capstone-Final1-BuğraBalantekin.docx

by BUGRA BALANTEKIN

Submission date: 05-Sep-2021 04:34PM (UTC+0300)
Submission ID: 1641700425
File name: capstone-Final1-BuğraBalantekin.docx (26.53M)
Word count: 3137
Character count: 18293

MEF UNIVERSITY

CUSTOMER SEGMENTATION & CHURN PREDICTION FOR CUSTOMER RELATIONSHIP MANAGEMENT

Capstone Project

Buğra Balantekin

İSTANBUL, 2021

MEF UNIVERSITY

CUSTOMER SEGMENTATION & CHURN PREDICTION FOR CUSTOMER RELATIONSHIP MANAGEMENT

Capstone Project

Buğra Balantekin

Advisor: Asst. Prof. Dr. Evren Güney

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: Customer Segmentation & Churn Prediction For Customer Relationship Management
Name/Last Name of the Student: Buğra Balantekin
Date of Thesis Defense: dd/mm/yyyy

I hereby state that the graduation project prepared by Buğra Balantekin has been completed under my supervision. I accept this work as a "Graduation Project".

dd/mm/yyyy
(Asst. Prof. Evren Güney)

I hereby state that I have examined this graduation project by Buğra Balantekin which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

dd/mm/yyyy

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Buğra Balantekin and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Your Advisor's Name	
2. Your Advisor's Name	……..

ACADEMIC HONESTY PLEDGE

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Buğra Balantekin	12.07.2021	

EXECUTIVE SUMMARY

CUSTOMER SEGMENTATION & CHURN PREDICTION FOR CUSTOMER RELATIONSHIP MANAGEMENT
Buğra Balantekin

Advisor: Asst. Prof. Evren Güney

JULY, 2021, (26 pages)

This study examines the principal components of Customer Relationship Management; Customer Segmentation and Churn Prediction. The research has used a Kaggle Dataset for Telecom Industry containing 100 variables and 100.000 observations. For churn prediction classifier algorithms are compared. According to the results LightGBM algorithm outperforms every algorithm compared. Also, results are statistically significant for XGBoost and LightGBM algorithms accuracy score comparison. In the customer segmentation part of research, Kmeans clustering algorithm is used with three different clusters, named Regular, Gold and Platinum.

Key Words: Customer Segmentation, Churn Prediction, Classifier Algorithms

ÖZET

MÜŞTERİ İLİŞKİLERİ YÖNETİMİ İÇİN MÜŞTERİ SEGMENTASYONU VE CHURN TAHMİNİ

Buğra Balantekin

Tez Danışmanı: Dr. Öğr. Üyesi Evren Güney

TEMMUZ, 2021, (26 sayfa)

Bu çalışma, Müşteri İlişkileri Yönetiminin temel bileşenleri olan Müşteri Segmentasyonu ve Terketme Tahminini incelemektedir. Araştırmada, Telekom Endüstrisi için 100 değişken ve 100.000 gözlem içeren bir Kaggle Veri Kümesi kullanılmıştır. Terketme tahmini için sınıflandırma algoritmaları karşılaştırılmıştır. Sonuçlara göre LightGBM algoritmasının, karşılaştırılan her algoritmadan daha iyi performans gösterdiği gözlemlenmiştir. Ayrıca sonuçlar, XGBoost ve LightGBM algoritmalarının doğruluk puanı karşılaştırması için istatistiksel olarak anlamlıdır. Araştırmanın müşteri segmentasyonu bölümünde, k-means kümeleme algoritması kullanılmış olup, tespit edilen kümelere Regular, Gold ve Platinum isimleri verilmiştir.

Anahtar Kelimeler: Müşteri Segmentasyonu, Terketme Tahmini, Sınıflandırma Algoritmaları

TABLE OF CONTENTS

ACADEMIC HONESTY PLEDGE vi

EXECUTIVE SUMMARY vii

ÖZET viii

TABLE OF CONTENTS ix

1. INTRODUCTION 1

2. LITERATURE REVIEW 2

3. DATA 4

3.1. Features 4

3.2. Exploratory Data Analysis 4

4. METHODOLOGY 10

5. RESULTS 12

REFERENCES 17

.

1. INTRODUCTION

In some sectors of economy, due to harsh competition, profit margins are narrowed. To expand the profits, businesses need to understand not only their rival's pricing attitude but also their own marketing campaigns expenditures and these expenditures effects on profits. One way to achieve this is by targeting correct products with correct customers with correct prices. To target customers, customer segmentation should be considered to understand group dynamics. Most of the companies' segment and label their customers according to their consumption behaviors. By labeling the customers, businesses can manage their campaign – marketing budgets effectively.

Also, businesses need to understand which customer have tendency to churn so that they can make a final touch to keep these customers onboard. Sharma and Panigrahi (2011) defined churn as the leave or abandon of a company of the customer. In telecom industry, churn can be defined as closing the cellular line of Company A and starting to use the plan of Company B. In banking industry churn can be defined as closing the account of customer in Bank A and move the funds/deposit to another Bank. Churn cannot be defined in just using Company A or B, it can also be defined as stopping to use a specific product of a Company. If a customer closes the credit card of a Bank but keeps using the other product of same Bank, this habit of credit card closure can be defined as churn as well.

It's important for companies to predict which customer have tendency to churn so they can send customized marketing campaigns to these customers to keep them onboard. This campaign may be assisted calls from call center of company, sending text messages – e-mails, sending push notifications from mobile application if the company has a mobile application, defining gift cards. From a Telecom Industry perspective, this can be defining of free minutes/ data plan or defining a completely customized usage plan to customer that tend to churn.

This project aims to understand the fundamental aspects of Marketing Analytics, Customer Segmentation and Churn Prediction.

This report is organized as follows: In Section 2 literature review about churn prediction and customer segmentation is examined. Section 3 explains the Features of Dataset and Exploratory Data Analysis known as EDA. Section 4 presents methodology of the study, along with the algorithms used for prediction of churn and customer segmentation. Last section explains the results with figures and tables.

2. LITERATURE REVIEW

Tripathi et al. (2010) has examined the role of customer segmentation in Customer Relationship Management. Authors has compared the main clustering algorithms which are K-means and Hierarchical Clustering. Their findings are not specifically performed on a dataset, but both algorithms are thoroughly inspected for their major drawbacks and advantages.

Nie et al. (2011) has gathered data for Chinese banks from a data warehouse, which included 60 million observations and 135 features. Authors have randomly sampled observations and limit their features according to their correlations. They've also compared the performance of logistic regression and decision trees, which is also included in this project. According to their measurement logistic regression has yielded better results for churn prediction.

Huang et al. (2012) has also gathered data from a data warehouse in Ireland about telecom customers churn prediction. Their data has more than 800K observations with more than 700 features for each observation, which is a more likely scenario that we can observe in real life, especially in finance sector in terms of feature count. Their project has compared seven different algorithm including linear and non-linear ones.

Keramati et al. (2016) has developed a model for churn prediction with CRISP-DM methodology using Decision Tree algorithms. Kumar (2008) compared six different algorithms for churn prediction in the research and also used SMOTE (Synthetic Minority Oversampling Technique) for overcoming the imbalance problem in dataset.

Ke et al. (2017) has deeply examined the performance of LightGBM algorithm and found that LightGMB is 6 to 20 times faster than compared algorithms in certain datasets. Their study has also revealed that XGBoost is consuming too much memory so that it couldn't run effectively on their dataset.

Segmentation of customers are handled to find which customers are loyal or not in the research of Syakur et al. (2018). Researchers had worked with K-Means clustering algorithm and elbow method to find the optimal number of cluster size "k". With the help of elbow method, it has been found that the ideal size of clusters is 3 for their dataset.

Ezenkwu et al. (2015) has collected data from a retail shop in Nigeria and normalized the features with z-score technique. After this step, with the Forgy method, k is selected as 4 in centroid initialization and assignment of data points to clusters has started. Their result showed that after 100 iteration, K-means algorithm works with %95 accuracy.

Rahman et al. (2020) has compared the performance and accuracy of boosting classifiers; XGBoost, Light GBM, AdaBoost, CatBoost and Gradient Boosting. Researchers dataset has consisted the data of daily activity that has collected through wearable technology devices and found out that 94 pct accuracy can be achieved with boosting algorithms for 6 class classification (walk, sit, stand, lie, up-stairs,down-stairs)

As it can be seen there are plenty of resources about Churn prediction and Customer Segmentation. What distinguish this research from the above-mentioned researches is mixed using of the methods in papers and performance comparison of classification algorithms. Also, Data used in this paper is from Telecom industry and has 100 features and 100.000 observations like the research of Huang et al. (2012). Similar to the studies above, Customer Segmentation of this research is created with k-means clustering algorithm.

3. DATA

Dataset for this study is obtained from Kaggle Datasets. [1] It has mainly organized for predicting customer churn in Telecom Industry.

3.1. Features

The project dataset has 100 features and 100.000 observations. It has both continuous and categorical variables. It has also lots of missing values especially for categorical variables.

Features of the dataset can be categorized into two segments;

1. Customer's Demographic Features
2. Customer's Behavioural Features

List of all variables are not mentioned in this part due to its size. Instead, list of all variables is explained in Appendix A.

3.2. Exploratory Data Analysis

Churners and Non-Churners are almost equally distributed in the dataset as shown in Figure 1 below.

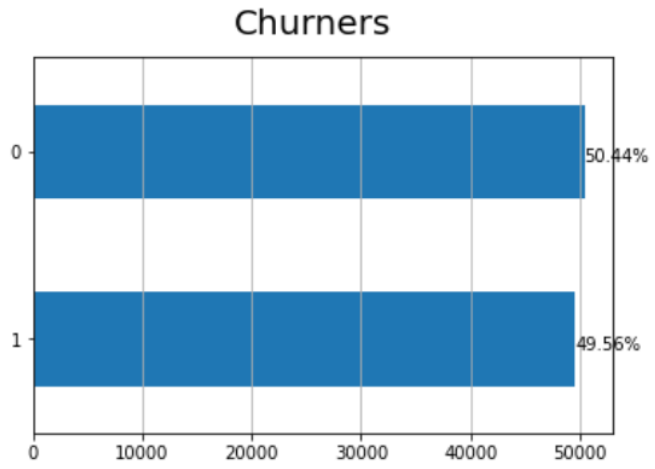

Figure 1 – Distribution of churners and non-churners in the dataset

[1] https://www.kaggle.com/abhinav89/telecom-customer

In the Figure 2, missing value ratios of features of raw dataset is shown. Almost 50 percent of Numbcars feature contains missing data.

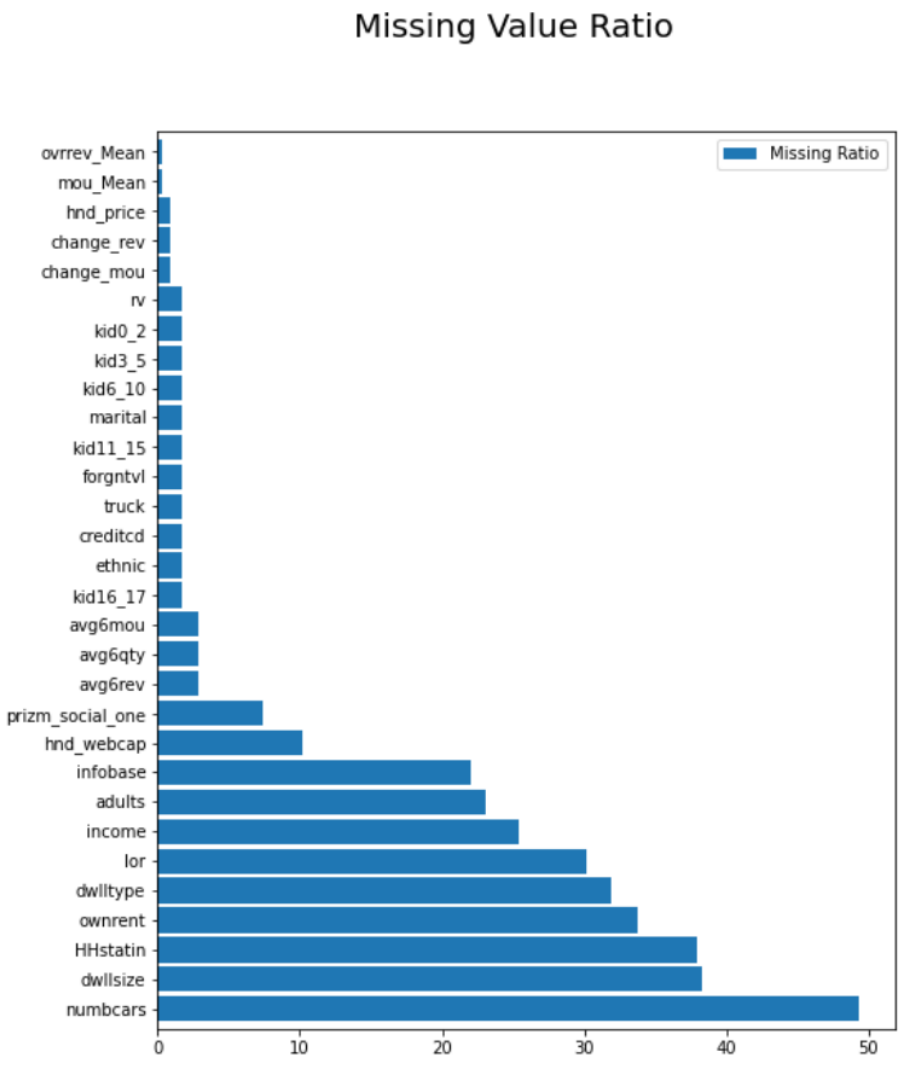

Figure 2 – Missing Value Ratio of Features

Figure 3 shows the missing value ratios of features after dropping 21 categorical variables. Also a boolean mask is created to eliminate features that has less than 40% missing values. Feature named "lor" kept purposely because it has been discovered that has a significant importance in modeling. Missing values will later be imputed with MICE method.

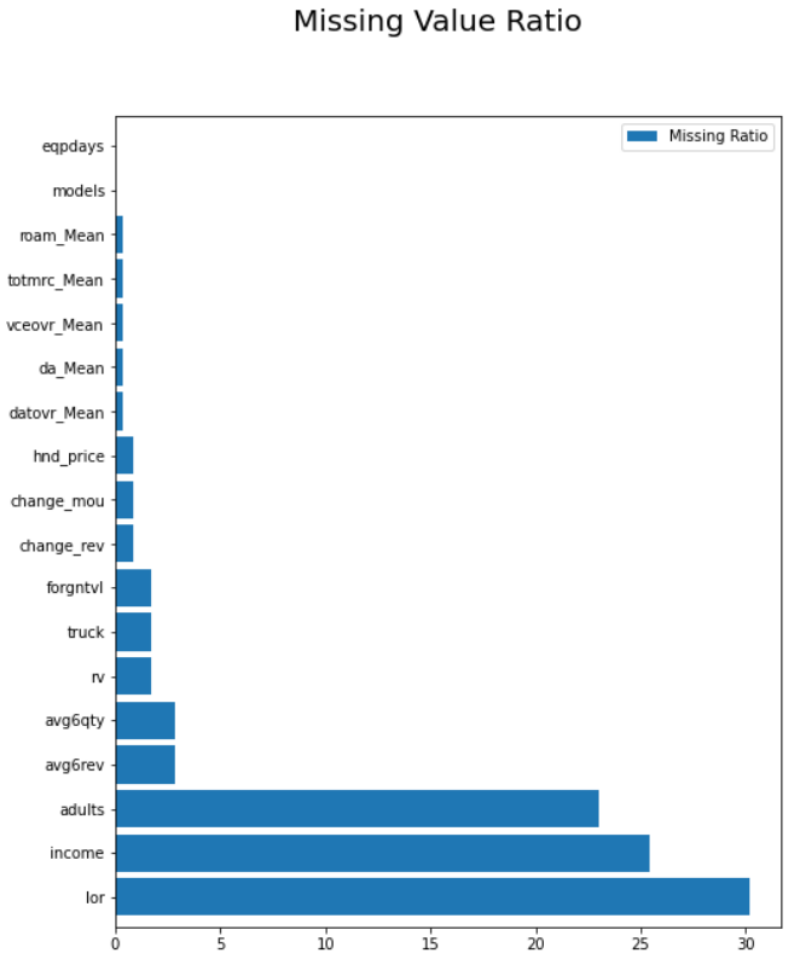

Figure 3 – Missing Value Ratio of Features after dropping

Figure 4 has visualized the correlation heatmap of final features. In the final stage there're 42 features left after dropping highly correlated values to avoid multicolinearity problem. Another boolean mask is created to drop columns that has more than 80 % correlation to achieve this result.

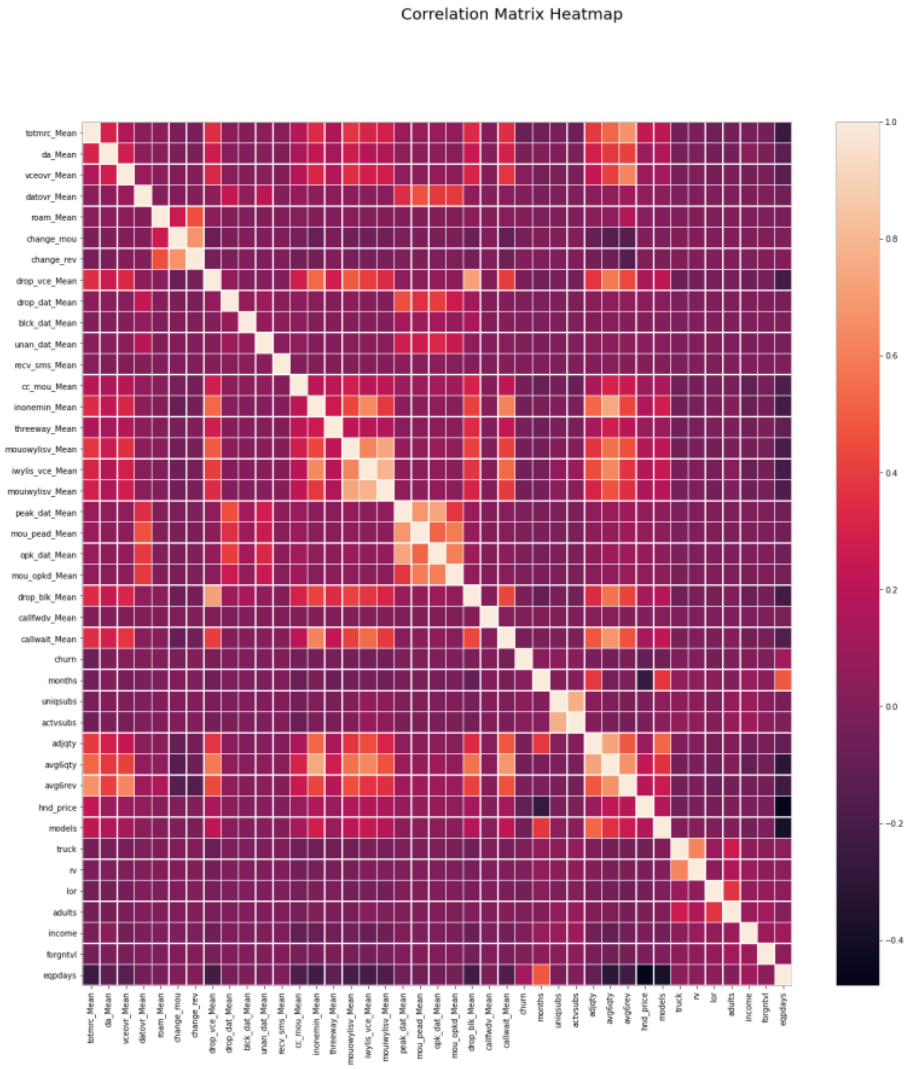

Figure 4 – Correlation Heatmap of Final Features

Figure 5 shows the total calls per area. Great Lakes area, North Florida, Ohio and Houston has significantly less calls comparing to other areas. Also in some areas there're extreme values.

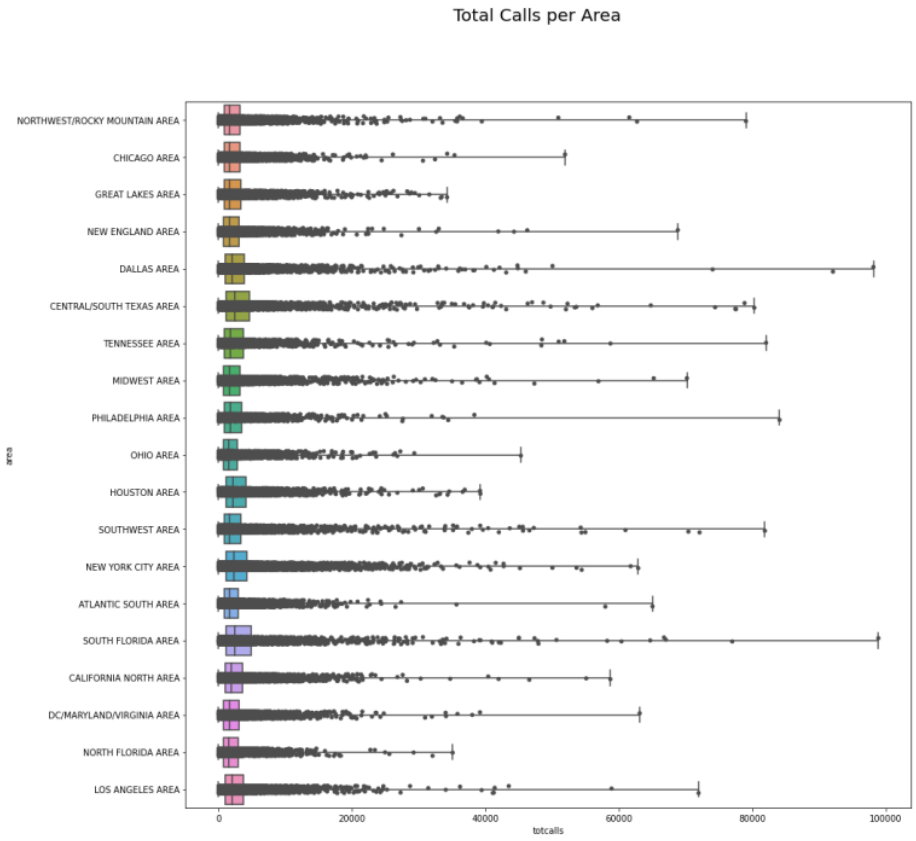

Figure 5 – Total Calls per Area in Raw Dataset

Statistical distributions of final features are shown in Figure 6. It has been observed that most of the variables are not normally distributed.

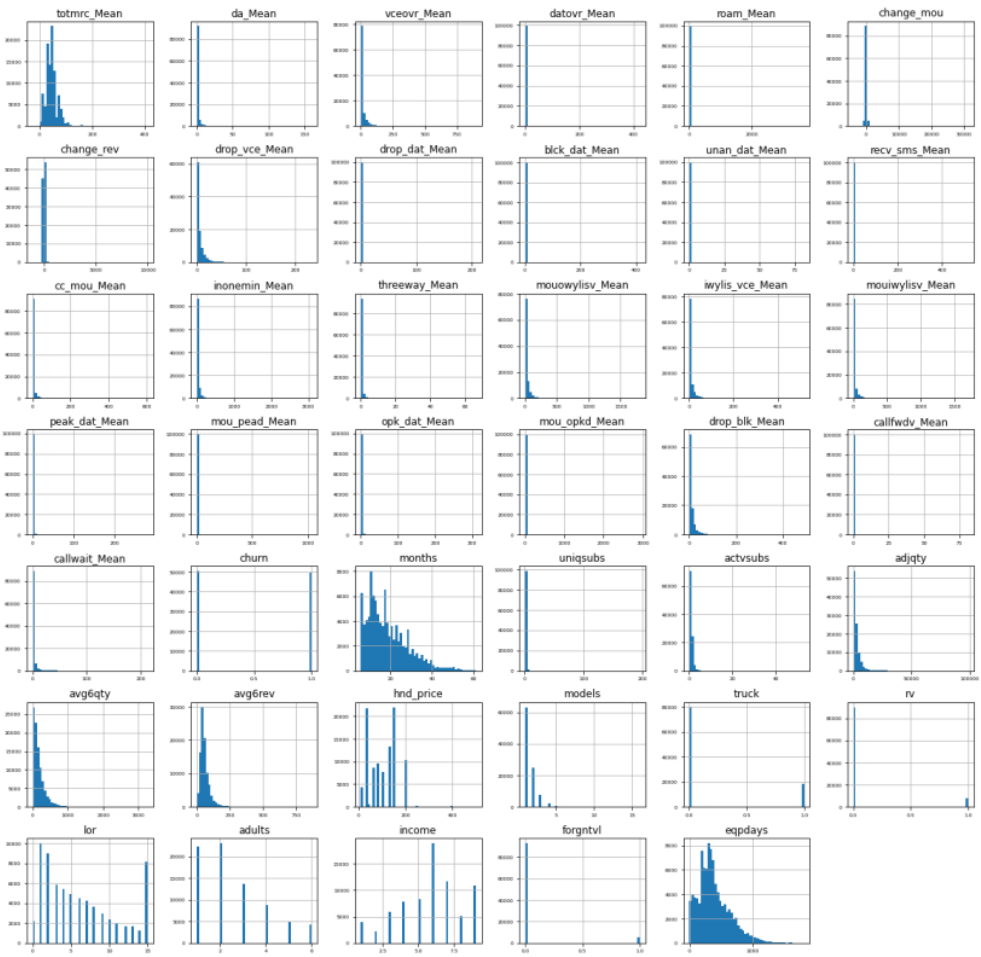

Figure 6 – Statistical Distributions of Final Features

4. METHODOLOGY

Since churn prediction is a binary classification problem, multiple classification algorithms are compared. Here's a list of compared algorithms;

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gaussian NB Classifier
- KNN classifier
- XGBoost Classifier
- Adaboost Classifier
- Light GBM

After dropping demographic features from dataset, as they're heavily categorical and has less explain ability for churn prediction from business perspective, FancyImputer package of Python is used to fill these values. Main method to fill missing values is MICE (Multiple Imputation by Chained Equation). KNN imputation method could not be used because of the Google Colab or Local Python IDE (VS Code) kept crashing due to RAM size.

To select the best variables for predicting churn, couple of methods are used;

- Categorical features are dropped
  - 21 features eliminated
- Features that have more than 40 % missing values are dropped
  - 1 feature is eliminated
- Features that are highly correlated (greater than 80 pct) are dropped to avoid multicollinearity
  - 36 features eliminated

In the modeling phase 42 features is used after dropping according to the above methods.

Even though the dataset is not constructed for customer segmentation, Churn column will be dropped to segment the customers using the given features. Effective Customer

Relationship Management for any company is based on differentiation of customers regarding their behavior. For customer segmentation, K-means clustering will be used with elbow method for selecting the optimal K number. After the execution of algorithm, cluster analysis will be made with visualizations.

5. RESULTS

For churn prediction, couple of algorithms are compared and results are shown below. Naïve Bayes is the fastest but worst performing one in terms of accuracy. Light GBM has outperformed every algorithm in terms of accuracy and the time elapsed is 3 seconds for 100.000 observations and 42 features as shown in Table 1.

Table 1. Comparison of churn prediction algorithms

Algorithm	Accuracy	Time Elapsed
'Logistic Regression'	0.5703	0:00:01.625048
'Naive Bayes'	**0.5092**	0:00:00.091600
'Decision Tree'	0.6226	0:00:02.156740
'Random Forest'	0.6750	0:00:26.804315
'Nearest Neighbors'	0.5501	0:01:19.980486
'XGBoost Classifier'	0.6616	0:00:09.226688
'AdaBoost Classifier'	0.6324	0:00:08.169018
'Light GBM'	**0.7002**	0:00:03.380557

Paired sample t-test is executed for comparing XGboost and Light GBM algorithm to check if there's a statistically significant difference between the accuracy scores of each algorithm while using 5-fold cross validation.

P value of the scores is 1.0543283138542076e-05, so we reject the null hypothesis because the p value is smaller than 0.05.

This result shows us that we have statistically significant evidence that performance of XGBoost and Light GBM models' accuracy is different.

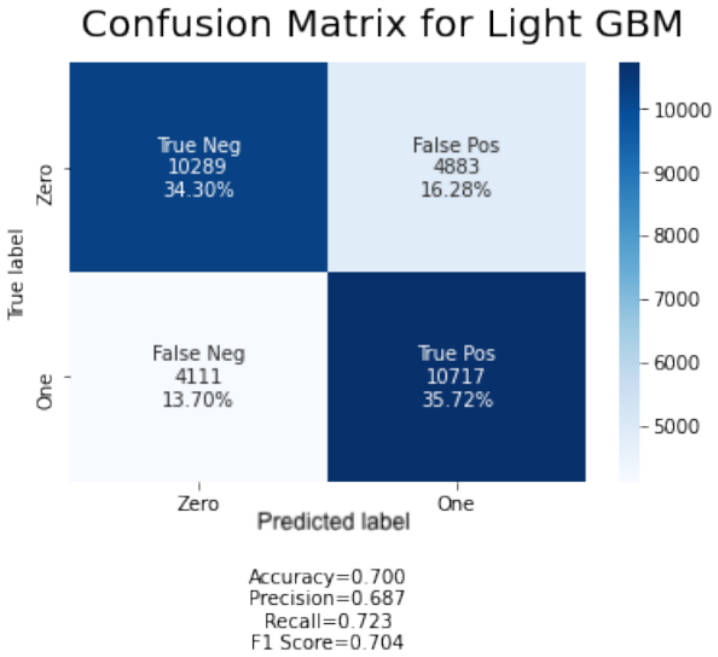

Figure 8 – Accuracy result for Light GBM algorithm

Figure 8 depicts the confusion matrix of Light GBM model. Accuracy, Precision, Recall and F1 score of the model is also shown below the figure

It has been mentioned before that the dataset for this study is obtained from Kaggle. Notebooks for this dataset in Kaggle is also examined. Most of the notebooks final score for churn is around 60-65 percent. It can be seen from the Figure 8 and Table 1 that with the help of Light GBM scores has risen to 70 percent in terms of accuracy.

While the Light GBM model has direct effect in this rise, this study has unique methods that distinguish from most of the Kaggle Notebooks. First, all the categorical variables are dropped. Secondly, features that has more than 40 percent missing value ratio are dropped. Thirdly, highly correlated variables are dropped to avoid multicollinearity. Fourth, MICE imputation method is used for filling remaining missing values. By the end of these steps, 100 features are dropped to 42. Moreover scaling – normalizing the features is also tried but didn't contribute to churn prediction accuracy.

Lastly recursive feature selection (RFE) of features with stepwise method is also executed for Light GBM algorithm. This method weights the features and eliminates them

iteratively with the parameter we define as 2, which means that it starts with 42 features and fits model, then drops 2 features and fits the model again until the number we set for number of features to select, in our case 16. The result of this method has also 69,4 percent accuracy score.

In Figure 9, importance of features is listed for Light GBM model, as it has the highest accuracy. Customer's length of residence, "lor" and number of adults in household, "adult" features surprisingly had a high importance for explaining churn prediction in Light GBM model.

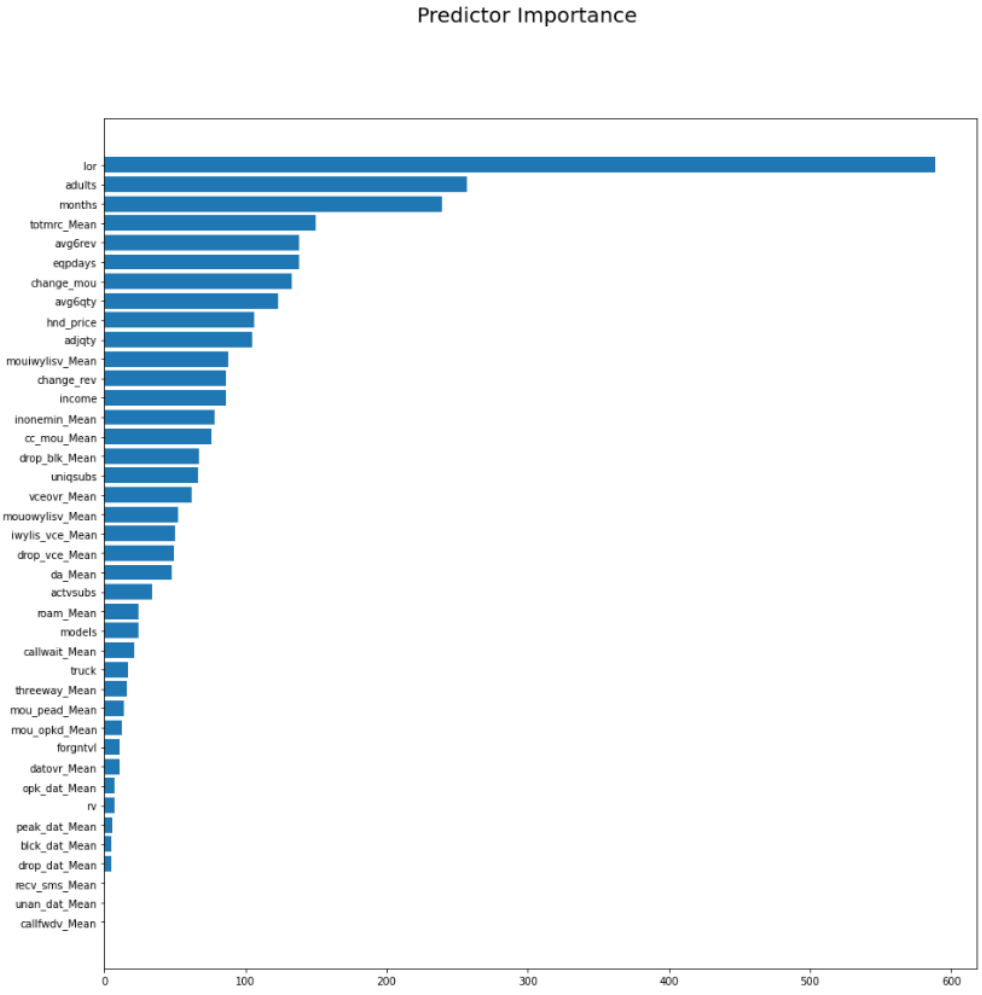

Figure 9 – Feature Importance for Light GBM Model

For customer segmentation K-means algorithm is used with elbow method to determine the cluster size. Cluster size is selected as 3 according to elbow method as shown in Figure 10.

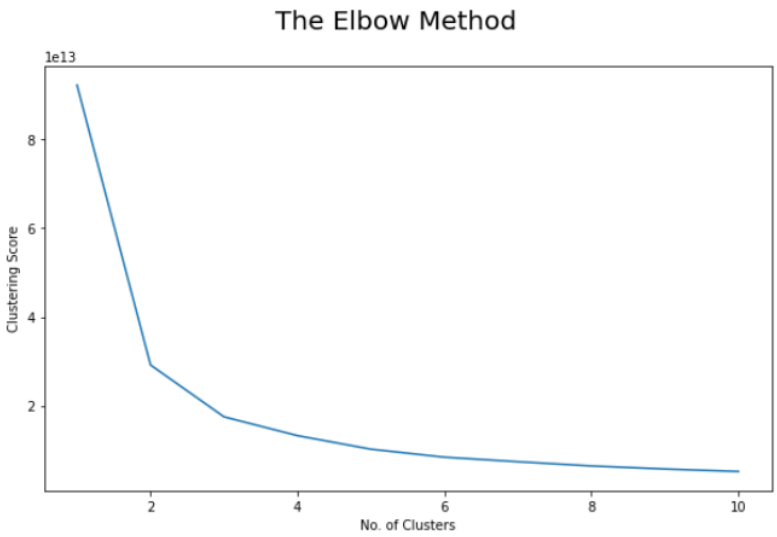

Figure 10 – Number of Clusters – Elbow Method

After selecting the cluster size as 3, model is fitted, and average of every feature is grouped by cluster numbers. Just like a RFM analysis, 3 features (Quantity, Monthly Use, and Revenue) have been selected from dataset for Kmeans clustering.

Table 2. Cluster Properties

ClusterID	Quantity_mean	MonthlyUse_mean	Revenue_mean
Gold	2.480.926.416	6.836.508.437	849.253.349
Regular	1.780.631.268	5.288.332.895	565.564.424
Platinum	4.340.945.152	10.710.937.360	1.499.144.614

Table 2 shows the average values of every feature used for clustering customers of Telecom Industry.

It's been observed that Cluster 2 brings the highest monthly use and revenue, so it has been named as Platinum. Cluster 1 has the lowest Revenue and Monthly Use, so it has been named as Regular. Lastly Cluster 0 stands between Regular and Platinum Users, so it has been named as Gold.

Names of Clusters are selected according to the real-world scenarios for Telecom Sector.

In the Figure 11 below, blue bar stands for Gold Users, orange for Regular and Green for Platinum User.

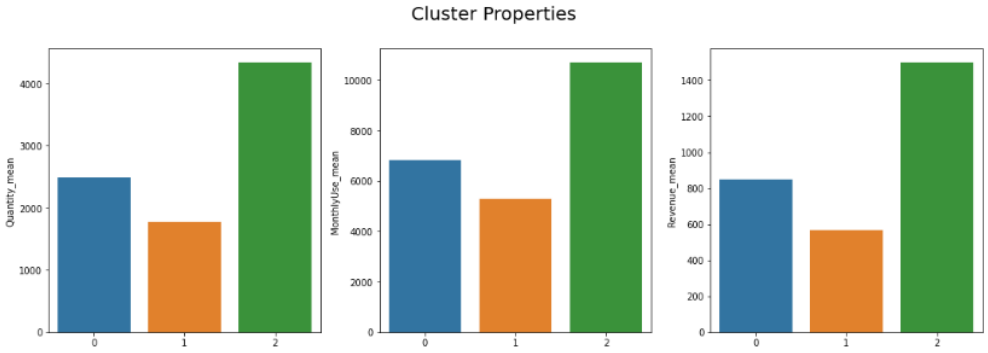

Figure 11 – Average of Features as Cluster Properties

REFERENCES

Tripathi, S., Bhardwaj, A., & E, P. (2018). Approaches to Clustering in Customer Segmentation. International Journal of Engineering & Technology, 7(3.12), 802-807. doi:http://dx.doi.org/10.14419/ijet.v7i3.12.16505

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 38(12), 15273–15285. https://doi.org/10.1016/j.eswa.2011.06.028

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation, 2(1). https://doi.org/10.1186/s40854-016-0029-6

Kumar, D. A., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies, 1(1), 4. https://doi.org/10.1504/ijdats.2008.020020

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conference Series: Materials Science and Engineering, 336, 012017. https://doi.org/10.1088/1757-899x/336/1/012017

Pascal, C., Ozuomba, S., & Kalu, C. (2015). Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. International Journal of Advanced Research in Artificial Intelligence, 4(10). https://doi.org/10.14569/ijarai.2015.041007

Rahman, S., Irfan, M., Raza, M., Moyeezullah Ghori, K., Yaqoob, S., & Awais, M. (2020). Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living. International Journal of Environmental Research and Public Health, 17(3), 1082. https://doi.org/10.3390/ijerph17031082

Ishantha, A.(2021, march). Mall customer segmentation using clustering algorithm. Paper presented at the LNBTI machine learning conference. (PDF) MALL CUSTOMER SEGMENTATION USING CLUSTERING ALGORITHM (researchgate.net)

Ke,G., Meng, Q., Finley, T., Wang, T., Wei, C. , Ma,W. , Ye, Q.& Liu, T. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. LightGBM: A Highly Efficient Gradient Boosting Decision Tree (nips.cc)

Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications, 27(11), 26–31. https://doi.org/10.5120/3344-4605

VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data (1st ed.). O'Reilly Media.

McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.

Hypothesis testing in Machine learning using Python. (2019). Towards Data Science. https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

1.13. Feature selection — scikit-learn 0.24.2 documentation. (n.d.). Sci-Kit Learn. Retrieved August 29, 2021, from https://scikit-learn.org/stable/modules/feature_selection.html#rfe

Telecom customer. (2017, August 27). Kaggle. https://www.kaggle.com/abhinav89/telecom-customer

capstone-Final1-BuğraBalantekin.docx

ORIGINALITY REPORT

10%	8%	2%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	8%
2	"Brain Informatics", Springer Science and Business Media LLC, 2020 Publication	<1%
3	core.ac.uk Internet Source	<1%
4	en.wikipedia.org Internet Source	<1%
5	pp-rai.pwr.edu.pl Internet Source	<1%
6	www.altexsoft.com Internet Source	<1%
7	"Advances and Trends in Artificial Intelligence. From Theory to Practice", Springer Science and Business Media LLC, 2019 Publication	<1%
8	Abbas Keramati, Hajar Ghaneei, Seyed Mohammad Mirmohammadi. "Investigating	<1%

factors affecting customer churn in electronic banking and developing solutions for retention", International Journal of Electronic Banking, 2020
Publication

9	unitpapers.com Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		