2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 978-1-6654-9552-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CSDE53843.2021.9718375

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

I. 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.

II. 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.

III. 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.

IV. 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.

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.

V. 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.

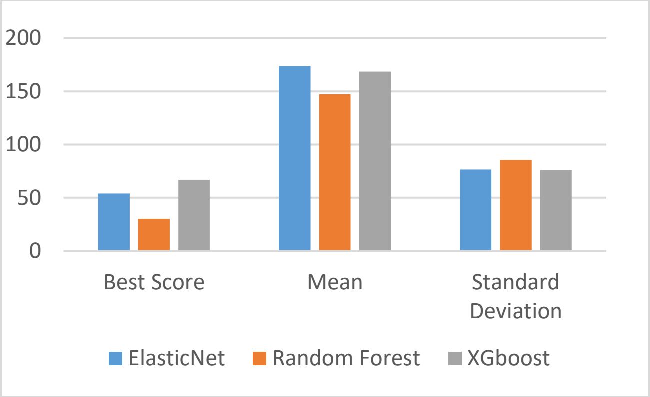

Figure 1: CV Performance Result

VI. 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.

VII. 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.

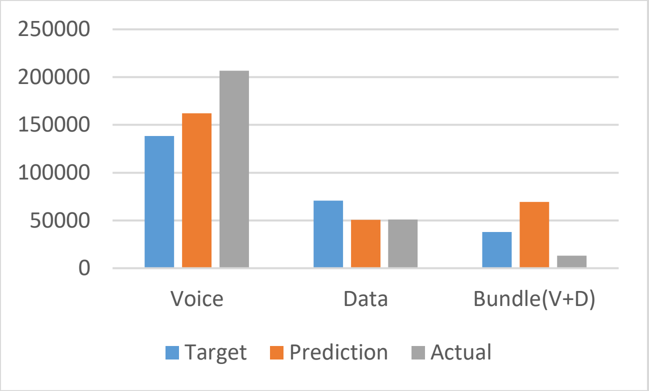

Figure 2: RF Predictions Compared to Target and Actual values

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.