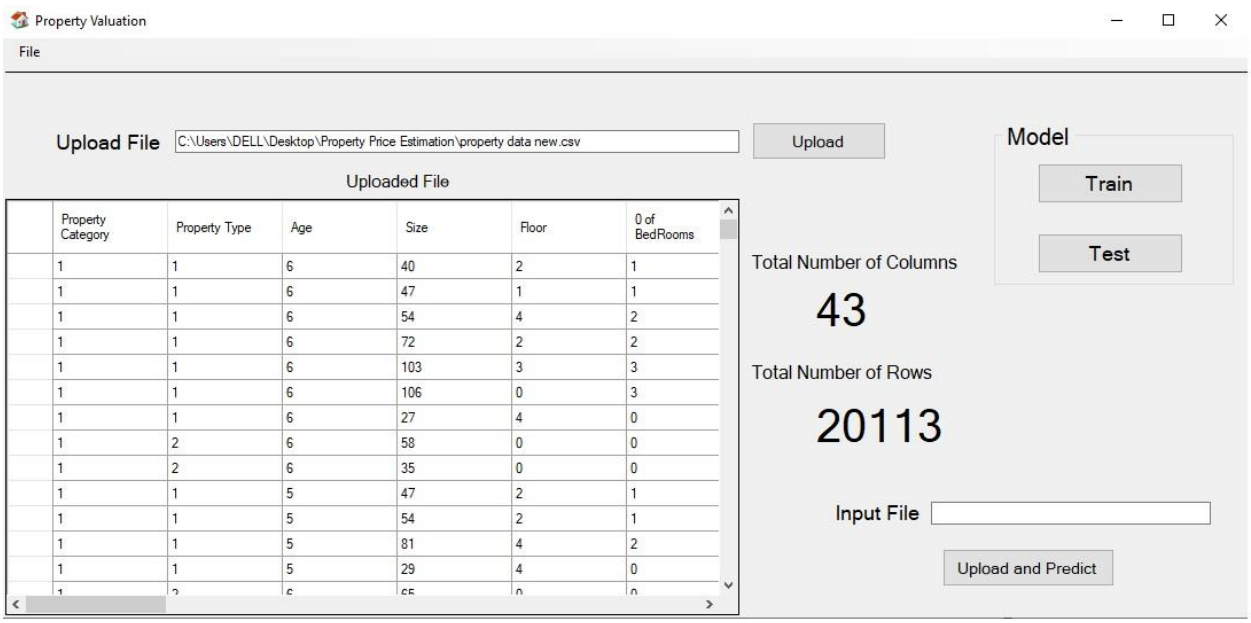

Figure 5.1: Running Prototype

After the previous sales data is inserted in to our SQL Server database it is then imported in to the system using panda's package which is an open source library providing high-performance, easy-to-use data structures and data analysis tools for the python programming language.

Another end-to-end open source platform for machine learning used is TensorFlow. It has a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets us push the state-of-the-art in machine learning and used to build and deploy our machine learning powered model easily. Tensor flow considers networks as directed graph of nodes with data flow computation and dependencies encapsulated in it. TensorFlow APIs are arranged hierarchically, with the high-level APIs built on the low-level APIs. We use the low-level APIs to create and explore new machine learning algorithms. In this class, we used a high-level API named keras to define and train machine learning models and to make predictions.

The performance and result of the developed model is graphically plotted using paython 2D visualization library called Matplotlib. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

5.4 Test Result

As we discussed in Section 4.6, apart from Regularized Linear Model we used Xgboost. XGBoost is an open-source optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. In addition, the model is tasted with only features that have the highest impact on determining the property value according to Section 4.4. In the experiment, we tried to test the accuracy, loss and performance of the developed model using Sigmoid and Softmax activation function, Rmsprop and SGD optimizer, Mean squared error, categorical_crossentropy and binary_crossentropy loss functions.

Sigmoid activation function

Activation functions are just functions that we use to get the output of node and also known as transfer function. The activation functions can be basically divided into 2 types, linear activation function and Non-linear activation functions. Linear activation functions are lines or they are linear and the output of the functions will not be confirmed between any ranges and doesn't help with the complexity or various parameters of usual data that is fed to the neural networks whereas Nonlinear Activation Functions are the most used activation functions and makes it easy for the model to generalize or adapt with variety of data and to differentiate between the output.

Sigmoid activation function is a nonlinear activation function which is easy to work with and has all the nice properties of activation functions. The Sigmoid Function curve looks like a S-shape as seen in Figure 5.2.

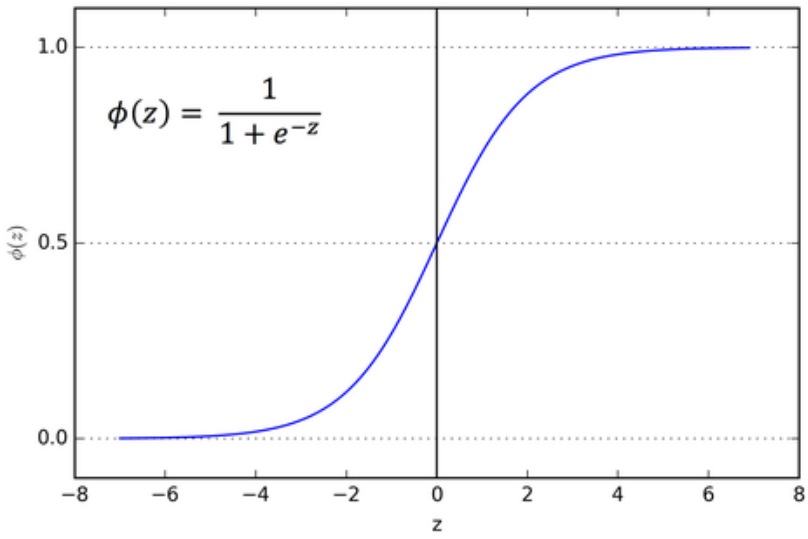

Figure 5.2: Sigmoid Function

Softmax activation function

The Softmax regression is a form of logistic regression that normalizes an input value into a vector of values that follows a probability distribution. The function is usually used to compute losses that can be expected when training a data set. Known use-cases of softmax regression are in discriminative models such as Cross-Entropy and Noise Contrastive Estimation. softmax is not ideally used as an activation function like Sigmoid but rather between layers which may be multiple or just a single one. The Sigmoid Function curve looks like a S-shape as seen in Figure 5.3.

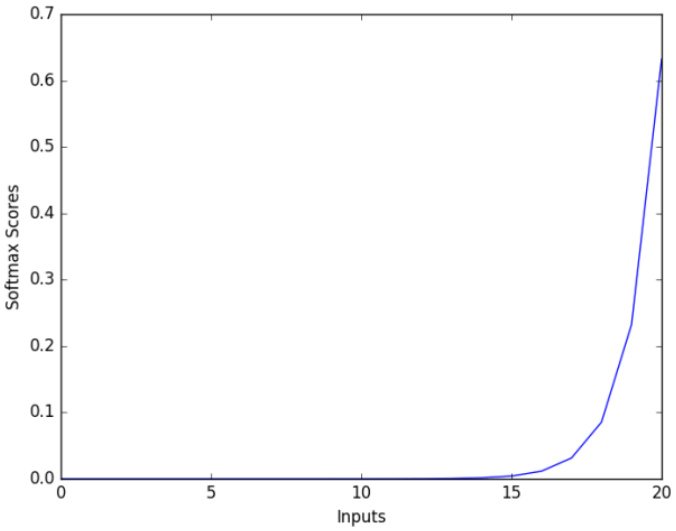

Figure 5.3: Softmax Function

RMSprop optimizer

Optimizers are a crucial part of the neural network, understanding how they work would help us to choose which one to use for our purpose. Picking the right optimizer with the right parameters, can help us squeeze the last bit of accuracy out of our neural network model.

Rmsprop is an optimization algorithm designed for neural networks which lies in the realm of adaptive learning rate methods, which have been growing in popularity in recent years. It is an adaption of rprop algorithm and its central idea is to keep the moving average of the squared gradients for each weight and then divide the gradient by square root the mean square, which is why it is called RMSprop (root mean square).

RMSprop is good, fast and very popular optimizer. It also takes away the need to adjust learning rate, and does it automatically. More so, RMSprop choses a different learning rate for each parameter.

SGD optimizer

Gradient descent is a very popular optimization technique in machine learning and deep learning and it can be used with most, if not all, of the learning algorithms. A gradient is basically the slope of a function; the degree of change of a parameter with the amount of change in another parameter. Gradient descent is a convex function and can be described as an iterative method which is used to find the values of the parameters of a function that minimizes the cost function as much as possible.

Stochastic gradient descent (SGD) calculates the cost of one example for each step. Hence, computationally much less expensive than typical Gradient Descent and this speeds up neural networks greatly. In most scenarios, SGD is preferred over Batch Gradient Descent for optimizing a learning algorithm.

MSE Loss Function

All the algorithms in machine learning rely on minimizing or maximizing a function, which we call “objective function”. The group of functions that are minimized are called “loss functions”. A loss function is a measure of how good a prediction model does in terms of being able to predict the expected outcome. There is not a single loss function that works for all kind of data. It depends on a number of factors including the presence of outliers, choice of machine learning algorithm, time efficiency of gradient descent, ease of finding the derivatives and confidence of predictions. Loss functions can be broadly categorized into 2 types: Classification and Regression Loss. Regression functions predict a quantity, and classification functions predict a label.

Mean Square Error (MSE) is the most commonly used regression loss function. It is the preferred loss function under the inference framework of maximum likelihood if the distribution of the target variable is Gaussian. MSE is calculated as the average of the squared differences between the predicted and actual values. The mean squared error loss function can be used in Keras by specifying ‘mse‘ or ‘mean_squared_error‘ as the loss function when compiling the model.

Binary Cross-Entropy Loss Function

It is also called Sigmoid Cross-Entropy loss. It is a Sigmoid activation plus a Cross-Entropy loss. Unlike Softmax loss it is independent for each vector component (class), meaning that the loss computed for every ANN output vector component is not affected by other component values. That's why it is used for multi-label classification, were the insight of an element belonging to a certain class should not influence the decision for another class. So when using this Loss, the formulation of Cross Entropy Loss for binary problems is often used.

Based on these the validation test examined the performance of ANN in forecasting property prices. The validation process was performed by comparing the actual and forecasted values for validation dataset. Table 5.2 and 5.3 shows the accuracy and loss of the model using different accuracy and loss functions.

Table 5.2: Accuracy result of the model using different activation, optimization and loss functions

Type of Attribute	Activation Function	Optimizer	Loss Function	Accuracy
All attributes	Sigmoid	Rmsprop	Mean squared error	85%
Important attributes	>>	>>	>>	89%
All attributes	Softmax	>>	>>	62%
Important attributes	>>	>>	>>	71%
>>	Softmax	SGD	>>	50%
>>	Sigmoid	>>	>>	63%
>>	>>	Rmsprop	categorical_crossentropy	85%
Important attributes	Sigmoid	Rmsprop	binary_crossentropy	55%

As seen the result in Table 5.2, our developed model have achieved the highest accuracy using mean squared error loss function, Root Mean Square Propagation Algorithm optimizer and Sigmoid activation function. Using these functions we plotted training loss and the validation loss over the number of model iterations. To display some nice graphs, we used matplotlib package and visualized the training loss and the validation loss as seen in Figure 5.4.

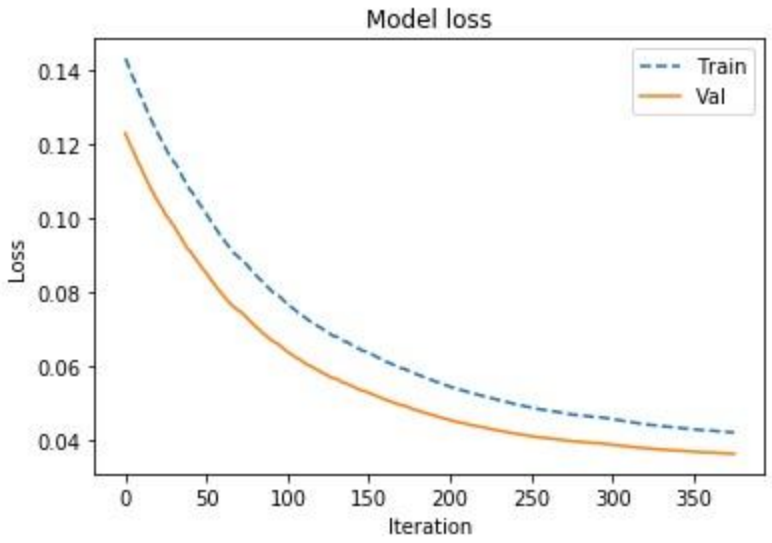

Figure 5.4: Model Loss graph

As we can see in Figure 5.4 the loss started to decrease from first iteration until 362nd iteration and becomes constant for the next five iterations, so the training algorithm stopped the learning processes at 367th iteration. Since the reduction of loss in our model in the training set looks somewhat matched up with reduction of loss to the validation set, over-fitting is not a problem in our model.

Using the same way we plotted training accuracy and the validation accuracy over the number of model iterations to visualize the training accuracy and the validation accuracy as seen in Figure 5.5.

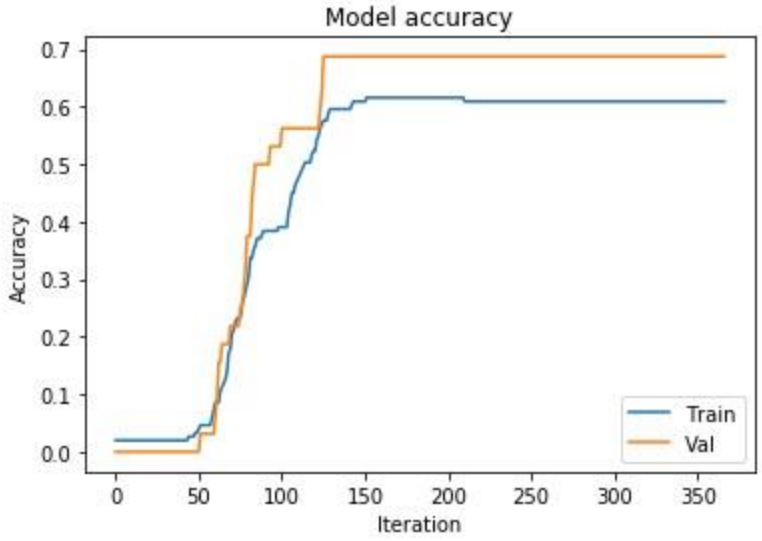

Figure 5.5: Model Accuracy graph

As we can see in the Figure 5.5 the accuracy started to increase from first iteration until 203rd iteration for training data and until 139th for validation data and becomes constant after that. Since the loss decreased until 362nd iteration, so the training algorithm stopped the learning processes at 367th iteration. Since the growth of accuracy in our model in the training set looks somewhat matched up with growth of accuracy to the validation set, over-fitting is not a problem in our model.

Using XGBoost which is an open-source optimized distributed gradient boosting library we increased the accuracy of our model both in the training and test phase as seen in Table 5.3 below.

Table 5.3: Accuracy result of the model using XGBoost

Dataset	Activation Function	Optimizer	Loss Function	Accuracy
Training	Sigmoid	Rmsprop	Mean squared error	98.45%
Test	>>	>>	>>	97.47%

5.5 Summary

We have shown that our proposed artificial neural network model predicts property values and achieved high accuracy. This is tested by using different activation, loss functions and using different optimizers. Moreover, it was shown that there are specific attributes especial attributes related to location are the ones which affect property prices significantly.

Chapter Six: Conclusion and future work

This chapter is conclusion of observations from our research. It also contains future works to show further researches that can be done in the future.

6.1 Conclusion

In this research work, property valuation model is developed in order to valuate a given property at a given time. A total of 43 attributes are identified to model ANN based property valuation and from those attributes which have high impact on the valuation are identified. Geospatial database is also used to get detail locational values of a given property in addition to the historical sales data collected.

MinMaxScaler is used to transform features by scaling each feature individually to a given range, e.g. between zero and one by preserving the shape of the original distribution. Search and Trace Anomaly algorithm, is also used to detect anomalies. Clustering is done by K-means method and finally to implement the models Keras is used because keras uses the idea of adding layers to the network instead of defining the entire network at once. This opens us to quick alter the number of layers and type of layers, which is handy when optimizing the network.

Results show that the overall success rate property price prediction using significant attributes is 97.47%. The success rates of predicting property price using significant attributes is higher than using all attributes. Moreover, these results show that, the proposed ANN based property price prediction algorithms and system architecture is effective in predicting the value of a given property. Hence, it is feasible to predict a given property price at current time for specific property valuation techniques using ANN. Therefore, it is practically possible to avoid the negative aspects of the manual work.

6.2 Contribution to Knowledge

This research work has contributed the following to the area of deep learning in relation to property valuation.

- We proposed novel generic property valuation model which can value a property based on the known valuation techniques.
- We used a total of 43 features that are essential for property valuation and we identified 24 features out of them which are significant for a successful property valuation. More of the significant features are locational values.
- We proposed system architecture which integrated with geo-database for getting additional locational values to add to our historical sales data
- We designed a model (using artificial neural network) that can successfully predict property price.

6.3 Future Work

Though this study has been able to predict property prices successfully, few works still remain unsolved. Based on our observation we recommend the following future research areas.

- In this study, sales comparison property valuation approach is mostly used and cost property valuation approach is implemented in small amount. For complete property valuation, the model should use fully all approach of property valuation data. For this, material cost data and income data of properties have to be collected in large amount and also have to be updated regularly. In the future, we aim to test our approach with all those data's in large data.
- We believe the deep neural network we used can be further refined by changing the network type and architecture.
- Our approach can be extended to rural area property valuation too. We currently focused on developing a model for property valuation of urban areas by collecting sales data of urban properties. This approach can be extended to rural property valuation by collecting historical sales data for the rural areas.

- Rather than adding new data in the future for using this model, it will be developed to save the predicted data's as in put data for the next time prediction. By making this the model will have latest data every time.
- Our approach can be customized for other prediction areas which have similar nature like properties. By only changing the training data and prediction approach of the specific area we can develop prediction model for other areas too.

REFERENCES

[1]. Khedkar, Bonissone. “Method and system for automated property valuation”, United States Patent, Patent No.: US 6,748,369 B2, Jun. 8, 2004.

[2]. French & Gabrielli, "Pricing to Market - Property Valuation Models in the context of Regulatory Definitions," ERES eres2018_285, European Real Estate Society (ERES), 2018.

[3]. French, Sloane, "Property valuation in the UK: Implicit versus explicit models–the baby and the bathwater", Journal of Property Investment & Finance, Vol. 36 Issue: 4, pp.397-406, Available: https://doi.org/10.1108/JPIF-04-2018-0024, 2018.

[4]. Abidoye, Chan, "Artificial neural network in property valuation: application framework and research trend", Property Management, Vol. 35 Issue: 5, pp.554-571, 2017. Available: https://doi.org/10.1108/PM-06-2016-0027.

[5]. Cheung, Simon KC, and Sahminan. "A Localized Model for Residential Property Valuation: Nearest Neighbor with Attribute Differences." *International Real Estate Review*20.2: 221-250, 2017.

[6]. Chiarazzoa, Caggiania, Marinellia and Ottomanellia. "A Neural Network based model for real estate price estimation considering environmental quality of property location." 17th Meeting of the EURO Working Group on Transportation, EWGT2014, Sevilla, Spain, 2-4 July 2014.

[7]. V.Yakubovskyi, Bychkov, Dimitrov, Panayotova. "Combined neural network model for real estate market range value estimation." *Proceedings of the Fourth International Conference on Artificial Intelligence and Pattern Recognition (AIPR2017), Lodz, Poland, 2017*.

[8]. Weldegebriel AMBAYE, Daniel, Land Valuation for Expropriation in Ethiopia: Valuation Methods and Adequacy of Compensation, 2018.

[9]. CBE, *Real Property Valuation manua*l, Commercial Bank of Ethiopia, July 2006.

[10]. Belachew Yirsaw Alemu, "Expropriation, valuation and compensation practice in Ethiopia: The case of Bahir Dar city and surrounding", Property Management, Vol. 31 Issue: 2, pp.132-158, Available: https://doi.org/10.1108/02637471311309436, 2013.

[11]. Belachew yirsaw Alemu, “Expropriation, Valuation and Compensation in Ethiopia”, Real Estate Planning and Land Law, Department of Real Estate and Construction Management, School of Architecture and the Built Environment, Royal Institute of Technology (KTH), Stockholm, Sweden, 2013.

[12]. Dambala Gelo, F.Koch, “Contingent valuation of community forestry programs in Ethiopia: Controlling for preference anomalies in double-bounded CVM”, University of Pretoria, March 2012.

[13]. Goodfellow, Tom. "Taxing property in a neo-developmental state: The politics of urban land value capture in Rwanda and Ethiopia." *African Affairs* 116.465 (2017): 549-572.

[14]. Negarit gazeta, “Civil code of the Empire of Ethiopia”, Proclamation No. 165 of 1960, 5th may 1960.

[15]. Abay bank, *Real Property Valuation manua*l, May 2016.

[16]. Blackledge, “Introducing property valuation”, ISBN-13: 978-0415434775.

[17]. International valuation standards. “General valuation concepts and principles”, international

valuation standards committee, 2003.
[18]. Calhoun, “Property Valuation Models and House Price Indexes for the Provinces of Thailand”, 2000.
[19]. Glumac and Rosiers, “Real estate and land property automated valuation systems: A taxonomy and conceptual model”, Luxembourg institute of socio-economic research, April 2018.
[20]. Belniak, Wieczorek, “Property valuation using hedonic price method”, Institute Construction and Transportation Engineering and Management, Faculty of Civil Engineering, Cracow University of Technology, June 2017.
[21]. Dr. N. B. Chaphalkar, Sandbhor, “Use of Artificial Intelligence in Real Property Valuation”, International Journal of Engineering and Technology (IJET), 2013.
[22]. Commerce Division, Lincoln University**,** “House Price Prediction: Hedonic Price Model vs. Artificial Neural Network”, NZARES Conference, June 25-26, 2004.
[23]. MELICHAR, VOJÁCEKb, RIEGERc, JEDLICKAd, “Application of Hedonic Price Model in the Prague Property Market”,
[24]. Owusu-Ansah, Anthon, “A review of hedonic pricing models in housing research”, A Compendium of International Real Estate and Construction Issues. 1. 17-38, 2013.
[25]. Fletcher M., Gallimore P. and J. Mangan, “The Modeling of Housing Sub-markets”, Journal of Property Investment & Finance, 18(4), 2000.
[26]. Feng, Yingyu, and Jones. "Comparing Methods: Using Multilevel Modelling and Artificial Neural Networks in the Prediction of House Prices based on property, location and neighbourhood characteristics." School of Geographical Sciences, University of Bristol, 2015.
[27]. Yacim, Awoamim, and Boshoff. "Impact of Artificial Neural Networks Training Algorithms on Accurate Prediction of Property Values." Journal of Real Estate Research 40.3 (2018): 375-418.
[28]. Sandbhor, Sayali, and N. B. Chaphalkar. "Effect of Training Sample and Network Characteristics in Neural Network-Based Real Property Value Prediction." Proceedings of the 2nd International Conference on Data Engineering and Communication Technology. Springer, Singapore, 2019.
[29]. Birkeland, Kristoffer, and D'Silva. "Developing and Evaluating an Automated Valuation Model for Residential Real Estate in Oslo." Master's thesis, NTNU, 2018.
[30]. Rahman, Abd, et al. "the artificial neural network model (ann) for malaysian housing market analysis." planning malaysia journal 17.9, 2019.
[31]. Chan, Albert PC, and Abidoye. "Advanced property valuation techniques and valuation accuracy: Deciphering the artificial neural network technique." RELAND: International Journal of Real Estate & Land Planning 2, 2019.
[32]. Swamynathan, Manohar. "Step 3–Fundamentals of Machine Learning." Mastering Machine Learning with Python in Six Steps. Apress, Berkeley, CA, 2017. 117-208.

[33]. Bulut, Bahar & Allahverdi, Novruz & Kahramanli, Humar & Yalpir, Sukran. “A Residential Real-Estate Valuation Model with Reduced Attributes.” International Journal of Mathematical Models and Methods in Applied Sciences. 5. 2011.
[34]. Oloke, C. Olayinka, R. F. Simon, and Ayotunde F. Adesulu. "An examination of the factors affecting residential property values in Magodo neighbourhood, Lagos state." International Journal

of Economy, Management and Social Sciences 2.8 (2013): 639-643.
[35]. Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.
[36]. Hall, Andrew. "Correlation-based feature selection for machine learning." (1999).
[37]. Reitermanova, Z. "Data splitting." WDS. Vol. 10. 2010.
[38]. May, Robert J., Holger R. Maier, and Graeme C. Dandy. "Data splitting for artificial neural networks using SOM-based stratified sampling." Neural Networks 23.2 (2010): 283-294.
[39]. Rojas, Raúl. Neural networks: a systematic introduction. Springer Science & Business Media, 2013.
[40]. KristinStock, HansGuesgen. “Geospatial Reasoning with Open Data”. School of Engineering and Advanced Technology, Massey University, New Zealand (Albany, Auckland campus).
[41]. Kieu, Tung, et al. "Outlier detection for time series with recurrent auto encoder ensembles." 28th international joint conference on artificial intelligence. 2019.
[42]. Kuck, Jonathan, et al. "Query-based outlier detection in heterogeneous information networks." Advances in database technology: proceedings. International Conference on Extending Database Technology. Vol. 2015. NIH Public Access, 2015.
[43]. Zhuang, Xinhua, et al. "Gaussian mixture density modeling, decomposition, and applications." IEEE Transactions on Image Processing 5.9 (1996): 1293-1302.
[44]. Cousineau, Denis, and Sylvain Chartier. "Outliers detection and treatment: a review." International Journal of Psychological Research 3.1 (2010): 58-67.
[45]. Slifer, Russell D. "Geospatial and temporal data system." U.S. Patent No. 10,216,760. 26 Feb. 2019.
[46]. Li, Yang, et al. "Efficient Spatial Nearest Neighbor Queries Based on Multi-layer Voronoi Diagrams." arXiv preprint arXiv:1911.02788 (2019).
[47]. Sharifzadeh, Mehdi, and Cyrus Shahabi. "Vor-tree: R-trees with voronoi diagrams for efficient processing of spatial nearest neighbor queries." (2010).
[48]. Oh, Eunjin, and Hee-Kap Ahn. "Approximate range queries for clustering." arXiv preprint arXiv:1803.03978 (2018).

Appendix A
Data PreProcessing

```
#!/usr/bin/env python3
import tensorflow as tf
import keras
import pandas
import sklearn
import matplotlib
import datetime
import pyodbc
import numpy as np

conn = pyodbc.connect('Driver={SQL Server};'
                      'Server=TEDO\SRV;'
                      'Database=valuation;'
                      'Trusted_Connection=yes;')

cursor = conn.cursor()
cursor.execute('SELECT * FROM valuation.dbo.historicalData')

dataset = cursor.fetchall()
dataset = np.asarray(dataset)

from datetime import datetime

for row in dataset:
   data = datetime.strptime(row[18],'%m/%d/%Y')
   data = data.date()
   data = data.strftime("%Y%m%d")
   cursor.execute("""UPDATE valuation.dbo.historicalValuationData SET Date =? WHERE Date =?""",[data,row[18]])
   conn.commit()
   dataset[index,18] = datetime.strptime(row[18],'%m/%d/%Y')
   dataset[index,18] = dataset[index,18].date()
   dataset[index,18] = dataset[index,18].strftime("%Y%m%d")
   dataset[index,28] = dataset[index,28].replace(",", "")

X = dataset[:,0:28]
Y = dataset[:,28]
from sklearn import preprocessing
min_max_scaler = preprocessing.MinMaxScaler()
X_scale = min_max_scaler.fit_transform(X)
from sklearn.preprocessing import StandardScaler
Y = min_max_scaler.fit_transform(Y.reshape(len(Y),1))[:,0]
```

Appendix B
Correlation

```
from scipy import stats
from scipy.stats import norm, skew
from sklearn import preprocessing
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.linear_model import ElasticNetCV, ElasticNet
from xgboost import XGBRegressor, plot_importance
from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import StratifiedKFold

sns.distplot(Y, fit=norm);
#Get the fitted parameters used by the function
(mu, sigma) = norm.fit(Y)
print('\n mu = {:.2f} and sigma = {:.2f}\n'.format(mu, sigma))

#Now plot the distribution
plt.legend(['Normal dist. ($\mu=$ {:.2f} and $\sigma=$ {:.2f})'.format(mu, sigma)],loc='best')
plt.ylabel('Frequency')
plt.title('Sale Price distribution')

#Get also the QQ-plot
fig = plt.figure()
res = stats.probplot(Y, plot=plt)
plt.show();

data_heatmap = data;
import matplotlib.pyplot as plt222
pandas.set_option('precision', 10)
plt222.figure(figsize=(14, 8))
sns.heatmap(data_heatmap.drop(['Price'],axis=1).corr(), square=True)
plt222.suptitle("Pearson Correlation Heatmap")
plt222.show();

import matplotlib.pyplot as plt2
data_corr = pandas.read_csv('propertydatarelated.csv')
corr_with_price = data_corr.corr()
corr_with_price['Price'].sort_values(ascending=False)
plt2.rcParams['figure.figsize'] = (14,6)
corr_with_price.drop('Price').plot.bar()
plt2.show();
import matplotlib.pyplot as plt3
```

```
data_corr_mat = pandas.read_csv('propertydatarelated.csv')
corr_mat = data_corr_mat.corr()
cols = corr_mat.nlargest(31, 'Price')['Price'].index
cm = np.corrcoef(data_corr_mat[cols].values.T)
f, ax = plt.subplots(figsize = (12,10))
sns.heatmap(cm, ax = ax, cmap ='YlGnBu', linewidths =0.1, yticklabels = cols.values, xticklabels = cols.values)

sns.pairplot(data_corr_mat[['Price', 'Size', 'Electrical fixtures', 'Sanitary fixtures', 'Veranda Width', 'Property Category',
                    'Parking', 'Place', 'No of BedRooms', 'No of Bathrooms', 'Number of Salon', 'No of kitchens']])
plt.show();

feature_importance = pandas.Series(index = data.columns, data = np.abs(cv_model.coef_))
n_selected_features = (feature_importance > 0).sum()
print('{0:d} features, reduction of {1:2.2f}%'.format(n_selected_features,(1-

n_selected_features/len(feature_importance))*100))
feature_importance.sort_values().tail(30).plot(kind = 'bar', figsize =(12,5));
```

Appendix C
Model Implementation

```
from keras.models import Sequential
from keras.layers import Dense
from keras.optimizers import SGD

model = Sequential([
   Dense(60, activation='relu', input_shape=(17,)),
   Dense(60, activation='relu', input_shape=(17,)),
   Dense(1, activation='sigmoid'),
])
opt = SGD(lr=0.5, momentum=0.9)
model.compile(optimizer=opt,
           loss='mse',
           metrics=['acc'])
#sgd = stochastic gradient descent


hist = model.fit(X_train, Y_train,
             batch_size=60, epochs=367,
             validation_data=(X_val, Y_val))

model.evaluate(X_test, Y_test)[1]

import matplotlib.pyplot as plt
plt.plot(hist.history['loss'], linestyle='dashed')
plt.plot(hist.history['val_loss'])
plt.title('Model loss')
plt.ylabel('Loss')
plt.xlabel('Iteration')
plt.legend(['Train', 'Val'], loc='upper right')
plt.show()

plt.plot(hist.history['acc'], linestyle='dashed')
plt.plot(hist.history['val_acc'])
plt.title('Model accuracy')
plt.ylabel('Accuracy')
plt.xlabel('Iteration')
plt.legend(['Train', 'Val'], loc='lower right')
plt.show()
```

Appendix D
Attributes Used

Property Category	Property Type	Age	Size
Floor	No of Bedrooms	No of kitchens	No of Bathrooms
Number of Salon	Parking	Place	Construction Phase
Access to Transportation	Access to schools	Access to shopping	Access to health center
Visibility	Date Sold	Veranda Width	Made Of
Distance From Main Road(KM)	Elevator	Sanitary fixtures	Electrical fixtures
Access to Cinema House	Access to Green area	Population Density	Noise Level
Access to Work Place	Access to religious places	Access to Banks	Access to Restaurants
Access to police stations	Access to Metro Station	Access to Airport	Polluted Area
Roofing and Ceiling	Roof terrace	Basement	Crime Rate
Close to the city center	Access to sewage disposal	Price	

Signed Declaration Sheet

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: **Tedros Fekadu Gebregziabher**

Signature: _____

Date: _____

Confirmed by Advisor:

Name: **Fekade Getahun (PhD)**

Signature: _____

Date: _____