# Customer Churn Prediction using Machine Learning

Customers are an asset to a company. There are two types of customers a company/organization can have. One is new and the other is existing customers. While increasing the number of both new and existing customers are important, keeping existing customers is far more profitable than acquiring new customers. So, it has become extremely important to understand customer behaviors to recognize possible leavers as soon as possible and take necessary measures to retain them within the company. Customer churn, also called customer attrition, is when a customer chooses not to use products and/or services of a certain company/organization.

## Data

The Telco Customer Churn dataset from Kaggle is being used for this project. This dataset contains information about a fictional Telco company based in California, in their third quarter. This company provides home phone and internet services to 7043 customers. Dataset includes 21 features including the target variable "Churn".

## Data Preparation

"TotalCharges" field was originally an object type, while we expect it to be numerical. Converting this field into a numerical field created 11 missing values, and these missing values had the tenure term as zero indicating that total charges for these records should also be zero. "Customer ID" and "gender" fields were removed from the dataset since they do not seem to be valuable to be further analyzed.

## Data Exploration

Nearly 74% of the customers we have in this dataset, have not churned. This signals that we have an imbalanced data set here.

Table 1: Churn and Not Churn Summary

| **Status** | **Percent** |
|---|---|
| Not Churn | 73.5% |
| Churn | 26.5% |

Out of those who churned, 64% have no partners, 83% have no dependents, and 75% are not senior citizens. With this information, one can conclude that a young, single person is more likely to churn than an older person with more responsibilities. This makes sense since younger people tend to have time and the interest in inspecting pros and cons of products and services being provided. Hence they may be switching from one company to another more frequently, after observing the quality and expenses of the services they receive.

Two main services this company provides are home phone and internet services. During the third quarter, of those who left the company 69.4% have used Fiber Optic Internet. On the other hand, only 41.9% of those who are using fiber optic have left the services, while others are still active customers. If we rank the customers who are still using the Internet through the company, 37.9% are using DSL, 34.8% are using Fiber Optic, and 27.3% are not using their Internet service.

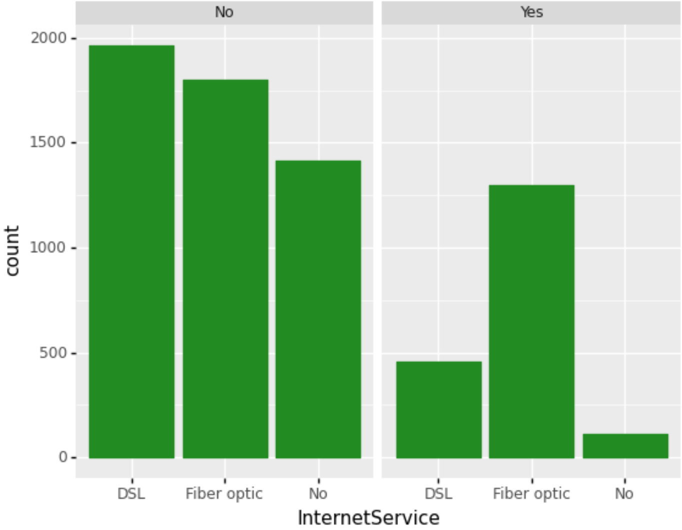

Figure 1: Number of customers using Internet Service, sectioned by the Churn Status

Of those who churned, 91% have received phone services and a similar percentage (90%) of active customers are also receiving phone services.

83.5% of churned customers have used both phone and internet services.

Table 2: Summary of Customers using Phone and Internet Services

| **Internet** | **Did Not Churn** | | | **Churn** | | |
|---|---|---|---|---|---|---|
| | **Phone** | | **Total** | **Phone** | | **Total** |
| | **No** | **Yes** | | **No** | **Yes** | |
| **DSL** | 512 | 1450 | **1962** | 170 | 289 | **459** |
| **Fiber Optic** | 0 | 1799 | **1799** | 0 | 1297 | **1297** |
| **No** | 0 | 1413 | **1413** | 0 | 113 | **113** |
| **Total** | **512** | **4662** | **5174** | **170** | **1699** | **1899** |

Percent of Churned customers using a month-to-month contract is 89%, one-year contract is 9%, two year contract is 3%. Differences between these groups are much higher when compared with active customers. Of active customers, 43% have month-to-month contracts, 32% have two-year contracts, and 25% have one-year contracts.

Average monthly charge of churned customers is $74 while the average monthly charge of active customers is 61%. But, average could be misleading since we have two imbalanced groups here.

For all the contract types, monthly charges distributions of churned customers seem to be left skewed with a higher median monthly charge than that of active customers.

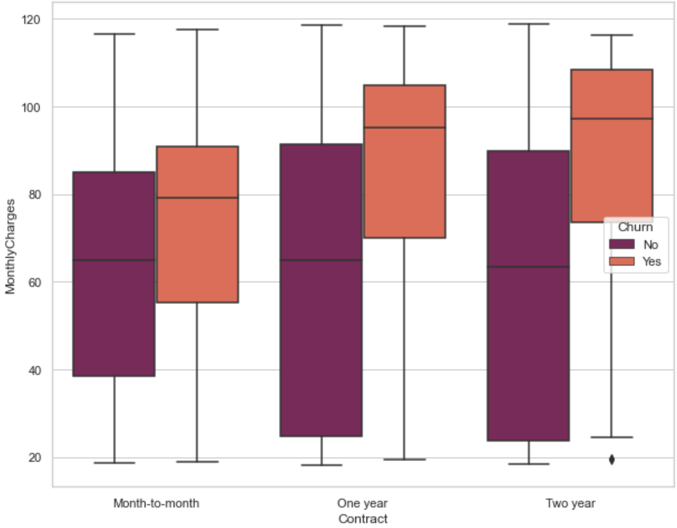

Figure 2: Monthly Charges with Contract for Churned and Active Customers

It can be observed from the below figure that most of the customers who have left the services of the Telco company, have used them for only a shorter period: majority of them have used the services less than 20 months. Tenure distribution of active customers shows that there are many who have stayed less than 2 months in this dataset and many who have stayed between 60 and 80 months. It exhibits that first 10 months is critical because that seems to be the time frame a customer decides whether to leave or stay.

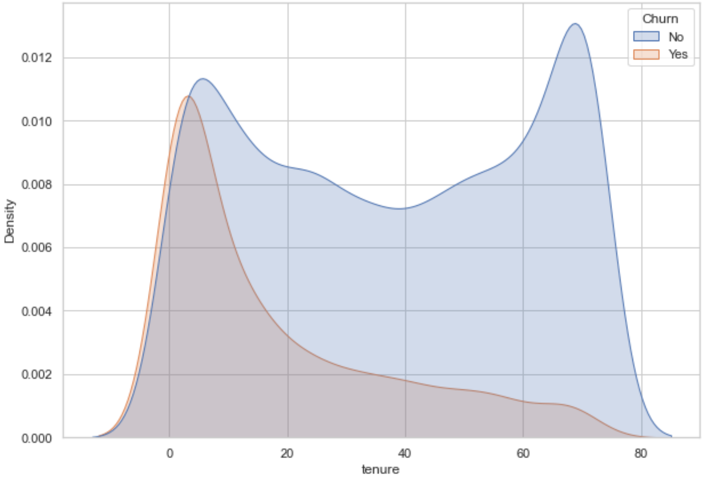

Figure 3: Number of months Churned and Active customers have been using the services provided by the Telco Company

Higher the number of months the customer has stayed with the company, higher the range of total charges. Ranges of the total charges of churned customers are less than that of customers who did not churn, and the charges of churned customers are mostly at the upper level. Also this signals that tenure is highly correlated with total charges. Correlation coefficient is used to verify the high correlation between tenure, total charges, and monthly charges; total charges is highly correlated with tenure and slightly correlated with monthly charges.

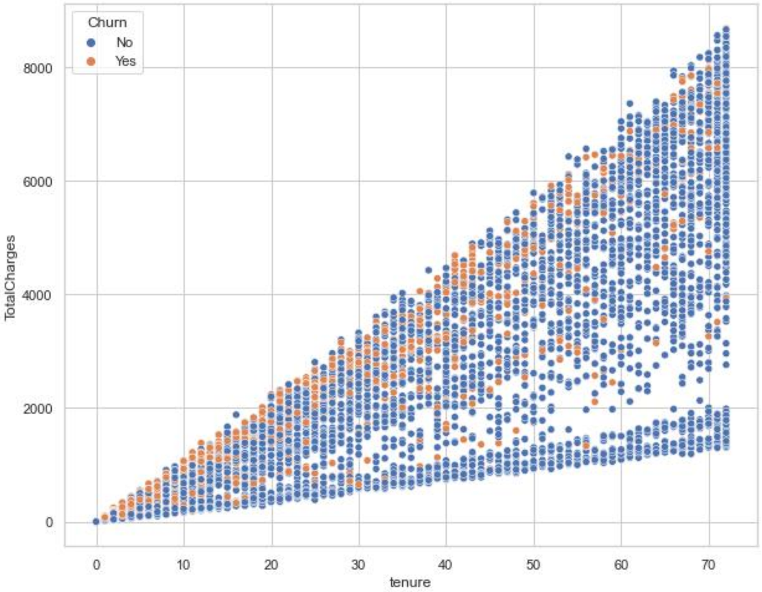

Figure 4: Total Charges for Different Tenures

Additionally, hypothesis testing was used to compare the means of monthly charges of the two groups: churned customers and customers who did not churn. Same techniques were used to identify if there is a significant difference in the mean of tenure between churned and active customers. However, according to data there is not enough evidence to support that there is a significant difference between the group means.

As for the categorical variables, chi square tests are being used to identify any dependencies between some selected variables with the churn variable. Those selected variables are, “InternetService”, “OnlineBackup”, “PaymentMethod”, and “Contract”. There was no indication of dependencies.

## Data Preprocessing and Modeling

As the baseline model, Logistic Regression is being used. Before feeding the variables into the model, we have to convert categorical variables into numerical variables. Some of the transforming techniques are,

1. Label encoding: Assign numbers to each category in categorical variables. This can be harmful in predictions, if there is no hierarchy present in categorical variables.
2. Target Encoding: Replace a categorical feature with the average target value of all data points belonging to the category.
3. One-hot encoding: Convert each categorical value into a new categorical column and assign a binary value of 1 or 0.

As we have 15 categorical variables and these are all low cardinal variables, one-hot encoding is used.

Then generate train and test data sets using the 80/20 ratio: 80% of the data will be the training dataset while the other 20% is the test dataset. Then standardize the numerical variables ("tenure", "MonthlyCharges", "TotalCharges"). The reason for standardization is that these variables have different scales and hence will not contribute equally to model fitting and possibly will create biases. To avoid this, it is recommended to standardize non-binary numerical variables.

Other Models (Supervised Learning Algorithms) Used:

1. **Random Forest Classifier**: It is an ensemble algorithm, meaning it combines more than one algorithm of same or different kind for classifications. Random Forest classifier creates a set of decision trees from randomly selected subsets of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

   One hot encoded categorical variables are being used without the feature scaling step: No standardization of non-binary numerical variables as it not important when decision tree algorithms are being used.

2. **CatBoost Algorithm**: CatBoost (Categorical Boosting) is known as one of the latest boosting algorithms. It is known that this performs better than other many boosting algorithms such as XGBoost, LightGBM etc. It is designed to work on categorical data well. Other advantages of using CatBoost are high quality without parameter tuning, reduce overfitting, predict fast, and work well with small datasets.

   Feature preprocessing (one-hot encoding) or feature scaling (standardization) is not being used for this algorithm. At the time of model fitting, categorical variables are being introduced separately for this algorithm.

3. **LightGBM Algorithm:** LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model.

   LightGBM does not accept categorical variables in object format or string format. So, converted object format into category type before splitting data into train and test sets. And used them straightly into modeling.

# Model Selection

Evaluation metrics used to select the models are,

1. **Accuracy**: This measures how often the classifier correctly predicts. This is the ratio of the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

   Accuracy might be good when the target class is well-balanced, but not a good choice for unbalanced classes.

Table 3: Accuracy Scores

| **Model** | **Accuracy Score** |
|---|---|
| **Logistic Regression** | 0.797019 |
| **Random Forest Classifier** | 0.792051 |
| **CatBoost Algorithm** | 0.803407 |
| **LightGBM Algorithm** | 0.802697 |

According to the accuracy, CatBoost is the best model. But need to be careful, not make select the final model based on this as we have an imbalanced dataset here.

2. **Confusion Matrix:** It is a table with combinations of predicted and actual values. This is extremely useful for measuring the Recall, Precision, Accuracy, and AUC-ROC curves.

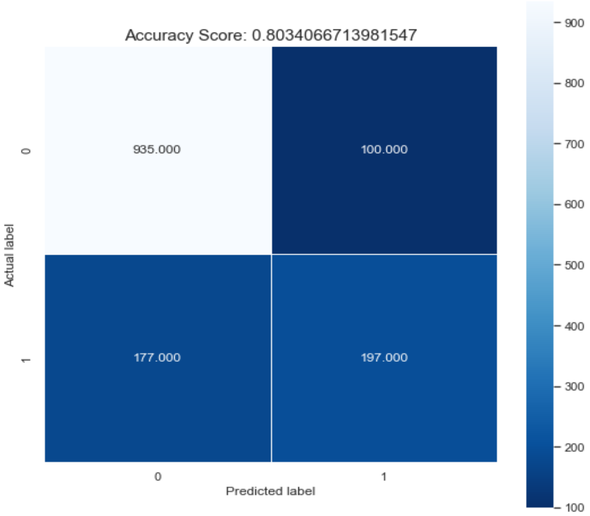

Figure 5: Confusion Matrix of CatBoost

Some metrics of the Confusion Matrix:

- $Precision = \frac{TP}{TP + FP}$
  Precision is the ratio between the number of true positives and the number of total predicted positives. Precision is useful in cases where FP is a higher concern than FN.

- $Recall = \frac{TP}{TP+FN}$

  Recall is useful in cases where False Negative is higher concern than False Positive. This is a good metric to use in Customer Churn, because predicting not churn when there is a churn is harmful for the business.

- $F1\ Score = 2.\frac{Precision*Recall}{Precision+Recall}$
  F1 Score is the harmonic mean of precision and recall. It gives a combined idea about precision and recall. F1 score is an effective metric when FP and FN are equally costly.

Table 4: F1 Scores for Churn

| **Model** | **F1 Score** |
|---|---|
| **Logistic Regression** | 0.58 |
| **Random Forest Classifier** | 0.56 |
| **CatBoost Algorithm** | 0.59 |
| **LightGBM Algorithm** | 0.59 |

- AUC-ROC – The ROC (Receiver Operator Characteristic) is a probability curve that plots the true positive rate against the false positive rate at various threshold values.
  The AUC (Area under the Curve) is the measure of the ability of a classifier to separate between classes.

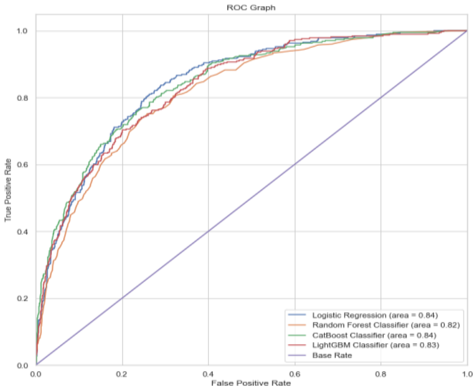

Figure 5: ROC curves of all models

F1 scores for churn are better with CatBoost and LightGBM. Observing the ROC curves AUC values for all the classifiers, CatBoost Classifier seems to be the model with the best performance.

Future Work

- Use resampling techniques (Under-sampling, Over-sampling, SMOTE) minimize the effect on imbalance data issue.
- Hyper parameter tuning to increase the accuracy and efficiency of modeling.