Predicting Customer Churn in the Telecommunications Industry

Executive Summary and Problem Definition

The Telecommunications Industry makes up a large part of our daily lives and is the backbone of long-distance communication whether it be across the globe or with friends. Almost everyone today owns a cellphone that they use to contact friends and family when they need help or to share long-lasting memories with those they are close to. However, companies in this industry suffer from a concept known as *customer churn*, which is a situation in which customers leave a company's service for a competitor out of dissatisfaction or simply outweighing benefits. Throughout this analysis, I will focus on service providers who offer (primarily) cellphone plans (contracts) to their customers. I will consider factors such as a customer's monthly data usage amount, call times per day on average, and their average monthly bill.

Based on a customer's average monthly bill for service, their monthly data usage (Gb), how long they spend on calls per day(min), and the number of calls a customer makes in a day, I will be predicting the churn status of new customers. For example, a new customer is one who recently joined with a particular service provider. This information will help companies in the Telecommunication Industry determine how they can improve their services to mitigate the number of customers who leave and at the same time, maximize their profits.

Data Cleaning (Data Preprocessing)

After the data file was read, I checked to see if there were any missing values in the dataset. Additionally, I looked at a summary of the dataset which showed preliminary summary statistics for each feature such as the mean, median, min, and max. The summary statistics are useful as preliminary information to check the distribution for each feature and potential changes I may need to make in terms of scaling the feature values. Based on the predictions I want to make for new customers for this hypothetical company, several features in the dataset were not important for making predictions. As a result, These features were dropped from the dataframe before any further analysis was done.

After creating a boxplot of Average Monthly bill vs Churn Status, I noticed that there were several outliers which were values of monthly charges greater than $100. I considered these as outliers in my dataset because many of these values were irregular compared to the overall distribution of the values. Therefore, I removed these outliers from my dataset before continuing with my analysis.

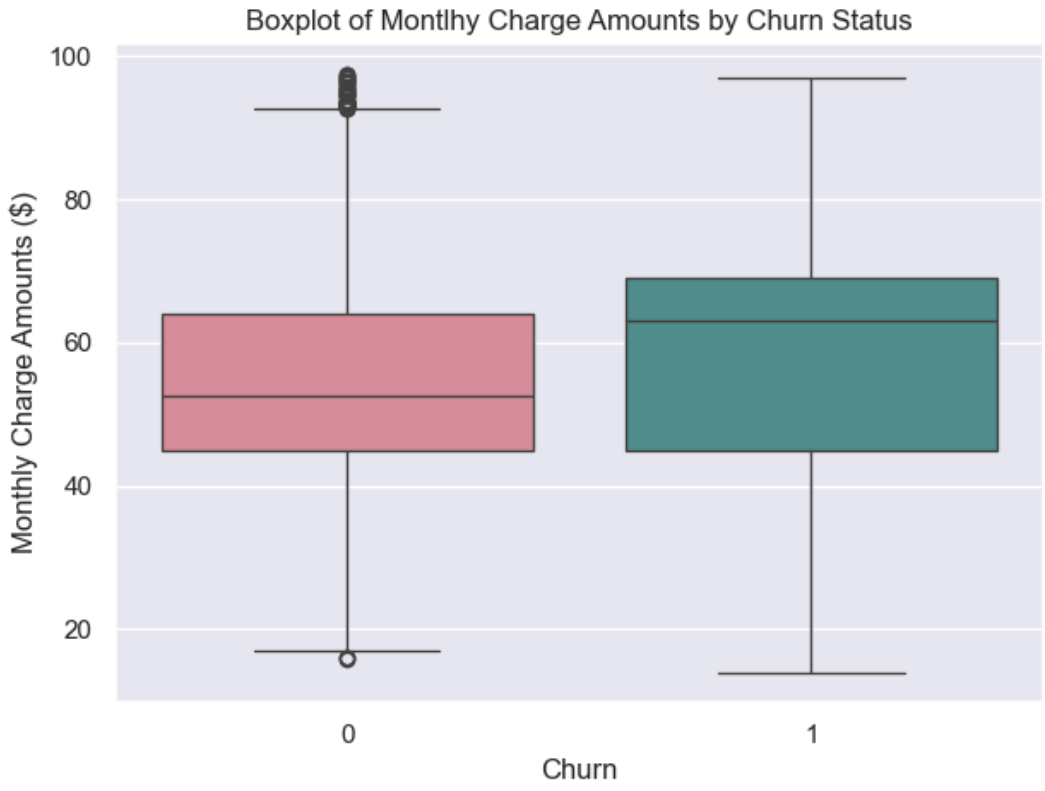
Visualizing the Data

I will be comparing two groups of consumers in this analysis: customers who stopped using the service (churned) and customers who are still using the service (non-churned). I will be comparing information such as monthly bill charges given to customers who have recently stopped using the service and customers who are currently a member of the telecommunications company. I will be assuming that charges were applied to churned customers when they were still members of the service provider.

I will focus on the following 2 features when comparing the distributions for churn vs non-churn customers: average daytime minutes and monthly bill amounts. These visualizations will be helpful to compare the differences in the distribution of these two features between churn vs non-churn customers.

Boxplot of Monthly Bill for Churn vs Non-Churn:

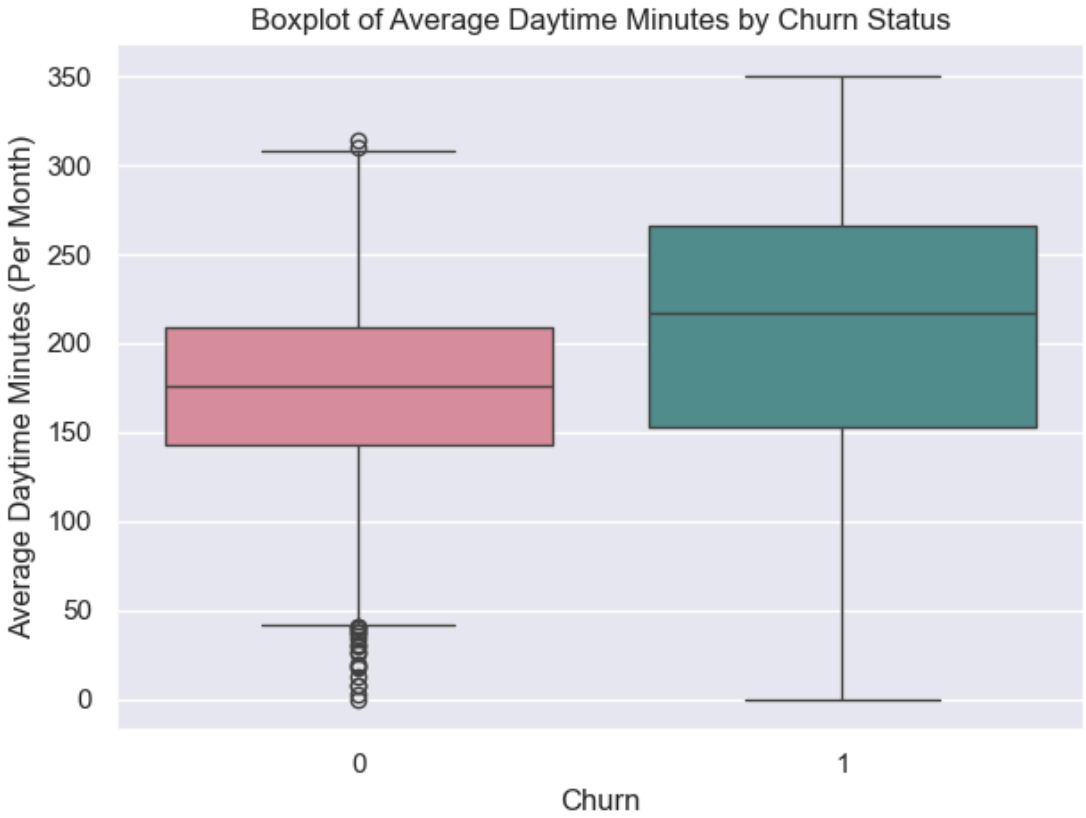
I created a boxplot to compare the distribution of the average monthly bill charged to customers who were of the churn category (Churn = 1) and of those who were not of the churn category (Churn = 0):



From this boxplot, customers who cancelled their service (Churn = 1) seemed to be charged a higher amount than customers who did not cancel their service. This suggests that customers who stopped using the service likely had a larger data usage compared to customers who were still members of the service. Other reasons could be that they were on call longer during the day or had higher daytime minutes on average.

Boxplot of Average Daytime Minutes Per Month for Churn vs Non-Churn:

I created a boxplot to compare the distribution of the average monthly daytime minutes for customers who were in the churn category and of those who were not:



From this boxplot, customers who cancelled their service were on call longer on average than customers who did not cancel their service. This means that customers who left the telecom service were charged a higher amount due to potential differences in data usage and as is seen here, a difference in average daytime minutes per month.

The findings from these plots are as expected since it makes sense to charge customers more depending on how much data they use or how long they spend on calls internationally, for example.

Performing the Relevant Statistical Tests

Chi-Square: I would like to determine whether or not the churn status of a customer is independent of their monthly bill average or not. One way to test for independence with categorical data is by using a *Chi-Squared test*. By using this test, we can obtain a p-value that will help us answer this question of independence. I performed a Chi-Squared test by considering the relationship between churn status and monthly bill averages:

Null Hypothesis of this Test: Churn status is independent of monthly charge amounts
Alternative Hypothesis of this Test: Churn status is dependent on monthly charge amounts
P-Value of this Test: **4.32*10^-36**

Based on a predefined significance level of 0.05, since the p-value is less than our significance level, we reject the null hypothesis of independence. Therefore, we can conclude that the monthly bill amounts charged to customers using the service depend on their churn status.

Two-Sample T-Test on Churn Status and Monthly Bill Averages: I would like to know if the average monthly bill amounts charged to customers are the same or different for churn vs. non-churned customers. An initial intuition of this can be drawn from the boxplot above but a statistical test can be done to gain more confidence in our results.

Before this test can be done, our data must satisfy the assumptions of a two-sample T-test. One of the main assumptions is normality in the two groups. Upon getting a histogram of churn and non-churn customers vs monthly charges, the resulting plots seemed relatively normal, this then satisfied the assumption of normality I needed before continuing with a two-sample T-test.

Null Hypothesis of this Test: The average monthly charge for churned customers is the same as the average monthly charge for non-churned customers.
Alternative Hypothesis of this Test: The average monthly charge for churned customers is different from that of non-churned customers.
P-Value of this Test: **4.87*10^-5**

Based on a predefined significance level of 0.05, since the p-value is less than our significance level, we reject the null hypothesis that the means are the same. Therefore,

we can conclude that the mean monthly bill amounts charged to customers using the service is different for churn and non-churned customers.

Mann-Whitney U-Test on Churn Status and Data Usage: To determine if the data usage amounts were different for churn and non-churn customers, I used a *Mann-Whitney U-Test.* A Mann-Whitney U-test does not assume anything about the distribution of the underlying data and since a histogram of data usage for both groups was not normal, this test was suitable. This test returned a value of **4.96*10^-10**. This implies that the data usage does differ significantly for churn vs non-churned customers.

Comparing and Contrasting Different Classification Models

Data Splitting Procedure: Before any classification models can be fit, first I will split my data into a training and testing set. This can easily be done using the train_test_split() function from the *Scikit-Learn* package.

Splitting the data is useful before fitting any model as the validation set (testing set) provides us with an independent set of data that has not been used in the training phase. Thus, the testing set is representative of real-world/unseen data. Additionally, a model's performance can be evaluated using the validation set, this tells us how well the model we chose is likely to perform on unseen data. Lastly, when comparing the performance of various models, the validation score serves as a benchmark of comparison.

Revisiting the Goal: I will predict the churn status of customers who register for the telecommunication service (new members). To do this, I will be using various classification models and comparing relevant scores on validation (test) data.

1. Naive Bayes Classifier: This classifier was fit to the data using the default parameters. To check for overfitting, I compared the scores for the training and validation set and as they are similar, no overfitting was done in this case.

Training Score	Validation Score
0.871	0.876

2. K-Nearest Neighbours Classifier: To use the KNN classifier, the data must be scaled so that the minimum value is 0 and the maximum value is 1. This scaling can be done by using the *MinMaxScaler.* Tuning on the parameters needed to be done to determine the optimal value of '*n_neigbour*s' in the function. After several iterations of tuning, I found that n = 25 was best.

Training Score	Validation Score
0.890	0.899

3.) Decision Tree Classifier: When using this classifier, there is a high risk of overfitting the model. In particular, if the '*max_depth*' parameter of the decision tree classifier is too large, then I am allowing for the decision tree to have a much more complex hierarchy, this can lead to overfitting. Therefore, I needed to tune this parameter such that the training score and validation scores were similar and not too distinct.

Training Score	Validation Score
0.925	0.876

4.) Random Forest Classifier: This is a very useful and powerful classifier that can be used to make predictions. There are mainly two parameters that I needed to tune when fitting this model. '*n_estimators*' represents the number of trees used in the model and '*max_depth*' represents the max depth of each tree. I found that using 500 trees with each having a max depth of 7 produced the highest validation score after several iterations of tuning.

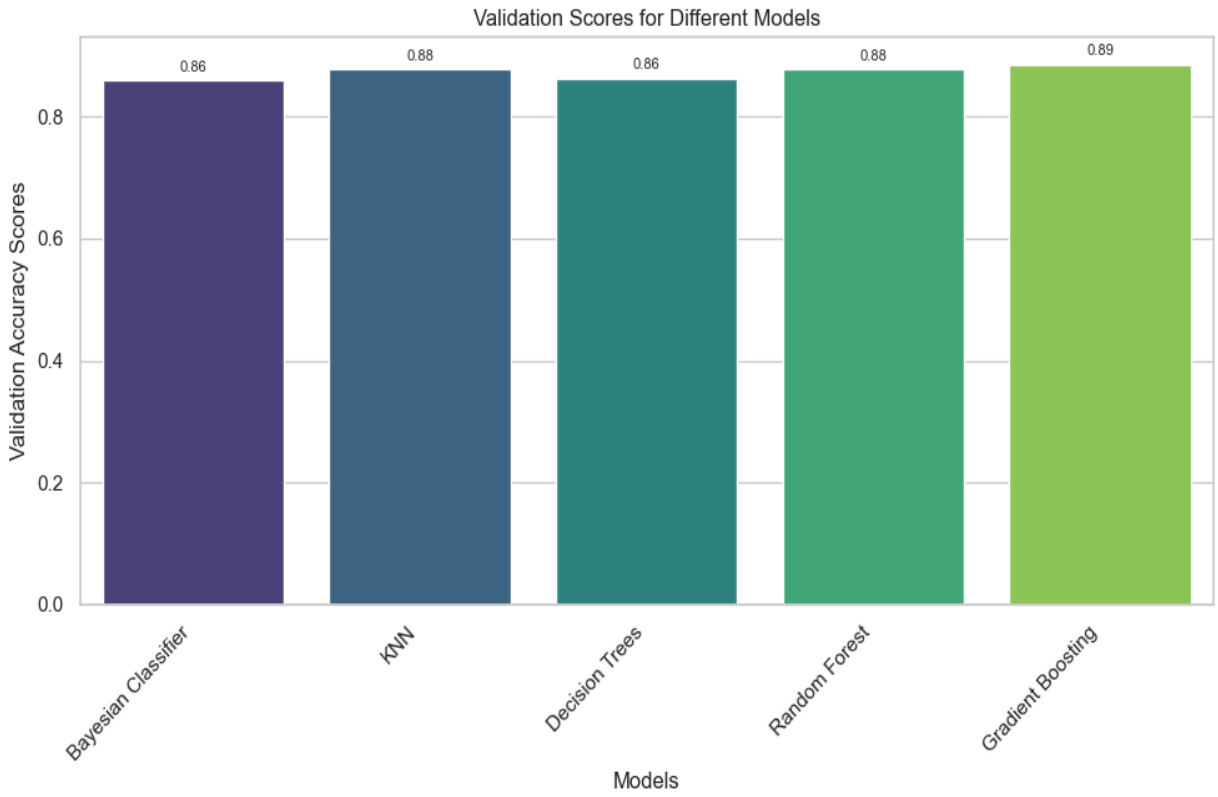
Training Score	Validation Score
0.909	0.897

5.) Gradient Boosted Trees Classifier: I fit this model using the training data and also performed several iterations of tuning for its parameters.

Training Score	Validation Score
0.903	0.896

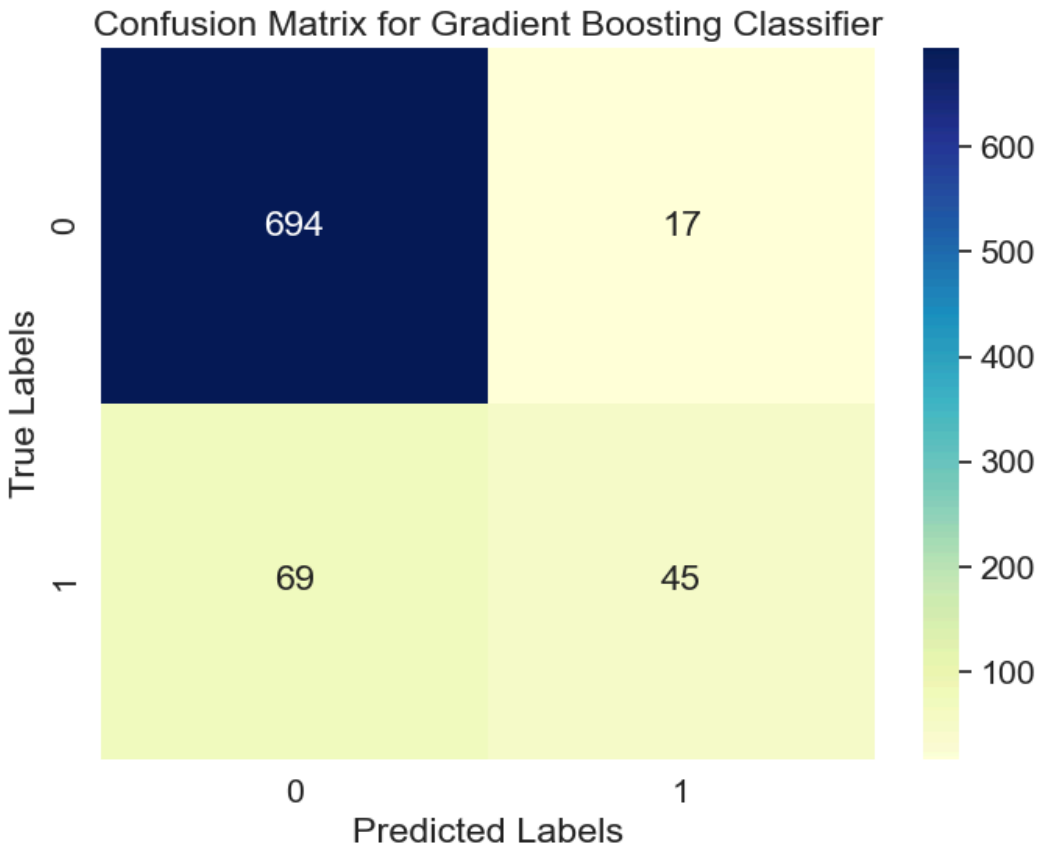
Which Model Performs Best?

After fitting each classifier, I created a barplot to compare the validation scores that were produced by each model. Based on the barplot shown below, I concluded that the Gradient Boosted Trees classifier performed the best for my specific problem.



Confusion Matrix for Gradient Boosted Trees Classifier

One way to compare the predicted values of the best-performing model and the true values of the test set is to create a *Confusion Matrix.* The confusion matrix shown below tells me that the gradient boosted trees classifier misclassified people with churn status = 1 (leaving the service) as churn status = 0 (still a member of the service) 69 times. Also, it misclassified people with churn status = 0 as churn status = 1 17 times. Overall, the misclassification rate from this classifier was around 10.4%. Therefore, 89.6% of predictions were correct using this classifier.



Conclusion

Based on a customer's average monthly bill for service, their monthly data usage (Gb), how long they spend on calls per day(min), and the number of calls a customer makes in a day, the chosen model was successfully able to predict the churn status for a new customer with a probability of around 90%. Using this model and the findings from this analysis, the company will be able to gain insight into what might happen with a new customer once they join their service. This will allow the company to provide a more seamless and high-quality service to their customers to minimize the company's churn rate and also maximize their revenue.