Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Customer Size Prediction using Machine Learning Approach for Mobile Package Development in ethio telecom

By

Desalegn Medhin Firdu

Advisor

Dr. Rosa Tsegaye Aga

A thesis submitted to the School of Electrical and Computer Engineering in Partial Fulfillment of the requirements for The Degree of Master of Science in Telecommunication Engineering (Telecommunication Networks Engineering Track)

December, 2021

Addis Ababa, Ethiopia

Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Customer Size Prediction using Machine Learning Approach for Mobile Package Development in ethio telecom

By: Desalegn Medhin Firdu

Approval by Board of Examiners

Dr. Rosa Tsegaye Aga	_____
Advisor	Signature
_____	_____
Examiner	Signature
_____	_____
Examiner	Signature

Declaration

I, the undersigned, declare that the thesis comprises my own work in compliance with internationally accepted practices. I have fully acknowledged and referred all materials used in this thesis work.

Desalegn Medhin Firdu _____

Name Signature

Place: **Addis Ababa University, Ethiopia**
Submission date: December, 2021

This master thesis has been submitted for examination with my approval as a university advisor.

Dr. Rosa Tsegaye Aga _____

Advisor Signature

Abstract

Nowadays the telecom market is competitive and telecom operators launch various new service packages to meet customer needs. New package and tariff preview is important to ensure business continuity for telecom operators. Hence, scientific and reasonable analysis of prediction is highly needed before new service package is introduced. In the case of ethio telecom, there is no automated method for package preview. To address this gap, Machine Learning (ML) approach has been followed to predict customer size for new mobile packages in the thesis work.

In this study, three ML algorithms, ElasticNet regression, Extreme Gradient Boosting and Random Forest (RF) regression have been used to train models. For this purpose, mobile package dataset is formed from the available data in ethio telecom. The model training has been conducted using the scikit learn Python library functions. Model evaluation is executed to calculate the error between the actual and the predicted values using two method: Root Mean Squared Error and Cross Validation. An optimal subset of hyper-parameters for the algorithms was selected through the grid search function for the best prediction.

The RF model has performed better than the other algorithm in terms of smaller prediction error and be better suited as a solution model for our purpose. The prediction error of the RF model is 1.3% of the average daily mobile package purchase rate.

Key words- Machine Learning, Mobile Package, Customer Size, ethio telecom

Acknowledgments

First of all, I give thanks to the Almighty God for helping me to finish the thesis work at this difficult time. I would like to extend my gratitude to my advisor Dr. Rosa Tsegaye Aga, for her support and guidance throughout the thesis work.

Sincere thanks to all ethio telecom colleagues for providing me the necessary supports during my research work. Finally, I would like to give special thanks to my beloved family (Abi, Miky and Abity).

Table of Content

List of Figures

List of Tables

List of Acronyms

B2B	Business to Business
BICP	Business Intelligence Communication Platform
CBS	Convergent Billing System
CV	Cross Validation
DT	Decision Tree
GridSearchCV	Grid Search Cross Validation
ML	Machine Learning
RF	Random Forest
RMSE	Root Mean Squared Error
SMS	Short Message Service
XGBoost	Extreme Gradient Boosting

Chapter 1: Introduction

1.1. Background

As communication technologies develop persistently, the competition advantages based on technology start to vanish. Now the telecom market is complex and telecom operators launch various sorts of new service packages. Every operator is trying to launch products and service packages to meet customer needs. Promotion of new telecom service packages concerns not only product orientation and operating income for the company but also service perception and acceptance for customers. Hence, scientific and reasonable analysis of prediction is highly needed before new service package is introduced [1].

With the globalization of the telecom market and diversification of users, competitions between telecom operators become fierce. As a result, telecom operators introduce new service packages to seize market opportunities, attract more new customers and increase business revenue. Telecom enterprises can reduce operational risk by updating service packages timely through previewing new package accurately and estimating the number of users, cost, revenue, etc. [2]. New package and tariff preview is important to ensure business continuity for telecom operators and they should employ some mechanism for this purpose.

The vast volume of data telecom companies collect and possess could be effectively utilized for solving their business problems. Data Mining can be utilized to automatically generate knowledge from the available data. Data Mining and Business Intelligence applications play a significant role in the telecom industry to overcome the hard competition in the sector [3], [4]. The available customer data can be used to profile customers for marketing and forecasting purposes. Hence, data mining is the most relevant tool to solve business and operational problems in telecom companies.

According to ethio telecom's official website, the company has above 54.3 million mobile customers [5]. The company has been offering various mobile service package options for Voice, Short Message Service (SMS) and Internet services. These service packages are provided at discount price with a fixed validity period and the goal of service packaging is to influence customer service usage and increase the company revenue [6]. Although revenue is the main concern of business entities, understanding you customer base is more important to develop products and services that satisfy their needs. As a result, the focus of this thesis work is to predict the customer size new mobile packages in ethio telecom.

Ethio telecom uses manual methods to design mobile packages and the market performance is evaluated after the package is released through post launch analysis. In this study, Machine Learning (ML) model that predict the customer size for new package is built. The model is trained using a dataset created by integrating existing packages information and purchase data from different sources. Similar research has not been done yet in ethio telecom and the research work will help the company to have good marketing plan and adjust itself for the upcoming competitive market in the country.

1.2. Statement of the Problem

Package is a service mode which is common in telecommunication service providers. It is designed by combining price plans, free resources (Voice, SMS and Data), preferential service tariffs and other value-added services according to users' need. When customers face many package options, their choice trend is unknown before the package launch. Therefore how to predict the consumer size for new package is crucial for telecom operators [7].

Ethio telecom's mobile package strategy is based on traditional customer classifications such as prepaid, postpaid and hybrid subscribers and similar package is designed for all customers [8]. Since packages are not designed for specific customer group, it is difficult to understand

whether it has meet its market share. In addition, there is no specific mechanism to predict the number of subscribers for a new package before release. This will have a negative impact on the effectiveness of the company's marketing decisions.

In ethio telecom, new package designing process is done based on market assessment survey results and manual customer usage analysis outputs [9]. Marketing principles and statistical methods are employed in the process. The newly designed package is released to market after the deployment process is finished. However, its market effect and the customer size is evaluated latter through post launch analysis. Corrective measures and other related decisions are conducted based on the post launch analysis results.

Within the Marketing Division of ethio telecom, package design is done by Product and Services Department while the post launch analysis is handled by Marketing Research and Intelligence Department. The post launch analysis is conducted to assess the impact of new packages on the company market trend. The analysis result includes number of subscriber for the new package, gained revenue, its influence on existing similar packages, etc. The analysis task may take weeks or months after launch depending on the urgency of the requesting department and the package type. Hence, there is a time delay to have the analysis result for further decisions.

In this thesis work, ML model has been built to predict customer size for new package using the available mobile packages information and purchase reports. This model will shorten the time needed for mobile package designing process and the post launch activities in the company. The prediction result of the ML model can help to set a target for new package purchase performance evaluation. The model will be used as a new package preview tool in the company and helps to predict the post launch analysis results before package release in a short time. As a result, the post launch analysis can be replaced by this model and some new

package related business effects can be included within the regular business review activities. In addition the model can be used to customize different packages under designing process and produce optimal packages.

1.3. Objective

1.3.1. General Objective

The general objective of this work is to build ML model that predicts customer size for new package before launch to improve the mobile package development process in ethio telecom.

1.3.2. Specific Objectives

The specific aims of the study are:

- To integrate data from different sources to form quality dataset
- To build a high performing ML regression model
- To compare and identify appropriate algorithm for the customer size prediction

1.4. Methodology

To meet the general and specific objectives of this research, we have followed the below methods as depicted in Figure 1.1.

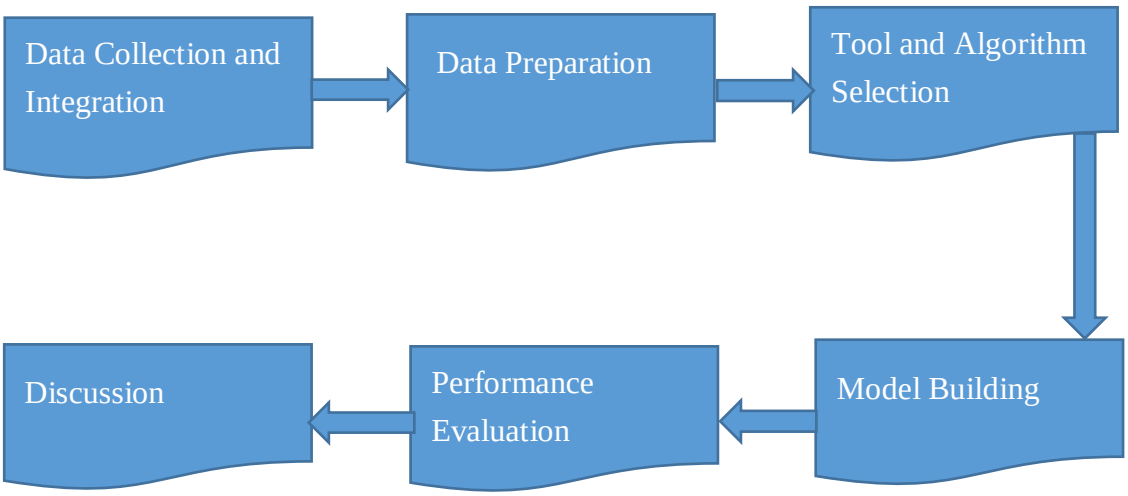

Figure 1.1: Research Methodology

Data Collection and Integration

For the thesis work, the main input data is mobile packages information and customer purchase report data. The package information has been collected from the Convergent Billing System (CBS) and the purchase report from Business Intelligence Communication Platform (BICP). In addition, marketing product catalog and post launch analysis reports have been considered as supporting data. The collected data is integrated using statistical tools to form complete dataset for the model training.

Data Preparation

In this step data cleaning, normalization and encoding tasks have been conducted. The collected raw data is transformed into a form that can be trained using ML algorithms. Moreover, important features have been selected and additional fields are constructed to enhance the data quality.

Tools and Algorithms Selection

Based on the available data and the state of the art, appropriate algorithms and tools have been selected. As the research focuses on customer size prediction, regression algorithms and other supporting python functions and libraries are employed.

Model Building

ML model is built by learning and generalizing from training data, then applying that acquired knowledge to new data to make predictions. For this thesis work, the prepared dataset has been split for training, validation and test sets. A regression model has been trained on the training dataset and package customer size is the target attribute. The training task is conducted for each algorithm and the model performance has been enhanced by tuning algorithm parameters.

Performance Evaluation

The model performance has been evaluated using standard methods such as Root Mean Squared Error (RMSE) and Cross Validation (CV). RMSE is the most popular evaluation metric used in regression problems with an assumption that errors are unbiased and follow a normal distribution. CV is a technique which uses different portions of the dataset to test and train a model on different iterations. CV helps to evaluate how accurately a predictive model will perform in practice. Comparing evaluation result of both methods used, the best model has been selected as the final solution model.

Discussion

Finally, the selected solution model and its prediction results have been analyzed. The contribution of the model to enhance the existing package development process is also discussed. Moreover, a real-time test is conducted to illustrate the deployment part and the prediction result is compared with the target and actual customer size.

1.5. Related Work

Related research works and journal papers in the area of data mining and customer prediction have been reviewed. Besides, marketing documents and reports were analyzed for a better understanding of the marketing domain as well as mobile service package development process. Some selected literature are discussed as follows.

In [10], the objective is forecasting the number of subscribers to telecom services and a choice-based substitutive and competitive model has been designed. The model is suggested to describe an environment in which substitution and competition occur simultaneously. The choice-based model is useful in that it enables the description of such complicated environments and provides the flexibility to include marketing variables in regression analysis.

The study is done based on marketing and statistical principles and it helps to understand the customer choice base for new service subscription.

In [1], the study focuses on the impact of new telecom services tariff on customers and the company revenue. Measurement model has been built through multi-nominal logit choice rule to predict the impact. Impact indicators like utility of service packages, transfer probability of the customers and expected change of revenue have been obtained. These are useful for market orientation, revenue prediction and optimization management of the new telecom services tariff. The results are based on customer behavior analysis which cannot be addressed through data mining methods.

In [2], the study has used statistics and data mining methods for the prediction of the number of new customers and transfer customers in telecom package preview. Linear Regression analysis and Exponential Smooth method have been used and compared in the establishment of prediction model for the number of new customers. Decision Tree algorithm has been used to set the transfer rules for customer changing service packages. The study has proposed that the key point of telecom tariff preview is to calculate the possible users of the new package. The possible customers are further divided into the new customers and the transfer customers. Finally, the authors have suggested that the prediction model in the study needs further improvement and validation owing to the limited sample data used.

In [11], the study has presented a brief analysis of the reliability of machine learning techniques for telecom Business to Business (B2B) sales prediction. Based on the performance assessment, a best-adapted predictive model for the B2B sales trend forecast has been suggested. Projection, estimation and analysis findings are summarized in terms of reliability and consistency of efficient prediction and forecasting techniques. The research has concluded that Gradient Boost Algorithm has good accuracy in B2B sales prediction for telecommunication companies.

In [12], a detailed study and analysis of comprehensible predictive models that improve future sales predictions has been carried out. The various techniques and methods for sales predictions are also described. On the basis of a performance evaluation, the Gradient Boost Algorithm has been suggested as a best suited predictive model for the sales trend forecast. Moreover, the research has concluded that an intelligent sales prediction system is essential for business organizations to utilize big data for business decisions.

1.6. Scope and Limitation

1.6.1. Scope of the Thesis

The scope of this thesis work is to predict customer size for new mobile packages in ethio telecom using ML approach. Moreover, three regression algorithms have been employed for model training and comparing their performances the best algorithm is identified.

1.6.2. Limitation of the Thesis

Mobile packages purchase depends on different factors which can be extracted from customer profile; like age, economic status, educational background, etc. As there is no fully developed customer profile data in ethio telecom, only package attributes and purchase reports are considered for the customer size prediction.

1.7. Contribution of the Study

The research has contributed ML solution for new mobile package customer size prediction in ethio telecom. This helps to improve the existing mobile package development process and reduce related workloads in the company. The dataset formed in this thesis work is also another contribution as it assists further studies on mobile packages.

1.8. Thesis Organization

The research paper in consists seven chapters. Chapter 1 is the introduction part and Chapter 2 presents an overview for the business domain understanding. Chapter 3 presents the detail on dataset formation, data description and preprocessing of the dataset. Chapter 4 introduces ML techniques and algorithms applied in this research work. Chapter 5 is the main part of the thesis work and focuses on the experimental analysis of ML model training for the customer size predilection using different algorithms. Chapter 6 focuses on model evaluation and interpretation of the prediction results and Chapter 7 covers conclusion and future work.

Chapter 2: Business Domain Understanding

2.1. Mobile Service Packages

Mobile services are provisioned through mobile phone and SIM card, which may move around freely within the service network coverage. It is supported by different mobile technologies such as, 2G, 3G and 4G, which incorporate better quality and feature advancement on the mobile business. Mobile technology supports different mobile services like; mobile voice, mobile internet, mobile SMS and other value added services [13].

Telecom network operators can combine mobile service usage price plan, discount price plan and free unit price plan or some of them together to define a package. Service packaging enables subscribers to enjoy preferential usage charging, discount charging or free unit by paying a certain rental fee. A typical mobile package charges the rental fee, giving free units and/or a favorable tariff or discounts on certain services [14]. Ethio telecom provides one-time and periodical/recurring mobile packages.

2.1.1. Periodical Package

After subscribing to a package, a subscriber is periodically presented with preferential tariff for particular services or free units and periodically charged a rental. The free units are immediately available to consume once presented to the subscriber and the validity period of the free unit is configurable at package level. The periodic package's cycle can be:

- Daily or several days
- Weekly or several weeks
- Monthly or several months
- Yearly or several years
- Bill cycle or several bill cycle

Ethio telecom provides monthly recurring voice and data packages and long validity periodical packages for customers through ethio gebeta (*999#), My Ethiotel app and customer relation management channels. The long validity packages are limited to quarter, semi-annual and a year validity periods.

2.1.2. One-time Package

For one-time package the subscriber will be charged a subscription fee for the application of the package, not periodically charged. The network operator can define the free units present to the subscriber and the validity period of the free units. The free units are immediately available to consume once presented to the subscriber.

The validity period of one time package can be:

- Daily
- Weekly
- Monthly

Daily, weekly and monthly one time packages of voice, SMS and data services are available for ethio telecom customers for self and gift purchases through the available channels.

2.2. Package Preferences

Package preferences is the options subscriber would like to choose and use. Package preferences allow a subscriber to choose a package based on subscription fee, discount or free unit amount. A package can support the following preferences or the combination of the preferences, including:

- Preferential usage tariff - lower tariff for the package owner
- Percentage discount preference - percentage discount for the package owner
- Reward preference - free resources for the package owner

All mobile services charges for voice, SMS and data are normally contained in basic tariffs of the primary offering plans. These can be redefined to promoted tariffs in package to form new

offering. The new tariffs can be combined according to time schema, service type, customer level, and other conditions.

Ethio telecom has developed different package options for a specific time schema (night and morning) and service type (voice, SMS, data or bundle). For illustration, mobile voice packages are presented here. Mobile voice package offers contain various voice plans to help mobile customer make a voice call with cheaper tariff compared to the normal usage tariff [13]. The available voice packages through USSD/ethio gebeta and MY Ethiotele App are:

- Night/Off peak Voice package
- Morning /Maleda Mobile Package
- Daily, Weekly and Monthly Voice Package
- Voice Plus Data Mobile Package

2.3. Package Development

Package development is regularly done by the Product and Service Department in ethio telecom. Based on marketing factors and tariff revisions, either existing packages are modified or new packages are released. In addition some promotional or event driven packages are becoming familiar in the company. Holiday package can be categorized in this group. Nowadays for every public holiday ethio telecom releases new brand packages that increase customer size.

2.4. Post Launch Analysis

Post launch analysis is conducted by Marketing Research and Intelligence Department for every released package. The analysis is done based on package purchase report and customer feedback on social media outlets. The analysis result is compared with the customer size forecast and revenue target given by the Product and Service Department. Finally, recommendations and remarks are given for further decisions and corrective actions.

Chapter 3: Data Analysis

The purpose of this thesis work is to build ML model for new mobile package customer size prediction. Hence, the available package data in ethio telecom is collected and analyzed in this part. The data preparation process is also discussed in this section.

3.1. Data Collection and Integration

Collecting data is the first step to achieve a machine learning model development. The more and relevant data that we collect, the better our model will be [15]. In this step all related data in ethio telecom has been collected form the responsible sections.

Main data sources for the mobile package dataset formation are CBS and BICP systems. Moreover marketing product catalog and package post launch analysis results are used. From CBS, mobile package information is collected and from the BICP system, package purchase report of two months is collected. Some additional package attributes are collected manually from the product catalog. In Table 3.1, each data source with corresponding collected package attribute is shown.

No.	Source	Attributes
1	CBS	Offer name, Offer ID, Price, Payment mode and Rent type
2	BICP	Offer name, Offer ID, Daily purchase and Revenue
3	Product Catalog	Offer name, Free resource amount (Voice, SMS and Data), Validity period

Table 3.1: Data Source

Most of the packages available in CBS are active and have purchase report in BICP for the collected months (February and March 2021) but some are out of market and their purchase and other related data is collected from post launch analysis reports. The collected data from CBS and BICP is integrated based on Offer ID and the remaining features are feed manually

from product catalog and post launch reports. Microsoft Excel tool is used for the dataset integration and its functions like pivot and look up are mostly employed.

3.2. Data Description

The newly formed dataset has (339x11) size and included every possible attributes that can help to characterize a given mobile package. A new feature 'Package_Type' is added to categorize similar packages and make clear distinction with the others based on the objective and time schema of the packages. Based on domain expert recommendations, five categories are formed for the Package_Type which are; Regular, Morning, Night, Weekend and Event Packages.

Regular packages are the most common packages that are available any time for purchase and usage such as the daily, weekly and monthly mobile packages. Whereas, the morning, night and weekend packages are designed for a specific usage time and their price is relatively cheaper. The last package type, event package, is promotional and event driven that is available for limited period only especially during public holidays. All the attributes in the dataset are described in Table 3.2.

No.	Attributes	Description	Data Type
1	OFFER_ID	Package unique ID / used as index	N/A
2	Price	Paid amount in birr	Numerical
3	Voice_Min	Package voice free resource	Numerical
4	SMS_Item	Package SMS free resource	Numerical
5	DATA_MB	Package data free resource	Numerical
6	VALIDTY_days	Package usage period	Numerical
7	PAYMENT_MODE	Prepaid/Postpaid/Hybrid	Categorical
8	PACKAGE_ownership	Self/Gift	Categorical
9	PACKAGE_TYPE	Regular/Morning/Night/Weekend/Event	Categorical
10	Rent_Type	One-time/ Recurring	Categorical
11	Daily_Purchase	Two months average purchase/ target feature	Numerical

Table 3.2: Package Attributes Description

To fully understand the dataset and the relationship between the attributes python data visualization methods are used. The correlation analysis of the numerical features in Figure 3.1, shows that the target feature (Daily_Purchase) has weak correlation with the other input features. On the other hand, the input features have relatively strong correlation to each other.

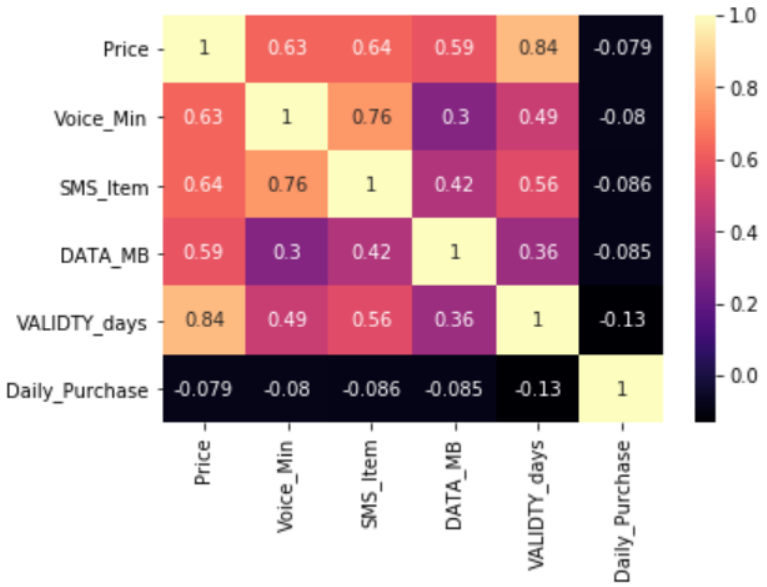

Figure 3.1: Feature Correlation

The categorical features relationship with the target feature is also shown below in Figure 3.2. Bar plot represents an estimate of central tendency for numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars. In our case, the bar plots show only the mean values and standard deviation of the observations is represented by the vertical line at the center of each bar.

As of Figure 3.2, it is observed that some of the categorical feature have high contribution in the package purchase amount while the others have less. For some features like Rent type and Payment mode, one value is dominant as most of the mobile packages have these specific categorical values. This is due to their acceptance at the market as customers prefer packages with these feature values. It is clear that customer size of a package is highly dependent on the package attributes and optimal attribute selection is important to have good customer size.

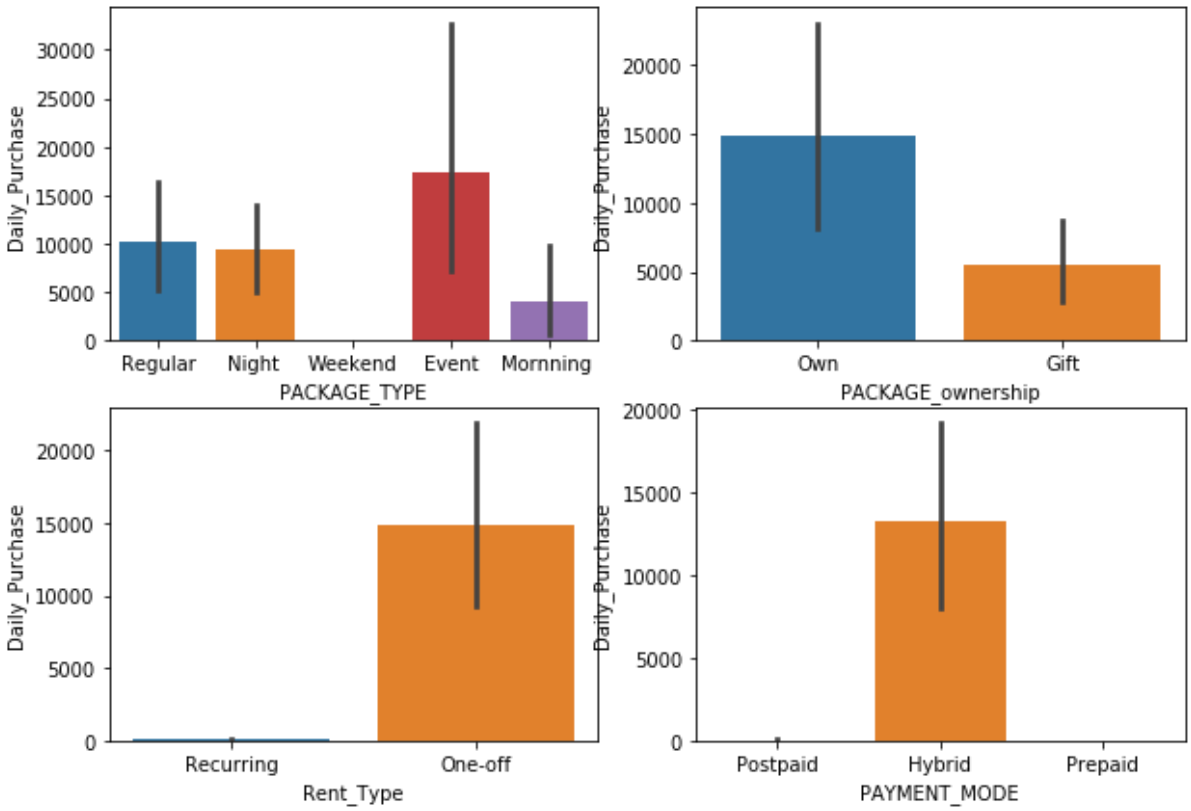

Figure 3.2: Categorical Features

From the collected mobile package features, Offer name and Package revenue are excluded as they are not relevant to the customer size prediction. OFFER_ID is not considered as an input feature it is only used for indexing. Daily_Purchase of a package is the target feature and the other numerical features have been used as input features. In addition, all the categorical features have been used as input features, since they help to differentiate similar packages for better customer size prediction.

3.3. Data Preparation

Data preparation (Data Preprocessing) is to transform our data in a way that can be feed into an ML model [15]. The most common tasks in data preprocessing are listed below and they are discussed in the next subsections.

- Dealing with missing data
- Handling categorical data
- Feature scaling
- Selecting meaningful features

3.3.1. Data Cleaning

The tasks in this stage include removing incomplete, erroneous, unnecessary and redundant values manually. Moreover, some packages with similar feature values but different Offer ID have been aggregated and their purchase is combined. For example, similar data packages prepared for 3G, 4G and data only users have been summarized together.

As some missed values in the dataset are feed manually, human errors are corrected thorough repeated assessments. Test packages formed during development stage and duplicated offers are removed for dataset quality. Some old package features are updated according to the current product catalog and missing values are fitted based on business rules and the product catalog inputs. Hybrid and business mobile voice packages that have higher tariff compared with other similar packages are removed. In addition long validity mobile packages with outlier feature values have been excluded in this step.

3.3.2. Data Normalization

The dataset contains numerical features with different value ranges and categorical features with nominal labels. In this phase normalization and encoding techniques have been used to transform the data.

Most of the ML algorithms perform better when their input features are normalized [15]. Normalization is re-scaling features to a specific range, which is convenient for the purpose at hand. To normalize our data we have applied the min-max scaling method to each numerical feature column and the values are scaled in the range between 0 and 1.

As ML models require all input and output variables to be numeric, there are two common categorical feature encoding techniques: Ordinal Encoding and One-Hot Encoding. For categorical variables with no ordinal relationship the latter encoding is appropriate. In reality, using ordinal encoding may result in poor performance or unexpected results [16]. As a result, one-hot encoding is applied to the categorical features in the dataset.

3.3.3. Feature Selection

Feature selection is the process of retrieving a subset of relevant features from the available features. The aim is to remove redundant or irrelevant features to simplify the model, shorten training times, and reduce dimensionality and the chance of over-fitting. For this thesis work, all the available features in the dataset are considered to be important and feature selection is not conducted. The number of available sample packages and their features is limited and eliminating some features due to selection may result in further sample reduction as of redundancy. Since some package samples differ each other only by a single feature, we have arranged to use all samples with complete feature values for the model training.

Chapter 4: Machine Learning Techniques and Algorithms

Machine learning is defined as the process of solving problems by collecting data, and algorithmically building a model based on that dataset. That ML model is assumed to be used to solve the practical problem. ML can be classified in to supervised, semi-supervised, unsupervised and reinforcement learning types [17].

Based on our input data type and objective of the thesis work, our focus is on supervised learning. Supervised learning is further classified in to regression and classification methods. Here, we have employed the regression analysis methods to build models for customer size prediction.

4.1. Regression Analysis

Regression analysis is a prediction modelling technique that considers the relationship between dependent and independent variables. It is used for forecasting, time series modelling and identifying cause and effect relationship between variables [18]. Some of the benefits of using regression analysis are:

- It indicates the significant relationships between dependent and independent variables.
- It indicates the strength of the impact each independent variable has on the dependent variable.

Various regression methods are available to make predictions. These mostly differ by the number of independent variables, type of dependent variables and purpose of the regression analysis. In the following subsections, we have discussed three regression algorithms which are employed in the thesis work. These are:

- ElasticNet Regression
- Random Forest Regression
- Extreme Gradient Boosting

4.1.1. ElasticNet Regression

ElasticNet Regression is a combination of L1 and L2 regularization techniques. It is used when there are multiple correlated features [18]. L1 regularization weights errors at their absolute value and results in models with fewer coefficients, as some coefficients can become zero. On the other hand L2 regularization weights errors at their square to punish higher errors more. L2 regularization is used to reduce model complexity. In general, ElasticNet regression encourages group effect for correlated variables and it has no limitations on variable selection.

The objective function of ElasticNet is given by the following equation:

$$\min_{w} \frac{1}{2n_{samples}} \|X_w - Y\|_2^2 + \alpha\rho\|W\|_1 + \frac{\alpha(1-\rho)}{2} \|W\|_2^2 \qquad (4.1)$$

ElasticNet produces the best solution by combining L1 and L2 regularization methods. Alpha (α) is the tuning factor in both methods that controls the strength of the penalty.

- If α = 0, the objective becomes similar to simple linear regression, achieving the same coefficients as simple linear regression.
- If α = ∞, the coefficients will be zero because of infinite weightage on the square of coefficients. Anything less than zero makes the objective infinite.
- If 0 < α < ∞, the magnitude of α decides the weightage given to the different parts of the objective.

In addition to α parameter, ElasticNet has another parameter, L1_ratio (ρ), a measure of how 'mixed' the L1 and L2 regularizations should be. The mixing factor ρ, determines how much of L1 and L2 regularization should be considered in the loss function. The value range of ρ is [0, 1].

- If ρ = 1, the penalty would be L1 penalty.
- If ρ = 0, the penalty would be L2 penalty.
- If 0 < ρ < 1, the penalty would be the combination of L1 and L2 regularizations.

4.1.2. Random Forest Regression

Random Forest (RF) is one of the supervised ML algorithms which is effective in regression as well as classification tasks. The RF regression is an ensemble learning method with multiple decision trees and predicts the final output based on the average of each tree output. The combined decision trees are considered as base models, and it is represented formally as:

$$G(x) = f0(x) + f1(x) + f2(x) +.... \quad (4.2)$$

RF algorithm builds many decision trees based on random subsets of samples and features which then vote. The outcome of a vote by weak learners is less overfitted than training on all the dataset to generate a single strong learner.

RF has hyper-parameter inputs including, the number of trees, tree depth, and how many features and observations each tree should use. While building random forest model, the main parameters used are [19] :

- **max_features:** the size of features to consider when splitting a node. If we choose this parameter's value to 'None' then it will consider all the features rather than a random subset.
- **n_estimators:** the number of trees in the forest. The higher the number of trees, the better the result will be. But it will take longer processing time.
- **max_depth:** the maximum depth of the tree. If 'None', then nodes are expanded until all leaves are pure.
- **min_samples_split:** the minimum number of samples required to split an internal node.
- **min_samples_leaf:** the minimum number of samples required to be at a leaf node.

4.1.3. Extreme Gradient Boosting

Gradient boosting is an ensembling method that usually involves decision trees. Boosting is a sequential technique involving a set of weak learners and delivers improved performance. Extreme Gradient Boosting (XGBoost) uses the gradient boosting framework at its core which is an optimized library. Nowadays, XGBoost is popular ML algorithm regardless of the type of prediction task [20].

Gradient boosting algorithms have a very large number of hyper-parameters, and tuning is an important part of using them. The most common tuning parameters for tree-based learners in XGBoost are [20]:

- **learning_rate:** step size shrinkage used to prevent overfitting in the range [0,1].
- **max_depth:** determines how deeply each tree is allowed to grow.
- **subsample:** percentage of samples used per tree.
- **colsample_bytree:** percentage of features used per tree.
- **n_estimators:** number of trees required to build the model.
- **objective:** determines the loss function to be used, like reg:linear for regression problems.

4.2. Tuning Model Hyper-parameters

After an appropriate algorithm is identified, hyper-parameters tuning is done to obtain the best possible performance. The most common method to find the best combination of hyper-parameters is Grid Search Cross Validation (GridSearchCV). Its implementation process is as following [15]:

- Set the parameter grid for tuning; by creating a dictionary of all the parameters and their corresponding set of values that we want to test for best performance.
- Set the number of folds and the random state and a scoring method.

- Build a K-Fold object with the selected number of folds.
- Build a Grid Search Object with the selected model and fit it.

The GridSearchCV function returns a set of hyper-parameter values that fits best with the validation dataset. For the thesis work, 5 fold CV is used for the GridSearchCV implementation with 'neg_mean_squared_error' scoring method. Based on the working principle of the selected scoring method, the parameter set with the lowest mean_squared_error will be identified as the best parameter set.

4.3. Evaluation Metrics

The performance of a regression model is evaluated by the error rate of the predictions made. A good regression model has small difference between the actual and the predicted values and it is unbiased [21]. Here we will discuss two selected metrics that we will use for our model performance evaluation. These are:

- Root Mean Square Error
- K Fold Cross Validation

4.3.1. Root Mean Square Error

Root Mean Square Error is a common performance evaluation method for regression problems [22]. The root mean squared error is used to check the performance of the trained model on the test set. RMSE is measured by taking the square root of the average of the squared difference between the prediction and the actual value. It is calculated with the following formula [21]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \tag{4.3}$$

RMSE is the default evaluation metric of many algorithms as the loss function defined in terms of RMSE is smoothly differentiable and easier for mathematical operations. RMSE squares the errors before taking the averages as a result, large errors receive higher punishment. It performs particularly well when large errors are undesirable for model performance [21].

4.3.2. K Fold Cross Validation

Cross-validation is a statistical method used to compare and evaluate the performance of ML models. In K fold CV, the training set is randomly split into K (usually between 5 to 10) subsets known as folds. Where K-1 folds are used to train the model and the other fold is used to test the model [15]. For the thesis work, we are have used 10 fold cross validation to evaluate the model performance.

Chapter 5: Experimental Analysis

The ultimate goal of training a prediction model is that it can generalize well on unseen data. As a result, the model could predict accurate results from new data based on the internal parameters adjusted through training and validation. One important problem during model training is the tension between optimization and generalization [15].

- Optimization is the process of adjusting a model to get the best performance possible on training data.
- Generalization is how well the model performs on unseen/test data. The goal is to obtain the best generalization ability.

When training starts, optimization and generalization are correlated. After some iterations generalization stops to improve and the validation metrics freeze, and then start to degrade. In this case, the model is overfitting and there are two ways to avoid it; getting more data and regularization. Having more data is usually the best solution as a model trained on more data naturally generalizes better [15].

5.1. Implementation

Python, a general-purpose high-level programming language, has been used for implementing the selected ML algorithms. It is used throughout the ML community due to its many libraries that contain various predictive analytics algorithms. One of the most known libraries is Scikit-learn. It provides state-of-the-art implementations of many ML algorithms with simple user interface; therefore it is well suited for the thesis work.

For the selected three algorithms, base models have been trained first using default parameter values. Then parameter tuning is done for performance enhancement and the models have been trained with the best parameter sets. On the test dataset, RMSE has been used to evaluate the base models and measure performance changes due to the parameter tune. CV method has been employed for performance comparison and the solution model selection.

5.2. Data Splitting

The dataset has been split into two parts for the modeling process, training and test datasets. The training dataset contain 80% of the total dataset and the test dataset will contain 20% of the total data. For each selected algorithm, a base model has been trained on the training dataset and evaluated on the test dataset created.

The training dataset is further split for validation, 20% of the training dataset is used for parameter tuning purpose. The other portion is used to train the model with the best parameters identified.

5.3. Model Fitting

Model fitting involves running an algorithm on a dataset to build an ML model. The trained model's outcomes are compared with the actual values to determine the model performance. Then, algorithm parameters are adjusted to increase the model performance. In this part, we have the algorithms and dataset ready, and we proceed to train and validate our models.

5.3.1. ElasticNet Model

The main purpose of ElasticNet Regression is to find the coefficients that minimize the sum of error squares by applying a penalty to these coefficients. With ElasticNet regression, we set up the model on the train set as a base model with default parameter values.

To increase the model performance, parameter tuning is done using GridSearchCV method on the validation dataset. The below parameter value ranges have been used to find the best set.

Alpha = [0.01, 0.1, 1, 10, 100]

l1_ratio = [0.93, 0.95, 0.97, 0.99]

The GridSearchCV result has been plotted in Figure 5.1 to show the effect of parameter changes on the model performance.

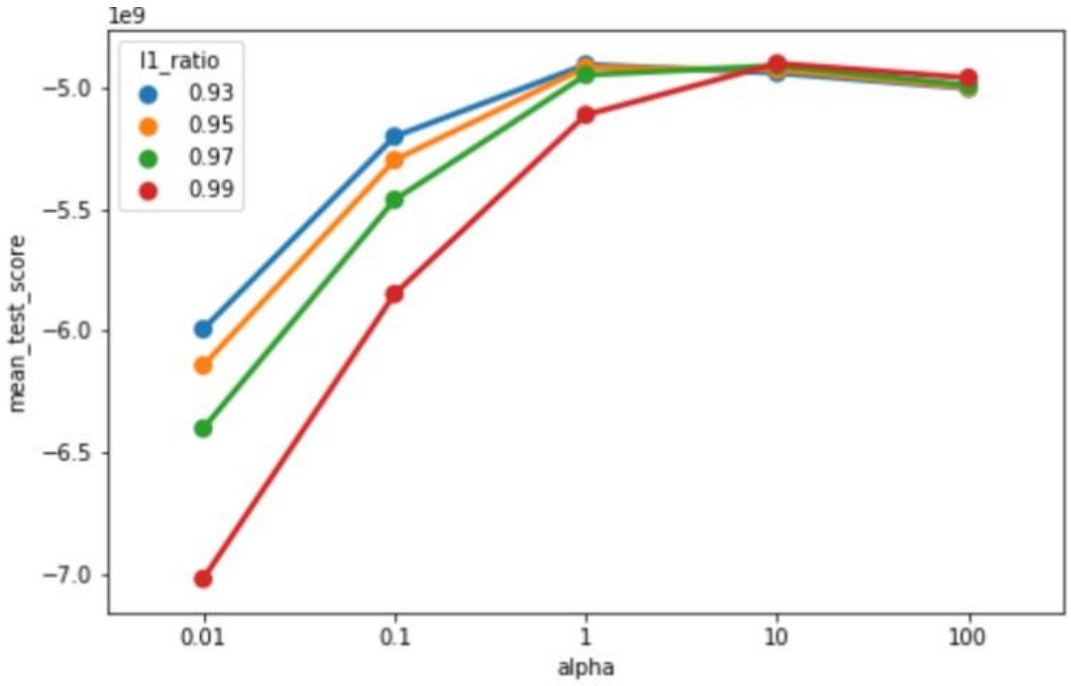

Figure 5.1: Grid Search Plot for ElasticNet Model

As shown in the above figure, the parameter value set (Alpha=10 and l1_ratio=0.99) has been identified to have the best score. Setting these parameter values the model is trained again and tested for performance evaluation. The model has shown a 3% performance increment through the parameter tuning process.

5.3.2. Random Forest Model

To implement RF Regression, we need RandomForestRegressor class from Scikit-Learn library. The based model has been trained and tested with default parameter values set. To improve the model performance we have applied parameter tuning by GridSearchCV method on the parameter value ranges listed in Table 5.1.

No.	Parameter	Value Range
1	max_depth	5,6,7
2	max_features	auto, sqrt, none
3	min_samples_leaf	3,4,5
4	min_samples_split	2,3,4
5	n_estimators	20,25,50

Table 5.1: Parameter Ranges for RF Model

To illustrate the effect of parameter value changes on the model performance, GridSearchCV result of two sample parameters (max_depth and n_estimators) is plotted in Figure 5.2.

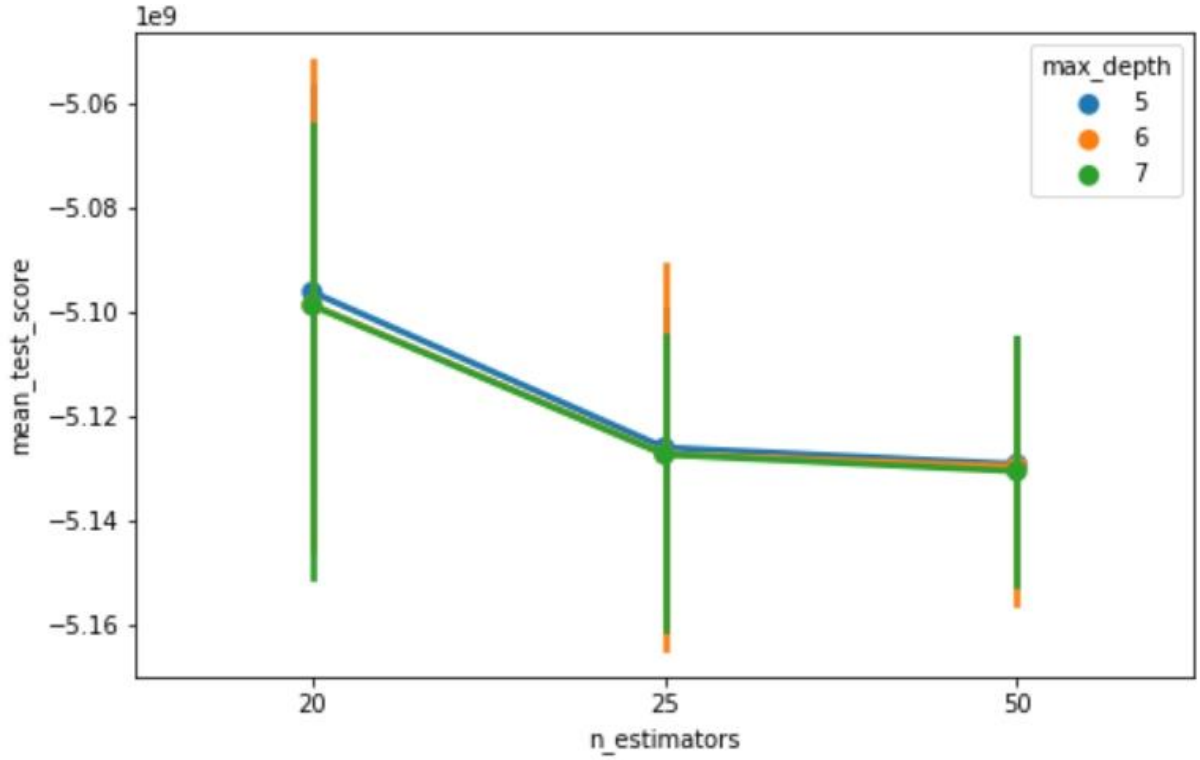

Figure 5.2: Grid Search Plot for RF Model

It is observed in Figure 5.2 that the model performance is affected by the parameter value changes and the best score is found at (6, 20). Generally, the following best parameter values result is found by the GridSearchCV tuning method on the validation dataset.

max_depth = 6

max_features = 'sqrt'

min_samples_leaf = 5

min_samples_split = 2

n_estimators = 20

Using the best parameter set the model is trained again and evaluated by the RMSE method. The model has shown better performance after the parameter tuning as the RMSE value has decreased by 26.74%.

5.3.3. Extreme Gradient Boosting Model

To build the base model, we have used the XGBoost library and import the XGBRegressor. The model is trained on the dataset with the default parameter values and tested. The base model performance of XGBoost algorithm is the least compared to the other algorithms used in the thesis work.

Similar to the other models, GridSearchCV is used for parameter tuning on the parameter value ranges listed in Table 5.2.

No.	Parameter	Value Range
1	max_depth	1,2,3
2	colsample_bytree	0.4,0.5,0.6
3	learning_rate	0.1,0.15,0.2
4	subsample	0.001,0.01,0.1
5	n_estimators	20,25,50

Table 5.2: Parameter Ranges for XGBoost Model

The GridSearchCV result of two sample parameters (learning_rate and n_estimators) is plotted in Figure 5.3. For the sample parameters, the model has better performance at the pair value of (0.15, 20).

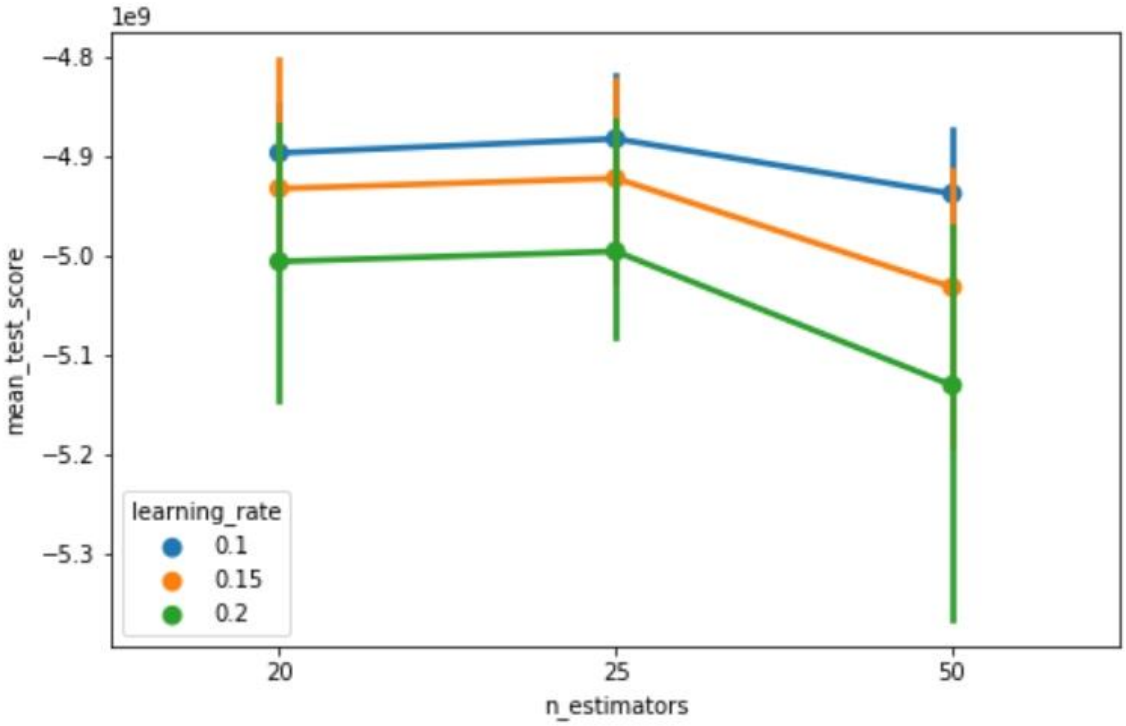

Figure 5.3: Grid Search Plot for XGBoost Model

After the GridSearchCV is applied on the validation dataset, the best parameter set found is:

colsample_bytree = 0.5

learning_rate = 0.15

max_depth = 1

n_estimators = 20

subsample = 0.01

Using the above values, the model is trained and has achieved high performance improvement. The RMSE result of the model has decreased 62.75% and it has the best result compared with the other models based on the RMSE method after parameter tuning is done.

Chapter 6: Model Evaluation and Interpretation

Three prediction models have been built using different methods as discussed in Chapter Five, here the models performance are evaluated for best model selection. The performance results have been interpreted for better understanding using visualization techniques. In addition feature importance and other related results of the best model have been discussed. Finally the selected model has been tested with new packages which are launched after the dataset formation. The prediction result is compared with the customer size forecast given by Marketing Division and the actual purchase report.

6.1. Model Evaluation

For model evaluation, we have employed RMSE and CV methods. Each model has been evaluated and the performance is compared to determine which model is most effective. The performance result of each model is presented in Table 6.1 and Table 6.2 below for RMSE and CV methods respectively. RMSE evaluation is conducted on the test dataset after each model training step. On the other hand, the CV evaluation is applied on the whole dataset using 10 folds for model training and testing with the best parameters.

Model	**ElasticNet**	**Random Forest**	**XGBoost**
Base	14614.126	20425.682	33022.152
Tuned	14162.212	14963.748	12300.827

Table 6.1: RMSE Evaluation Results

It is observed form the RMSE result of the three models in Table 6.1 that the models performance has increased through parameter tuning. Even though the RMSE values for the tuned model is still large, it is acceptable considering the business domain. The amount of mobile packages purchase is from single digits to hundreds and thousands per day and it is highly dynamic market. Hence, customer size prediction errors in the range of ten thousands for new packages is a good start for further improvements.

Model	Best Score	Mean	Standard Deviation
ElasticNet	54.114	173.731	76.459
Random Forest	30.174	147.187	85.52
XGboost	66.831	168.54	76.428

Table 6.2: CV Evaluation Results

For the CV evaluation method, the scoring metric used is 'neg_root_mean_squared_error'. As a result, smaller score values indicate good performance of the evaluated model. The evaluation result has three output values for each model to be used as a comparison criteria for best model selection. These are:

- **Best Score:** the smallest RMSE from the 10 fold scores of a model
- **Mean:** the average RMSE of all fold scores, this value can represent the real performance of a model
- **Standard Deviation:** is a measure of the amount of variation in the score values of each fold

Based on the above evaluation results, we will handle the model performance comparison in the next sections and further analysis is given on the selected model.

6.2. Performance Comparison

We have used RMSE and CV methods for performance evaluation of the trained models. For the best model selection, we have compared the models base on the CV performance results. The CV method employs all the samples as a training and testing inputs for the model training. CV evaluation result is more general and represents model's performance in real scenarios. The CV performance comparison of our models is shown below in Figure 6.1.

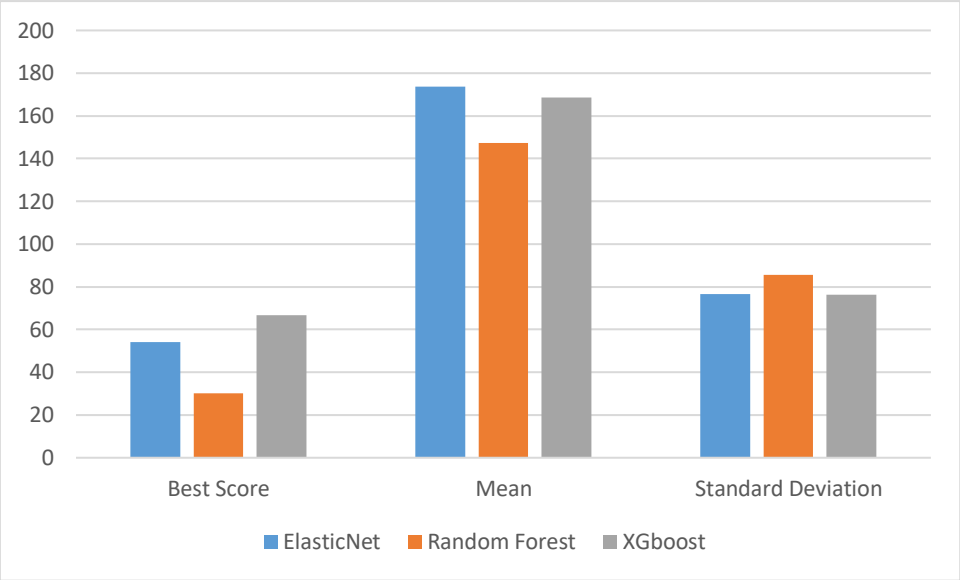

Figure 6.1: CV Performance Result

As of Figure 6.1, the RF model has best performance. As the scoring metric used in the CV method is 'neg_root_mean_squared_error', the best model will have lower value results. The RF model has better results for the best score and mean values and its standard deviation is nearly equal with the other models. In general the RF model has better performance and it is selected as a solution model for the customer size prediction purpose.

6.3. Result Analysis

In this section, result analysis is done for the selected solution model and some model features have been discussed. For the feature importance analysis two methods are used; RF feature importance and permutation feature importance. The merits and demerits of each method with related definitions is discussed in detail. Moreover, to visualize the RF model sample decision trees is presented.

As presented in the model evaluation part, the performance results for the RF model based on the CV method are:

Best Score: 30.174

Mean: 147.187

Standard deviation: 85.52

The least error (Best score) the model achieved is a promising result whereas the mean and standard deviation values are relatively higher. Considering the mobile package daily purchase rate's high value range distribution, the result is satisfactory. Mobile package purchase rate varies from single digit to thousands in a day for different packages. The mean error of the new RF model is 1.3% of the average daily mobile package purchase rate (11,446.664).

6.3.1. Feature Importance

The feature importance describes which features are more relevant to the solution model. It can help for better understanding of the solved problem and sometimes lead to model improvements by employing the feature selection. In our case, the feature importance is presented here for the purpose of the solution model understanding only. Feature selection is not considered here as the number of features are limited and all the features in the dataset are assumed to be helpful for the solution model training.

We have used built-in method in the RF algorithm to compute feature importance from scikit-learn package in Python. The feature importance chart of our RF model is shown below in Figure 6.2.

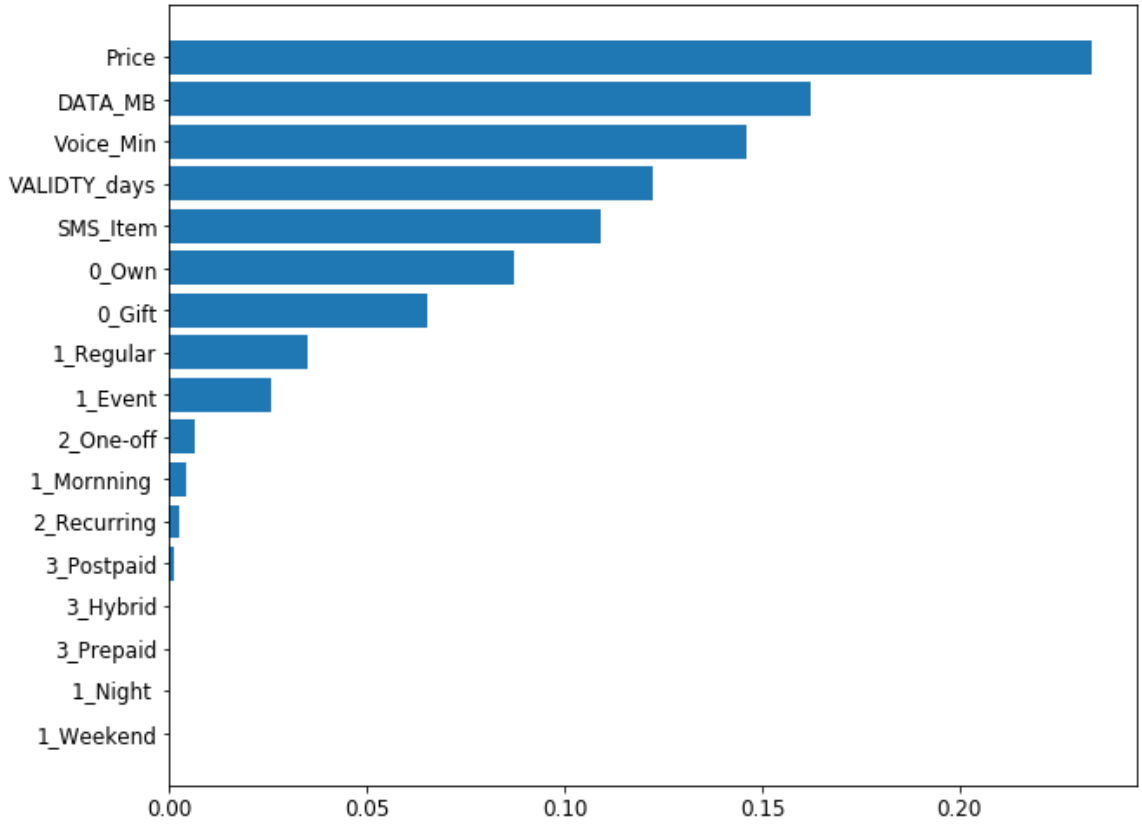

Figure 6.2: RF Model Feature Importance

As of the feature chart above, Price is the most relevant one for the target prediction and the other numerical features have significant importance. In general the numerical features are more relevant to the solution model compared to the categorical features. Out of the categorical features, 'Package_ownership (Own/Gift)' has higher importance in the solution model.

However, there are some limitations of the random forest feature importance function. It ranks the numerical features to be the most important features and the categorical values are presented as normal features. This problem stems from the following limitations [19]:

- Categorical values are represented as separate features due to one-hot encoding
- RF feature importance is biased towards high cardinal features
- RF feature importance is computed on training set statistics and lacks the ability of generalization to the test set.

6.3.2. Permutation Feature Importance

Permutation feature importance is defined as the decrease in a model performance when a single feature value is randomly shuffled. This procedure evaluates the relationship between the feature and the target. The drop in the model score indicates how the model depends on the feature [19]. The permutation_importance function of scikit-learn calculates the feature importance of a model for the given dataset.

For the trained RF model permutation feature importance is shown in Figure 6.3. The box plot shows the feature importance values across the iterations of the algorithm. The x-axis shows the impact that permuting a given feature had on the model's prediction score. The y-axis shows the input features in the relative importance order. The top being the most important, and the bottom being the least important. The minimum, first quartile, median, third quartile, and a maximum of the feature importance values across different iterations of the algorithm are shown by each box. Moreover, outliers are shown in small circles.

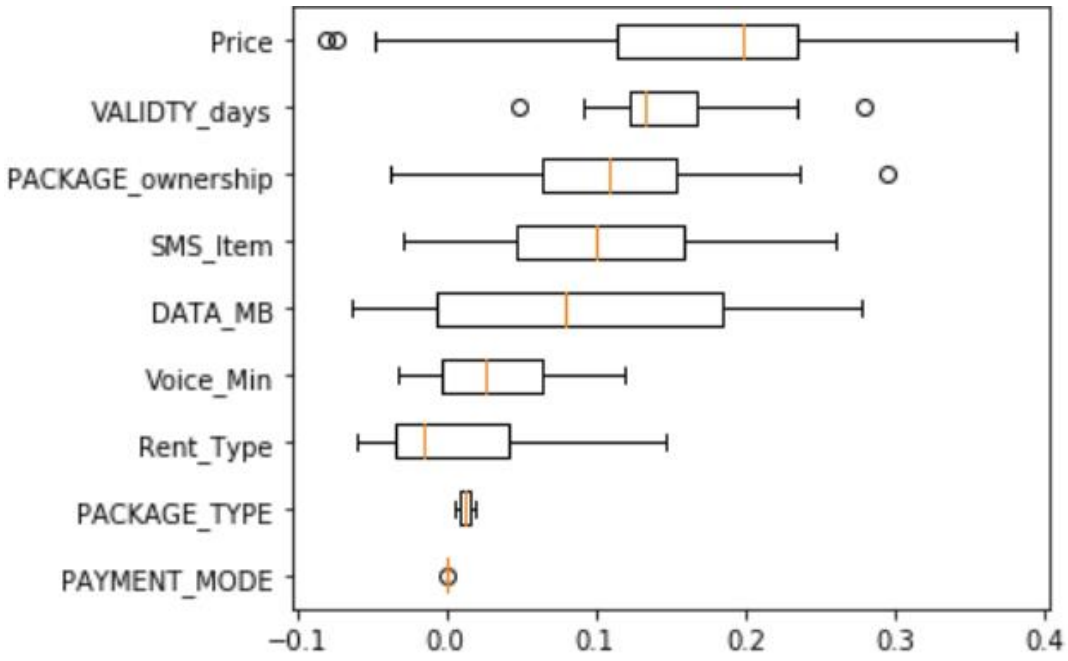

Figure 6.3: Permutation Feature Importance

From the above figure, we can observe that the permutation feature importance of the categorical features is set at feature level, not for individual values. Price is still the most

important feature and from the categorical feature 'Package_ownership' has better importance compared with the RF importance results. Based on the importance results of both methods, we can conclude that the numerical features highly affect the model performance rather than the categorical features.

6.3.3. Decision Tree Visualization

As the RF model is assemble of Decision Trees (DT), it is possible to visualize each DT from the Random Forest. To make visualization readable it is recommended to limit the depth of the tree. For this purpose the RF model is trained with max_depth=3. The first DT for the RF model built (estimators [0]) is shown in Figure 6.4.

We have used dtreeviz python package to visualize the first random forest DT. It renders better looking and intuitive visualizations while offering greater interpretability options. The dtreeviz helps to visualize how features split up at decision nodes, how samples get distributed in leaf nodes and how the tree makes predictions [23]

In general, dtreeviz is useful in understanding how classification or regression decision trees work. For better understanding of dtreeviz figures, the following points are important [23]:

- The horizontal dashed lines indicate the target mean for the left and right buckets in decision nodes;
- A vertical dashed line indicates the split point in feature space.
- The black wedge highlights the split point and identifies the exact split value.
- Leaf nodes indicate the target prediction (mean) with a dashed line.

We can observe from the figure that subset of the features are used for the first DT formation. As there are more similar trees in the RF model which employ the other features, the final prediction result will be the average output of all the available DTs.

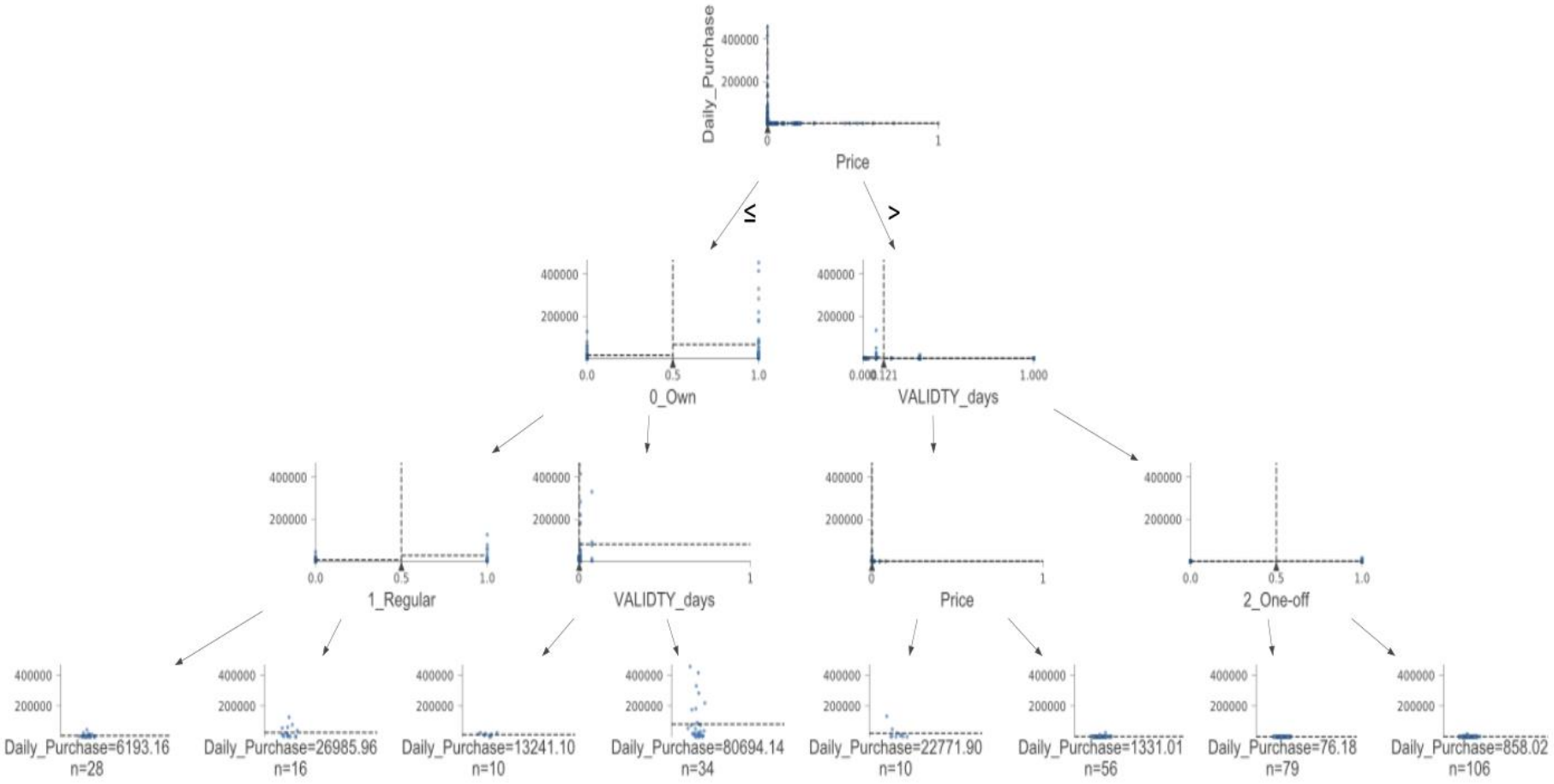

Figure 6.4: First Decision Tree

6.4. Real-time Test Scenario

To illustrate the model deployment part, we have tested the RF model using recently developed packages. The model prediction result is compared with the actual and target customer size. For this case we have used 18 packages which were launched for the Easter and Ed Al-fitr holidays (2021). In Table 6.3 the aggregated customer size based on service type is shown for the target, predicted and actual results.

Daily_purchase	Target	Prediction	Actual
Voice	138362.6429	162045.1067	206806.9107
Data	70579.71429	50699.90106	50948.71429
Bundle (V+D)	37839.92857	69255.516	13030.375

Table 6.3: Customer Size for Test Scenario

For the purpose of visualization, the prediction result of the RF model is depicted with the target and actual customer size in Figure 6.5.

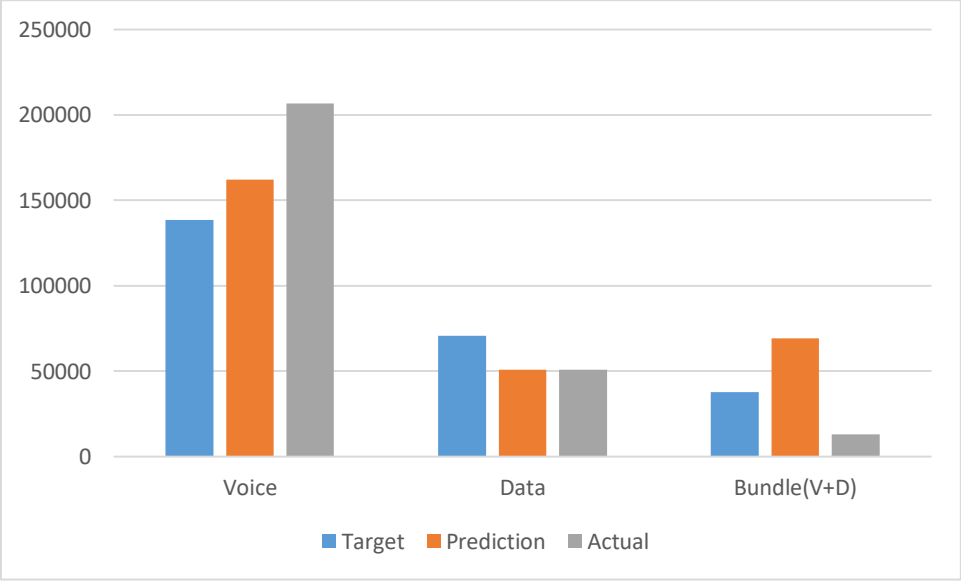

Figure 6.5: RF Predictions Compared to Target and Actual values

As of Figure 6.5, the solution model has better prediction results than the forecasted target for voice and data only packages but for the bundle packages the result is not satisfactory. It is because the purchase rate of bundle packages has dropped sharply for the holiday packages compared to the normal trend. Moreover, most mobile subscribers are either voice or data intensive users and the number of sample bundle packages in the dataset used is limited.

In general, the RF model can help to anticipate the customer size for new mobile packages and file the gap between the target and actual values. Usually the target value is either underestimated or overestimated. Hence, this model has predicted better customer size than the existing manual methods used in ethio telecom for the voice and data packages. The model will help a lot to foresee the customer size for decision making and other related activities in the company.

Chapter 7: Conclusion and Future Works

7.1. Conclusions

In the thesis work, we have identified the existing customer size forecast problems in ethio telecom for mobile packages. These have a negative impact on the package development process and after launch correction measures. To improve the package development process, ML approach is used for customer size prediction. Mobile package dataset has been formed by integrating available data from different sources. Most important mobile package attributes and purchase reports are included in the dataset.

In this study, three ML regression algorithms, ElasticNet, XGBoost and RF have been used to train possible solution models. The RF model has better performance and it is selected to be the solution model. This ML model will improve the existing customer size forecasting method in ethio telecom. Furthermore, there will be fast decision making as the post launch analysis result could be replaced by the model prediction.

As of the real-time test scenario results, the ML model can be easily employed to predict customer size for new packages under development process. This will support the marketing team to evaluate its packages before launch and suggest reasonable recommendations for the higher management. The model's limitation on bundle packages prediction can be improved by including related samples on the dataset to fully utilize the model for any mobile package customer size predictions.

7.2. Future Works

This thesis work can be a starting point for the following recommended future works:

- Expand the mobile package dataset to include all possible features and make it ready for further mobile package related studies.
- Improve the model performance using different techniques and more advanced algorithms such as Neural Networks.
- Use similar ML approach to predict customer size for all packages and services in ethio telecom as the scope of this thesis is mobile packages.

References

[1] J. Xin-kuang and C. Xu, "Research on Prediction Model of the Impact of New Telecom Services Tariff Based on the Customer Choice Behavior," *Advanced Materials Research,* Vols. 765-767, pp. 3249-3252, 2013.

[2] J. Danhua, Z. Xiaogeng and W. Runrun, "Reaserch on the Amount of Customers in Telecom Package Preview Based on Data Mining," in *International Conference on Computer Science and Service System*, 2012.

[3] M. V. Joseph, "Data Mining and Business Intelligence Applications in Telecommunication Industry," *International Journal of Engineering and Advanced Technology (IJEAT),* vol. 2, no. 3, pp. 525-528, 2013.

[4] H. H. Darji, "Data Mining in Telecommunication Industry," *IJSRD - International Journal for Scientific Research & Development,* vol. 2, no. 8, pp. 7-9, 2014.

[5] "Ethio telecom," [Online]. Available: https://www.ethiotelecom.et. [Accessed 18 August 2021].

[6] B. Demelash, *Usage Based Clustering of Customers for Mobile Service Packaging,* Addis Ababa, 2019.

[7] Y. Miao, J. Tang and G. Qu, "Behavior Prediction of Telecom Consumers' Choice with Packages Based on Improved Nested Logit Model," in *25th Chinese Control and Decision Conference (CCDC)*, Guiyang, 2013.

[8] M. Tajudin, *Customers Segmentation for Profitability Enhancement Using Data Mining Technique: The case of ethio telecom,* Addis Ababa, 2020.

[9] Y. Bishaw, *Market Segmentation of Mobile Internet Customers Using Clustering Algorithms: The Case of Ethio Telecom,* Addis Ababa, 2020.

[10] D. B. Jun, S. K. Kim, Y. S. Park, M. H. Park and A. R. Wilson, "Forecasting Telecommunication Service Subscribers in Substitutive and Competitive Environments," *International Journal of Forecasting,* vol. 18, pp. 561-581, 2002.

[11] O. Wisesa, A. Adriansyah and O. I. Kahlaf, "Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm," in *2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*, Yogyakarta, Indonesia, 2020.

[12] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," in *International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, Shouthend, UK, 2018.

[13] Ethiotelecom, *Product & Service Marketing Catalog, V 1.6,* 2020.

[14] Ethiotelecom, *CBS Basic Feature-Function Requirement Specification,* 2014.

[15] V. Roman, "How To Develop a Machine Learning Model From Scratch," 23 December 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af. [Accessed 25 May 2021].

[16] J. Brownlee, "Why One-Hot Encode Data in Machine Learning?," 30 June 2020. [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. [Accessed 12 August 2021].

[17] A. Burkov, The Hundred-Page Machine Learning Book, 2019.

[18] C.-S. Wu, P. Patil and S. Gunaseelan, "Comparison ofDifferent Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," in *9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018.

[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research (JMLR),* vol. 12, pp. 2825-2830,, 2011.

[20] M. Pathak, "Using XGBoost in Python," 8 November 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/xgboost-in-python. [Accessed 21 April 2021].

[21] H. Parvez, "4 Best Metrics for Evaluating Regression Model Performance | Machine Learning," [Online]. Available: https://www.aionlinecourse.com/tutorial/machine-learning/evaluating-regression-models-performance. [Accessed 2 June 2021].

[22] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, USA: O'Reilly, 2017.

[23] P. Pandey, "towards data science," 18 May 2021. [Online]. Available: https://towardsdatascience.com/a-better-way-to-visualize-decision-trees-with-the-dtreeviz-library-758994cdf05e. [Accessed 15 Augst 2021].

Appendix A: Conference Paper (8th IEEE CSDE 2021)

Customer Size Prediction using Machine Learning Approach for Mobile Package

Desalegn Medhin Firdu
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
desalegn.mf@gmail.com

Rosa Tsegaye Aga
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
rosatsegaye@gmail.com

***Abstract*— Nowadays the telecom market is competitive and telecom operators launch various new service packages to meet customer needs and attract more customers as well. Ethio telecom is the only telecommunications service provider in Ethiopia. In the case of ethio telecom, as there is no an automated method for package preview, Machine Learning (ML) approach has been studied to predict customer size for new mobile packages. Three ML algorithms that are, ElasticNet regression, Extreme Gradient Boosting and Random Forest regression (RF) have been used to train the prediction models. To train the model, mobile package dataset has been constructed by integrating data from three different sources in ethio telecom. The sources are business support systems, marketing product catalog and mobile package post launch analysis results. As the study has showed, the RF model has outperformed the ElasticNet regression and Extreme Gradient Boosting models.**

Keywords—Machine Learning, Mobile Package, Customer Size, ethio telecom

INTRODUCTION

With the globalization of the telecom market and diversification of users, competitions between telecom operators become fierce increasingly. As a result, telecom operators introduce new service packages to seize market opportunities, attract more new customers and increase business revenue [1]. New package and tariff previews are important to insure business continuity for telecom operators and some mechanisms have to be employed for this purpose.

The objective of this study is to build ML model that predicts customer size for new package and improve the mobile package development process in ethio telecom. Ethio telecom is one of the oldest telecom service providers in East Africa Ethiopia which is established in 1894. Currently, the company has above 54.3 million mobile customers. It has been offering various mobile service packages with different free resource options for voice, Short Message Service (SMS) and Internet services. These service packages are provided at discount price with a fixed validity period. In ethio telecom, the goal of service packaging is to influence customer service usage and increase the company revenue.

The vast volume of telecommunication data can be effectively used to improve telecom business. Data Mining can be utilized to automatically generate knowledge from the available data. Data mining and Business Intelligence applications play a significant role in the telecom industry to overcome the hard competition in the sector [2]-[3]. The available customer data can be used to profile customers for marketing and forecasting purposes. Hence, data mining is the most relevant solution to improve business and operations in telecom companies.

Ethio telecom uses manual methods to design mobile packages and the market performance is evaluated after the packages are released through post launch analysis. In this study, three regression ML models that predict the customer size of the new mobile packages have built. The models have been trained using a dataset that has created by integrating existing mobile package information and purchase report data from the Information Systems (IS) in the Telecom. The models have been evaluated using appropriate evaluation techniques. In addition, the outperformed model has been validated on the real-time scenario using some new mobile packages shortly introduced in ethio telecom.

This study contribute to ethio telecom a better marketing plan and helps to adjust itself for the upcoming competitive market in the country. Moreover, the mobile package dataset that has been constructed for this study will be available for further studies in this area. The prediction results of the selected ML model can help to set an appropriate customer

size target for new packages. The model will be used as a mobile package preview tool in the company and helps to predict the existing post launch analysis results before package release. In addition the model can be used to customize different packages under designing process and produce optimal mobile packages.

The research paper has been organized in seven sections. In Section II, background of the research has been explained and Section III the related works. Section IV presents the details on the dataset formation, data description and preprocessing of the dataset. Section V focuses on the ML algorithms and methods that have applied in the research work model training and evaluation. Section VI presents the result analysis and discussion part. Finally, section VII conclusions of the research work.

RESEARCH BACKGROUND

Telecom network operators combine mobile service usage price plan, discount price plan and free unit price plan or some of them together to define a package. Mobile operators have introduced several innovative price plans to attract and retain their customers [4]. Service packaging enables subscribers to enjoy preferential usage charging, discount charging or free unit by paying a certain rental fee. A typical mobile package charges the rental fee, giving free units and/or a favorable tariff or discounts on certain services.

Ethio telecom provides one-time and periodical/recurring mobile packages. All mobile services, voice, SMS and data basic tariffs are normally contained in primary offering plans. These can be redefined to promoted tariffs in package to form new offering. The new tariffs can be combined according to time schema, service type, customer level, and other conditions.

Ethio telecom has developed different package options for a specific time schema (night and morning) and service type (voice, SMS, data or bundle). Package development is regularly done in ethio telecom to activate customer consumption. Usually, based on marketing factors and tariff revisions, either existing packages are modified or new packages are released. In addition some promotional or event driven packages are becoming familiar in the company. Holiday packages are worth mentioning in this regard. Nowadays, for every public holiday, ethio telecom releases new brand mobile packages.

Post launch analysis has been conducted for every new package released. The analysis is done based on package purchase report and customer feedback on social media outlets. The analysis result is compared with the customer size forecast and revenue target set. Finally recommendations and remarks are given for further decisions and corrective actions. The analysis task may take several days after the package launch depending on the package type. Hence, there is a time delay to have the analysis result for further decisions. The new ML model will reduce the work load in mobile package development and post launch activities in the company. Moreover, reasonable customer size is predicted before package launch and informed market decisions can be done on time.

RELATED WORKS

Jiang and Chen in [5] has assessed the impact of new telecom services tariff on customers and the company revenue. Impact indicators like utility of service packages, transfer probability of the customers and expected change of revenue have been obtained. These are useful for market orientation, revenue prediction and optimization management of the new telecom services tariff. The results are based on customer behavior analysis which cannot be addressed through data mining methods.

Danhua, Xiaogeng and Runrun in [1] have used statistics and data mining methods for the prediction of the number of new customers and transfer customers in telecom package preview. The study has proposed the key point of telecom tariff preview to calculate the possible users of the new package. The possible customers are further divided into the new customers and the transferred customers. The focus of the study is to predict the number of new customers for a given package based on monthly subscription trends. The transfer rules are also defined based on the customer service usage history. The study has not consider the attributes of the package for the analysis, which is the main focus of our research work.

In [6], the study has presented a brief analysis of the reliability of machine learning techniques for telecom Business to Business (B2B) sales prediction. The research has concluded that Gradient Boost Algorithm has better accuracy in B2B sales prediction for telecommunication companies. Detailed analysis is given for the Gradient Boost Algorithm, but no other ML algorithm is compared in the paper.

In [7], detailed study and analysis of comprehensible predictive models that improve future sales predictions has been carried out. On the basis of performance evaluation results, the Gradient Boost Algorithm has been suggested as a better suited predictive model for the sales trend forecast. The prediction is done based on historical sales data of each item in similar to the mobile package purchase history that has used in this study. In our case, we additionally have introduced the mobile package attributes in predicting the daily package purchase rate, which is equivalent to the number of customers subscribed for the new package.

DATA ANALYSIS

A. Data Collection and Integration

In this step all related data in ethio telecom has been collected from the responsible sections. The main data sources that have used for the mobile package dataset construction are the IS business support systems, marketing product catalog and mobile package post launch analysis results. As most of the packages are active, their attributes and purchase reports have been collected from the operational systems. For some packages that are out of market their data has been collected from post launch analysis reports. The collected data have been integrated based on Offer ID and missed features have feed manually from the marketing product catalog.

B. Data Description

The data has collected from three different sources that are business support systems, marketing product catalog and post launch analysis reports. The business support systems contain all mobile package attributes and related purchase reports. The marketing product catalog is a reference manual for any package development and has been used as a reference to construct a complete dataset. The post launch analysis reports also have similar information and they have used as a source for old packages.

The newly formed dataset has (339x11) size and every possible attributes that can help to characterize a given mobile package. A new feature 'package type' is added to categorize similar packages and label them based on the objective and time schema of the packages. All attributes in the dataset are described in Table I.

PACKAGE ATTRIBUTES DESCRIPTION

Attributes	Description	Data Type
Offere_ID	Package unique ID / used as index	N/A
Price	Rent amount in birr	Numerical
Voice_Min	Package voice free resource	Numerical
SMS_Item	Package SMS free resource	Numerical
DATA_MB	Package data free resource	Numerical
Validity_days	Package usage period	Numerical
Payment_Mode	Prepaid/Postpaid/Hybrid	Categorical
Package_Ownershp	Self/Gift	Categorical
Package_Type	Regular/Morning/Night/Weekend /Event	Categorical
Rent_Type	One-time/ Recurring	Categorical
Daily_Purchase	Two months average purchase/ target feature	Numerical

The correlation analysis of the numerical features shows that the target feature (Daily_Purchase) has a non-linear correlation with the other features. In contrary, the input features are correlated to each other with different degrees. Based on the correlation results of the input features, appropriate ML algorithms are selected.

C. Data Preparation

In this step, the data has prepared to be fitted in the machine algorithms for model training. The data cleaning has been conducted on the collected samples and packages with the appropriate feature values have been selected. Some packages with similar feature values have been aggregated and their purchase combined. For example, similar data packages that have prepared for 3G, 4G and data only users have been summarized together.

As the dataset contains numerical features with different value ranges and categorical features with nominal labels, normalization and encoding techniques have been applied to prepare the data. Normalization is re-scaling features to a specific range, which is convenient for the purpose at hand. To normalize our data we have applied the min-max scaling method to each numerical feature column and all the values have been scaled in the range between 0 and 1.

To make all the input values numeric, One-Hot encoding has been applied to the categorical features in the dataset. For categorical variables with no ordinal relationship the One- Hot encoding is an appropriate transformation method [8].

For the purpose of feature selection, Offer name and Package revenue have been excluded from the collected mobile package features as they are not relevant to the customer size prediction. Moreover, OFFER_ID has not considered as an input feature. It is only used for indexing before the model training. The rest of the features, Daily_Purchase of a package that is the target feature and the other numerical and categorical features have been used as input features to train the model.

METHOD

A. The Machine Algorithms

Machine Learning is defined as the process of solving problems by collecting data, and algorithmically building a model based on a dataset [9]. Then the model is used to solve practical problem. ML can be classified in to supervised, semi-supervised, unsupervised and reinforcement learning types [9]. Based on our input data type and objective of the study, we have used supervised learning approach. Supervised learning is further classified in to regression and classification techniques. For this study, regression technique has been considered. Three algorithms have been employed to build alternative models for customer size prediction. These are:

- ElasticNet Regression
- Random Forest Regression
- Extreme Gradient Boosting

ElasticNet Regression is a combination of L1 and L2 regularization techniques. It is used when there are multiple correlated features [10]. In our case, the numerical features are correlated. This algorithm has suggested for the study. L1 regularization weights errors at their absolute value and results in models with fewer coefficients. On the other hand L2 regularization weights errors at their square and reduce model complexity. ElasticNet produces the best solution by combining the two regularization methods. It encourages group effect for correlated variables and has no limitations on variable selection; But it may suffer double shrinkage due to the L1 and L2 effects [10].

Random Forest (RF) is one of the supervised ML algorithms which is effective in regression as well as classification tasks. In this study, this algorithm has selected because it has been used highly in the state of art. The RF regression is an ensemble learning method with multiple decision trees and predicts the final output based on the average of each tree output [11]. It builds many decision trees based on random subsets of samples and features which then vote. The outcome of a vote by weak learners is less overfitted than training on all the dataset to generate a single strong learner. RF has hyper-parameter inputs including, the number of trees, tree depth, and how many features and observations that each tree should use [11].

Gradient boosting is an ensembling method that usually involves decision trees. Boosting is a sequential technique involving a set of weak learners and delivers improved performance [6]. From this ML algorithm family, Extreme Gradient Boosting (XGBoost) is now popular for prediction tasks and we have selected it for this research work. The most common parameters for tree-based learners in XGBoost includes learning_rate, max_depth, subsample, and n_estimators.

For the identified algorithms hyper-parameters tuning has been done to obtain the best possible performance. The most common way to find the best combination of hyper-parameters is Grid Search Cross Validation (GridSearchCV) that has used in this study. The GridSearchCV function returns a set of hyper-parameter values that fits best with the validation dataset [11]. For this study, 5 fold CV has been used for the GridSearchCV implementation with the scoring Metric of 'neg_mean_squared_error'. Based on the working principle of the selected scoring method, the parameter set with the lowest mean_squared_error result has been identified as the best parameter set.

B. Model Training

The ultimate goal of training a prediction model is that it can generalize well on unseen data. As a result, the model could predict accurate results from new data based on the internal parameters adjusted through training and validation. Python, a general-purpose high-level programming language, has been used for implementing the selected ML algorithm with the most known python libraries Scikit-learn.

The dataset has been splitted into two parts for the modeling process, training and test datasets. The training dataset contains 80% and the test dataset 20% of the total data. The training dataset has further splitted for validation and 20% of the training dataset has been used for parameter tuning purpose.

For the selected three algorithms, base models have been trained first using the default parameter values. The trained model outcome has been compared with the actual values to determine the model performance. Then, the algorithm parameters values have been adjusted to increase the model performance using the GridSearchCV function. All models have a significant performance improvements due to the parameter tune as of the evaluation results discussed below.

C. Model Evaluation

The performance of a regression model is evaluated by the error rate of the predictions that has made. A good regression model has small difference between the actual and the predicted values and it is unbiased. For this study, two evaluation metrics have been selected; Root Mean Square Error (RMSE) and K Fold Cross Validation (CV). Each model has been evaluated and the performances have been compared to determine the most effective solution model.

RMSE is the default evaluation metric of many algorithms as the loss function defined in terms of RMSE that is smoothly differentiable and easier for mathematical operations. RMSE squares the errors before taking the averages as a result, large errors receive higher punishment. RMSE has been used to evaluate the base models and measure performance changes due to the parameter tune. As a result, the ElasticNet model has improved its performance by 3% and the RF and XGBoost models have improved by 26.74% and 62.75% respectively.

The CV method employs all the samples as a training and testing inputs for the model training. CV evaluation result is more general and represents model's performance in real scenarios. 10 fold CV has been employed for the final model performance comparison and the best model selection. The CV performance comparison result of our models is shown in Fig. 1.

The CV evaluation result has three output values for each model to be used as a comparison criteria for the best model selection. These are:

- **Best Score:** the smallest error value from the 10 fold scores of a model.
- **Mean:** the average value of all fold scores, this value can represent the real performance of a model.
- **Standard Deviation:** is a measure of the amount of variation in the score values of each fold.

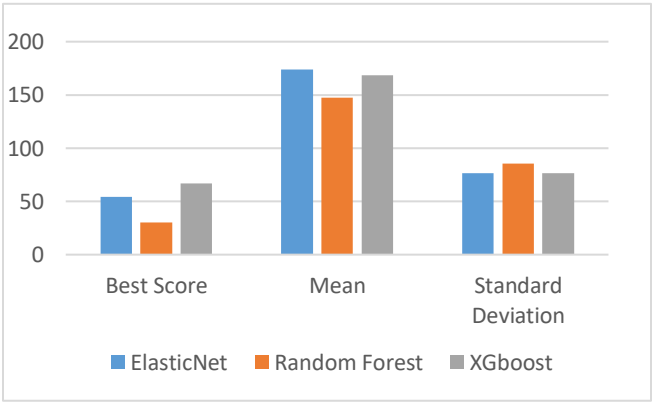

Figure 1: CV Performance Result

RESULT AND DISCUSSION

Based on the evaluation result of the CV method, the RF model has outperformed. As the scoring metric used in the CV method is 'neg_root_mean_squared_error', the best model has lower error value results. The RF model has better results for the best score and mean values and its standard deviation is nearly equal with the other models. In general the RF model has better performance and it is selected as a solution model for the customer size prediction purpose.

The performance results for the RF model are; Best Score: 30.174, Mean: 147.187 and Standard deviation: 85.52. The least error (Best score) that the model has achieved is a promising result whereas the mean and standard deviation values are relatively higher. Considering the mobile package purchase rate's high value range distribution, the result is satisfactory. The mean error of the new RF model is 1.3% of the average daily mobile package purchase rate. Now, having

the solution model we focus on the interpretation and analysis of the selected model.

Using built-in method in the RF algorithm that computes feature importance from scikit-learn package in Python, feature importance of the RF model is analyzed. The feature importance describes which features are more relevant to the solution model and helps in better understanding of the solved problem. As a result, 'Price' is the most relevant feature for the target prediction and the other numerical and categorical features have contributed as well. In general the numerical features are more relevant but some of the categorical features have significant importance values. Out of the categorical features, 'Package_ownership (Own/Gift)' has higher importance in the solution model.

To illustrate the solution model deployment part, we have tested the RF model using recently developed mobile packages as a real-time scenario. The model prediction result is compared with the actual and target customer size values. For this case we have used 18 packages which have been launched for two public holidays after our dataset has formed. The packages are aggregated by service type (Voice, Data and Bundle, a combination of both services) for the analysis purpose. The average daily purchase rate has been used for comparison based on the actual reports.

For the purpose of visualization, the prediction result of the solution model is depicted with the target and actual customer size in Fig. 2. As the figure shows, the model has good performance for voice and data individual packages but for the voice and data bundle packages the result is not satisfactory.

By the model, voice packages, 78.4% of the actual value has been predicted and data packages 99.5% has achieved. On the other hand, the predicted customer size for bundle packages is about 5 times of the actual value. This is because of the purchase rate of the bundle packages has dropped sharply for the holiday packages compared with the normal trend. Moreover, most mobile subscribers are either voice or data intensive users; As a result, the number of bundle package users is lower than the expected number.

In general, the RF model can help to anticipate the customer size for new mobile packages and fill the gap between the target and actual values. Usually the target value is either underestimated or overestimated. Hence, this model has predicted better customer size than the existing methods that has been used in ethio telecom. The model helps a lot to foresee the customer size for decision making and other related activities in the company.

CONCLUSIONS

In the study, we have identified the existing customer size forecast problems in ethio telecom for mobile packages. Then, to improve the package development process, ML approach has been studied for customer size prediction. Mobile package dataset has been constructed from the available data sources and integrated. Most important mobile package attributes and purchase reports have included in the dataset.

Three ML regression algorithms have been used to train the possible solution models. The RF model has outperformed and selected to be the solution model. This model improves the existing customer size forecasting method in ethio telecom. Furthermore, it help in fast decision makings as the delayed post launch analysis results could be replaced by the model results.

The trained model can help the telecom operators in general with same system like ethio telecom to improve customer size prediction and produce optimal service packages. Further studies are needed to improve the model performance using other techniques and more advanced algorithms. Moreover, the dataset can be expanded to include all the telecom packages and services.

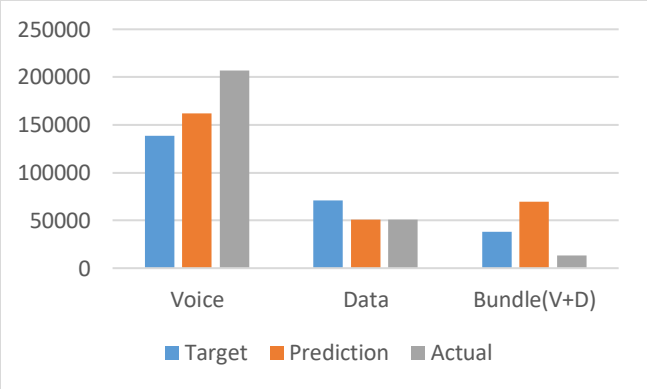

Figure 2: RF Predictions Compared to Target and Actual values

ACKNOWLEDGEMENT

The authors disclosed receipt of the financial support for the publication of this article under the IEEE CSDE 2021 scholarship support.

REFERENCES

[1] J. Danhua, Z. Xiaogeng and W. Runrun, "Research on the Amount of Customers in Telecom Package Preview Based on Data Mining," in *International Conference on Computer Science and Service System*, 2012.

[2] M. V. Joseph, "Data Mining and Business Intelligence Applications in Telecommunication Industry," *International Journal of Engineering and Advanced Technology (IJEAT)*, ISSN: 2249 – 8958, Volume-2, Issue-3, 2013.

[3] H. H. Darji, "Data Mining in Telecommunication Industry," *IJSRD - International Journal for Scientific Research & Development*, Vol. 2, Issue 08, 2014.

[4] C. Srinuan, P. Srinuan and E. Bohlin, "Pricing strategies and innovations in the Thai mobile communications market", *info*, Vol. 15 No. 1, pp. 61-77, 2013. https://doi.org/10.1108/14636691311296219

[5] X. K. Jiang and X. Chen, "Research on prediction model of the impact of new telecom services tariff based on the customer choice behavior," *Advanced Materials Research*, vol. 765-767, pp. 3249–3252, 2013.

[6] O. Wisesa, A. Adriansyah and O. I. Khalaf, "Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm," *2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*, Yogyakarta, Indonesia, pp. 101-106, doi: 10.1109/BCWSP50066.2020.9249397.

[7] S. Cheriyan, S. Ibrahim, J. Mohanan, and S. Treesa, “Intelligent Sales Prediction Using Machine Learning Techniques,” *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, Shouthend, UK, pp. 53-58. DOI: 10.1109/iCCECOME.2018.8659115.

[8] J. Brownlee, *Why One-Hot Encode Data in Machine Learning*, June 30, 2020. Accessed on: Aug. 12, 2021. [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machinelearning/.

[9] A. Burkov, *The Hundred-Page Machine Learning Book*, 2019.

[10] C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.

[11] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research (JMLR)*, Vol. 12, pp. 2825-2830, 2011.