Overview

The goal of this challenge is to predict if a client will stop running advertising campaigns (churn) based on the provided dataset. To do this, I will use the 9 variables included with the dataset, such as the length of time the client has been a client and the number of calls/ clicks the client received as a result of the advertising. I used various machine learning models to classify customer churn, using recall as an evaluation metric, and examined the best model to make business recommendations.

Data Visualization

The dataset contains 10,000 clients, and 20% of those customers stopped running advertising campaigns with us.

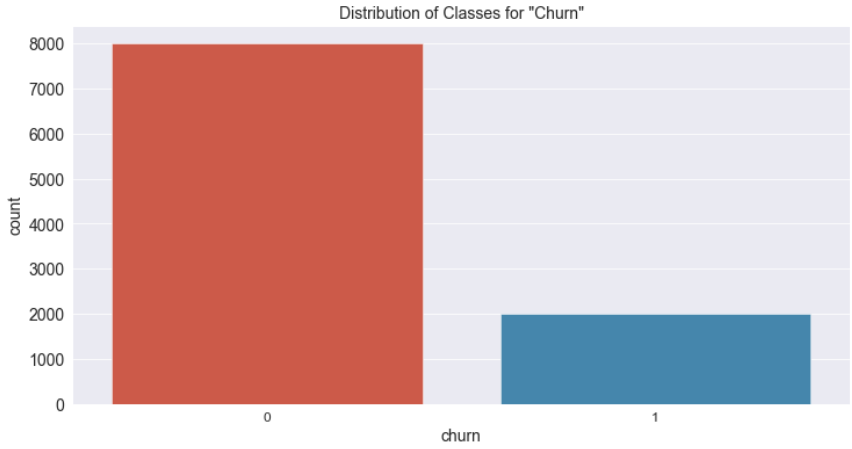

Figure 1: Customer Churn Distribution

While we were able to retain most of the customers for this time period, we want to be able to identify the customers who are going to churn and take measures to retain them in order to maximize our digital advertising revenue. In order to do so, we need a model that can identify as many clients that are about to churn as possible.

Model Evaluation

After cleaning the data and performing feature engineering, I fit the following machine learning models, which are typically used for classification problems similar to this: logistic regression, decision tree, SVM, random forest, gradient boosting machine. To evaluate the models' performance, we can use the following metrics:

1. Accuracy: how many clients did the model correctly predict?
2. Precision: of all the clients that the model predicted would churn, how many actually churned?
3. Recall: of all the clients that actually churned, how many did the model predict would churn?

As I noted earlier, we need a model that can identify as many clients that are about to churn as possible in order to prevent them from churning. Therefore, I chose recall as the evaluation metric because I want the model to be able to correctly identify all the clients who actually churned. As a counterexample, if I used precision instead, I could have a model that predicts every client will churn, resulting in 100% precision but not providing any useful information. By choosing to use recall, I'm willing to misidentify a few clients who actually didn't churn if it means I can correctly classify more clients that actually churned. This will allow the model to make better predictions for the customers that actually churned, which allows us to better target clients in order to reduce churn rate.

After fitting the models, the random forest model performed the best, with a recall of 68.8% and an overall accuracy of 82.6%.

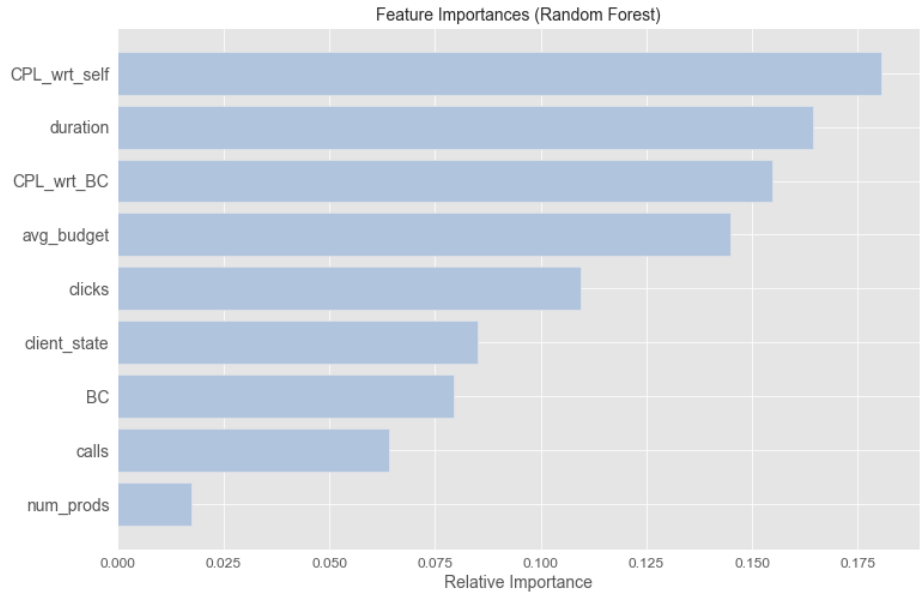

Figure 2: Feature Importance

Using the final model, I determined that the three most important factors for a client churning were the following (from most important to least important):

1. Change in the client's cost per lead in the past three months
2. The number of months the client has been our client
3. Difference in the client's cost per lead compared to their business category.

Recommended Next Steps

With a model to predict which clients will churn, we can increase retention by first identifying the clients that are likely to churn, and giving them a retention offer, such as a discount that temporarily lowers their cost per lead. While this will decrease our revenue in the short run, doing so will increase the length of time the client stays with the company, client, so that when the discount period ends, the increase in duration may be enough for the customer to stay, even without a retention offer. To test the effectiveness of such a campaign, we can conduct an A/B test to see if these offers actually increase retention.