DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING

FACULTY OF ENGINEERING
UNIVERSITY OF RUHUNA

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE DEGREE OF THE BACHELOR OF THE SCIENCE OF ENGINEERING HONOURS

19th APRIL 2023

Customer churn prediction in telco context

Group 14

LIYANAGE D.L.S.B. (EG/2019/3655)
WEERAWARDHANE W.A.S.V. (EG/2019/3775)

Contents

List of Figures

Acronyms

ML	-	Machine Learning

Chapter 1

Customer churn prediction in telco contex

1.1 Introduction

With a number of telecommunication providers in this vastly competitive telecommunication industry, customer retention is of utmost importance when it comes to sustaining business growth and profitability. Customer churn, which is the phenomenon of customers leaving a service provider, presents a significant challenge for telco companies worldwide. Recognizing the critical importance of mitigating customer churn, this project titled "Customer churn prediction in telco context" aims to leverage machine learning to anticipate and address this issue proactively.

The primary objective of this project is to develop a classification model capable of identifying customers at a higher risk of churn. By analyzing a comprehensive dataset [1] containing various customer attributes, service subscriptions, account information and demographic details, the project seek to uncover patters and insights that will help telco companies to implement retention strategies to minimize customer churn.

Through exploratory data analysis, feature selection, model building and interpretation, this project aims to provide proactive insights to telco companies. By understanding the factors that contribute to customer churn and predicting churn probabilities accurately, companies can design effective strategies to retain valuable customers in advance before they leave.

The report dives into the methodology, results, discussion and conclusion derived from the project.

1.2 Methodology

1.2.1 Data Loading and Preprocessing

1. **Initial Data Exploration**

   The initial exploration of the dataset involves examining its structure and contents. This includes displaying the first few rows of the dataset to get a glimpse of the data, checking basic statistics (such as mean, median, and standard deviation) to understand the distribution of numerical features, and examining the data types of each column to identify categorical and numerical variables [2].

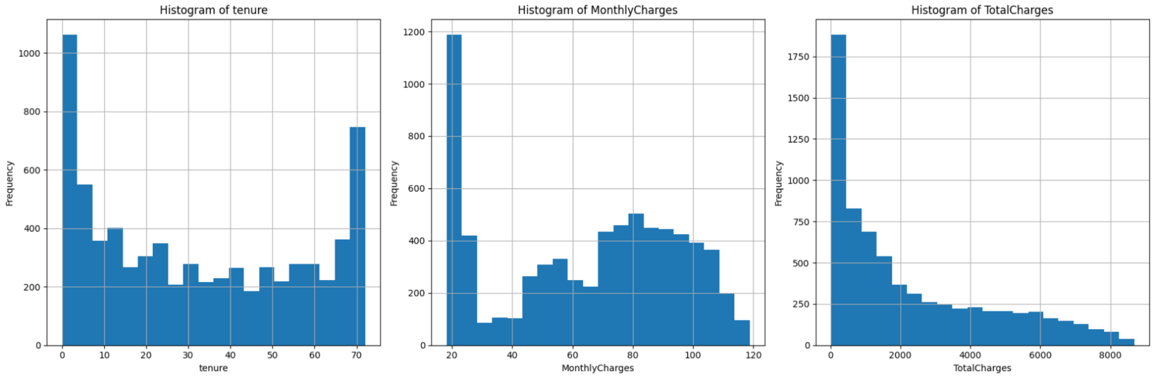


Figure 1.1: Histograms of numerical features

2. **Handling Non-numeric Values**

   Certain columns, such as 'tenure', 'MonthlyCharges', and 'TotalCharges', may contain non-numeric values initially. These values are identified and handled appropriately. In this case, non-numeric values are likely to represent missing or placeholder values, which need to be addressed before further analysis. In TotalCharges there were 11 non-numeric values, these were later handled to represent as missing values (NaN), so that the missing values can be handled later on [3].

3. **Handling Missing Values**

   Missing values are a common issue in datasets and need to be addressed before modeling. Depending on the extent of missing data and the nature of the problem, missing values can be treated using various methods such as imputation, mean replacement, median replacement (replacing missing values with a suitable estimate) or removal (excluding observations with missing values from the analysis). In this scenario the missing values of TotalCharges were replaced with the mean of the said column [3].

4. **Handling Outliers**

   Outliers are data points that deviate significantly from the rest of the observations in the dataset. They can skew statistical analyses and model predictions if not addressed properly. The Inter-quartile Range (IQR) method is used to identify outliers and subsequently treat them, often by removing or transforming extreme values [4].

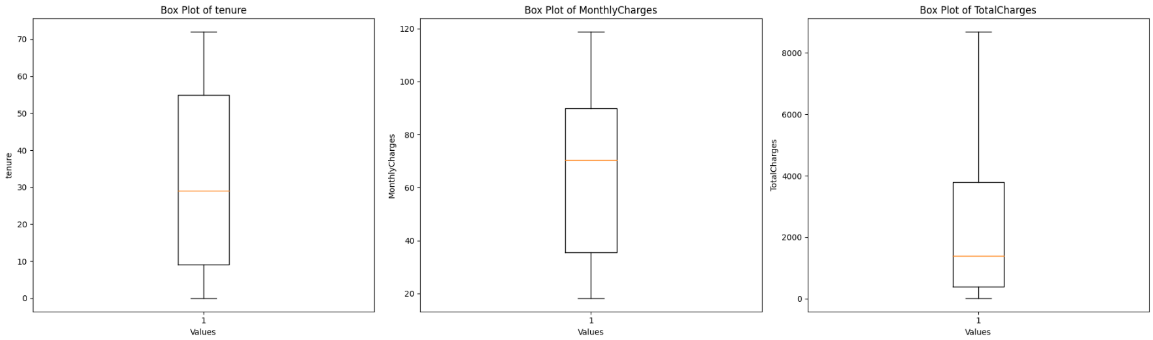


Figure 1.2: Box plots of numerical features

5. **Dealing with categorical features**

   Categorical features such as gender, Partner, Dependents, PhoneService, PaperlessBilling and Churn, had only two variables, either male/female or yes/no. These were mapped into 1 and 0.

6. **Creating Dummy Variables**

   For categorical variables with multiple levels (i.e., more than two categories), dummy variables are created to represent each category as a separate binary variable. MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection,StreamingTV, StreamingMovies, Contract, PaymentMethod were used to create dummy variables. This process, known as one-hot encoding, ensures that each category is appropriately represented in the modeling process without introducing ordinality [5].

7. **Standardizing Continuous Variables**

   Continuous variables are standardized to ensure consistent scales across features. Standardization involves transforming the data such that it has a mean of 0 and a standard deviation of 1. This helps prevent certain features from dominating others during model training and improves the stability of the modeling process [6].

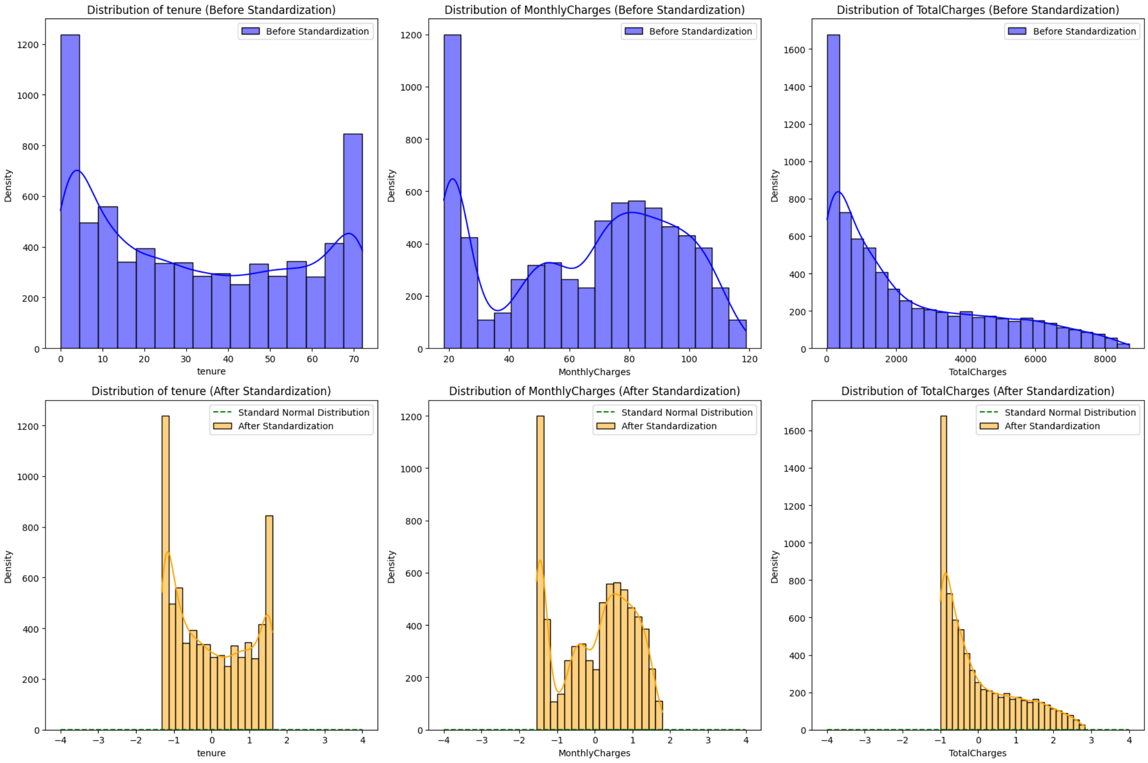


Figure 1.3: Before and after standardization for numerical features

8. **Splitting the Dataset**

   The dataset is split into training and testing sets to evaluate the performance of the predictive models. The training set is used to train the models, while the testing set is used to assess their performance on unseen data. This ensures an unbiased evaluation of model performance and generalization to new data.

9. **Addressing Class Imbalance**

   Class imbalance occurs when one class is significantly under represented compared to another class in the dataset. In the dataset, Non-churners - 0 count was 4113, while the churners - 1 count was 1521. To address this issue and prevent models from being biased towards the majority class, techniques such as Synthetic Minority Over-sampling Technique (SMOTE) are used to balance the distribution of target classes in the training data. SMOTE generates synthetic samples of the minority class to create a more balanced dataset for model training [7].

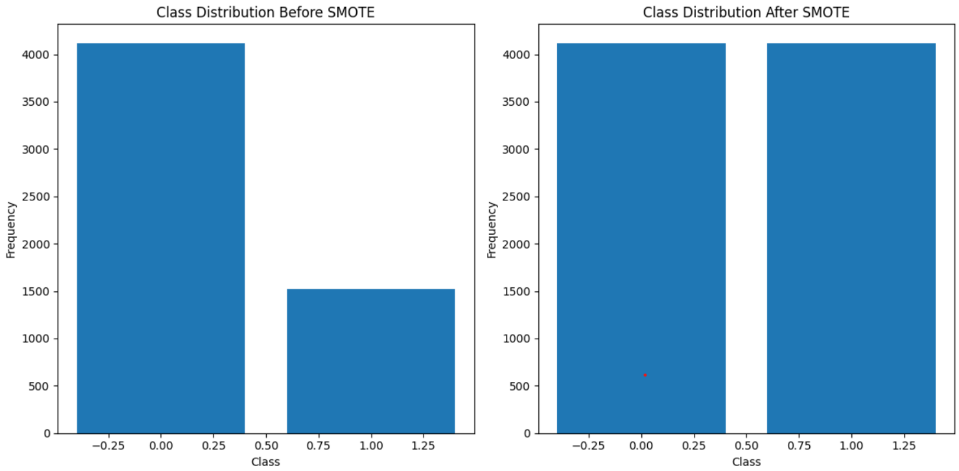


Figure 1.4: Addressing Class Imbalance

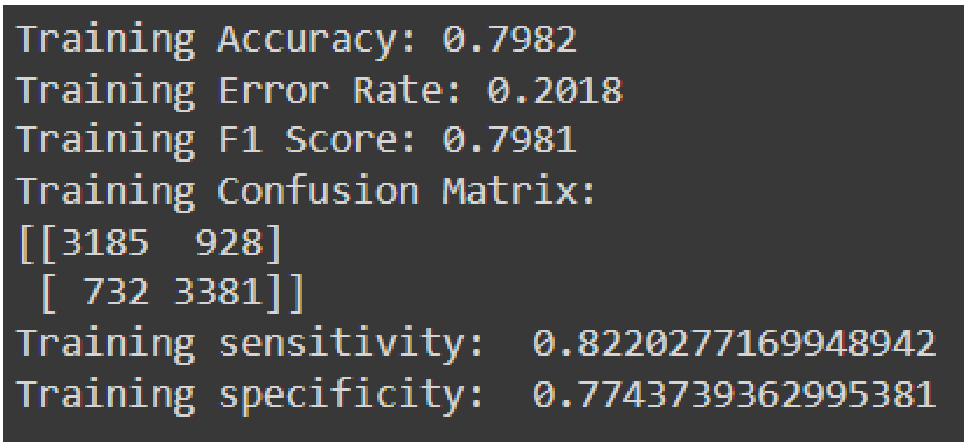
1.2.2 Model Training

Two classification algorithms, namely Logistic Regression [8] and Support Vector Machine (SVM) [9], are chosen for training on the preprocessed data. These algorithms are selected based on their suitability for binary classification tasks and their common usage in predictive modeling. The selected classification algorithms, Logistic Regression and SVM, are trained on the preprocessed dataset. During training, the models learn patterns and relationships between the input features and the target variable (churn) from the training data. This process involves adjusting model parameters to minimize prediction errors and optimize performance.

1.3 Results

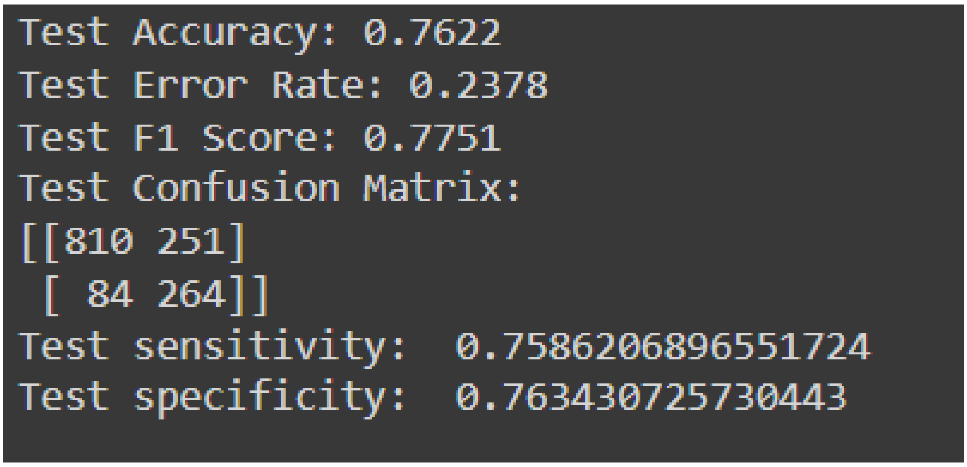
1.3.1 Performance Metrics Summary

- Logistic Regression



```
Training Accuracy: 0.7982
Training Error Rate: 0.2018
Training F1 Score: 0.7981
Training Confusion Matrix:
[[3185  928]
 [ 732 3381]]
Training sensitivity:  0.8220277169948942
Training specificity:  0.7743739362995381
```

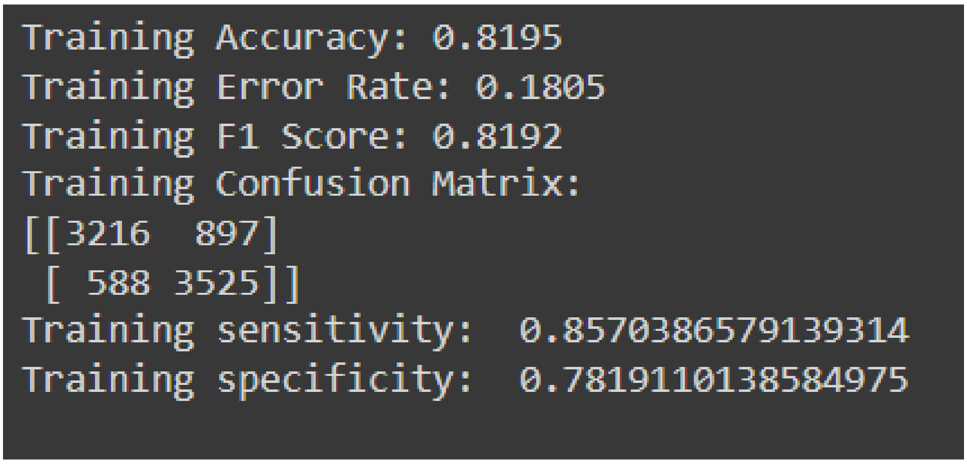
Figure 1.5: Performance metrics of training model of logistic regression



```
Test Accuracy: 0.7622
Test Error Rate: 0.2378
Test F1 Score: 0.7751
Test Confusion Matrix:
[[810 251]
 [ 84 264]]
Test sensitivity:  0.7586206896551724
Test specificity:  0.763430725730443
```

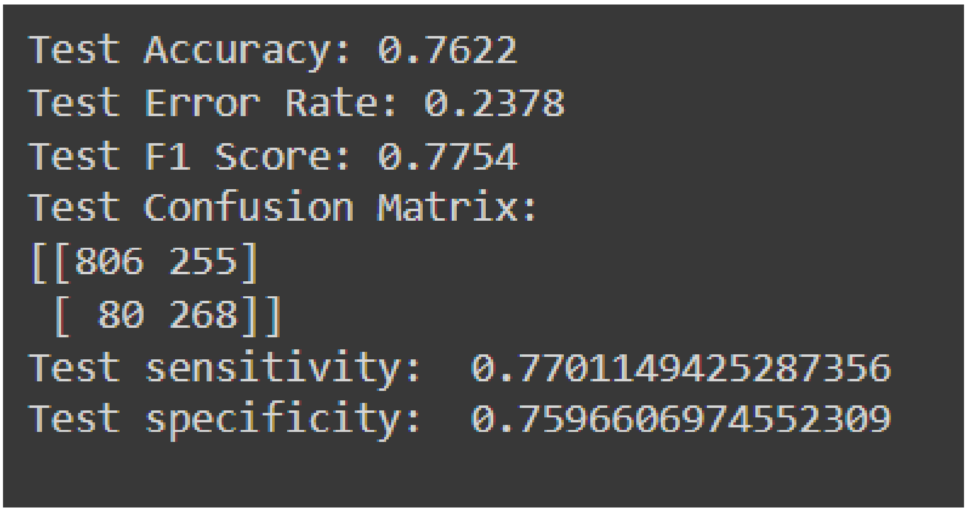
Figure 1.6: Performance metrics of testing model of logistic regression

- Support Vector Machine - SVM



```
Training Accuracy: 0.8195
Training Error Rate: 0.1805
Training F1 Score: 0.8192
Training Confusion Matrix:
[[3216  897]
 [ 588 3525]]
Training sensitivity:  0.8570386579139314
Training specificity:  0.7819110138584975
```

Figure 1.7: Performance metrics of training model of SVM

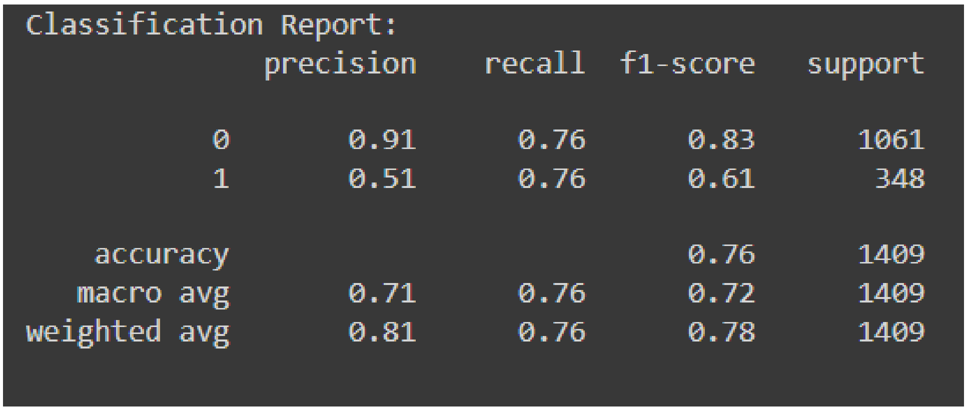


```
Test Accuracy: 0.7622
Test Error Rate: 0.2378
Test F1 Score: 0.7754
Test Confusion Matrix:
[[806 255]
 [ 80 268]]
Test sensitivity:  0.7701149425287356
Test specificity:  0.7596606974552309
```

Figure 1.8: Performance metrics of testing model of SVM

1.3.2 Classification report

- Logistic Regression



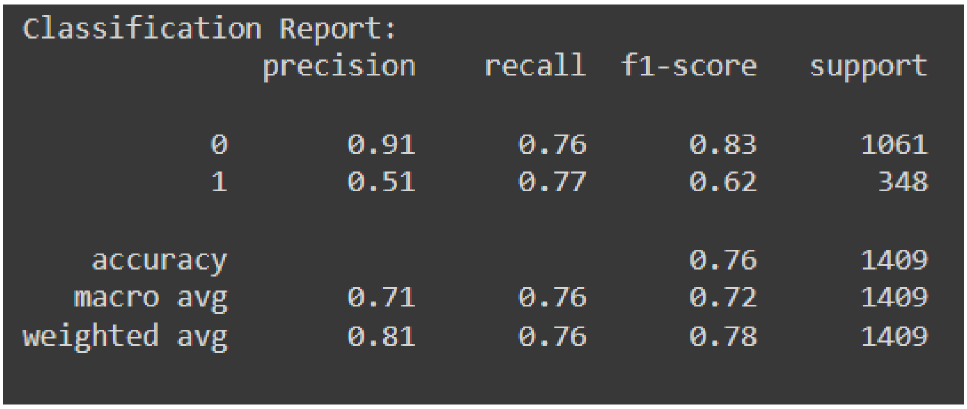
```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.76      0.83      1061
           1       0.51      0.76      0.61       348

    accuracy                           0.76      1409
   macro avg       0.71      0.76      0.72      1409
weighted avg       0.81      0.76      0.78      1409
```

Figure 1.9: Classification report of logistic regression

- Support Vector Machine - SVM



```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.76      0.83      1061
           1       0.51      0.77      0.62       348

    accuracy                           0.76      1409
   macro avg       0.71      0.76      0.72      1409
weighted avg       0.81      0.76      0.78      1409
```

Figure 1.10: Classification report of training model of SVM

1.3.3 Feature importance plot

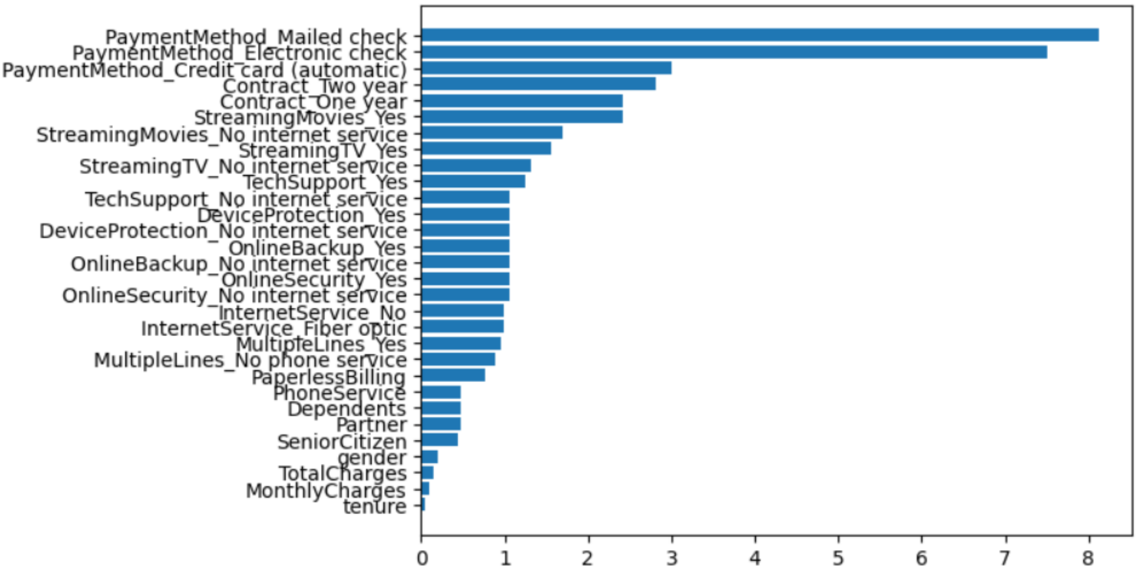


Figure 1.11: Feature importance plot

1.3.4 AUROC (Area Under the Receiver Operating Characteristic Curve).

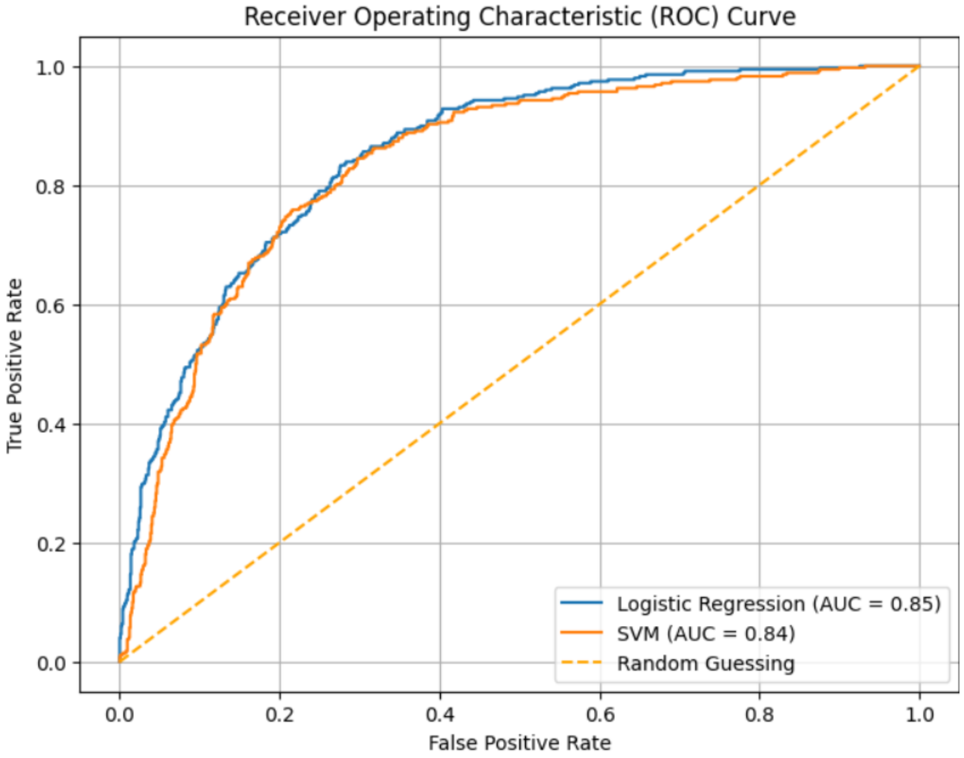


Figure 1.12: AUROC curve

1.4 Discussion and Conclusion

In this study, we trained and evaluated both Logistic Regression and Support Vector Machine (SVM) models for predicting customer churn in the telecom industry.

Our analysis revealed the following key findings:

- **Model Performance**

  Both Logistic Regression and SVM models exhibited satisfactory performance metrics on both the training and testing sets. Logistic Regression demonstrated marginally higher accuracy and F1 score on the testing set, while SVM also showed competitive results.

- **Feature Importance**

  Analysis of feature importance highlighted variables such as payment type and contract type as significant predictors of customer churn. These insights can guide targeted retention strategies and enhance customer satisfaction.

- **Model Comparison**

  While Logistic Regression showed a slight advantage in predictive performance, SVM offers an alternative modeling approach that may capture nonlinear relationships more effectively. The choice between the two models depends on specific requirements and constraints.

Overall, our findings underscore the importance of employing machine learning techniques to proactively address customer churn in the telco industry. By leveraging insights from predictive models, telco companies can optimize retention efforts, improve service offerings, and foster long-term customer loyalty in a competitive market landscape. Further research could explore ensemble methods or incorporate additional data sources to enhance predictive accuracy and robustness.

References

[1] “Telco Customer Churn-LogisticRegression — kaggle.com.” https://www.kaggle.com/code/farazrahman/telco-customer-churn-logisticregression/input. [Accessed 18-04-2024].

[2] “1. Supervised learning — scikit-learn.org.” https://scikit-learn.org/stable/supervised$_l$*earning.htmlsupervised* − *learning*. [*Accessed*18 − 04 − 2024].

[3] “6.4. Imputation of missing values — scikit-learn.org.” https://scikit-learn.org/stable/modules/impute.html: :text=Missing [Accessed 18-04-2024].

[4] “2.7. Novelty and Outlier Detection — scikit-learn.org.” https://scikit-learn.org/stable/modules/outlier$_d$*etection.html* : : *text* = *One* [*Accessed*18 − 04 − 2024].

[5] “sklearn.preprocessing.OneHotEncoder — scikit-learn.org.” https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. [Accessed 18-04-2024].

[6] “sklearn.preprocessing.StandardScaler — scikit-learn.org.” https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html: :text=Standard [Accessed 18-04-2024].

[7] “SMOTE &x2014; Version 0.12.2 — imbalanced-learn.org.” https://imbalanced-learn.org/stable/references/generated/imblearn.over$_s$*ampling.SMOTE.html.*[*Accessed*18 − 04 − 2024].

[8] “sklearn.linear$_m$*odel.LogisticRegression* − − − *scikit* − *learn.org."https* : *//scikit* − *learn.org/stable/modules/generated/sklearn.linear*$_m$*odel.Logisti* 04 − 2024].

[9] “sklearn.svm.SVC — scikit-learn.org.” https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html. [Accessed 18-04-2024].