Fall 2020 IST 718 Project Final Report

**Telecom Customer Churn Analysis**

Group 13 | Tian Jiang, Wanyue Xiao, Yiyuan Cheng

Table of Contents

[Abstract 1](#_Toc57544605)

[1. Data Exploration 2](#_Toc57544606)

[1.1. Data Description 2](#_Toc57544607)

[1.2. Data Cleaning and Preprocessing 2](#_Toc57544608)

[1.3. Exploratory Data Analysis 3](#_Toc57544609)

[2. Methodology 5](#_Toc57544610)

[3. Models 6](#_Toc57544611)

[3.1. Random Forest 6](#_Toc57544612)

[3.2. Gradient Boosting Tree 7](#_Toc57544613)

[3.3. Logistic Regression 8](#_Toc57544614)

[3.4. Linear Support Vector Machine 9](#_Toc57544615)

[3.5. K-Means Clustering 10](#_Toc57544616)

[4. Conclusion 10](#_Toc57544617)

**Abstract**

To achieve the goals of revenue increase and of existing customers retention in the competitive environment, this study aims to discover critical features influencing churn rate, and then to develop models that are capable of predicting customer churn after comprehensive data exploration and analysis. Therefore, using data about customer personal information, account, and registered telecom services to predict customer leaving is a forthright prediction goal. One classification algorithm is used for investigation of consumer distribution while four prediction techniques (Logistic Regression, Linear Support Vector Machines, Random Forest, Gradient Boosting Tree) are utilized in churning rate prediction.

To accomplish the prediction goal, five inference questions have been proposed, which could be divided into three categories. For the first one, which is to discover potential factors that might significantly influence the target variable, a feature importance table obtained by the Random Forest model and the Gradient Boosting model could be a crucial indicator. What’s more, by using the coefficients of features generated by the Logistic Regression model, the most valuable features in predicting loyal customers could be identified, which answer questions covered in the second category. For the inference question of data linearity, the efficiency and performance of Tree-based models and regression-based models have been compared, revealing the result that linear models are capable and competent.

This study successfully developed, concluded, and evaluated the four prediction models mentioned above, providing comprehensive suggestions to the proposed questions. The result indicates customers with Online Security and Tech Support Services prone not to default while that customer with characteristics of partnership and longer contract tend to stick with the company. Even though this study used coefficients as indication of feature importance, the result still has the possibility to varied until a significance test and t-test, which verify the significance level of each predictor, the have been conducted.

1. **Data Exploration**

## Data Description

Containing 7043 rows and 21 columns, the explanatory variables in telecom customer churn dataset could be separated to three categories, personal attributes, accounts details, and services registered up by customer. Customer attributes contains *gender*, *seniorCitizen*, *partner*, and *dependents*. For the customer account information, it includes *tenure*, *contract*, *MonthlyCharge*, *TotalCharges*, etc. The various services provided by company are included in the third group, ranging from information protection (such as online backup) to entertaining services (such as Streaming TV). Each service attribute possibly has a value of “Yes/No” denoting whether the customer has signed up for this service or not, or “No Service”, which stands for no such service provided.

Except the *tenure*, *MonthlyCharge*, and *TotalCharges* variables included in the account details group, all of the rest variables are categorical type. Specifically, *tenure* is a numeric variable which represents the number of months that a customer sticks to the company. Investigating the distribution of churning customer across different stages of tenure might be insightful for understanding the dispersion between churning customer and customer who left yet fail to find a new service provider.

## Data Cleaning and Preprocessing

The preprocessing process contain three steps, including index variable dropping, data type transformation, and missing values management.

1. The *CustomerID* has been eliminated since each row in the dataset represents a unique yet anonymous customer.
2. The data type of numeric variables has been transformed into float type since some variables have decimal values involved.
3. 11 missing values have been found in *TotalCharges* column with a zero value in *Tenure* coincidentally. Customer with zero in Tenure column should be treated as invalided. Therefore, all of those values have been imputed with zero.

Accomplishing preprocessing is not the end. The churn prediction dataset has the class imbalance problem. Besides, due to the large number of features, the technique of dimensionality reduction or feature selection, imbalance management are necessary for data preprocessing. Apart from that, considering that most of the machine learning algorithms only accept numerical values, it is ineluctable to apply encoding techniques, such as String Indexing and One Hot Encoding techniques, to categorical features.

1. A balancing ratio was been calculated based on the target object, which has 73.5% of ‘No’ against 26.5% of ‘Yes’. Then a weightCol parameter to under-sample the negative class and over-sample the positive class, indicating that algorithm should treat the negative class with lower weight.
2. The cumulative principal component analysis shows that, by containing 17 components inside, the cumulative sum of variance explained could reach to 0.95. The model with 19 components will yield a value of 0.98. Given that the abundant computational power and the negligible difference of values, it is less helpful to conduct PCA.
3. Apart of feature engineering mentioned above (encoding techniques), the dataset has been standardized and normalized to set all variables in the same unit measuring.

## Exploratory Data Analysis

The correlation analysis on data could be beneficial for key features identification. Accordingly, a strong correlation (which has absolute values above 0.3) could be found between target variable and *Contract*, *OnlineSecurity*, and *TechSupport* respectively. It is worthwhile to point out that *TotalCharges* and *MonthlyCharges* seems to have a rather strong correlation, which eventually will cause multicollinearity issue. Therefore, Variable Inflation Factors (VIF) could be used to detect the strength of the correlation of a variable with a group of other. What’s more, churn correlated with tenure and payment through electronic check.

A picture containing chart

Description automatically generated

Graph 1 Corelation Map of Customer Churn Dataset

To identify pattern that causes customer churn, conducting exploratory analysis in inevitable. The leftmost plot illustrated by feature *Tenure* and *Churn* shows how churn is distributed across the customer lifespan.

A picture containing histogram

Description automatically generated

Graph 2 Density plot of Tenure, Monthly Charges, and Total Charges by Churn

To identify pattern that causes customer churn, conducting exploratory analysis in inevitable. The leftmost plot illustrated by feature *Tenure* and *Churn* shows how churn is distributed across the customer lifespan. One can summarize that the majority of customers, regardless of churning result, chose to cancel their subscription in the first 10 month. Reasons for such a result might be unsatisfactory experience and trial periods. while the number of customers who did not switch to another platform after leaving increased in the 70 to 80 section. Combing the rest two plots, one could speculate that lower monthly charges would help in reducing churn while the total charges might have limited influence on customer churn, which is unexpected, since the trend of customers are similar regardless of churning.

Chart, bar chart

Description automatically generated

Graph 3 Bar plot of TechSupport, Contract, PaymentMethod, and OnlineBackup by Churn

Based on the correlation plot, those four plots show the features where high discrepancies between classes could be observed, providing insights regarding which option chosen by customers who are more likely to defeat.

* Accordingly, customer that do not have online backup and tech support services enabled are more likely to leave.
* For the contract type, the largest majority of customers with month-to-month contract type tend to cancel the subscription.
* Customers who have an Electronic Check Payment Method tend to leave the platform more when compared to other options.

# Methodology

After dataset preprocessing and visualization, some insightful patterns might be discovered. Therefore, one could use the information obtained above to adjust the input features, such as imbalance handling, data transformation, feature engineering. The next step is to develop model respectively and provide a detailed model summary. After having all the essential statistics, models comparison and evaluation could be conducted by using evaluation metrics. Finally, based on the evaluation result, all those inference questions could be answered specifically.

Diagram

Description automatically generated

Graph 4 Methodology Flow Chart

# Models

## Random Forest

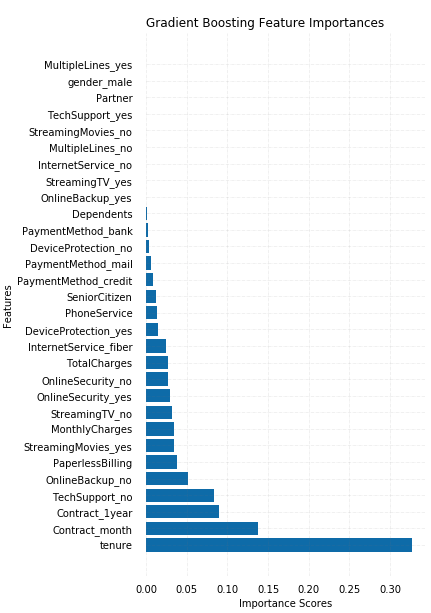
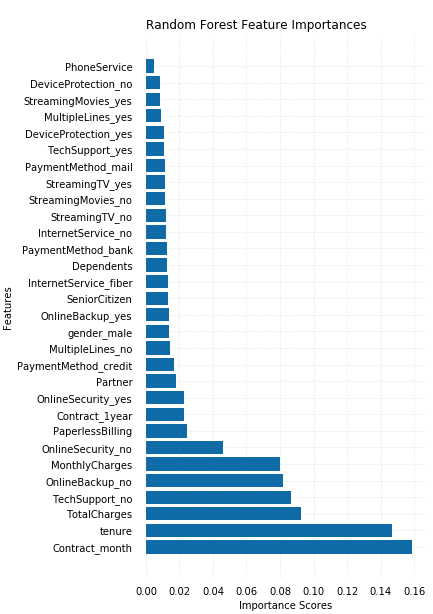
First of all, the leaving customers are classified by the Random Forest model. The model is predicting the target variable “Churn” based on all the independent attributes. Table 1 displays the main metrics used to evaluate the model performance. The majority of them approaches to around 0.8, which indicates that Random Forest model performs well. The number of the area under ROC curve reveals that the model could explain 84% variances of the test dataset.

Table 1 RF model performance summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrics | **Accuracy** | **AUC (PR)** | **AUC (ROC)** | **Precision** | **Recall** | **F1 Score** |
| Score | 0.776898 | 0.647922 | 0.837863 | 0.807569 | 0.875522 | 0.840174 |

Among these model evaluation indicators, the accuracy rate would be the most direct and efficient one. As for precision and recall, the former represents the percentage of the true positives among all the predicted positive values, and the later evaluates how the model performs regarding minimizing the false positives. F1 score measures the average performance of precision and recall. Specifically, since the project goal is to understand the churned customers so that the company can reduce customer leaves, the group wants to predict them (the positive group) more accurately. Thus, a fewer false negative and a higher recall would be expected. The group would also assess the proper metrics to understand the different aspects of the models.

According to the feature importance scores extracted from the best Random Forest Model (Graph 5), monthly contract and tenure are the two most important features, followed by total charges, not registering tech support, not registering online backup service, and monthly charges. Besides, personal characteristics such as partnership, gender, seniority as well as dependency have a faint influence on customers’ willingness to leave. Other service registration conditions regarding online security (yes), multiple lines (no), internet service (Fiber), and streaming TV (no) also exert a slight function on the prediction process.



Graph 5 Bar plot of RF importance scores Graph 6 Bar plot of GBT importance scores

## Gradient Boosting Tree

The other tree-based model used to predict a churned customer is the Gradient Boosting model. The performance of this model is also desired as its accuracy rate is nearly 75%. Noticeably, this model achieves a higher recall score of 0.908, which means it performs better in terms of minimizing the false negatives. In contrast, the precision of the Gradient Boosting Tree falls significantly. As precision and recall varies a lot, F1 score would provide a thorough view about the prediction outcomes of this model. Therefore, it can be concluded that the Gradient Boosting model has the strength in predicting churn but with the trade-off of classifying stayed customers.

Table 2 GBT model performance summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrics | **Accuracy** | **AUC (PR)** | **AUC (ROC)** | **Precision** | **Recall** | **F1 Score** |
| Score | 0. 749884 | 0. 665218 | 0. 846863 | 0. 728672 | 0. 9088 | 0. 808828 |

Compared to the Random Forest model (Graph 5), the Gradient Boosting model considers less features, assigning a much higher score to the variable tenure. Apart from that, GBT classifies the leaving customers based on having a monthly or one-year contract, not registering tech support, and not registering online backup. In addition, given that the effects of features such as paperless billing, registering the online security service, and paying by credit card (automatic) are relatively consistent in two models, they could also shed light on the inference analysis.

## Logistic Regression

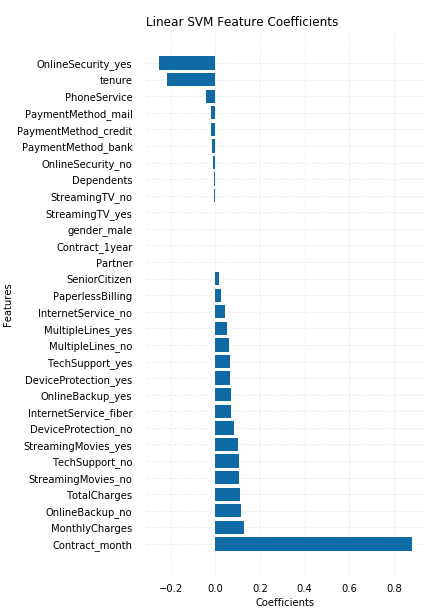
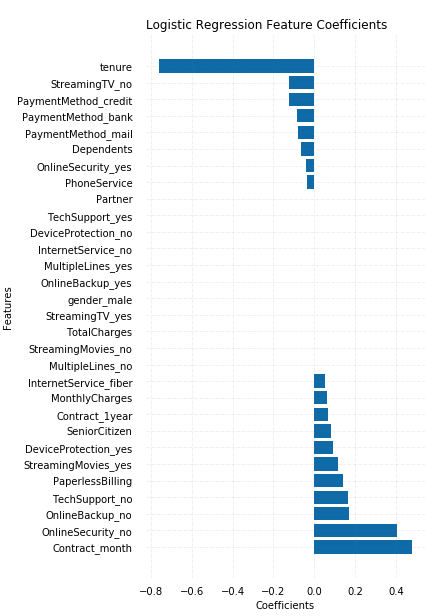
The table below shows performance of using Logistic Regression to predict whether a customer will stop using the services provided by the telecom company. The model accuracy is 75.83% and could explain 85% variances of the test dataset. Its evaluation is close to Gradient Boosting tree in every aspect. Again, there is a noticeable difference between precision and recall. So, the group values the F-measure more for its thoroughness.

Table 3 LR model performance summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrics | **Accuracy** | **AUC (PR)** | **AUC (ROC)** | **Precision** | **Recall** | **F1 Score** |
| Score | 0. 758267 | 0.668801 | 0. 8476436 | 0. 739576 | 0. 910742 | 0. 816283 |

The Pyspark Logistic Regression model (and the Linear SVM model to be mentioned in the next section) only provides coefficients for each feature in its inherent model summary. However, directly interpreting the coefficients can be risky if all explanatory variables are not standardized or without knowing the corresponding statistical significance. As for this project, even if there is no efficient way to calculate the statistics, all data has been scaled, and the coefficient outcomes are in line with the feature importance results. Providing these, the group decided to interpret coefficients for the features that are identified as critical by the Random Forest and Gradient Boosting models. In other words, the feature importance scores will be properly utilized as the significance statistics. Therefore, the coefficients can help the group to elaborate on the effects of the selected features.

It is illustrated that with a Logistic Regression model, tenure, not registering Streaming TV, and paying by credit card (automatic) exert a negative effect on customer churn, i.e., positive effect on customer loyalty. On the contrary, a monthly contract and not registering online security, online backup, or tech support services is predicted to increase the likelihood of leaving (Graph 7).



Graph 7 Bar plot of LR coefficients Graph 8 Bar plot of Linear SVM coefficients

## Linear Support Vector Machine

Linear Support Vector Machine is the last supervised machine learning model used to classify two categories of customers. It is indicated that the model performance is the lowest among the four models, with an accuracy of 0.707. The imbalance between precision and recall is dramatic, leading to a 0.764 F1 score. A linear SVM might not be such a suitable model because there are too many dimensions that need calculation, and the model cannot distinguish them effectively.

Table 4 Linear SVM model performance summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrics | **Accuracy** | **AUC (PR)** | **AUC (ROC)** | **Precision** | **Recall** | **F1 Score** |
| Score | 0. 707033 | 0. 662248 | 0. 846349 | 0. 652982 | 0. 920433 | 0. 763977 |

The inference results of Linear SVM are similar to the previous analysis (Graph 8). A monthly contract as well as a higher regular charge is more associated with churn while a longer tenure and having registered the online security service have a positive impact on customer staying.

## K-Means Clustering

Chart, scatter chart

Description automatically generatedPrior to conducting the K-means clustering analysis, principal components are calculated. It is identified that 10 principal components would explain more than 70% cumulative sum of variance. Silhouette score is used to find the optimal cluster number k, and the score is the highest when k is 3. Clustering is unsupervised learning, so the group is unable to extract the metrics evaluating the model performance. But through visualizations, the clusters are clearly distinguishable. Apart from that, clustering is not an effective tool to compare the contribution of different features. Since the principal components cannot point to the actual attributes, it is difficult to interpret the output cluster and thus understand which attribute makes a difference.

Graph 9 Clustering Plot

# Conclusion

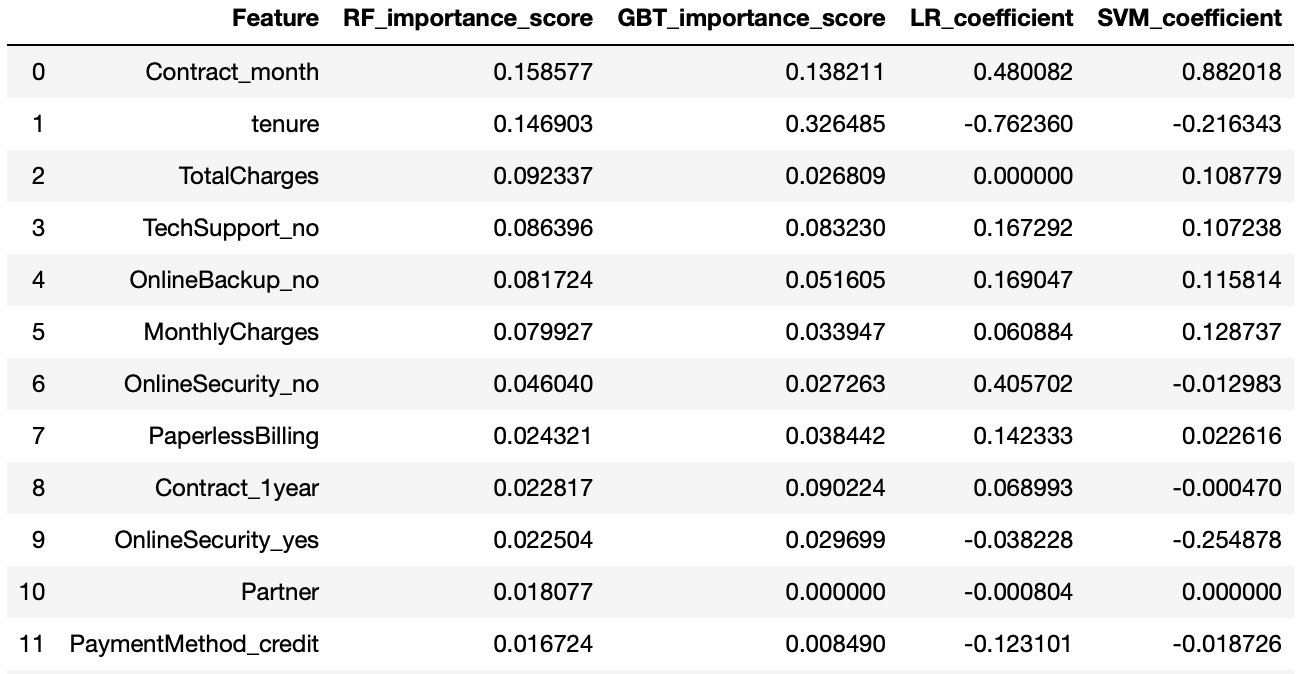
Chart, bar chart

Description automatically generated

Graph 10 Evaluation and Comparison of Random Forest, Gradient Boosting Tree, Logistic Regression, and Linear Support Vector Machine Models

Using the combination of grid search and three-fold cross validation, an average accuracy of 0.74 could be obtained across the four supervised models in order to predict churning customers. Accordingly, Random Forest model outperformed the others in terms of the accuracy rate (0.7769) and F-measure (0.8402), followed by Logistic Regression and Gradient Boosting model whose accuracy reaches 0.7583 and 0.7499, respectively. They also have more strength in minimizing the false negatives if considering recall while the telecom company prioritizes churning customer prediction. However, the increases in recall score are accompanied by sharply falling precisions, giving rise to the risk of biased prediction. So, the F-measure of Logistic Regression and Gradient Boosting is not as desired as the Random Forest model. Last but not least, given that the accuracy rates for Logistic Regression and for Linear Support Vector Machine are all above 70%, one can conclude that a linear model is capable of describing the data relationship.

Table 5 Inference summary



To sum up, the focal features for predicting customer churn are two account-related variables, tenure and contract type, whereas the importance of personal features is trivial. The reliable indicators of a loyal customer would be a longer tenure and a registration on the online security or tech support service, whereas a churned customer is more indicated by having a monthly contract. Services such as tech support, online backup, and online security is inferred to be an advantage to the company. Therefore, to prevent customers from leaving, a length contract, a recommendation in registering featured services, and an automatic payment would be suggested.

# 