Telco Customer Churn Prediction

Ying Wei [*]

Johns Hopkins University, Washington, DC, USA

[*] Corresponding Author Email: ywei42@alumni.jh.edu

Abstract. In recognizing the significance of retaining current consumers to thrive in this competitive landscape, this paper aims to predict customers' churn probability based on the background and behaviors of previous customers. The paper utilizes churn probability as a proactive means to identify customers at a high risk of leaving, serving as a reference for Telco to make informed decisions and take actions aimed at enhancing customer loyalty. This paper pre-target customers who have a high risk of leaving based on our churn probability with at least over 80% accuracy rate. And then use the churn probability as the reference to help Telco make better decisions and take actions to retain customers, such as providing a private discount or coupon. This paper limits the choice of customers in the services provided by telecom companies. However, customer attrition decisions may also be related to the quality of the service and the level of price, which the study was unable to quantify. The study addresses two primary challenges: determining models that accurately predict customer churn and identifying the characteristics of customers more prone to churn.

Keywords: Machine learning, Customer churn, Prediction.

1. Introduction

With the increasing dependence on and demand for mobile communication devices, such as mobile phones, the markets for internet service, streaming TV, and mobile communication have become highly competitive. To survive in the competitive market, the company should have strategies for both potential customers and existing customers. However, with the intensification of market competition, companies devote more time and effort to maintaining connections with existing clients rather than attracting new customers [1]. If Telco can predict customers who are likely to leave the company in the early stages, it could tap into a potentially significant additional revenue source [2]. This idea raised our business goal – prioritizing client retention over the acquisition of new customers, which is to predict customers' churn probability based on the background and behaviors of previous customers.

Currently, predictive analytics is commonly performed using machine learning methods. In their study, Kiran Dahiya and Surbhi Bhatia proposed a new churn prediction model framework, wherein they compared the efficiency and performance of decision tree and logistic regression techniques specifically in predicting churn within the Telecom industry [1]. The study highlights the advantages of machine learning methods in accurately predicting churn within the Telecom industry.

Therefore, this paper aims to explore the application of machine learning algorithms in predicting customer churn. The paper seeks to address two main questions: firstly, which model(s) can help the company accurately predict customer churn, and secondly, what characteristics define customers more likely to churn? The study employed a classification tree model, random forest, support vector machine (SVM) model, and logistic regression model. Subsequently, a comparison of these four models was conducted using ROC and AUC methods to determine the most effective deployment model.

2. Method and Data

In this paper, machine learning algorithm had been utilized to predict customer’s churn rate. Various models were employed to estimate the target-churn rate, including classification tree, random forest, support vector machine (SVM), and logistic regression. Given the binary nature of the problem, the study compared these models' using ROC and AUC methods to identify the most effective deployment model.

2.1. Data Description

The data comes from https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download. The data sorting routine began with establishing a foundational business understanding, then setting goals to predict customer churn rate. The original dataset comprised 7043 rows (customers) and 21 columns (features), with the target variable being 'Churn,' indicating whether a customer had churned or not. (A detailed description of each variable is listed in Appendix). Subsequently, the data underwent cleaning and preprocessing. Afterwards, the study gained valuable insights into the data, revealing nuances not captured by a simple statistical summary. Several models were constructed to predict the outcome, and the best-performing model was selected based on prediction results.

2.2. Data cleaning and Pre-processing

Since the business is to help the market to predict behavior to retain customers, the study analyzes all relevant customer data and develop focused customer retention programs. Thus, the project will be a classification and the "Churn" column is the target variable. The target variable is unbalanced: The result showed 5174 customers will not churn, and 1869 customers will churn.

After the initial analysis, the dataset contains 7043 instances. There are 11 missing values and 22 duplicates. The study dropped all missing values. 11 is relatively small compared to 7043 instances, so deleting the missing value will have a subtle impact on the whole dataset. Then deleted the duplicates to make the dataset clean and compact for the same reason. Also, the customer ID is just an index, which is not helpful for classification, so the study decided to drop it as well. They are "customer ID" which is just an index. Finally, 7010 instances are used for further analysis.

Then, the study converted character data into factors to fit the classification model. There are three numeric variables. The study checked their distribution and found that they were not in the normal distribution. Roughly speaking, there is no outlier for three numerical variables. But, while relating three numerical variables to the target variable, outliers exist for tenure and total charges when people churn. And the total charges have some outliers.

2.3. Model

This paper employed four models to assess accuracy: Classification Tree Model, Random Forest Model, Support Vector Machine (SVM) Model, Logit Regression Model. Classification tree is a binary decision structure used for categorizing objects, primarily serving as a classifier. Decision-makers often prefer less complex decision trees since its more understandable [3]. The random forest is also a tree-based model that consolidates the results of multiple decision trees to produce a single outcome [4]. The Support Vector Machine (SVM) classifies based on a linear function of features, aiming to maximize the margin between classes. It identifies the widest gap, and the linear discriminant is the center line through this margin [5]. The Logit Regression Model is also a classification model and can be utilized for predictive analytics. It proves to be a more efficient method for both binary and linear classification problems [6].

3. Result

The study first plots the distributions of 3 numeric variables as shown in Fig. 1. In view of the fact that, based on the "tenure" distribution of the churn graph, most people churn in the range of 0-20. In the "Monthly_charges" distribution graph, It is apparent that more people choose to churn between the price range of 60 to 120. By looking at the total charges of distribution, the study finds out that $0 to around $2000 has the most people churn.

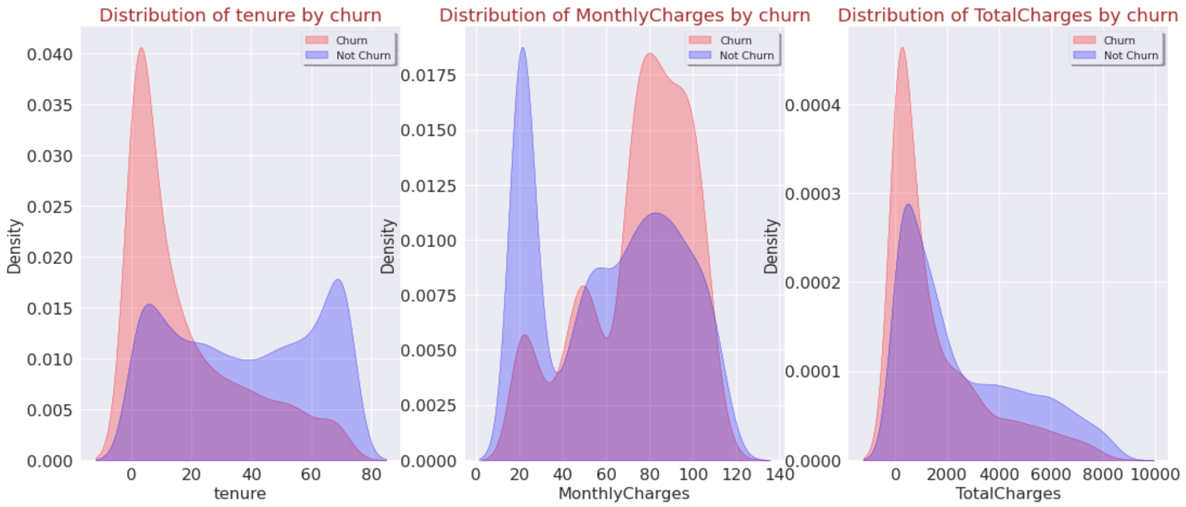


Figure 1. Numerical Account Features Distributions by Churn

Furthermore, the objective is to understand how these variables are interconnected, leading to the creation of a correlation map for the numeric variables, as shown in Fig. 2. There is a weak correlation between “Tenure” and “MonthlyCharges”, but the study discovers that there is a strong correlation between “TotalCharges” with both “MonthlyCharges” and “Tenure”.

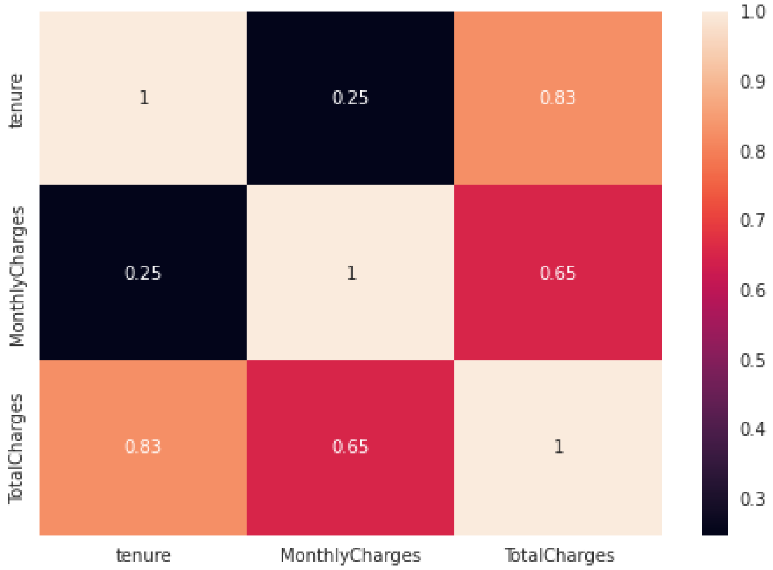


Figure 2. Correlation matrix of the Telco Customer Churn dataset

The study also plots some bar charts for the categorical features which is shown in Fig. 3. The count of males and females is the same for churn customers and the study views customer churn is not affected by gender. There are more churn customers in the single relationship category compared to married people. Customers who have Fiberopic internet service are more likely to churn. Customers with no internet service have a very low churn rate. The customer who takes the contract month on month is more likely to be churn. The customer who pays the bill through an electronic check is more likely to churn which is shown in Fig. 4. The below bar chart shows which variables are leading to more customer churn rate.

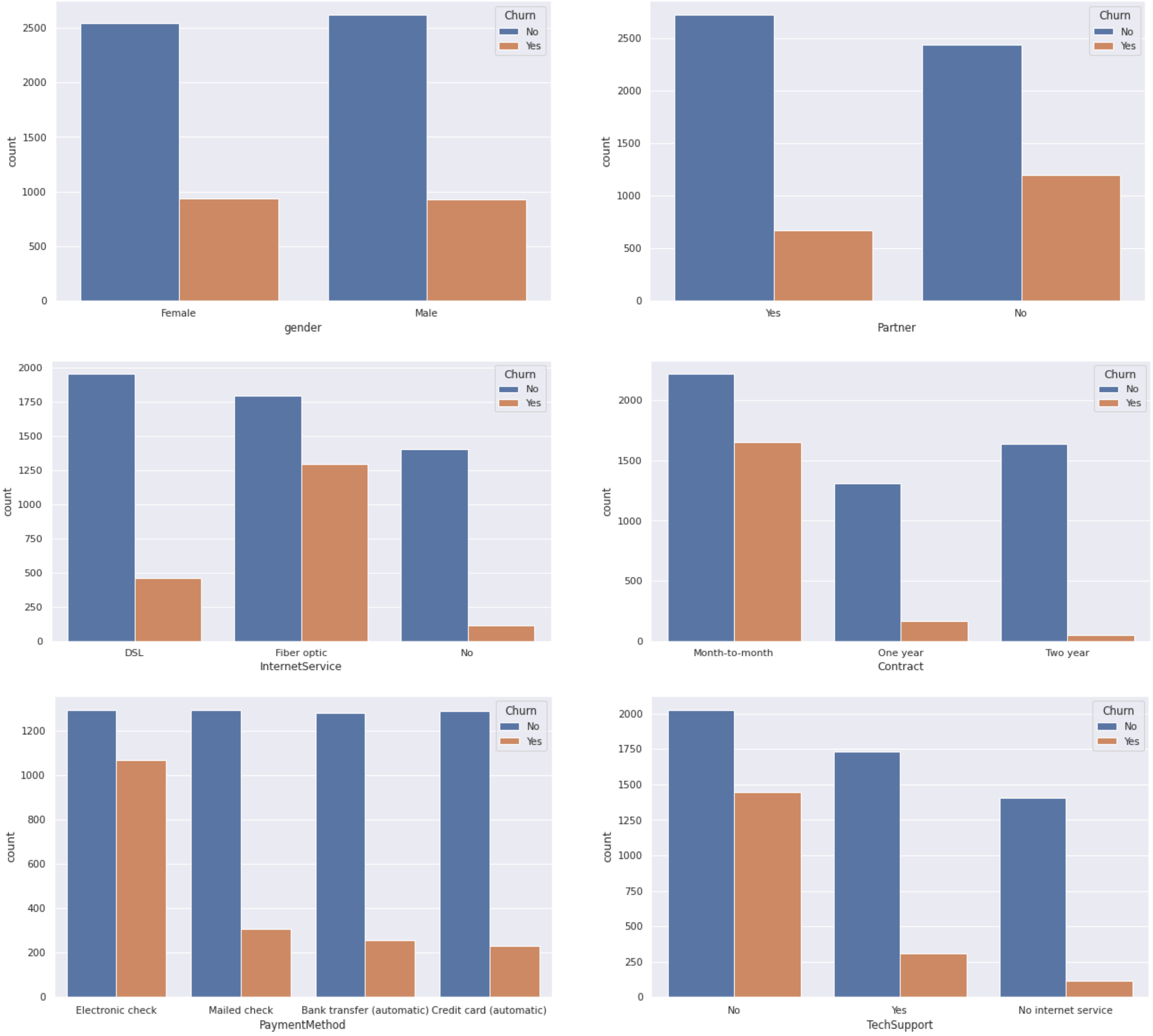


Figure 3. Bar chart

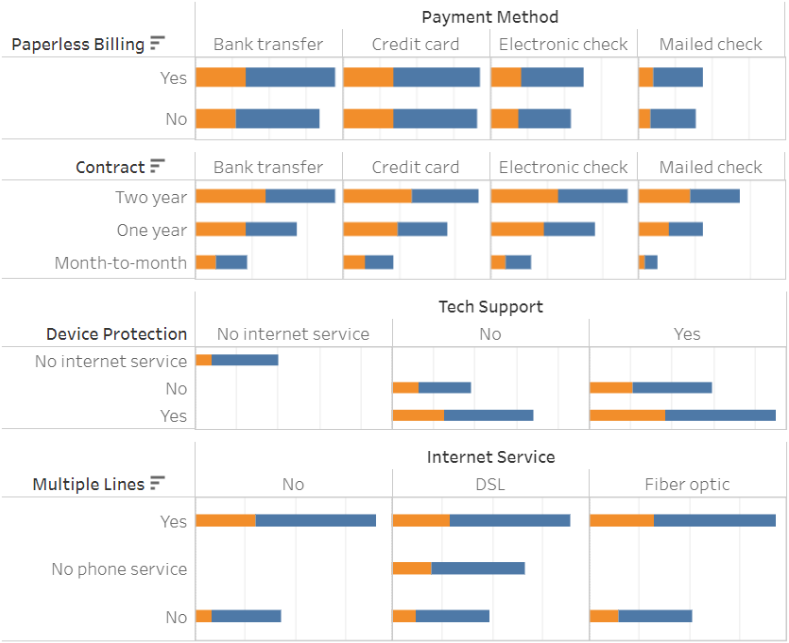


Figure 4. Payment method, tech support and internet service related to churn.

In the case of paying Bank transfer & Credit card which is automatic have longer tenure, regardless of their churn and paperless billing. For customers who use paper billing and leave the service, payment using a credit card has a slightly longer tenure than bank transfer. But customers who use paperless billing and leave the service, payment using bank transfer had a slightly longer tenure than a credit card.

On every payment type, customers with longer contract terms have longer tenure regardless of their churn. Focusing on a 2-year contract term, people who churn have longer tenure than those who didn't with every payment type. Focusing on the 1-year contract term, people who churn and use credit cards & electronic checks have longer tenure than those who didn't.

Focusing on Tech support & Device protection, people who actively use these services have longer tenure regardless of their churn. And customers who didn't leave the service have longer tenure in every case.

Focusing on Internet service & Multiple lines, there are some interesting results. Customers who use multiple lines have longer tenure. But customers who use DSL, they didn't use phone service have longer tenure than those who did regardless of their churn. The graph is shown in Fig. 5.

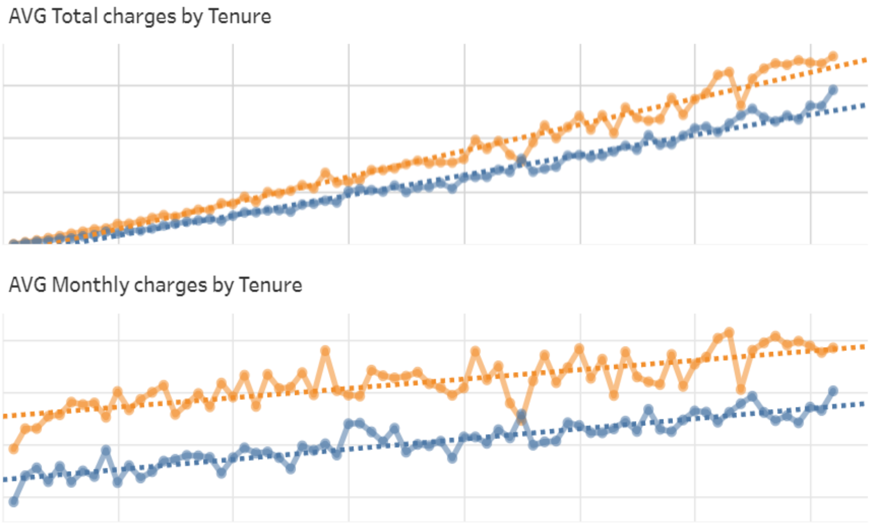


Figure 5. Tenure trend graph

Focusing on average monthly & total charges, if customers have the same tenure, the group who left the service have a bigger value compared to the opposite group at ALMOST every case. Interestingly, customers who have 45 months of tenure, and groups that didn't leave the service have bigger value and this is the only case.

4. Feature Importance

The study employs Principal Component Analysis (PCA) to identify predictor strength, as illustrated in the Fig.6. This technique extracts significant variables, filtering out those with little correlation to the target variable to address overfitting issues. The most crucial variables, highlighted in the feature importance graph, include 'MonthlyCharges' (0.28), 'TotalCharges' (0.20), 'Tenure' (0.17), 'Contract_Two year' (0.15), 'Contract_One year' (0.09), 'PaymentMethod' (0.06), 'PaperlessBilling_Yes' (0.03), and 'PhoneService_Yes' (0.01). Additionally, Random Forest classifiers are employed to identify the best features, ultimately incorporating 'tenure,' 'PaymentMethod,' 'MonthlyCharges,' 'TotalCharges,' 'PhoneService_Yes,' 'Contract_One year,' 'Contract_Two year,' and 'PaperlessBilling_Yes' into the model for optimal accuracy.

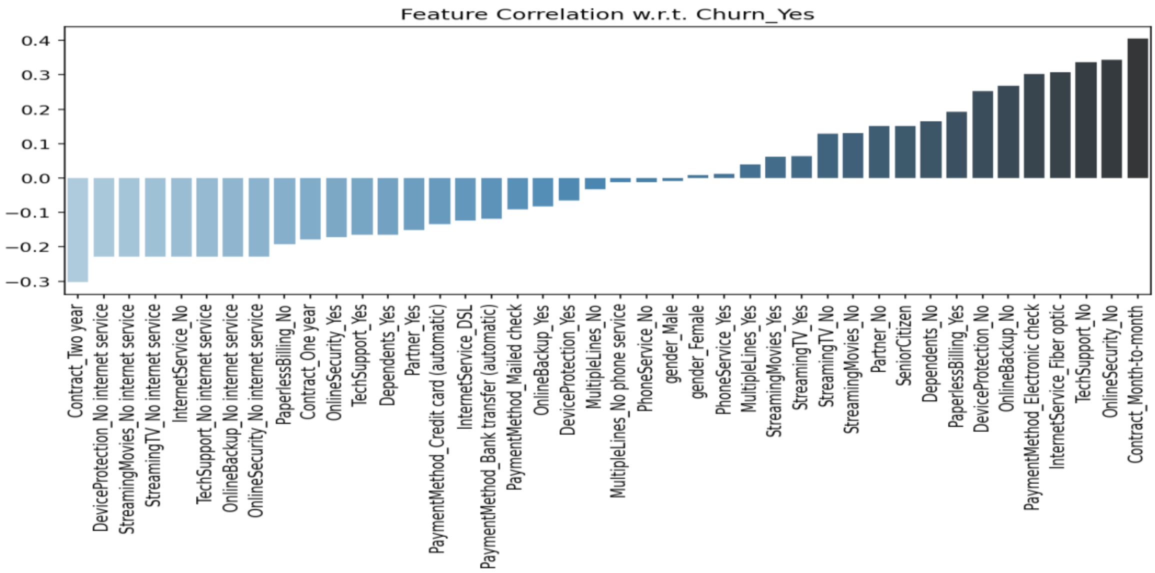


Figure 6. correlation matrix and use bar charts to visualize correlation

5. Modeling & Evaluation

Since the evaluation of training data cannot assess how well each model generalizes to new data, the whole data has been splitted into two parts: a training set and a test set with a ratio of 80% over 20%. The study uses the training set, 5608 instances, to build models. Then, predict the values of the test set, 1402 instances, on each model and compare them with true values to see how accurate each model predicts the customer churn and which one should be the best model for customer churn prediction.

5.1. Classification Tree Model

With the result of feature selection, the study builds the classification tree with ‘PhoneService’, ‘Contract’, ‘Paperless Billing’, ‘Payment Method’, ‘tenure’, ‘MonthlyCharges’, and ‘TotalCharges’. Also, the study set the model complexity to 0. The model was tested with a 50 % cut-off, resulting in an accuracy rate of 78.60%, false positive rate of 13.33%, and true positive rate of 55.86%.

Next, the study prunes the tree by using the model complexity (cp) with the lowest xerror rate. The pruned tree is with cp = 0.002684564 and xerror = 0.793959732. The tree is pruned to have 19 nodes, so it is still hard to interpret a large tree. The pruned classification tree gives a better performance: accuracy rate = 79.32%, false positive rate = 12.95%, and true positive rate = 57.49%.

5.2. Random Forest Model

The Random Forest model is built as a development of the Classification Tree model. First set the seed to 1. Then use the same predictors and the same cut-off value (i.e. If more than 50% of trees vote for ‘YES’, the instance will be classified as ‘YES’.) as the tree model and set the number of trees as 500 to build the Random Forest model. Random Tree model produces the scores of variable importance: Three numerical variables (‘MonthlyCharges’, ‘TotalCharges’, and ‘tenure’) receive the highest scores among 7 variables. The test result gives an accuracy rate of 80.81%, false positive rate of 10.34%, and true positive rate of 55.86%.

Then, the study apply TuneRF() function to find the best Random Forest model, The function indicates that when mtry = 2, the result showed that OOB is the lowes. The tuning process improves the Random Forest model a little: accuracy rate = 81.17%, false positive rate = 10.24%, and true positive rate = 56.95%. Overall, Random Forest has a better prediction performance than single trees.

5.3. Support Vector Machine (SVM) Model

The SVM model is built with the same features, and the study set the kernel as ‘linear’. There are 2769 support vectors. As for the model’s performance on the test set, it gives an accuracy rate of 81.03 %, false positive rate = 9.76%, and true positive rate = 55.04%.

When it comes to tuning for SVM, set the kernel to ‘radial’ and test a range of cost values. The output shows the best SVM model is with cost = 1. The SVM model has an accuracy rate = 81.95%, false positive rate = 6.09%, and true positive rate = 48.22%.

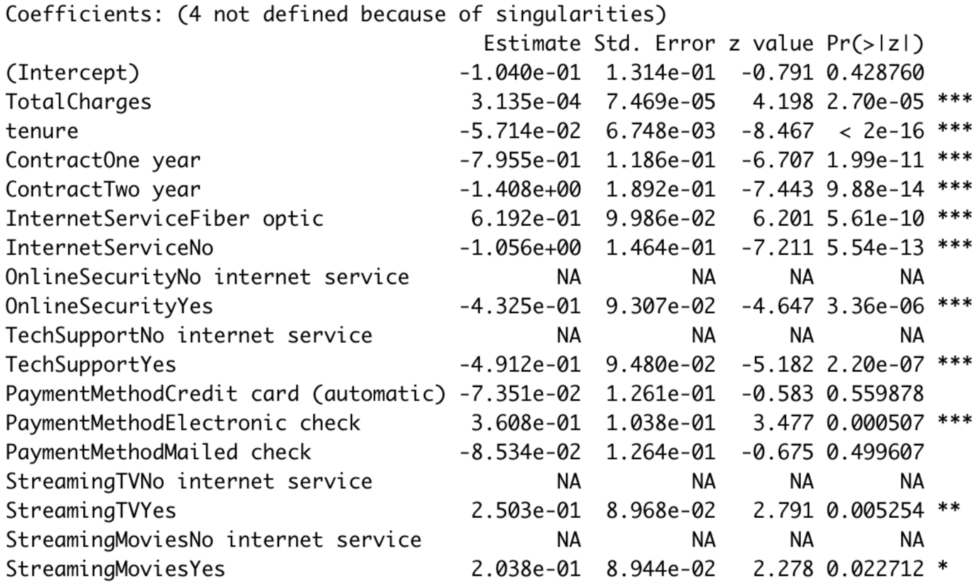
It is noticeable that the tuning process decreases the false positive rate a lot while the true positive rate also gets lower. Also, the number of false negative cases increases from 165 to 190, which is a big problem because the study does not want to fail to monitor customers who churn. Therefore, the result cannot conclude that tuning helps improve the SVM model.

5.4. Logit Regression Model

The logistic regression model is built first with variables that have correlation scores with Churn outside the range of [-0.3, 0.3]. The study used the train data to construct the model and the test data to predict the Churn probability and then compared it with the true Churn outcomes. It provides an accuracy rate of 81.4%, a true positive rate of 55.6%, and the false positive rate of 9.4%. The model indicates the p-value of Monthly Charges is larger than 0.05. Hence, the study conclude that this variable will not affect the prediction of Churn outcomes.

For the second logit model, deleted the variable Monthly Charges and added two variables Streaming Movies and Streaming TV which also have a high correlation with Churn. In this model, the accuracy rate is 81.7%, the true positive rate is 58%, and the false positive rate is 9.9%. The true positive rate in this model increases by 2.4%, even though the false positive rate increases by 0.5%. Therefore, the study found that this model performs better than the first logit model.

After the two trials, the study used the stepwise regression by forward and backward methods, which includes the nine variables in the second logit model. The two methods yield the same accuracy rate, true positive rate, and false negative rate as the previews model. Also, the coefficients in these two models are the same as those in the second logit model. Hence, the most significant coefficients in logistic regression model as below (Fig.7).



```
Coefficients: (4 not defined because of singularities)
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -1.040e-01  1.314e-01  -0.791 0.428760
TotalCharges                        3.135e-04  7.469e-05   4.198 2.70e-05 ***
tenure                             -5.714e-02  6.748e-03  -8.467  < 2e-16 ***
ContractOne year                   -7.955e-01  1.186e-01  -6.707 1.99e-11 ***
ContractTwo year                   -1.408e+00  1.892e-01  -7.443 9.88e-14 ***
InternetServiceFiber optic          6.192e-01  9.986e-02   6.201 5.61e-10 ***
InternetServiceNo                  -1.056e+00  1.464e-01  -7.211 5.54e-13 ***
OnlineSecurityNo internet service          NA         NA      NA       NA
OnlineSecurityYes                  -4.325e-01  9.307e-02  -4.647 3.36e-06 ***
TechSupportNo internet service             NA         NA      NA       NA
TechSupportYes                     -4.912e-01  9.480e-02  -5.182 2.20e-07 ***
PaymentMethodCredit card (automatic) -7.351e-02  1.261e-01  -0.583 0.559878
PaymentMethodElectronic check       3.608e-01  1.038e-01   3.477 0.000507 ***
PaymentMethodMailed check          -8.534e-02  1.264e-01  -0.675 0.499607
StreamingTVNo internet service             NA         NA      NA       NA
StreamingTVYes                      2.503e-01  8.968e-02   2.791 0.005254 **
StreamingMoviesNo internet service         NA         NA      NA       NA
StreamingMoviesYes                  2.038e-01  8.944e-02   2.278 0.022712 *
```

Figure 7. logit- model- best- fit

5.5. Comparison of Four Models

In order to have a more accurate prediction of customer churn, the study compares four models by checking ROC curves and AUC scores. When it needs a single number to summarize the performance of a model, AUC is the best choice among all metrics because it considers classifier performance comprehensively. The result of ROC and AUC is shown in Fig. 8, where the color blue represents the Classification Tree Model, green represents Random Forest Model, black shows the Support Vector Machine Model, and red stands for Logistic Regression Model. Among these four models, Logistic Regression has the highest AUC score, which is 0.852. Support Vector Machine has the lowest AUC score, 0.811, which,it can be supposed due to the increasing number of false negative cases. Also, judging by the data, Random Forest Model has a higher AUC score than the Classification Tree Model, which are 0.837 and 0.818 respectively. Overall, the result suggests that Logistic Regression is the best model to predict customer churn.

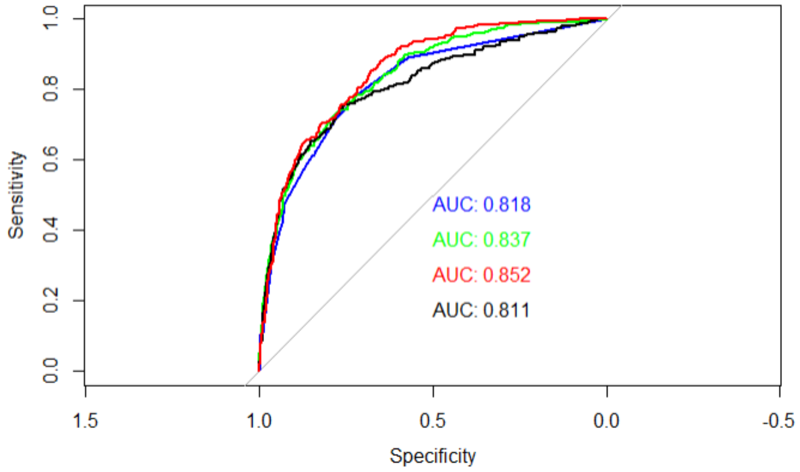


Figure 8. ROC - AUC curve

6. Conclusion

This paper mainly explores the application of different machine learning algorithms in customer churn prediction. The conclusions were drawn are as follows:

According to the logistic regression model, tenure and total charges are two numerical variables that significantly affect the churn probability. The rise of total charges increases the churn probability, whereas the increase in tenure decreases the churn rate. From this information, Telco can create a more attractive pricing plan for different lengths of multiple services. For customers who have longer-time service, Telco can charge less money proportionally each month to decrease the total charges. The customers' features of contract-one-year and contract-two-year affect the churn rate in a similar way as tenure. Telco is expected to form a price competitive contract for customers with longer service and to maintain the quality of service for a longer time, resulting in the enhancement of customer loyalty.

Besides the price and service time of customers, customers' services they signed up for like technology support, online service, and internet service are also significant for the churn decision. The internet service of fiber optic increases the churn decision by 0.62 units, while no internet service decreases the churn rate. Customers with online-security services and tech-support services are less likely to churn. It could deduce that the fiber optic internet service needs to be improved significantly by Telco if they intend to continue this service, otherwise they would lose more customers. Additionally, the company is supposed to keep its advantages on the other two services and improve its service quality sustainably. There might be a price issue of fiber optic service, which cannot be detected from the model due to the lack of data. Hence, other than considering the service quality, Telco might also need to collect more data about prices and to see how customers react to a variety of prices.

For more personal preference services, customers who have streaming TV services are more likely to churn by 0.25 units more than those who do not. Similarly, streaming movies service increases the churn rate by 0.2 units. This information indicates that Telco should improve its streaming TV or movie services or decrease the price of these services. Also, customers might not be interested in these services and the company can delete these services to lower their costs. However, it requires more investigation and market research to decide whether streaming TV or movie services are meaningful for Telco to build up customer loyalty. The decision-making would be complicated.

The risk of this project is the limit of scope in customers' characteristics. It can be assumed variables in the dataset are variables possible to affect the churn decision, but there might be other variables we did not detect and include in the project. Moreover, we also limited customers' choice of services which Telco provides to be yes or no. However, customers' churn decisions might be also related to the quality and price levels of the services, this project does not allow for the quantification of these factors. These risks can be mitigated by conducting more comprehensive marketing research, which includes customer surveys and the grading of customer satisfaction with Telco services. by following these approaches, obtaining a closer look at the specific service problems the company has, leading to a more precise prediction of the customers' churn decision.

References

[1] K. Dahiya and S. Bhatia.Customer churn analysis in telecom industry. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2015, pp. 1-6, doi: 10.1109/ICRITO.2015.7359318.

[2] Ahmad, A. K., Jafar, A., & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data,2019, 6 (1), 1 - 24.

[3] Rokach, L., Maimon, O. Classification Trees. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. 2009.https://doi.org/10.1007/978-0-387-09823-4_9.

[4] Rigatti, Steven J. "Random Forest." Journal of Insurance Medicine, 2017, 47 (1), 31 – 39, meridian.allenpress.com/jim/article/47/1/31/131479/Random-Forest, https: //doi.org/10.17849/insm-47-01-31-39.1.

[5] Jakkula, Vikramaditya. Tutorial on Support Vector Machine (SVM). 2006.

[6] Boateng, Ernest Yeboah, and Daniel A. Abaye. A Review of the Logistic Regression Model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing, 2019, 7 (4), 2019, 190 – 207, https: //doi.org/10.4236/jdaip.2019.74012. Accessed 6 Nov. 2023.