Customer Retention Analysis

Abstract

The rate at which consumers decide to discontinue doing business with a firm is known as customer churn. Churn is the rate at which subscribers abandon their services and switch to a rival in the telecommunications industry. Customer retention is critical to a company's success, especially in a competitive field like cellular services. Getting new clients is not only more difficult, but also more expensive for businesses than keeping existing ones. So for this project, we have used the dataset of an internet service provider **Telco** extracted from Kaggle. It helps to understand the links between the attributes and the goal variable, as well as identify variables that influence customer attrition prediction. We tried to answer some simple questions to understand the data and visualize the variables more clearly. We also developed a predictive model using various Analytical Algorithms such as Logistic Regression, Support Vector Machines, and Random Forest. This is to assist the organization in proactively reducing churn and using the model's findings to boost client retention tactics.

Keywords : Churn, visualization, Smote, EDA, Logistic Regression, SVM, Random Forest, ROC

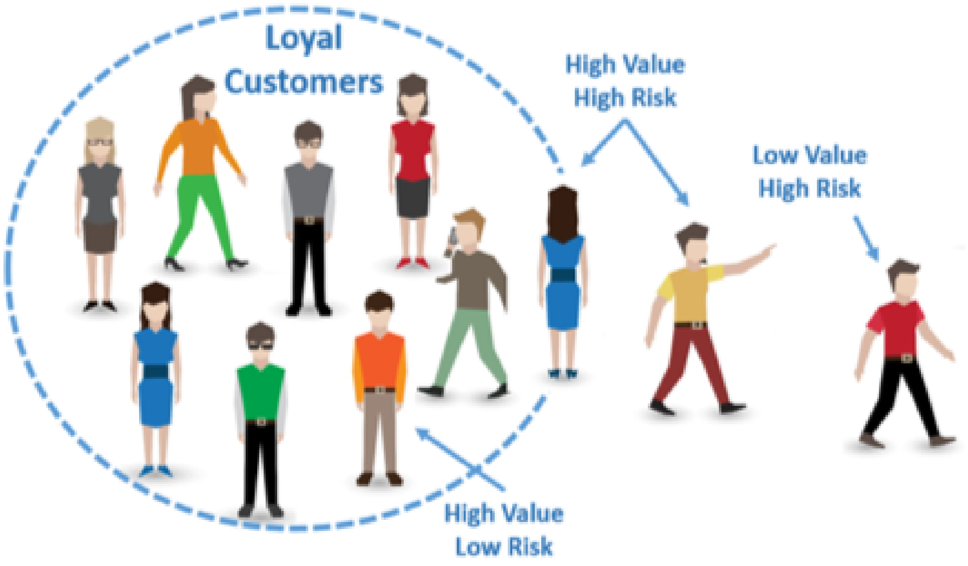
1. Introduction

Customer churn is the rate at which customers decide to stop doing business with a company. For a telecommunications company, churn would be the rate at which subscribers drop their services and leave for a competitor.

Retaining customers is key to a company's success, especially in an industry as competitive as wireless services. Acquiring new customers is not only more difficult but also much more costly to companies than maintaining existing customer relationships. A high churn rate could adversely affect profits and impede growth. The churn rate is an important factor in the telecommunications industry. In most areas, many of these companies compete, making it easy for people to transfer from one provider to another. The churn rate not only includes when customers switch carriers but also includes when customers terminate service without switching. This measurement is most valuable in subscriber-based businesses in which subscription fees comprise most of the revenues. It is an important metric in business, especially in the telecommunications industry, since it is more expensive to sign on new customers rather than retain current ones given marketing budgets aimed toward non-customers.

An analysis and continuous monitoring of customer churn can help companies pinpoint weaknesses and strengths in the customer attrition strategy. An analysis of who is more likely to leave can also help generate creative solutions for customized services and service packages.

What is considered a good or bad churn rate can vary from industry to industry



CUSTOMER SATISFACTION
TIMELY RESPONSE
SERVICE WITH A SMILE
CONTINUOUSLY IMPROVE CUSTOMER EXPERIENCE
100% SERVICE
100% QUALITY
A HAPPY customer is a LOYAL customer
Value the customer`s FEEDBACK
FOCUS on the CUSTOMER

A company can compare its new subscribers versus its loss of subscribers to determine both its churn rate and growth rate to see if there was overall growth or loss in a specific time period. While the churn rate tracks lost customers, the growth rate tracks new customers.

If the growth rate is higher than the churn rate, the company experienced growth. When the churn rate is higher than the growth rate, the company experienced a loss in its customer base.

For example, if in one quarter a company added 100 new subscribers but lost 110 subscribers, the net loss would be 10. There was no growth for the company this quarter but rather a loss. This would be a negative growth rate and a positive churn rate.

2. Background Study

The literature on customers is presented in this section.

Don Jyh-Fu Jeng, Thomas Bailey, has used hybrid, multiple criteria decision-making (MCDM) method to inspect customer retention framework and they found that the most common response was to look at pricing and customer service [1]. Ali Tamaddoni Jahromi, Stanislav Stakhovych, Michael Ewing used models for churn predictions in a B2B context, and to increase the profitability of retention campaigns; he found that boosting model, logistic regression, cost-sensitive and straightforward decision tree is applied on tests [2].

The tweets were divided into three groups using a lexicon-based classifier, and the main churn component was discovered via association rule mining [3]. Ensemble classification approaches with hybrid fuzzy clustering, according to J. Vijaya, E. Sivasankar, and S. Gayathri, give more accuracy and performance than single classifiers and clusters [4]. A distance factor was utilised in classifier determination by Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., and Anwar, S. [5]. Using a model that was confined just four components (pricing, network, customer care, and brand image) to explain customer happiness, Hossain, M.A., Chowdhury, M.R., and Jahan, N. validated the importance of customer satisfaction in developing a relationship between buyer and seller [6].

Adnan, Sajid, Awais, M.Nawaz, K.Alawfi, Amir, and Kaizhu have suggested a realistic strategy to categorise, forecast, and extract crucial decision rules linked to customer churn or not, based on an intelligent rule-based decision-making technique (RST). Exhaustive Method (EA), Genetic Algorithm (GA), Covering Algorithm (CA), and the LEM2 algorithm are used to evaluate the performance of RST (LA). RST based on GA is the most efficient approach for extracting implicit information in the form of decision rules, according to the results [7]. J. Vijaya and E. Sivasankar, on the other hand, combined RST with additional approaches including bagging, random subspace, and boosting.

On a real-time dataset of prepaid telecom customers from south Asia, Azeem et al. [8] developed a churn prediction model for prepaid customers in telecom using fuzzy classifiers, NN, LR, SVM, AdaBoost, and RF techniques were compared with a fuzzy nearest-neighbor classifier to predict an accurate set of churners. The model was used to consider and improve parameters such as TP rate and AUC.Large data volumes, according to the authors, can increase prediction accuracy. It was also recommended that a churn prediction model be built for other applications such as e-commerce.

Baumann et al. [9] constructed an algorithm to extract information from a website's clickstream data and assessed the model's prediction power using data such as number of clicks, repeated visits, repetitive purchases, and so on. The surfing structure of a surfer on a website was discovered through correlation analysis of graphical data. The results of graph metrics and regression analysis showed that the models' predictive capacity increased. The authors proposed that a weighted graph be utilised to increase forecast accuracy and consumer surfing patterns on websites.

Jie et al. [10] used OWA and k-mean clustering to evaluate consumer purchase data in order to enhance customer service. The approach was also used to detect prospective loser candidates. Customers are valued depending on their lifetime worth. The authors proposed demonstrating the updated questionnaire can be used to predict outcomes accurately churners-in-waiting To keep customers, customer service may be enhanced. customers. More impacting elements can be explored influencing customer fulfilment.

In e-commerce, Fridrich [11] suggested an ANN-based optimization model for predicting client turnover using GA. The prediction model was created to identify clients who would leave.
On the basis of factors such as TP rate, FP rate, and accuracy, the suggested model increased customer churn prediction ability. Multiple goal optimizations, enhanced LR, DT, and fuzzy logic, according to the authors, can be applied. In addition, new criteria should be added to e-commerce for improved forecast accuracy.

For Italian online e-commerce clients, Gordini and Veglio [12] created a churn prediction SVM model. For performance comparison, criteria such as recency, frequency, duration, product category, failure, monetary, age, occupation, gender, request status, and so on were used. The suggested technique outperformed LR, NN, and SVM in terms of prediction power, notably for noisy, imbalanced, and nonlinear data. The authors say that the model's staying power cannot be forecast since it is vital to verify the customer's time frame for staying with the organisation, which will undoubtedly aid in taking more preventative efforts to avoid churn. The authors also suggested that the SVM kernel function be chosen more precisely and that additional prediction variables be included.

Machado and Ruiz [13] created a mobile app usage-based churn prediction model. For a UCI dataset concerning Portugal cab service, the suggested model was found to be more accurate and precise than earlier research. It was suggested that the enhanced algorithm be employed for other applications in the future, such as e-commerce. It was also proposed that mobile and online apps based on data mining may be developed to forecast client attrition.

Zhao [14] created an e-commerce customer churn prediction model based on GA, SVM, and NN. The proposed model's accuracy and precision were found to be superior than prior versions. It was proposed that a better model with additional features and parameters may be built in the future.

Hosseini et al. [15] suggested a clustering approach for an Iranian firm product dataset by combining the WRFM-method with the K-mean algorithm in data mining to categorise customer loyalty based on recency, frequency, and monitoring. The findings of statistical tests for model validation are satisfactory, according to the authors. The authors created a customer loyalty evaluation function to determine the value of client loyalty (CLV) of a customer.

3. Proposed Methodology

3.1 About the Dataset

This project uses an IBM sample set called Telecom Customer Churn available for the purpose of practicing data analysis on a ‘real-world’ type of business problem. The dataset contains 7043 observations (rows) and 22 variables (columns) that contain information about customer demographics (gender, senior citizenship status, children, and marital status), services they signed up for (phone line, multiple lines, online security, online backup, device protection, tech support, streaming TV and Movies), account information (type of contract, payment type, paperless billing, monthly and total charges) We then predict the feature churn (whether the customers left within the past month when the data was collected)

3.2 Methodology

01

We began with performing all the necessary data preprocessing by analyzing the dataset. We identified where any data imputation was required to replace the missing values. We then updated the column names and values were converted to 0 and 1 for easy modeling.

02

We will then perform Exploratory Data Analysis to understand the basic trends and answer some simple questions such as customer’s tenure with Telco and most common account services, customer demographics and so on

03

To predict which customers are most likely to churn, several different types of classification models will be evaluated, including logistic regression, support vector machines, and random forests. Since the numeric predictors, MonthlyCharges and Tenure, have skewed distributions and varying scales, normalization of the features to have a mean of 0 and a standard deviation of 1 was done. To fit the models, 10-fold cross-validation will be used and the model will be tested on the sample dataset. This set was held out of resampling and is more representative of the true class distribution.

4. Implementation

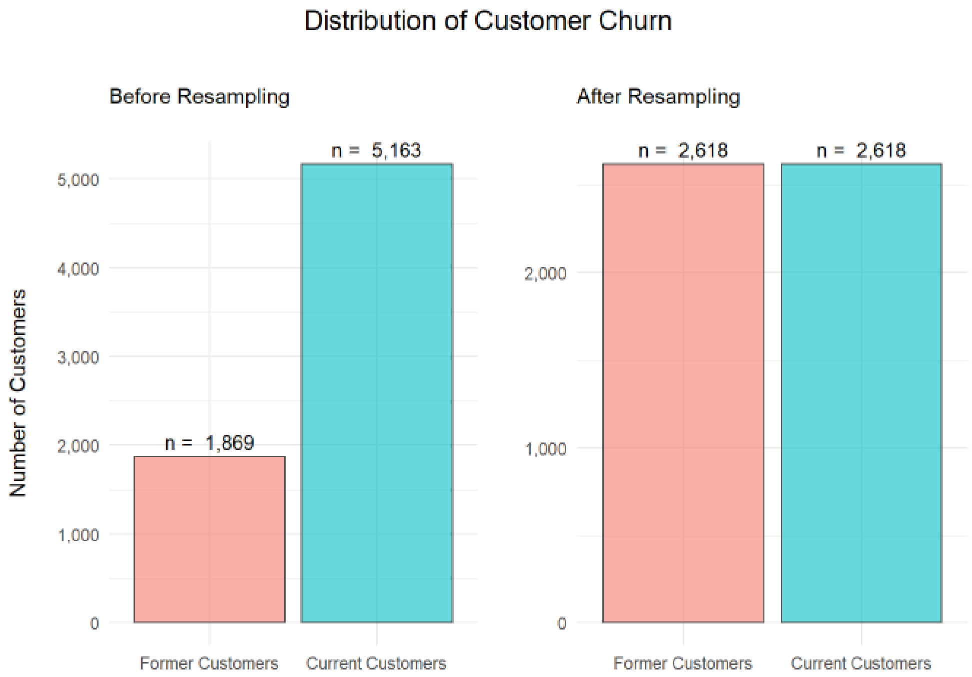
The implementation is done via R Studio Software in the R programming language. Various Libraries such as caret, tidyverse, and randomForest were included to perform the necessary preprocessing and modeling.

4.1 Data Preprocessing

There were 11 customers with missing TotalCharges. Since it is a fairly small amount, these observations will be removed prior to beginning the analysis, leaving 7,032 customers in the data set. In addition, several of the Yes/No categorical variables contained an additional group indicating that the customer had no phone or internet service. These were recorded and combined with the value No.

Our target variable, Churn, is quite imbalanced with a little over 26% (1,869 customers) leaving the company within the past month. Since class imbalance can negatively affect the precision and recall accuracy of statistical models, so we sued a **s**ynthetic **m**inority **o**ver-sampling **te**chnique known as **smote** to create a more evenly distributed training set.

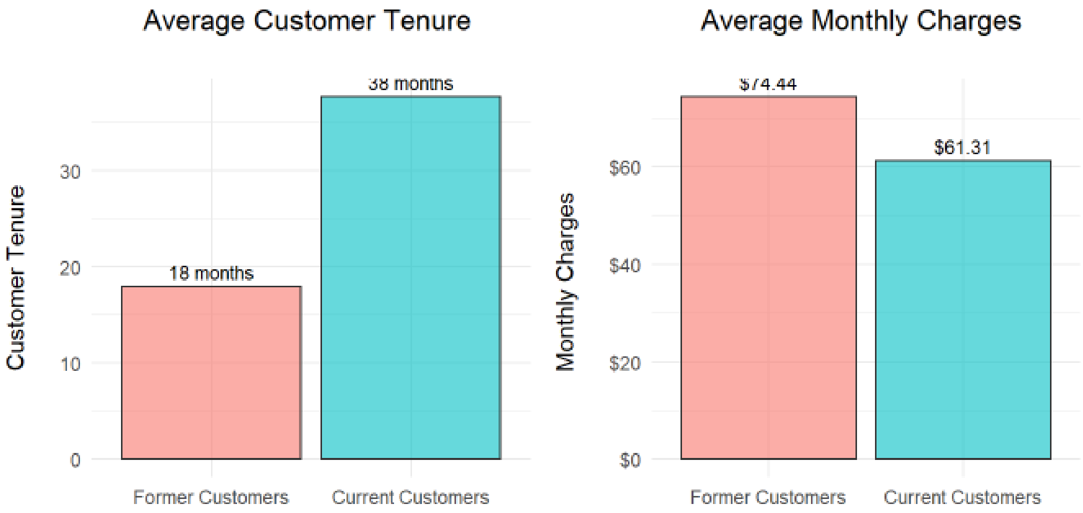
The smote algorithm artificially generates new instances of the minority class using the nearest neighbors of these cases and under-samples the majority class to create a more balanced data set. After applying smote, our training set now consists of an equal proportion of current and former customers.

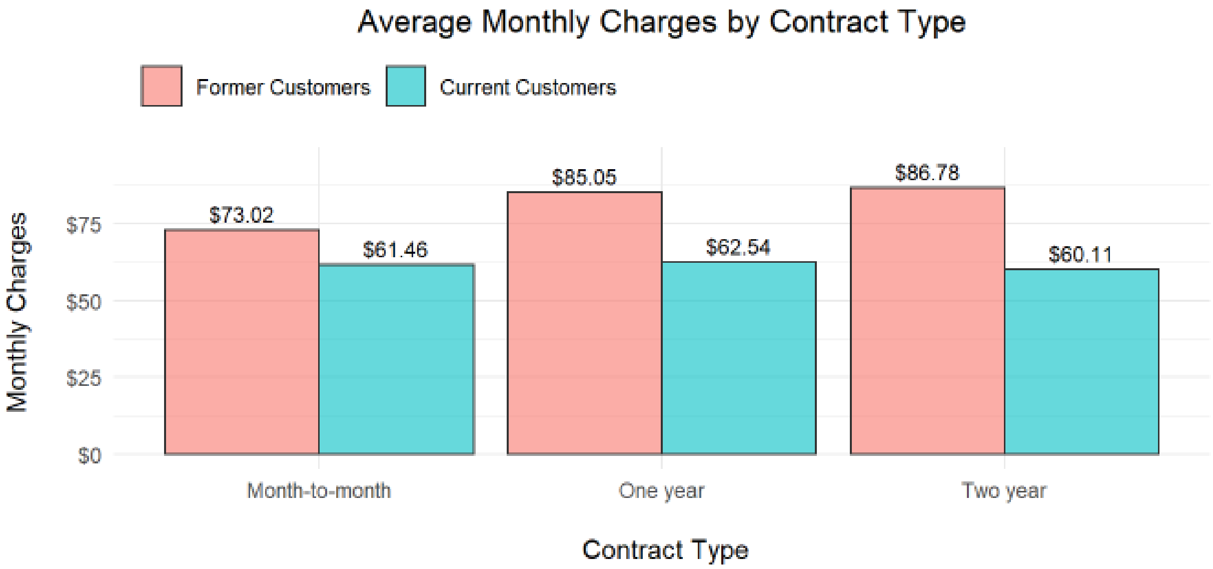


4.2 Exploratory Data Analysis

In this module, we tried to answer some questions to understand the data better and visualize the trends in the data.

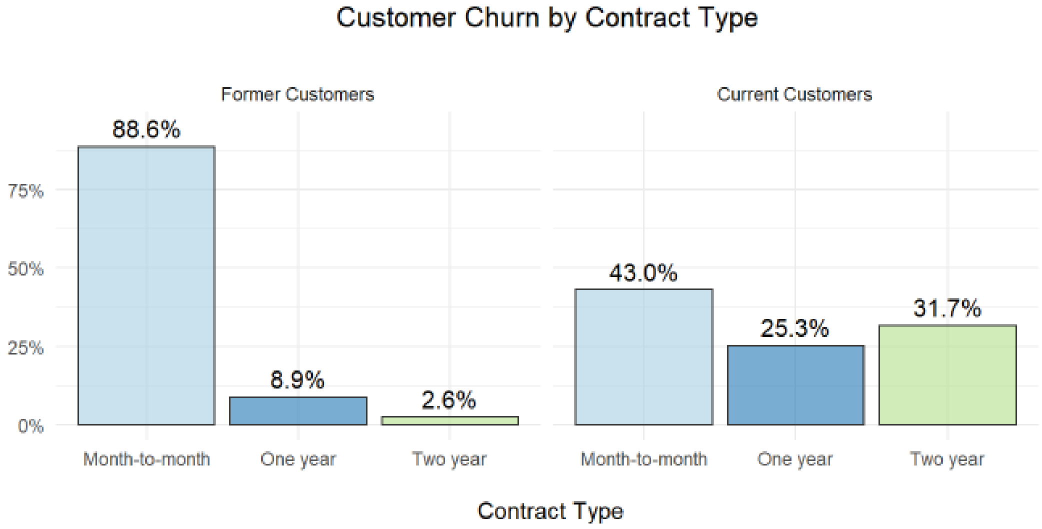
What is a customer's average tenure with Telco and what are their average charges?

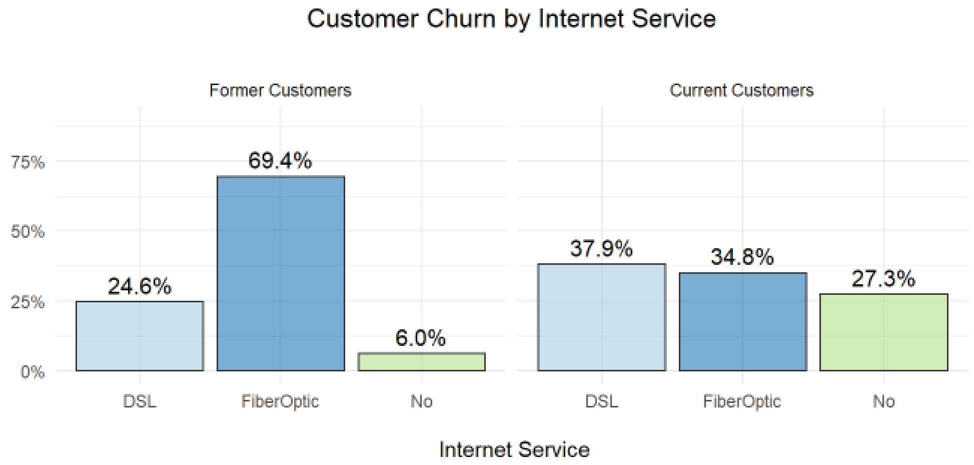




The graphs above show the average tenure of Telco's current and former customers and their monthly charges. Telco's current customers have been with the company for just over 3 years, while customers who left kept their services for about 18 months. Additionally, former customers had higher monthly charges on average by about $13. This holds true across each contract type.

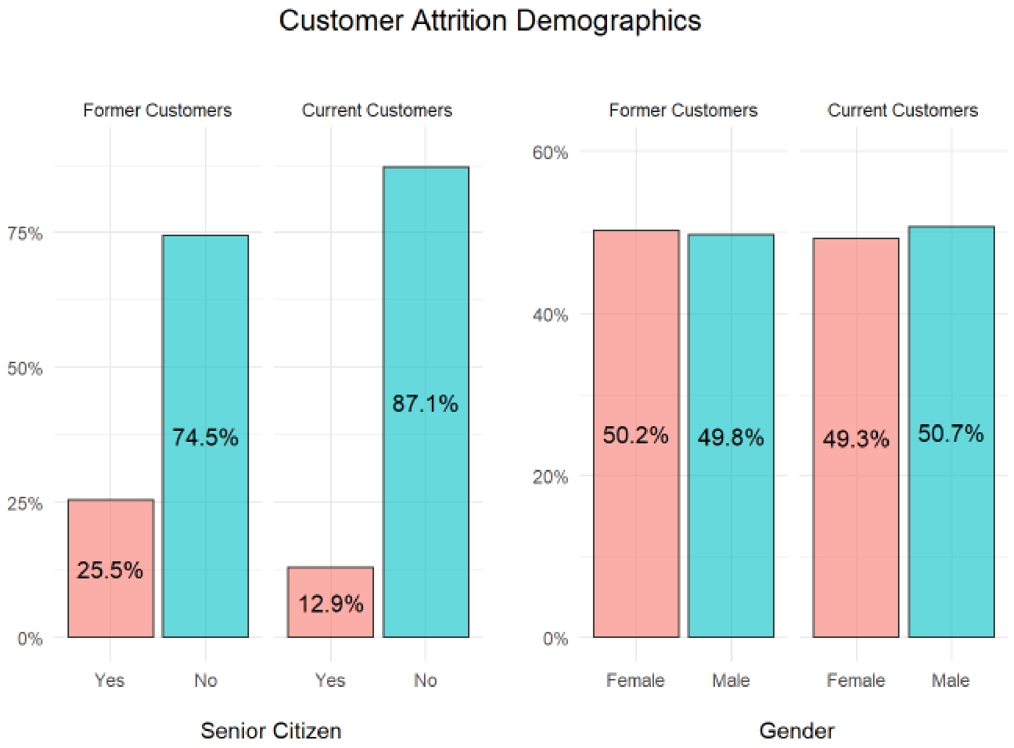
What type of account services do customers have?





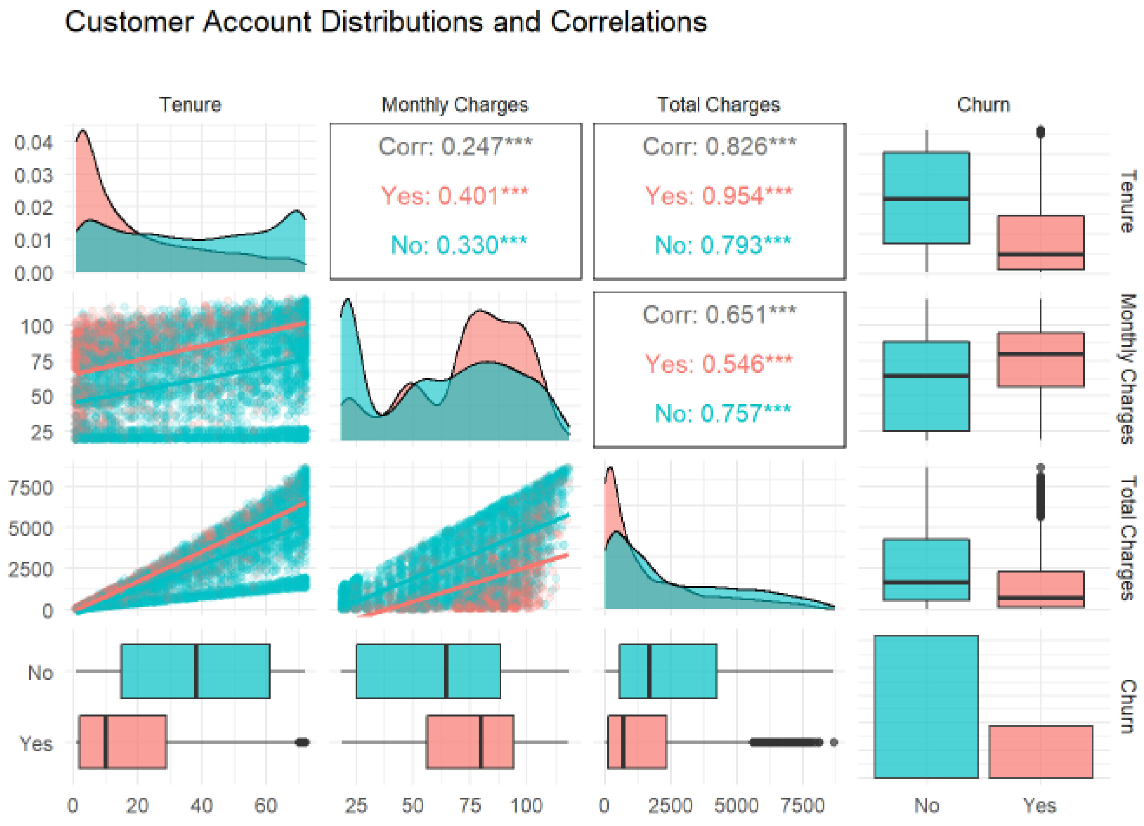
Nearly 89% of former customers were on month-to-month contracts, with a much smaller proportion in one or two-year contracts. Of customers who left, a little over 69% had Fiber Optic internet. This could be an indicator of potential dissatisfaction with the service and should be further reviewed by the company since currently over a third of their customers have this type of internet.

Customer Attrition Demographics



Based on the demographic attributes of Telco's customers, about a quarter of those who left were senior citizens, and just under 13% of their current customers are 65 years or older. The distribution of gender is proportional in both current and former customers, with an approximately equal number of men and women leaving within the last month.

Distributions and Correlations



The correlations between our numeric variables show that TotalCharges is strongly correlated with customer tenure, especially among customers who left (Churn = Yes), with a correlation of more than 0.95. There is also a slightly positive relationship between MonthlyCharges and Tenure of 0.25 and it is significant. The histogram of MonthlyCharges has a unique shape that appears to be multimodal, while the distribution of customer tenure is relatively uniform among current customers but skewed to the right in customers who left.

4.3 Feature Extraction

The Chi-Squared Test of Independence evaluates the association between two categorical variables. The null hypothesis for this test is that there is no relationship between our response variable and the categorical feature, and the alternative hypothesis is that there is a relationship. Looking at the results of the tests, Gender and PhoneService have very small chi-squared statistics and p-values that are greater than the significance threshold, α of 0.05, indicating they are independent of our target variable. The rest of the categorical features do have a statistically significant association with customer churn.

4.4 Predictive Analysis

4.4.1 Logistic Regression

Logistic regression is a parametric classification technique that estimates the probability of an event occurring, for instance, whether or not a customer will leave the company. Based on the size of the coefficients and the significance of the predictors, the model is able to quantify the relationships between our response and the input features.

The R function used to perform Logistic Regression is `glm()` and after which the confusion matrix was found. The results and summary of the model, including the accuracy, f1- score and other statistics were measured.

4.4.2 Support Vector Machine

Support vector machines (SVMs) are a commonly used statistical learning model. It is nonparametric, which means that it does not make any assumptions about the data like logistic regression does. SVMs involve finding a hyperplane that separates the data as well as possible and maximizes the distance between the classes of our response variable.

The R function used to develop a Support Vector Machine is `svm` with the method as `svmlinear` and after which the confusion matrix was found. The results and summary of the model, including the accuracy, f1- score, and other statistics were measured.

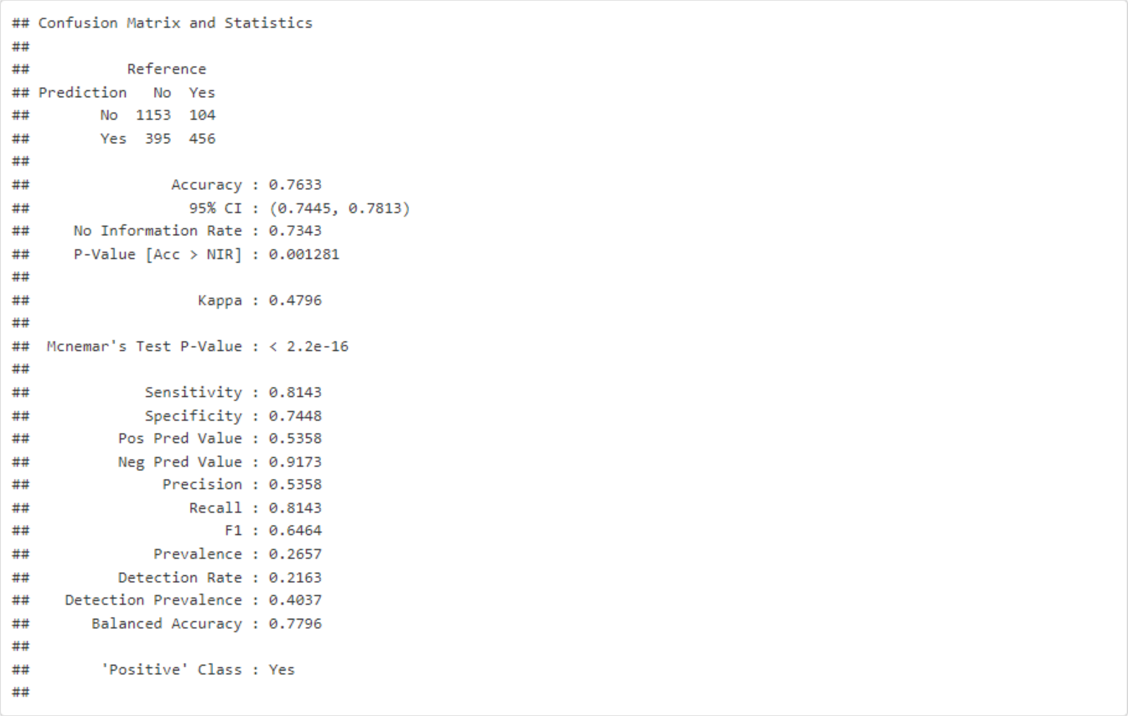
4.4.3 Random Forest

Random forest is a commonly used ensemble technique in machine learning. The model is built using a combination of many decision trees, where each takes a random sample of the data with replacement and selects a random subset of predictors. Each tree then makes a prediction and the class with the most votes becomes the model's final prediction.

The R function used to develop the Random Forest is `rf` after which the confusion matrix was found. The results and summary of the model, including the accuracy, f1- score, and other statistics were measured.

5. Results and Discussion

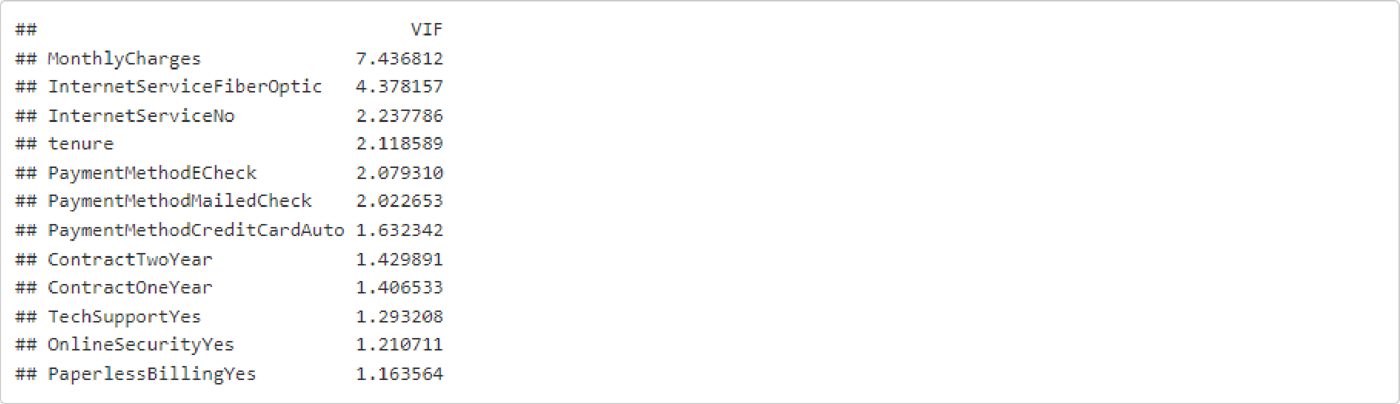
5.1 Summary of Logistic regression



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1153  104
##        Yes  395  456
##
##                Accuracy : 0.7633
##                  95% CI : (0.7445, 0.7813)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 0.001281
##
##                   Kappa : 0.4796
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8143
##             Specificity : 0.7448
##          Pos Pred Value : 0.5358
##          Neg Pred Value : 0.9173
##               Precision : 0.5358
##                  Recall : 0.8143
##                      F1 : 0.6464
##              Prevalence : 0.2657
##          Detection Rate : 0.2163
##    Detection Prevalence : 0.4037
##       Balanced Accuracy : 0.7796
##
##        'Positive' Class : Yes
##
```

Our logistic regression model has an overall accuracy of 76.3% and a precision of 53.6% on the test set. This means that when the model predicts a customer will leave, it is correct around 54% of the time. The recall of our model is 81.4%, which means that it correctly identified about 81% of all customers who left.

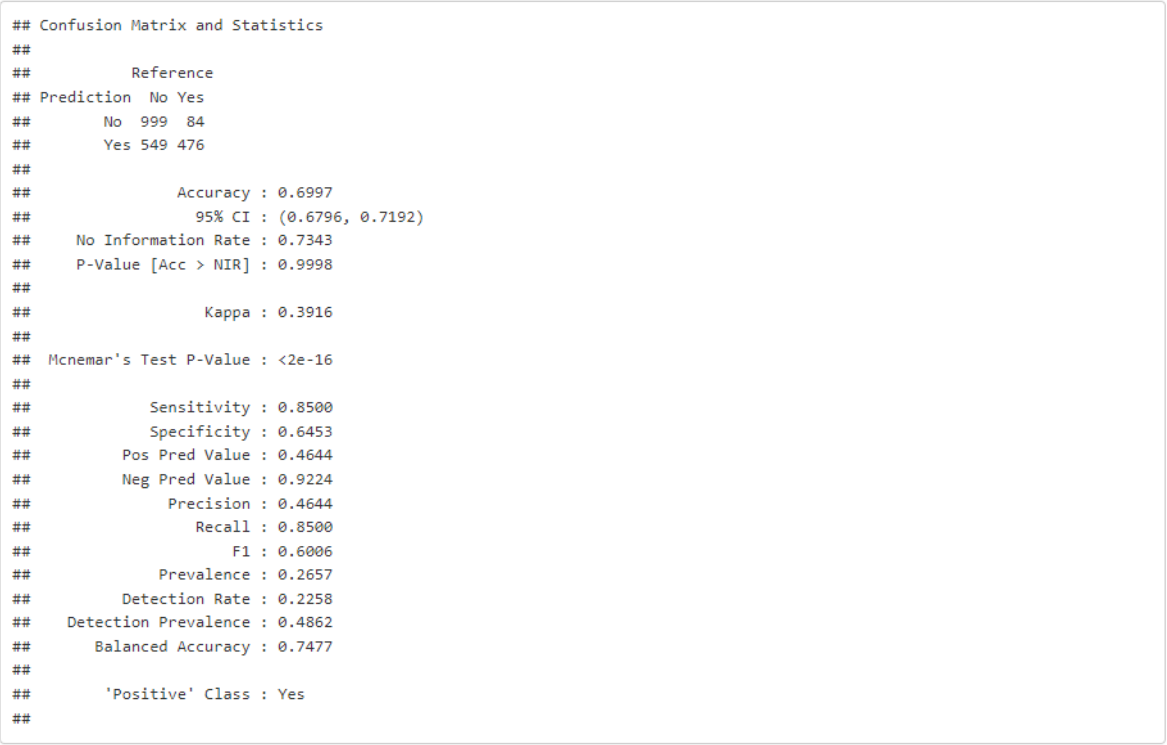
5.2 Multicollinearity



```
##                                  VIF
## MonthlyCharges             7.436812
## InternetServiceFiberOptic  4.378157
## InternetServiceNo          2.237786
## tenure                     2.118589
## PaymentMethodECheck        2.079310
## PaymentMethodMailedCheck   2.022653
## PaymentMethodCreditCardAuto 1.632342
## ContractTwoYear            1.429891
## ContractOneYear            1.406533
## TechSupportYes             1.293208
## OnlineSecurityYes          1.210711
## PaperlessBillingYes        1.163564
```

One of the assumptions of logistic regression is that the predictors are not too highly correlated with each other. The Variance Inflation Factor (VIF) measures the amount of multicollinearity between the features in the model. A general rule of thumb is a VIF score of no higher than between 5 and 10. Since the majority of the predictors have a VIF of less than 5 and none exceed 10, we are good.

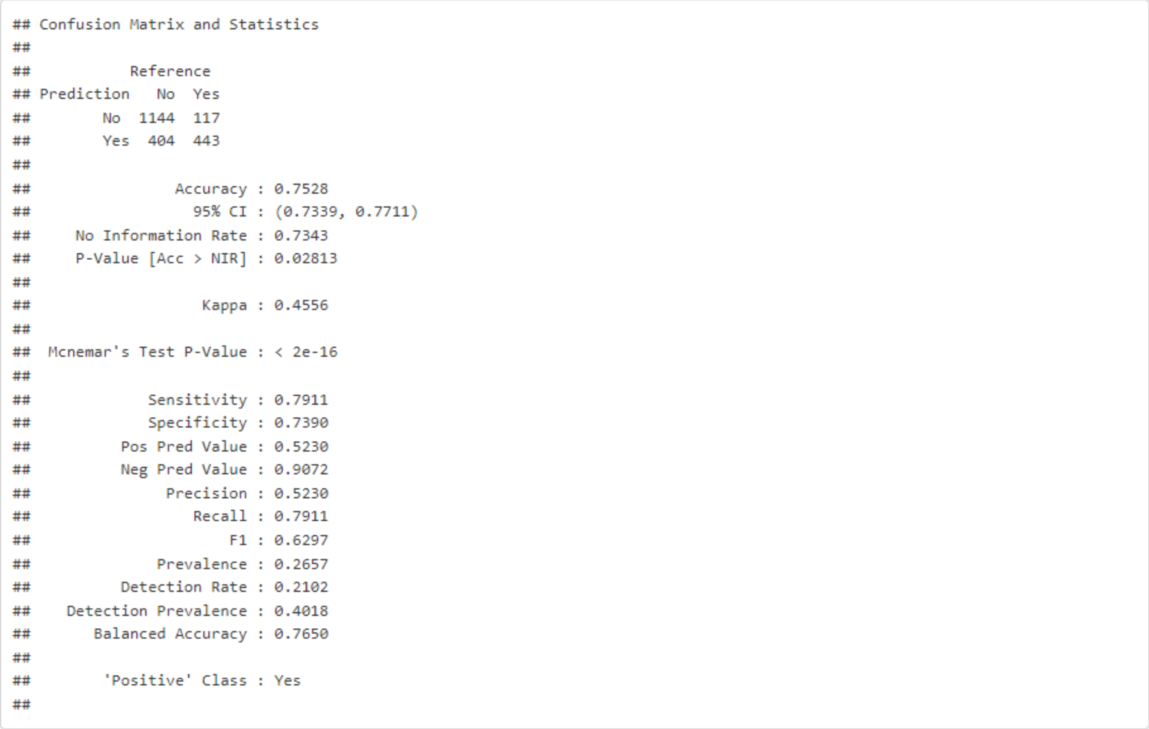
5.3 Summary of SVM



```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction  No Yes
##        No  999  84
##        Yes 549 476
## 
##                Accuracy : 0.6997
##                  95% CI : (0.6796, 0.7192)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 0.9998
## 
##                   Kappa : 0.3916
## 
##  Mcnemar's Test P-Value : <2e-16
## 
##             Sensitivity : 0.8500
##             Specificity : 0.6453
##          Pos Pred Value : 0.4644
##          Neg Pred Value : 0.9224
##               Precision : 0.4644
##                  Recall : 0.8500
##                      F1 : 0.6006
##              Prevalence : 0.2657
##          Detection Rate : 0.2258
##    Detection Prevalence : 0.4862
##       Balanced Accuracy : 0.7477
## 
##        'Positive' Class : Yes
## 
```

The accuracy of the linear support vector machine is about 70% and the precision is 46%, which is not an improvement from the previous models. The recall did increase to 85%, which is the highest so far.

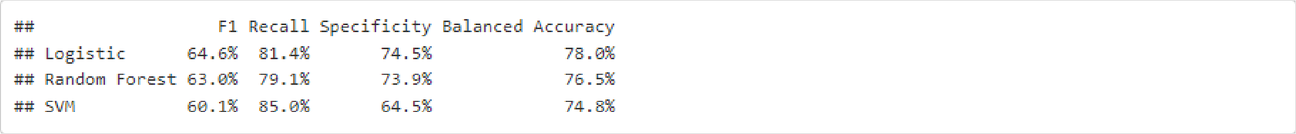
5.4 Summary of Random Forest



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1144  117
##        Yes  404  443
##
##                Accuracy : 0.7528
##                  95% CI : (0.7339, 0.7711)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 0.02813
##
##                   Kappa : 0.4556
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.7911
##             Specificity : 0.7390
##          Pos Pred Value : 0.5230
##          Neg Pred Value : 0.9072
##               Precision : 0.5230
##                  Recall : 0.7911
##                      F1 : 0.6297
##              Prevalence : 0.2657
##          Detection Rate : 0.2102
##    Detection Prevalence : 0.4018
##       Balanced Accuracy : 0.7650
##
##        'Positive' Class : Yes
##
```

The random forest classifier has an accuracy of 76% and a precision of 53%, higher than the SVM but just below our logistic model. The recall of the model is about 79%, the lowest overall.

5.5 Model Performance on the Test Set

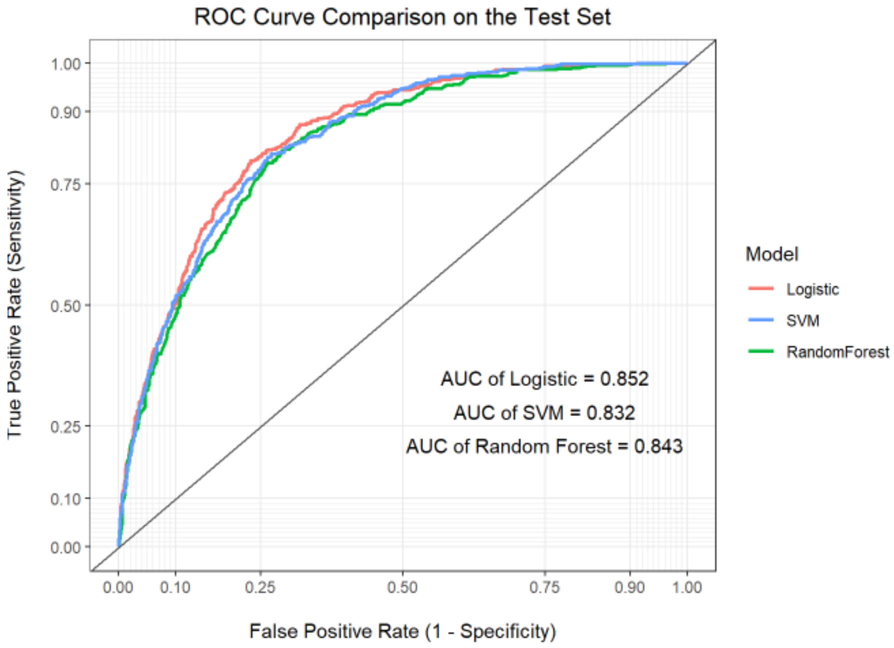


```
##                  F1 Recall Specificity Balanced Accuracy
## Logistic      64.6%  81.4%       74.5%             78.0%
## Random Forest 63.0%  79.1%       73.9%             76.5%
## SVM           60.1%  85.0%       64.5%             74.8%
```

Out of the three models, logistic regression produces the highest F1 - score, which represents the balance between precision and recall, as well as the highest specificity, which measures how well the model identifies negative cases correctly.

5.6 ROC Curves

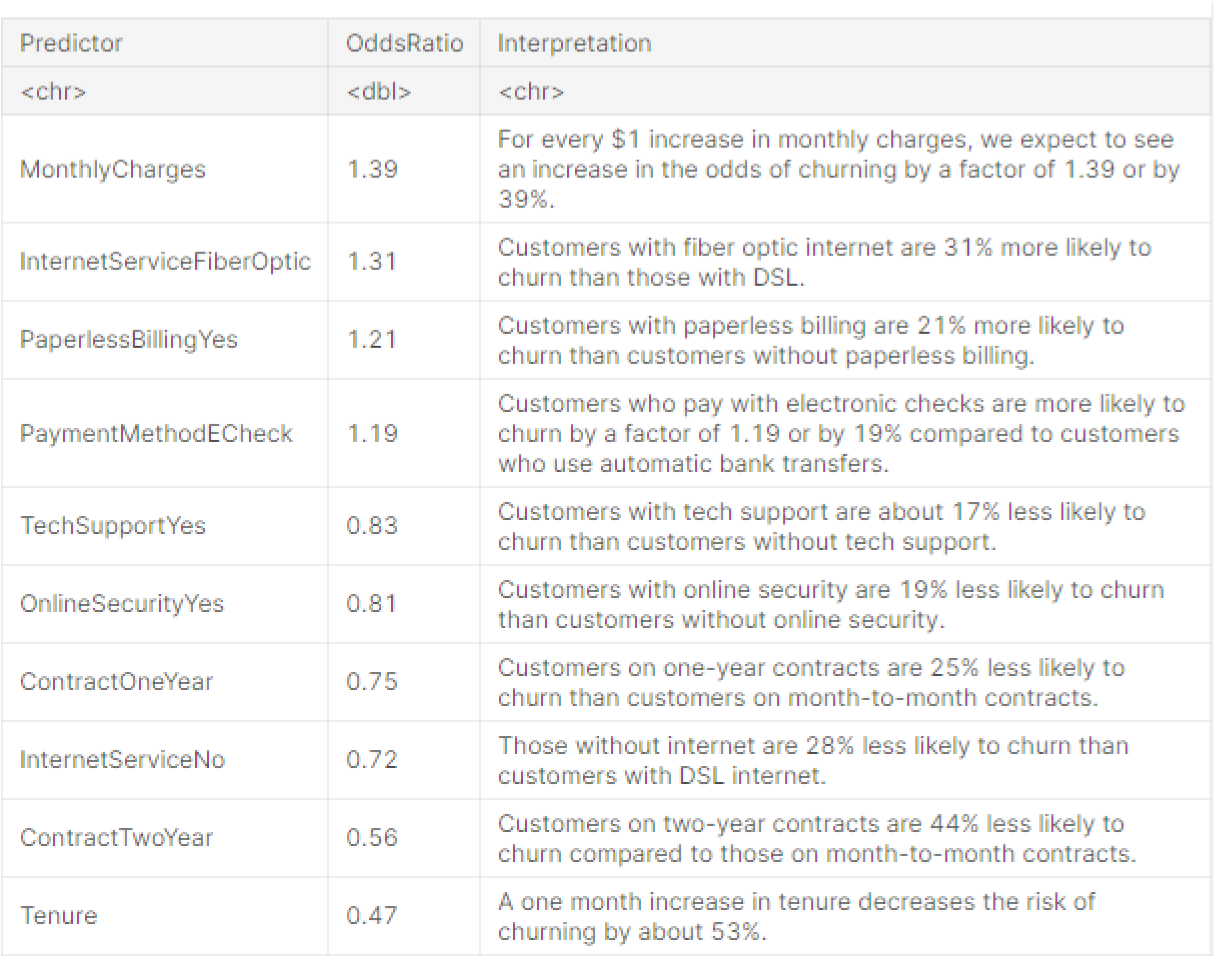
We then plotted the ROC curves of each model with their corresponding Area Under the Curve. The Area Under the Curve measures the model's performance across all possible classification thresholds. A higher AUC indicates the model is better able to distinguish between the classes.



Out of the three classifiers, the logistic model has the highest Area Under the Curve of 0.852 on the test set. This represents the probability that our model will rate or rank a randomly chosen observation from the positive class, Churn = Yes, as more likely to be from that class than a randomly chosen nonpositive observation, Churn = No

5.7 Final Discussions and Key Findings

Overall, the logistic regression model had the strongest performance on the test set. Based on the coefficients from the model, at least one category in all eight predictors has a significant association to customer attrition. A summary of the relationships of each, when all other variables are held constant, is listed in the table below.



Predictor	OddsRatio	Interpretation
<chr>	<dbl>	<chr>
MonthlyCharges	1.39	For every $1 increase in monthly charges, we expect to see an increase in the odds of churning by a factor of 1.39 or by 39%.
InternetServiceFiberOptic	1.31	Customers with fiber optic internet are 31% more likely to churn than those with DSL.
PaperlessBillingYes	1.21	Customers with paperless billing are 21% more likely to churn than customers without paperless billing.
PaymentMethodECheck	1.19	Customers who pay with electronic checks are more likely to churn by a factor of 1.19 or by 19% compared to customers who use automatic bank transfers.
TechSupportYes	0.83	Customers with tech support are about 17% less likely to churn than customers without tech support.
OnlineSecurityYes	0.81	Customers with online security are 19% less likely to churn than customers without online security.
ContractOneYear	0.75	Customers on one-year contracts are 25% less likely to churn than customers on month-to-month contracts.
InternetServiceNo	0.72	Those without internet are 28% less likely to churn than customers with DSL internet.
ContractTwoYear	0.56	Customers on two-year contracts are 44% less likely to churn compared to those on month-to-month contracts.
Tenure	0.47	A one month increase in tenure decreases the risk of churning by about 53%.

6. Conclusion and Further works

In predicting customer attrition, logistic regression produced the highest Area Under the Curve, F1 score, and specificity. Some of the most important predictors of customer attrition include Tenure, MonthlyCharges, InternetService, PaymentMethod, Contract, OnlineSecurity, TechSupport, and paperless billing. We also found that the most significant relationships from our logistic model are the customer's monthly charges, the type of internet service and contract they have, and the length of time they have been customers with Telco. To proactively reduce their churn rate, Telco could target customers who are on month-to-month contracts, use fiber optic internet, have higher monthly charges on average, and who have a shorter tenure of fewer than 18 months, which is the average tenure of their former customers.

7. References

1. SORIANO, Domingo Ribeiro; JENG, Don Jyh‐Fu; BAILEY, Thomas. Assessing customer retention strategies in mobile telecommunications. Management Decision, 2012.

2. JAHROMI, Ali Tamaddoni; STAKHOVYCH, Stanislav; EWING, Michael. Managing B2B customer churn, retention and profitability. Industrial Marketing Management, 2014, 43.7: 1258-1268.

3. VARSHNEY, Nitish; GUPTA, S. K. Mining Churning Factors in Indian Telecommunication Sector Using Social Media Analytics. In: International Conference on Data Warehousing and Knowledge Discovery. Springer, Cham, 2014. p. 405-413.

4 VIJAYA, J.; SIVA SANKAR, E.; GAYATHRI, S. Fuzzy Clustering with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector. In: Recent Developments in Machine Learning and Data Analytics. Springer, Singapore, 2019. p. 261-274.

5. AMIN, Adnan, et al. Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 2019, 94: 290-301.

6. HOSSAIN, Md Alamgir, et al. Customer Retention and Telecommunications Services in Bangladesh. International Journal of Asian Social Science, 2017, 7.11: 921-930.

7. AMIN, Adnan, et al. Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 2017, 237: 242-254.

8. Azeem, M.; Usman, M.; Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. Springer Telecommunication Systems, 66 (4), pp.603-614.

9. Baumann, A.; Haupt, J.; Gebert, F.; Lessmann, S. (2018). Changing perspectives: Using graph

metrics to predict purchase probabilities. Expert Systems with Applications, 94, pp.137-148

10. Jie, C.; Xiaobing Y.; Zhifei Z. (2015). Integrating OWA and data mining for analyzing customers churn in e-commerce. Springer Journal of Systems Science and Complexity, 28, pp.381–392.

11. Fridrich, M. (2017). Hyperparameter optimization of artificial neural network in customer churn prediction using Genetic Algorithm. Trends in Economics and Management, 28(1), pp.9–21.

12. Gordini, N; Veglio, V. (2016). Customer churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter selection technique in B2B e-commerce industry. Industrial Marketing Management. 8, pp. 1-8.

13. Machado, N. L. R.; Ruiz, D. D. A. (2017). Customer: A novel customer churn prediction method based on mobile application usage. IEEE Wireless Communications and Mobile Computing Conference, pp.2146-2151.

14. Zhao, X. (2014). Research on E-commerce customer churning modeling and prediction. The Open Cybernetics & Systemics Journal, 8, pp.800-804.

15. Hosseini, S. M. S.; Maleki, A.; Gholamian M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess customer loyalty. Expert Systems with Applications, 37, pp. 5259–5264.