UNRAVELLING THE DYNAMICS OF CUSTOMER LOYALTY

A Project Report Submitted to the
Department of Computer Science of
Ramakrishna Mission Vivekananda Educational and Research Institute, Belur,
in partial fulfilment of the requirements for the degree of
MSc in Big Data Analytics.

Submitted by
ARNAB MUKHERJEE, NIRMALYA NANDY AND SOURISH GHOSH

Guide-cum-Supervisor:
Dr. Raju Maity
Economic Research Unit
INDIAN STATISTICAL INSTITUTE

Department of Computer Science
Ramakrishna Mission Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India
January 7, 2024

UNRAVELLING THE DYNAMICS OF CUSTOMER LOYALTY

By

ARNAB MUKHERJEE, NIRMALYA NANDY AND SOURISH GHOSH

Declaration by students:

"We hereby declare that the present dissertation is the outcome of our project work under the guidance of Dr. Raju Maity and we have properly acknowledged the sources of materials used in our project report."

_____

(Arnab Mukherjee, Nirmalya Nandy and Sourish Ghosh)

A project report in the partial fulfilment of the requirements of the degree of MSc in Big Data Analytics

Examined and approved on

_____

by

_____

Dr. Raju Maity (supervisor)
Economic Research Unit
INDIAN STATISTICAL INSTITUTE

Countersigned by

_____

Registrar
Ramakrishna Mission Vivekananda Educational and Research Institute

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India

January 7, 2024

Acknowledgement

The present project work is submitted in partial fulfilment of the requirements for the degree of Master of Science of Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI). We express our deepest gratitude to our supervisor Dr. Raju Maity of Ramakrishna Mission Vivekananda Educational and Research Institute for his inestimable support, encouragement, profound knowledge, largely helpful conversations and also for providing us a systematic way for the completion of our project work. His ability to work hard inspired us a lot. We are also extremely grateful to the Vice-Chancellor of this University for his encouragement and support throughout the course. Last but not the least, this work would not have been possible without support of our fellow classmates.

Belur
January 7, 2024

(Arnab Mukherjee, Nirmalya Nandy and Sourish Ghosh)
Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute

Contents

Chapter 1

Introduction

In this project, we delve into the intricate landscape of customer churn, aiming to decipher the underlying reasons behind customer attrition. By adopting a realistic perspective, we seek to unravel the multifaceted factors influencing churn, leveraging advanced predictive analytics to classify customers into churn and non-churn categories.

Our endeavor is not just about foreseeing customer departure but comprehending the authentic triggers behind it, paving the way for strategic interventions to enhance customer retention and satisfaction.

Chapter 2

Objective

To predict classes of the customer churn (Binary Classification) and also find the reasons for which customers churn.

Chapter 3

Methodologies

Logistic regression

It is a statistical method used for binary classification, predicting the probability of an event occurring. It models the relationship between independent variables and the log-odds of the dependent variable, applying the logistic function to constrain predictions between 0 and 1. Parameters are estimated through maximum likelihood estimation, and the resulting model is useful for making predictions and understanding the influence of variables on the outcome. Despite its name, logistic regression is employed for classification tasks, not regression, making it a fundamental tool in machine learning and statistics for analyzing and modeling categorical outcomes. For more information, one can refer [5]

K Nearest Neighbors (KNN)

It is a simple yet effective machine learning algorithm used for classification and regression tasks. It works by assigning a data point to the majority class among its k-nearest neighbors in the feature space. The choice of k influences the model's sensitivity to local patterns. KNN is non-parametric, meaning it doesn't make assumptions about the underlying data distribution. While computationally expensive for large datasets, KNN is intuitive and easy to understand. It is particularly useful when the decision boundaries are complex and not easily captured by simpler models. For more information, one can refer [3]

Decision Tree

A Decision Tree is a predictive modeling tool that visually represents decisions and their possible consequences in a tree-like structure. Each node in the tree represents a decision or attribute, and branches represent possible outcomes. It is widely used

in machine learning for classification and regression tasks. The tree is constructed by recursively splitting data based on the most significant features, aiming to maximize information gain or minimize impurity. Decision Trees are interpretable and versatile, suitable for various domains. However, they may be prone to over-fitting, requiring techniques like pruning. Popular algorithms include CART (Classification and Regression Trees) and ID3 (Iterative Dichotomiser 3). For more information, one can refer [6].

Random Forest

Random Forest is an ensemble machine learning algorithm that builds multiple decision trees during training and merges them to achieve more accurate and stable predictions. It operates by constructing a multitude of trees and combining their outputs to reduce overfitting and enhance robustness. Each tree is trained on a random subset of the data and features, fostering diversity in the model. Random Forest is effective for both classification and regression tasks, offering high performance, scalability, and resilience to noisy data. Its versatility and ability to handle large datasets make it a popular choice in various domains, such as finance, healthcare, and ecology. For more information, one can refer [1].

Support Vector Machine (SVM)

It is a powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points into different classes, maximizing the margin between them. SVM is effective in high-dimensional spaces, handles non-linear relationships through kernel functions, and is robust against overfitting. It is widely used in various fields such as image recognition, text classification, and bioinformatics. SVM's ability to handle complex data and achieve high accuracy makes it a popular choice in machine learning applications. For more information, one can refer [4].

Multi Layer Perceptron

Multilayer Perceptron (MLP) is a type of artificial neural network characterized by multiple layers of interconnected nodes or neurons, including an input layer, hidden layers, and an output layer. It excels in non-linear mapping and complex pattern recognition tasks. Each neuron applies a weighted sum of inputs, followed by an activation function. MLPs are trained through backpropagation, adjusting weights to minimize prediction errors. Widely used in machine learning, they can handle various data types, making them versatile for tasks like classification and regres-

sion. However, their architecture requires careful tuning, and they may suffer from overfitting with insufficient data.. For more information, one can refer [4].

XG Boost

XGBoost, or eXtreme Gradient Boosting, is a powerful and efficient machine learning algorithm known for its speed and accuracy in predictive modeling. It is an ensemble learning method that combines the strengths of multiple weak learners, often decision trees, to create a robust and accurate predictive model. XGBoost employs a gradient boosting framework, optimizing the model's performance by minimizing loss functions. It incorporates regularization techniques to prevent overfitting and handles missing data effectively. Widely used in various domains for tasks like classification, regression, and ranking, XGBoost has become a popular choice in machine learning competitions due to its exceptional performance. For more information, one can refer [2].

Exploratory Data Analytics

Exploratory Data Analysis (EDA) is a crucial phase in data analysis where raw data is visually and statistically examined to discover patterns, trends, and anomalies. It involves techniques such as summary statistics, data visualization, and hypothesis testing to gain insights into the data's structure and characteristics. EDA helps analysts understand the distribution of variables, identify outliers, and make informed decisions about subsequent modeling or analysis. By revealing patterns and relationships within the data, EDA plays a pivotal role in uncovering valuable information that guides further investigation and aids in the formulation of data-driven strategies.

Problem of encoding of categorical columns

For classification task we often do require the categorical columns effect. One-hot encoding has issues like it increases number of columns and cause multicollinearity issue. Different categorical columns encoded can become highly correlated So, we have not used categorical columns for classification task.

Feature selection

Iterating over the thresholds of Variance Information Factor (VIF), we found approximately VIF of 50 as the threshold providing best validation accuracy. So, we considered only those columns which had VIF lesser than 50. 64 columns were selected based on this criteria..

Missing value handling

We have dropped the columns which had missing values more than 50 per cent. We finally got 163 columns on which we performed our EDA. Mean imputation when range is below 500 and rest median imputation. We used 41 and 87 columns on which we performed mean imputation and median imputation respectively. Modal class imputation for categorical columns. We used 23 columns on which we performed this.

Evaluation metric

Test Accuracy. Binary Cross-Entropy (BCE) loss is commonly used in binary classification tasks. Test accuracy, a key performance metric, measures the model's ability to correctly classify instances in the test set. BCE loss, calculated during training, quantifies the difference between predicted probabilities and actual labels. A higher test accuracy indicates better model performance, as it reflects the proportion of correctly classified samples. However, it's important to note that accuracy alone may not capture the full picture, especially in imbalanced datasets. Evaluating precision, recall, and the confusion matrix alongside BCE loss provides a more comprehensive assessment of the model's effectiveness.

Chapter 4

Data Description

Churn Data

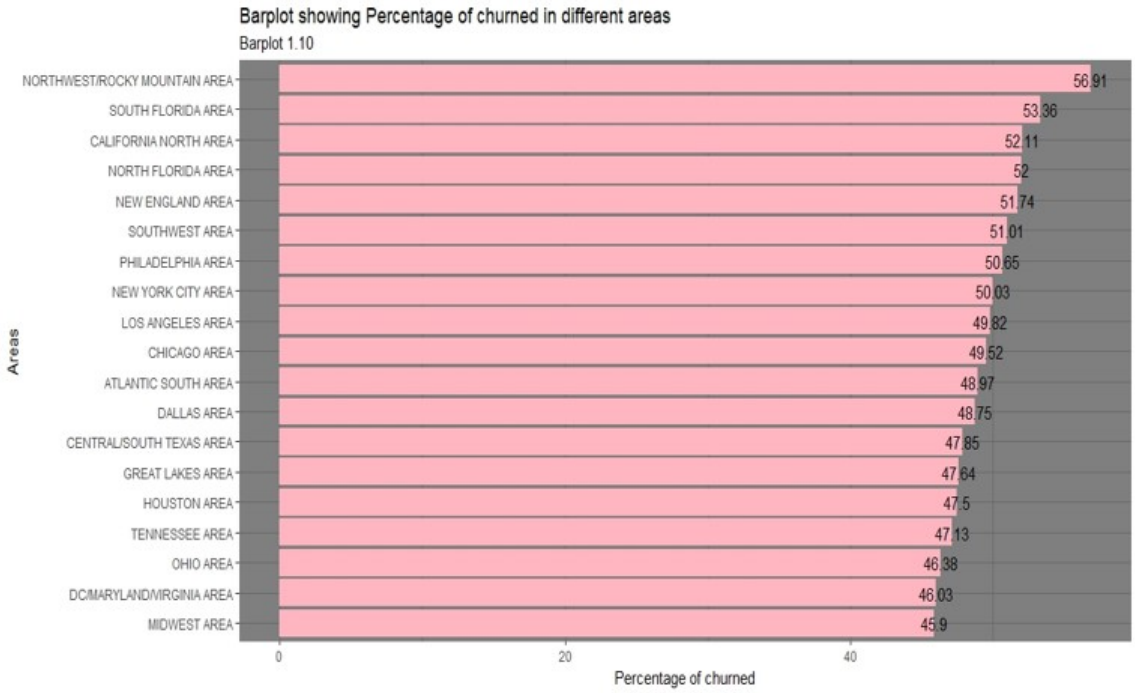
The dataset contains data of 100,000 customers of a telecom company. Per customer there are 173 variables available. From the 173 variables there are 128 variables that are numeric and 45 that are character values. Of the 100,000 people there are 49,562 people that have churned and 50,438 people are(still) a current customer (non-churned). Churn status 0 indicates non-churned and 1 indicates churned. There are missing cells in both numeric and categorical columns.

Data Source

Link for Data Source

Chapter 5

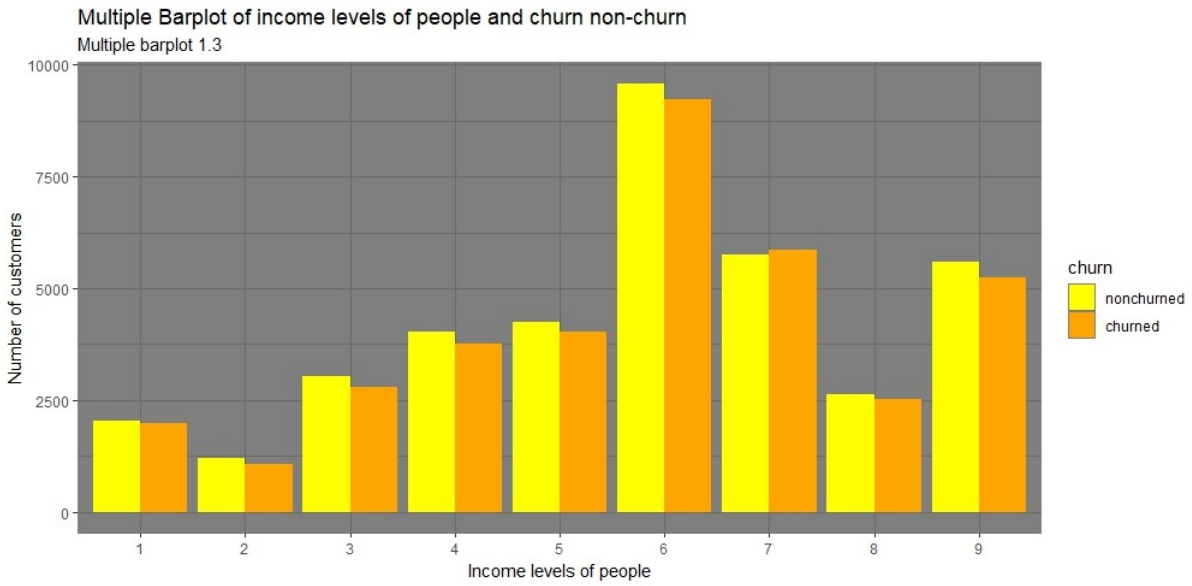
Results



Inference

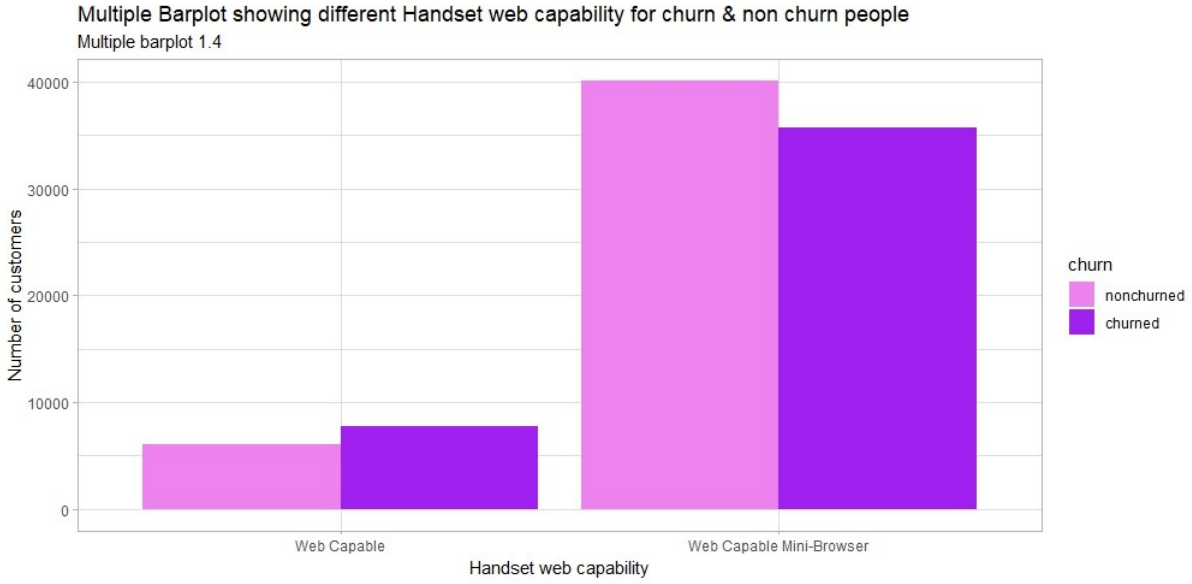
Highest proportion of churned : Northwest/Rocky Mountain area

Lowest proportion of churned : Midwest Area.



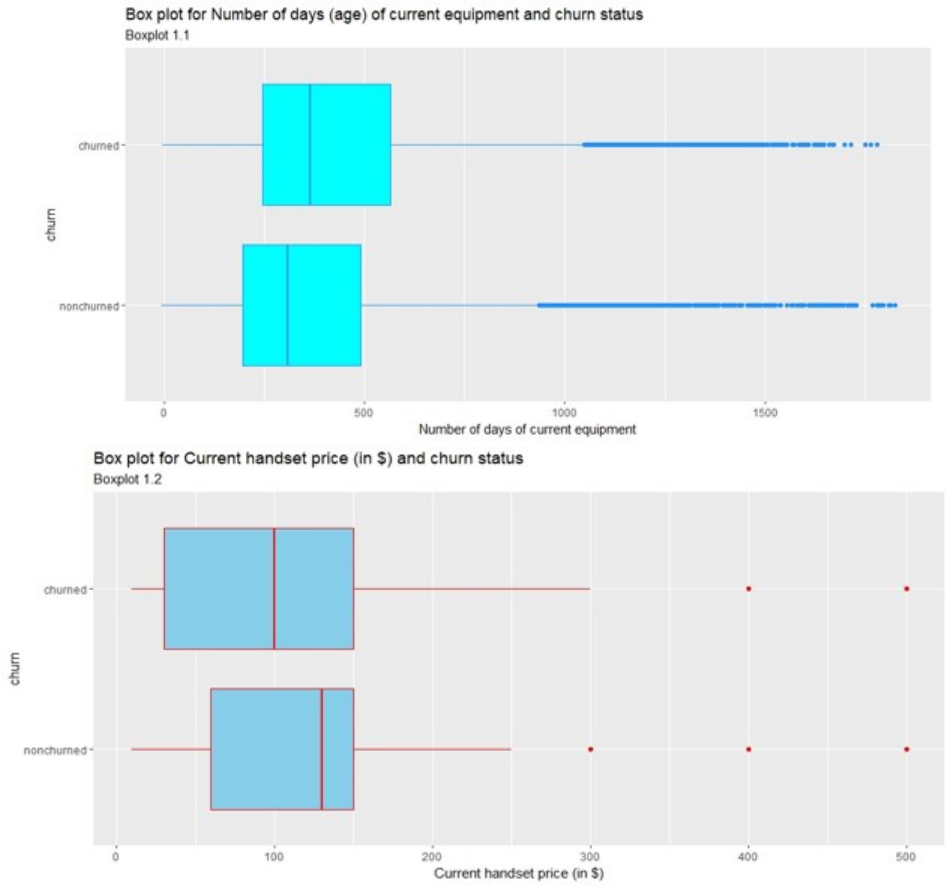
Inference

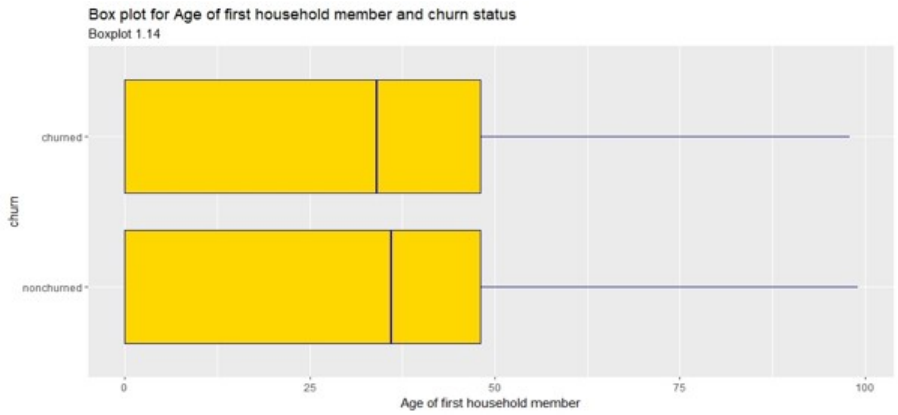
Higher number of non-churned in all income levels except income level 7 ($75,000-99,999$) than customers than churned.

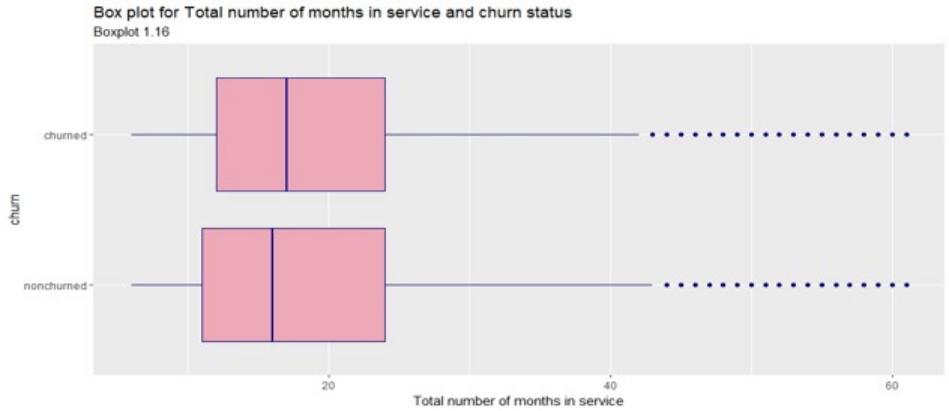


Inference

For Web Capable Mini-browser, non-churned customers are relatively more than Web Capable. Opposite.





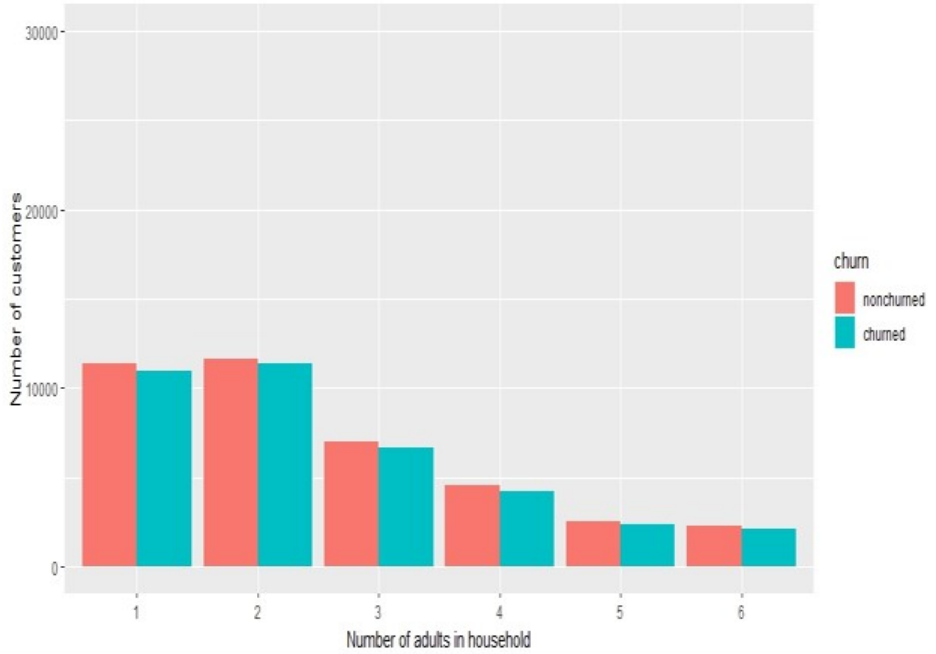


Inference

Some covariates have effect on churn status namely 'age of current equipment', 'current handset price', 'age of first household member' and 'total number of months in service'.

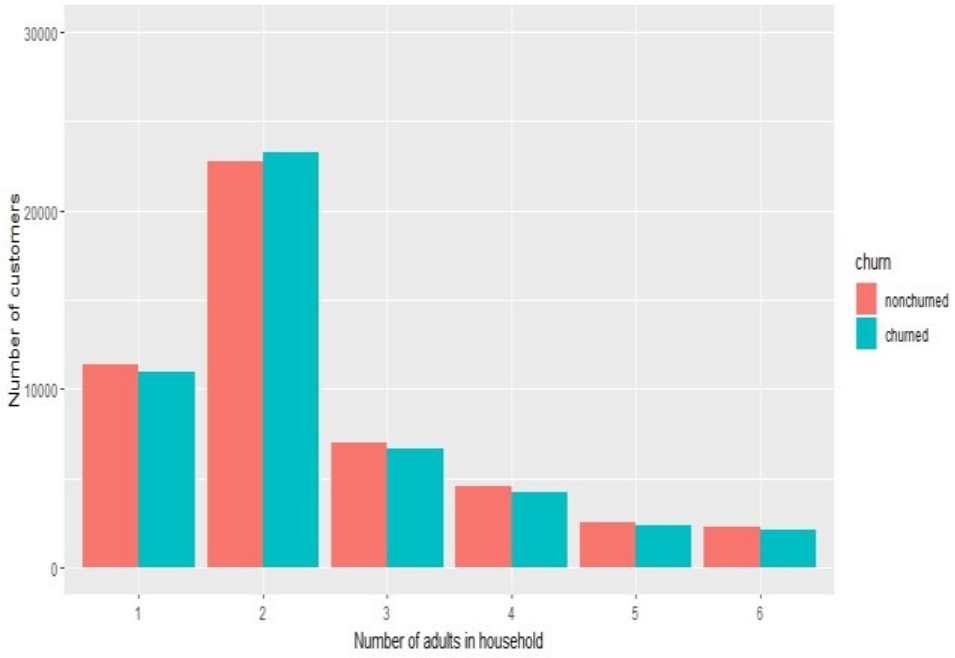
Multiple Barplot of household adults and churn non-churn

Multiple barplot 1.8



Multiple Barplot of household adults and churn non-churn after substitution of missing values

Multiple barplot 1.7



Inference

Missing category imputation has caused more churned in modal class

Note:

These EDA are not done here overall but can be found from another report provided as a supplement.

Prediction by different methods

Logistic regression

Test Accuracy: 0.5692

K Nearest Neighbour (KNN)

Test Accuracy: : 0.5725

Decision Tree

Test Accuracy: 0.5451

Random forest

Test Accuracy: 0.6153

Support Vector Machine (SVM)

Test Accuracy: 0.5881

Multi Layer Perceptron

Test Accuracy: 0.5891

XG Boost

Test Accuracy: 0.6205

Chapter 6

Conclusion & Future Work

Conclusion

We found that for classification task Logistic regression is a good probabilistic model but in this problem, XG Boost, MLP and Random forest has dominated it due to the fact that this models address higher complexities. Decision Tree has worst accuracy but note that it is non-parametric. Features which contributed towards churn are 64 in number. Overall, the data is very dirty with many missing values and outliers as well.

Future Work

Finding how categorical columns can be used for classification purpose. And Outlier analysis can be done for further study.

Bibliography

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[3] Pádraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers - a tutorial. *ACM Computing Surveys*, 54(6):1–25, July 2021.

[4] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[5] Marius-Constantin Popescu, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8, 07 2009.

[6] Lior Rokach and Oded Maimon. *Decision Trees*, volume 6, pages 165–192. 01 2005.