Analysis of Telco Customer Churn using Statistical and Machine Learning Techniques

by

Group 12:
Parth Keyur Gawande(774001701)
Sarthak Rajendra Bora(774003002)

Labor distribution among team members

Parth Keyur Gawande	**Sarthak Rajendra Bora**
Data Preprocessing	EDA
Encoding and Scaling	Clustering Model
Regression Models	Anomaly Detection
Classification Models	Advanced Method

Foundation of Data Science and Analytics DSCI: 633.
Semester: 2225 (Spring 2023)

Rochester Institute of Technology

B. Thomas Golisano College
of
Computing and Information Sciences

School of Information

05/01/2023

Abstract

The telecommunication industry has a high rate of customer churn, which can significantly impact revenue and profitability. The Telco Customer Churn dataset offers valuable insights into the factors that contribute to customer churn and provides an opportunity for developing effective retention strategies. This study aimed to develop and evaluate several machine learning models, including regression models and classification models, to predict monthly charges and customer churn in a telecommunications company. Statistical tests and unsupervised clustering algorithms were used to identify significant factors and gain insights into customer behavior and preferences. The effectiveness of LOF anomaly detection methods was also evaluated. The study highlighted the limitations of the dataset, such as its small size and class imbalance, and provided recommendations for future studies. Despite these limitations, the Telco Customer Churn dataset remains a valuable resource for studying customer churn in the telecom industry and developing effective retention strategies.

Table of Contents

List of Figures

List of Tables

Introduction

Background and context:

The telecommunications industry has reportedly grown extremely competitive as a result of technological advancement and an increase in operators[1]. Customer churn, or the rate at which customers transfer to other service providers, is a big risk for telecom firms. To effectively retain customers, telecommunications businesses must be able to predict customer attrition. Getting new clients is more expensive than keeping your current ones. To solve this issue, telecom firms analyze customer data to recognize customers who are at risk of leaving and take proactive steps to keep them. Companies have created a variety of techniques to increase income in order to survive in this market, including gaining new consumers, upselling to current customers, and extending client retention periods[1]. The most profitable method, according to studies, is to concentrate on extending client retention periods since it has a larger return on investment (RoI) than the other two techniques. This is because keeping a current client costs less money and is simpler to do than upselling than getting a new one[2,3]. Companies must try to reduce the chance of customer churn, or the switching of customers from one operator to another, in order to successfully apply the customer retention strategy[4].

Dataset Description:

In order to forecast customer churn—the possibility that a client would stop using the company's services—a telecommunications corporation has compiled a dataset called Telco Churn https://www.kaggle.com/datasets/blastchar/telco-customer-churn. 7,043 observations across 21 variables make up this dataset.: The features can be comprised as:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Project Goals:

The project aims to work on the dataset Telco Customer Churn to develop and evaluate several machine learning models, including regression models like Linear Regression,Decision Tree, K-NN using performance metrics like MAE, RMSE and R^2 score to predict monthly charge and classification models like logistic regression, decision trees, random forests, and support vector machines using Accuracy, Precision, Recall and F1 scores on customer churn in a telecommunications company. The study will identify significant factors that contribute to customer churn using statistical tests and feature engineering methods, and unsupervised clustering algorithms will be used to gain insights into customer behavior and preferences. The effectiveness of LOF anomaly detection methods will also be evaluated. The performance of the classification models will be improved using gradient boosting and scaling and transformation methods. The study will provide recommendations to reduce customer churn and improve customer retention in the telecom industry.

Methodology

- **Data preprocessing and EDA-** We started with data preprocessing on the Customer Churn Dataset , we filtered the duplicate and null values , converted the 'TotalCharges' column to numeric,plotted bor-plots and density graphs w.r.t the target variable Churn.

- **Clustering -** Used K- means clustering to get patterns on customer behavior based on the tenure and monthly charges features, we normalized the numeric columns for ease of calculation ,used the elbow method to find the optimal k value and plotted the clusters with effective insights by demographic analysis.

- **Encoding and Scaling -** Encoded the categorical features manually at first to get an understanding of the dataset and then used Label Encoding later for regression analysis and One-Hot Encoding for classification through pipelines.

- **Feature Engineering using statistical tests -** For categorical features used the chi-squared tests to get the most important features and for numerical features used the t-test to get features which were further used in the models.

- **Train-test split and k-fold cross-validation -** Splitted the dataset into a train-test split ratio of 80%-20% and 5-fold cross-validation for both regression and classification models.

- **Regression -** Used several Regression models for predicting the monthly charges values namely, linear regression, k-nn regression, and decision tree regression and performed hyperparameter tuning and got several performance metrics namely MAE, RMSE, MSE, and R^2 score for further comparing the models to get the best-fit models for prediction

- **Classification -** Used several classification techniques for Predicting whether a customer is more likely to churn or not , namely, Logistic Regression, K-NN Classification, Decision Tree Classification, SVM Classification. Performed hyperparameter tuning for each model and compared the models on the basis of their accuracy,precision, recall and F1 score.

- **Advanced method -** For the classification technique, we used pipelines, where we used column transformer to apply OneHotEncoder to categorical columns and StandardScaler to numerical columns and then defined hyperparameters for tuning via GridSearch, Furthermore to get better results we used Gradient Boosting classifier for improved accuracy.

- **Outlier/ Anomaly detection -** Used LOF method for anomaly and outlier detection and its effectiveness on the dataset.

Results

1) EDA

The churned customers tend to have a lower mean tenure and higher mean monthly charges compared to non-churned customers. This suggests that high monthly charges could be a contributing factor to customer churn.

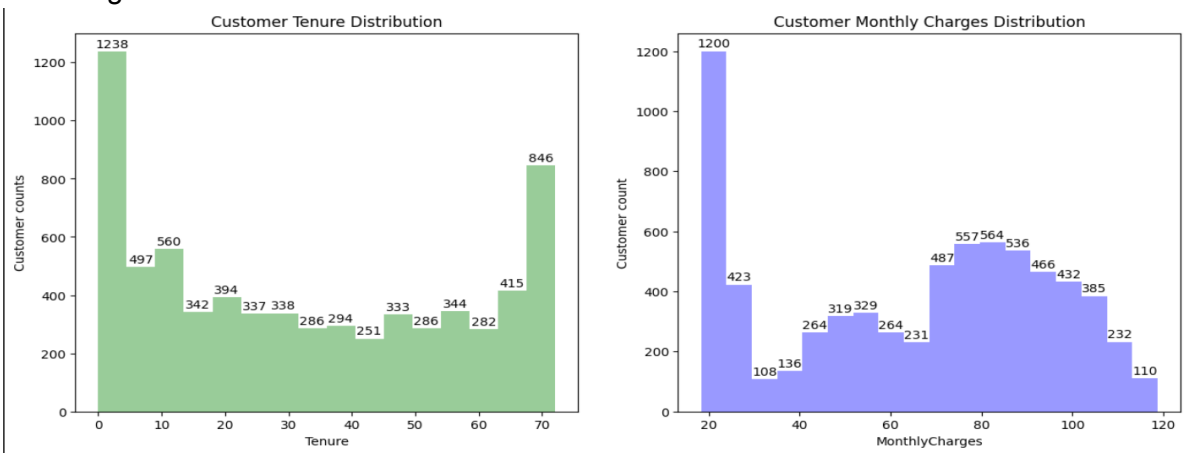


Figure 1: Tenure and Monthly Charges distribution for customers

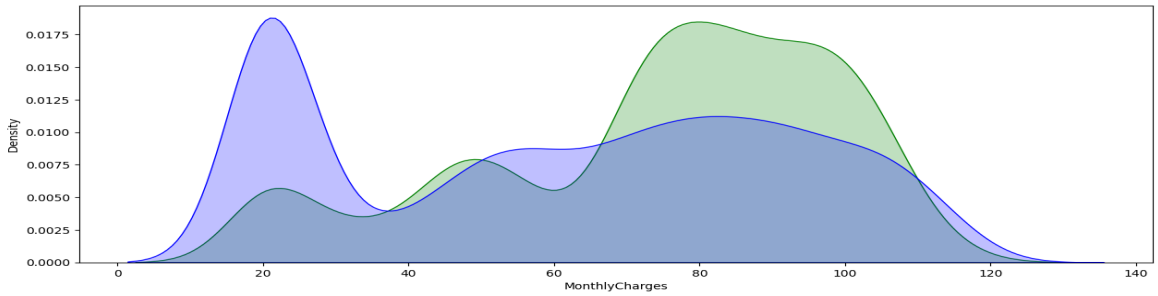


Figure 2: Density Plot distribution of churn and non churn customers on tenure and monthly charges

The distribution of monthly charges for both churned and non-churned consumers is seen in the density plot above. Customers who have churned are represented by the green distribution, while those who have not churned are represented by the blue distribution.It appears that customers who have churned tend to have higher monthly charges as compared to those who have not churned. Therefore, it could be a factor that triggers customers to leave the brand.

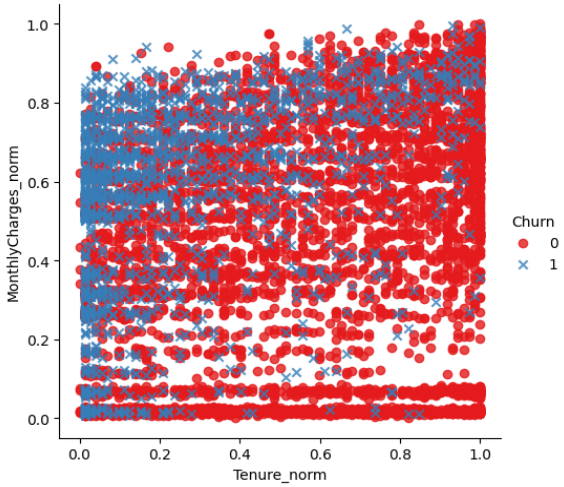


Figure 3: Churn and Non-churn scatter plot

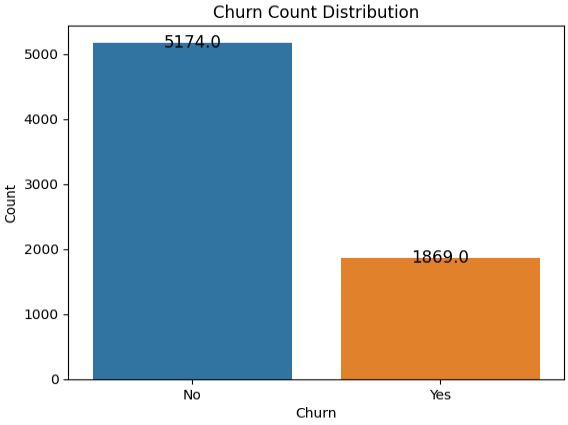


Figure 4 : Churn v/s non churn countplot

The plot above shows that customers who churn tend to have higher monthly charges and lower tenure compared to customers who do not churn. This indicates that customers who are paying higher monthly charges may not be satisfied with the services and are more likely to leave earlier. The above countplot shows the count of customers who churned and those who did not. The count of each category is displayed on top of the bars using annotations. Here we can see that out of the total 7043 customers, 5174 are non-churn customers and 1869 customers who are churn.

2) K-Means Clustering

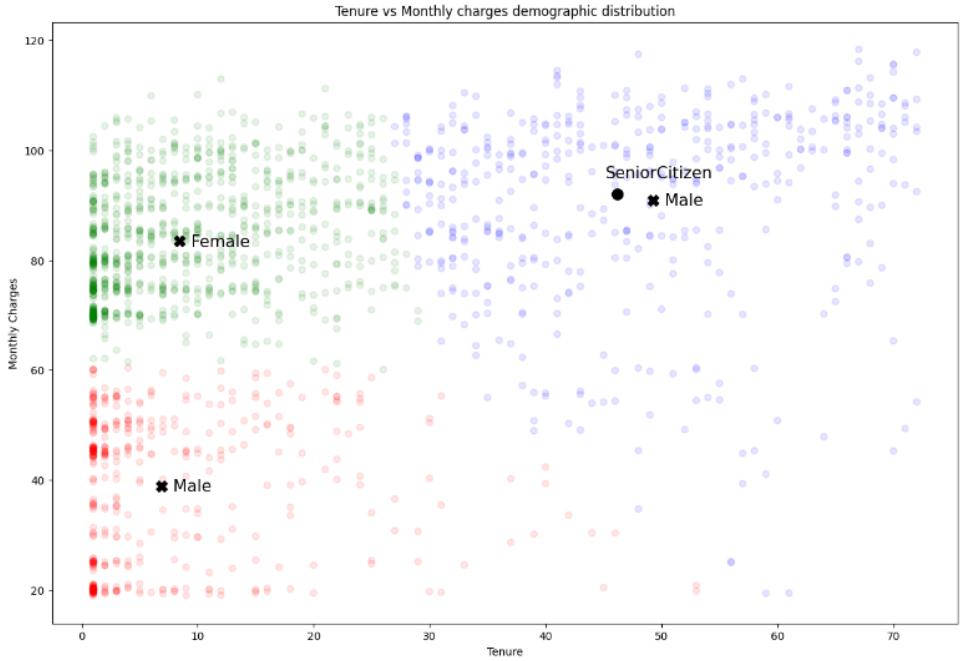


Figure 5: Tenure v/s Monthly Charges demographic distribution

- Cluster 1: Less tenure and high monthly charges - More likely to be Female
- Cluster 2: High tenure and High monthly charges - More likely to be male and senior citizen
- Cluster 3: Less tenure and low monthly charges - More likely to be male

We can see that 78.33% of the dataset's customers have access to internet service. It can be seen that more churned customers (93.95%) than active customers (72.69%) have access to the internet. This implies that customers who have internet service are more likely to leave than those who do not.

3) Feature Engineering using statistical tests

The chi-squared test is applied to each categorical feature in the dataset to evaluate the association between the feature and the target variable 'Churn'. The selected categorical columns are Contract, OnlineSecurity, TechSupport, InternetService, and PaymentMethod. Next, the T-test is performed on all numerical features to evaluate the difference in means between churned and non-churned customers for each feature. The features with the highest absolute t-value are selected as the most important numerical features for the prediction model. The selected numerical feature columns are Tenure, MonthlyCharges, and SeniorCitizen.

4) Regression

Models	MAE	MSE	RMSE	R^2
Linear Regression	0.4152	0.2715	0.5210	0.7286
K-NN Regression	0.1178	0.0353	0.1879	0.9647
Decision Tree Regression	0.0814	0.0157	0.1254	0.9843

Table 1: Performance metrics comparison of regression models

The decision tree regression model has the lowest mean RMSE and highest R^2 score, indicating the best performance among the three models.

5) Classification along with advanced method as Gradient Boosting

Models	Accuracy	Precision	Recall	F1
Logistic Regression	0.7945	0.8005	0.8084	0.8027
KNN Classification	0.7916	0.8032	0.8112	0.8052
Decision Tree Classification	0.7900	0.7899	0.8020	0.7868
Random Forest Classification	0.7966	0.7958	0.8070	0.7943
SVM Classification	0.7952	0.8018	0.8105	0.8037
Gradient Boosting	0.7989	0.7965	0.8062	0.7981

Table 2: Performance metrics comparison of classification models

We can conclude from the table that Gradient Boosting Classifier has the best accuracy of 79.89%.

6) Outlier/ Anomaly detection

There are 325 anomalies detected by the Local Outlier Factor (LOF) algorithm. The majority of the data points (6718 out of 7043) are classified as normal.

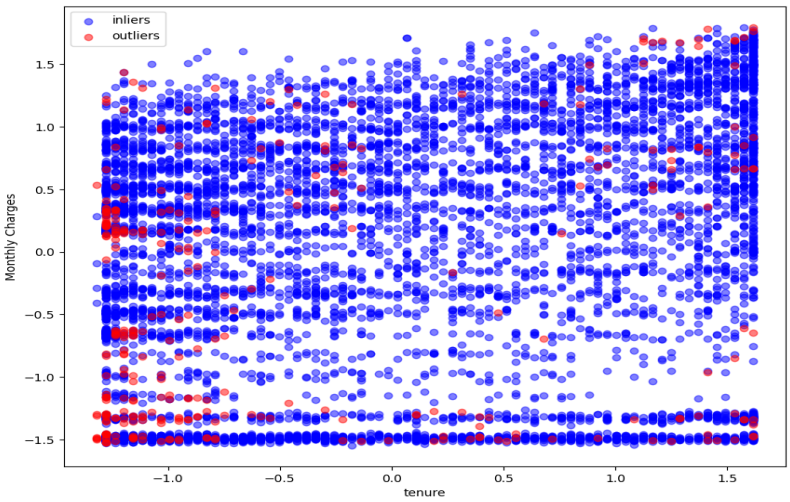


Figure 6: Scatter plot for inliers and outlier data points

Actionable Insights

- According to the data, it seems that high monthly costs have a big role in client attrition. Higher monthly fee customers may not be happy with the services and are more likely to quit the branch sooner.
- Furthermore, customers who have internet access are more likely to leave than those who do not. The analysis also revealed that the most important characteristics for predicting customer churn were Contract, OnlineSecurity, TechSupport, InternetService, PaymentMethod, Tenure, MonthlyCharges, and SeniorCitizen. This data can be used to create specific initiatives to reduce the loss of customers, such as offering discounts or incentives to consumers with high monthly prices or improving internet service quality to retain more users.
- Customers who have phone service are less likely to churn than those who have additional services such as multiple lines or internet service. This emphasizes the importance of knowing the individual demands of customers and modifying services accordingly.
- Senior citizens experience much more churn than non-seniors. Customers on month-to-month contracts churn at a substantially higher rate than those on other contract lengths. Customers without partners have a moderately higher churn rate than those with partners. Customers without children have a much higher churn rate compared to those with children. Customers with fiber optic internet as part of their contract have a much higher churn rate than those without.
- Customers who make automatic payments are less likely to churn, whereas those who pay by electronic check are more likely to churn. This suggests that offering convenient payment options can help retain customers.
- Customers who have a longer tenure with the company but are not on a long-term contract are more likely to churn. This indicates that customers may need more incentives to sign up for longer-term contracts.

Discussion and Critics

- Small dataset size: The dataset contains only around 7,000 records, which may not be representative of the entire population of telecom customers. This can limit the generalizability of the findings.
- Imbalanced class distribution: The dataset has a class imbalance, with only about 27% of customers in the churn category. This can lead to biased model performance and may require techniques such as oversampling or undersampling to address.
- Limited timeframe: The dataset was published five years ago and covers a limited time frame of about six years, which may not be sufficient to capture long-term trends or changes in customer behavior.
- Despite these limitations, the dataset still provides valuable insights into the factors that contribute to customer churn in the telecom industry and offers opportunities for developing effective retention strategies. Future studies could expand the dataset to include more recent data and a larger sample size to increase the robustness and generalizability of the findings.

Conclusion

In conclusion, this project analyzed the Telco Customer Churn dataset and developed several machine learning models to accurately predict customer churn and monthly charges in a telecommunications company. The study also identified significant factors that contributed to customer churn and gained insights into customer behavior and preferences using unsupervised clustering algorithms. Based on the findings, recommendations were provided to reduce customer churn and improve customer retention in the telecom industry.

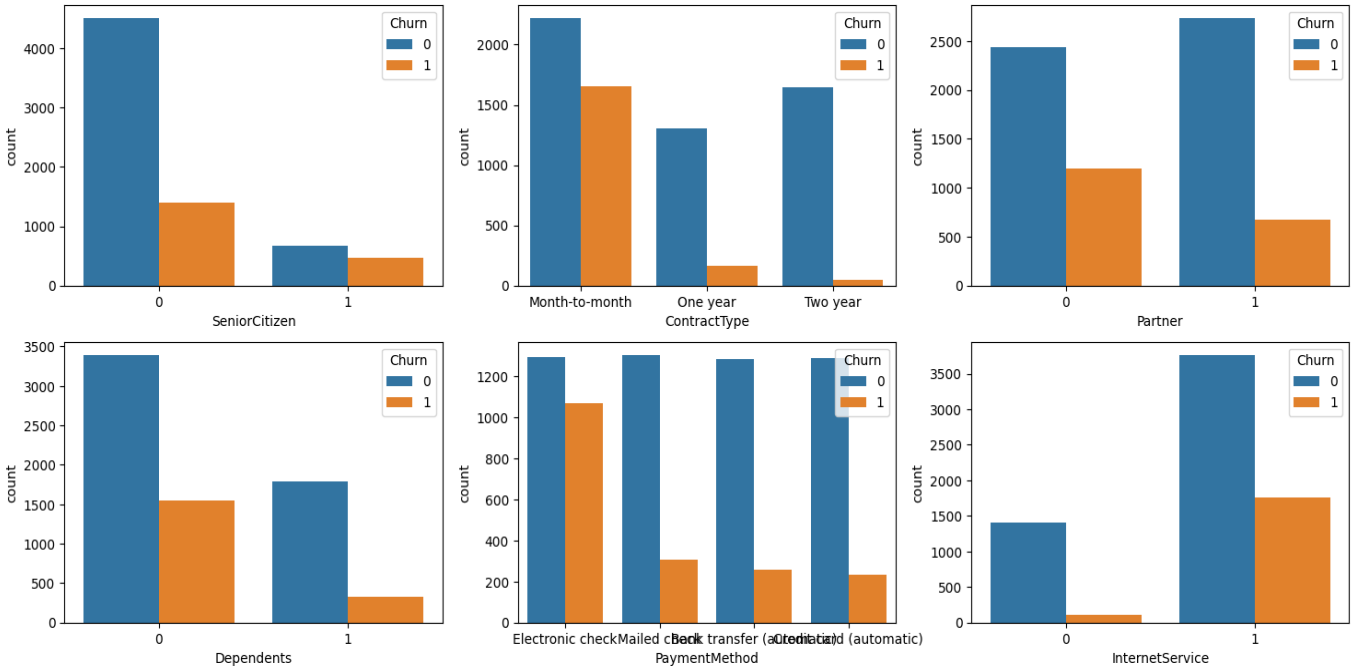
Future work for the study could include expanding to additional datasets or covering a longer time period to capture long-term trends and changes in customer behavior. Further investigation could also be conducted on the identified factors contributing to customer churn to determine the root causes and develop targeted interventions to improve customer retention. Additionally, neural networks and deep learning techniques could be explored to improve the performance of the models and provide more accurate predictions.

References

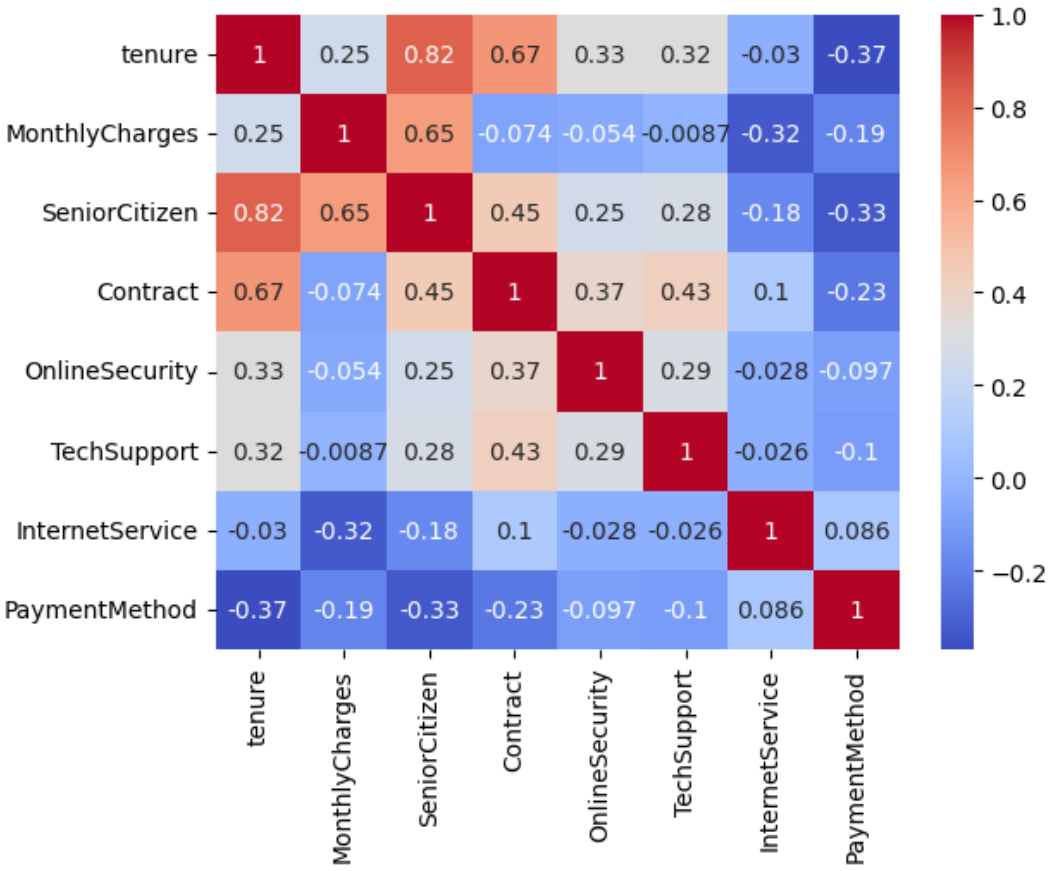
[1] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0191-6
[2] Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013). Telecommunication subscribers' churn prediction model using machine learning. In *International Conference on Digital Information Management*. https://doi.org/10.1109/icdim.2013.6693977
[3] Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment. *Journal of Marketing Research*, *53*(1), 46–60. https://doi.org/10.1509/jmr.13.0483
[4] Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer perceptron neural networks: Modeling and analysis. *Life Science Journal*, *11*(3), 75-81.
[5] Frohböse, F. (2022). Machine Learning Case Study: Telco Customer Churn Prediction. *Medium*. https://towardsdatascience.com/machine-learning-case-study-telco-customer-churn-prediction-bc4be03c9e1d

Appendices

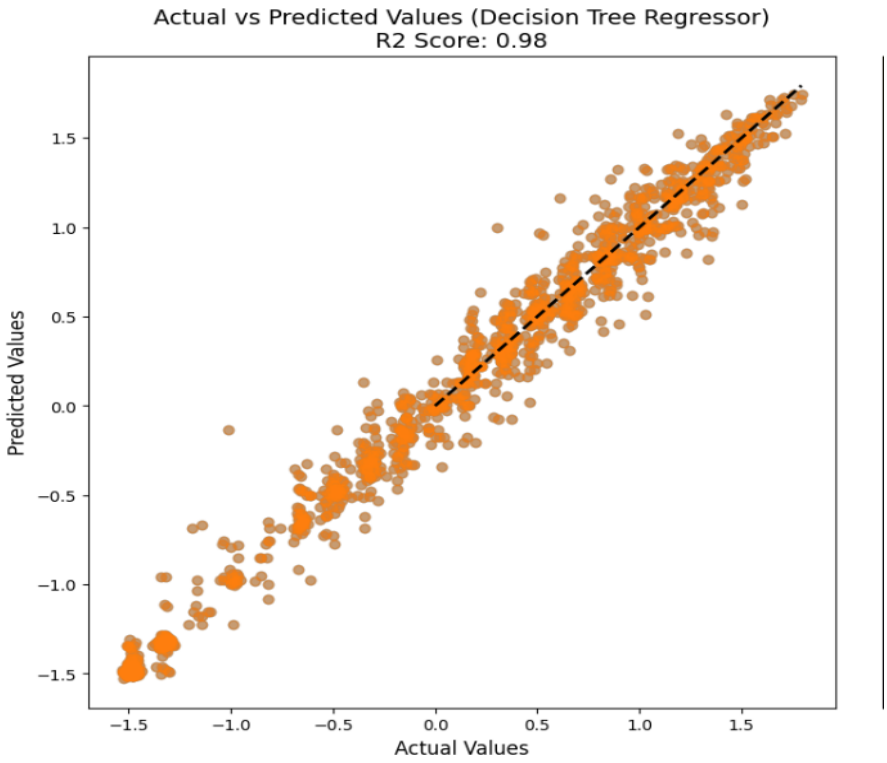
- Churn v/s Non-Churn customer distribution based on important features



- Correlation Heatmap of the 8 most important features



- Decision Tree Regressor as the best regression model



- Bar Graph to compare accuracies of various classification model