Telecom Customer Churn

IST 707 Group 6 | Jenny Cao, Jieer Chen, Xiaodan Yu

Presentation Video Link: https://youtu.be/vQrpZtirme4

Introduction

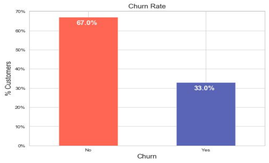
Churn is a one of the biggest problems in the telecom industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%. Through this project, we are trying to answer the questions about what are the significant factors that determine customer churn, which machine learning model can help us better predict customer churn, and what should telecom companies do to increase retention.

Data Description

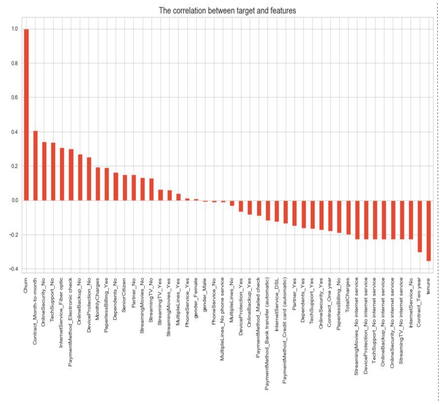
The dataset used in this project is from Kaggle.com. The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target. Each row represents a customer, each column contains the customer's attributes described on the column Metadata. The dataset can be used to analyze all relevant customer data and develop focused customer retention programs from the IBM Watson analytics community.

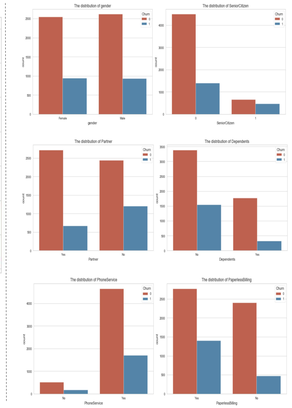
Data Preparation

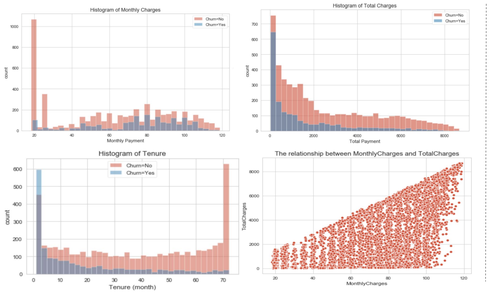
- Change data type
- Clean the missing value
- Get dummy variables (One-hot Encoding)
- Normalization

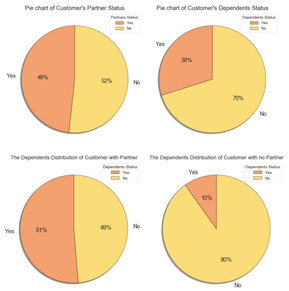


Exploratory Data Analysis (EDA)





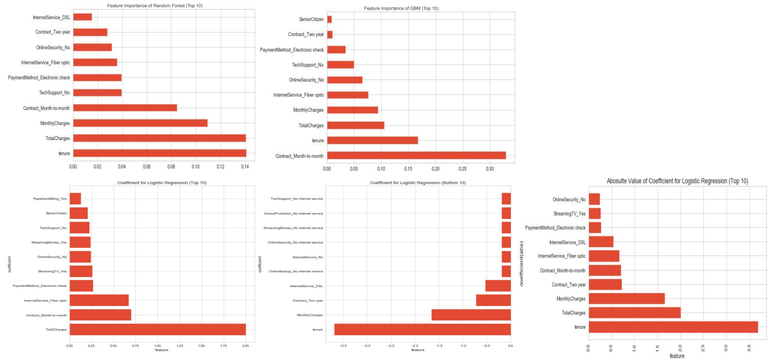




Model Highlights

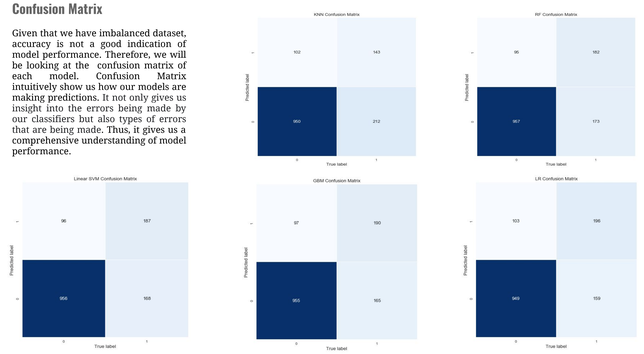
We identified the ten most important features by plotting feature importance in Random Forest and Gradient Boosting Machine. The important factors includes but not limited to "Contract Month to Month", "Total Charges", "Monthly Charges", and "Tenure". These four features are among the top 5 in both models. As previously discussed in the EDA section, consumers who have tenure with higher number of months are less likely to leave the company than consumers who have tenure with lower number of months. Based on the result from the correlation matrix, both "Contract Month to Month" and "Total Charge" has a positive correlation with our target variable Churn. Therefore, we can conclude, consumer with a month to month contract with Telecom company are more likely to churn. Consumer with higher total charge are less likely to leave the company. However, consumer with higher monthly charge are more likely to leave the company.

Feature Importance



Confusion Matrix

Given that we have imbalanced dataset, accuracy is not a good indication of model performance. Therefore, we will be looking at the confusion matrix of each model. Confusion Matrix intuitively show us how our models are making predictions. It not only gives us insight into the errors being made by our classifiers but also types of errors that are being made. Thus, it gives us a comprehensive understanding of model performance.



Model Summary

We generated a model performance master table that includes all the algorithms we applied to our data. We used five consistent model performance evaluation methods (i.e. use 10 fold cross validation to get the prediction accuracy). The table is sorted in the decreasing order of the model performance for the ROC_AUC Score metric. For model comparison, we decided to use AUC Score as the key primary metric. The reason is that AUC measures how well a parameter can distinguish between two diagnostic groups, which is the same as our goal: we want to ensure that the two target labels are separable.

Algorithm	Key hyperparamters	Precision	Recall	Accuracy	ROC AUC Score
GBM	learning_rate:0.06, loss:exponential, max_depth:2, n_estimators:110, subsample:0.5	66.20%	53.52%	81.38%	84.28%
Logistic Regression	C:6.158, penalty:l2	65.55%	55.21%	81.38%	84.08%
Linear SVM	C:0.5, penalty:l2	66.08%	52.68%	81.24%	83.55%
Random Forest	bootstrap:True, max_feature:auto, min_sample_leaf:2, min_sample_split:5, n_estimators:25	65.70%	51.27%	80.95%	82.84%
KNN	metric:manhattan, n_neighbors:6, weights:uniform	58.37%	40.28%	77.68%	78.31%

Conclusion

In this project, we have implemented 5 models in total. With the ROC_AUC curve as the key primary metrics, we identify Random Forest as the best prediction model. It seems like tenure is an important feature here, the longer the tenure, the less likely it is to lose customers. Additionally, charges and contract also have a vital impact on customer retention.