Telco Customer Churn

Project Overview

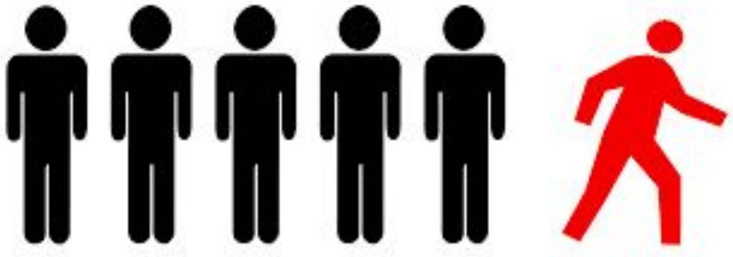
John Carlo Maula

Introduction

Background:

The Telco Churn dataset contains information about the customers of Telco, a telecommunications company that offers internet and phone service, and whether or not they "churned". Churning is defined as leaving the company (i.e. unsubscribing from their services) within the last month. The dataset also contains information about a customer's demographic and account information.

The goal of this project is to explore and analyze the Telco Churn dataset and build a logistic regression model to predict customer churning based on their features.



Description of the Dataset:

- The dataset has been split into a training and testing set, each with 2000 rows and 21 columns
- After data cleaning, the finalized training set contains 1995 rows and 20 variables
- 27.3% of the customers in the dataset churned.
- Variables include type of services (e.g. **phone service**, **online service**, etc.), demographics (e.g. **gender**, **age**, **dependents**, etc.), and account information (e.g. **contract**, **payment method**, etc.)

Exploratory Data Analysis

Summary of Findings:

- Customers who have a higher monthly charge and lower tenure tend to churn more.
- Customers who are not senior citizens, have partners, have dependents, or don’t use electronic checks are less likely to churn.
- Customers with online security, online backup, device protection, and tech support are less likely to churn regardless of the cost and type of internet service.
- Although customers with the *Fiber Optic* are more likely to churn, it’s due to the higher cost of that specific service rather than the quality of its services.

Based on these findings, I determined that **dependents** and **monthly charges** will be useful in predicting churning. In addition, I created 3 new features using variables from the dataset:

1. **Monthly Contract** - whether or not a customer has a *month-to-month* contract
2. **Has Service** - whether or not a customer has at least ONE of the following services: *online security*, *online backup*, *device protection*, or *tech support*
3. **Electronic Check** - whether ot not a customer’s payment method is electronic check

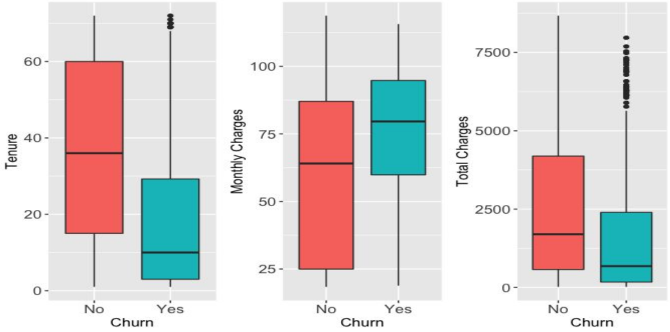


Figure 1: Box plots of tenure, monthly charges, and total charges.

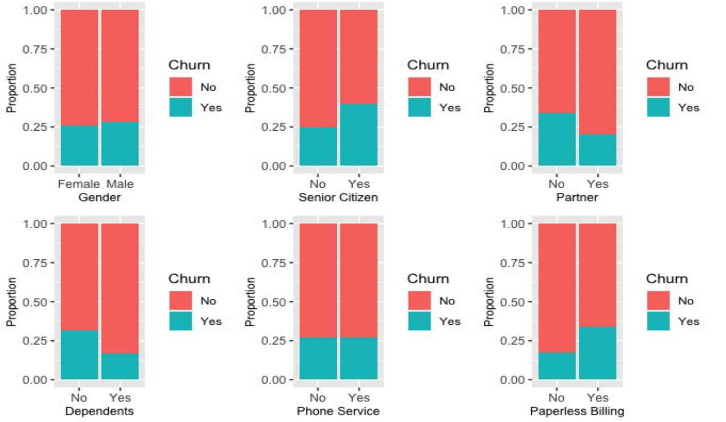


Figure 2: Barplots of gender, senior citizenship, partner, dependents, phone service, and paperless billing.

More Figures

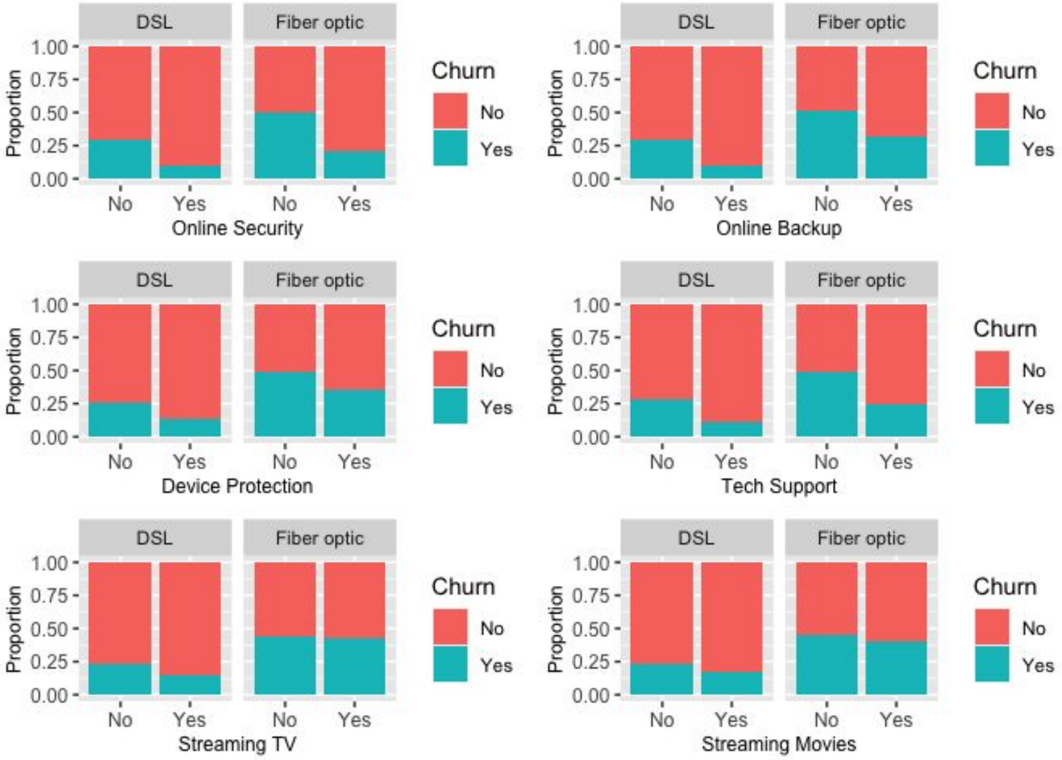


Figure 3: These bar plots depict the proportions of churning customers for each service based on their type of internet service.

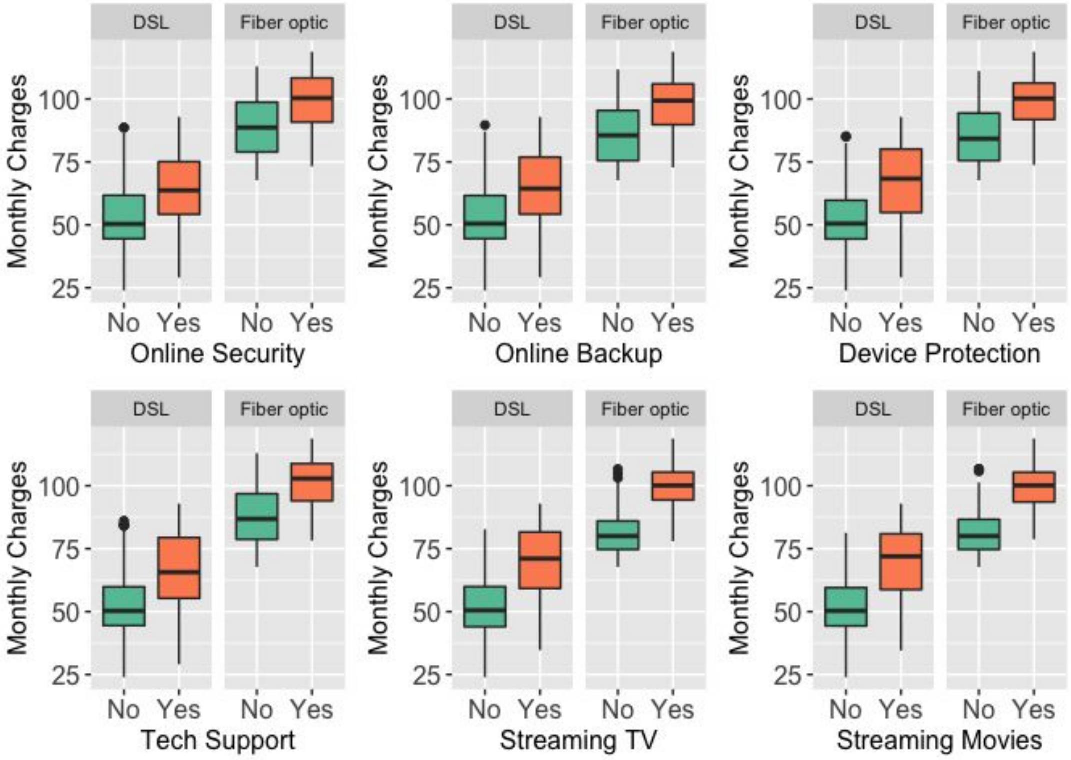


Figure 4: Box plots of monthly charges for each service based on the type of internet service.

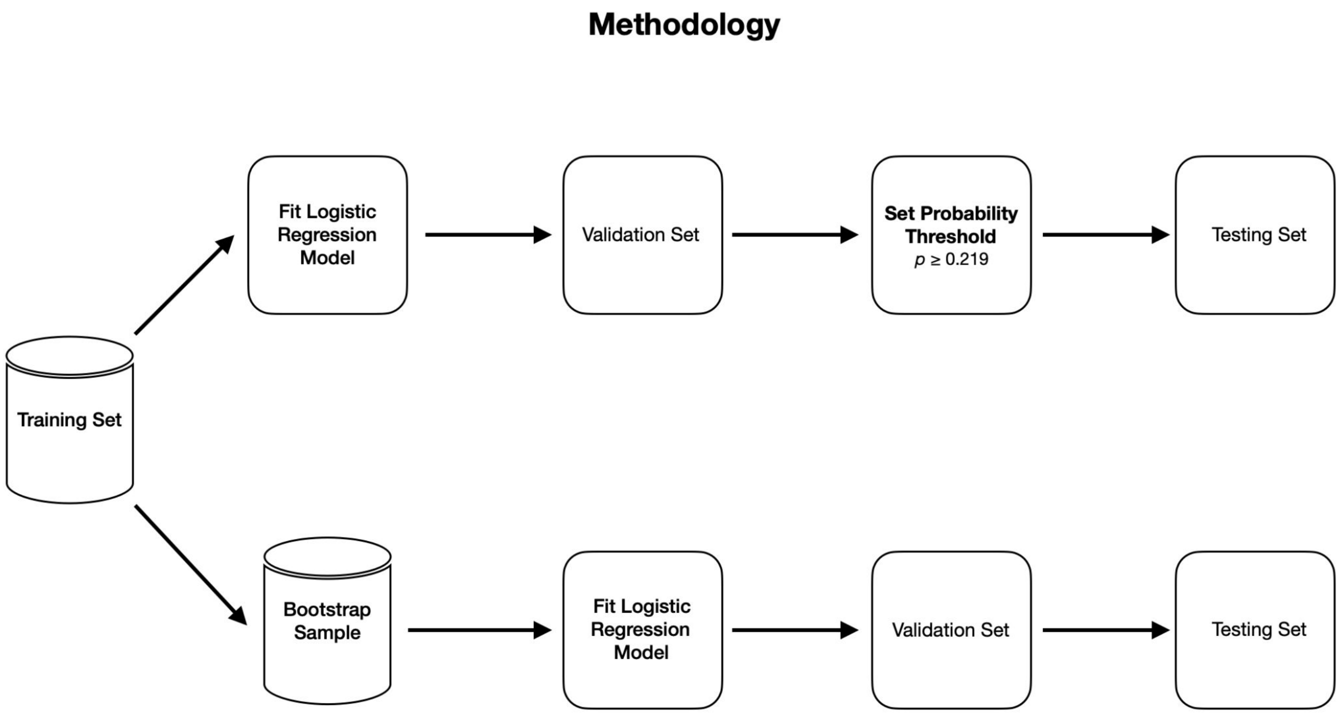
Data Modeling

Due to the binary nature of the response variable, I decided to use logistic regression to build my predictive model. I used the testing set to measure its performance, which resulted in a **79.0% accuracy**. While the accuracy is pretty good, the model has a low **sensitivity of 46.9%**. This means that more than half of customers who churned were misclassified.

Thus, I decided to improve the model in two different ways:

1. Changing the probability threshold of classification to maximize sensitivity and specificity.
2. Bootstrapping samples to achieve a balanced dataset (i.e. 50% prevalence in churning).

Then, I will use the testing set to measure its final performance. The methodology can be summarized by the figure to the right.

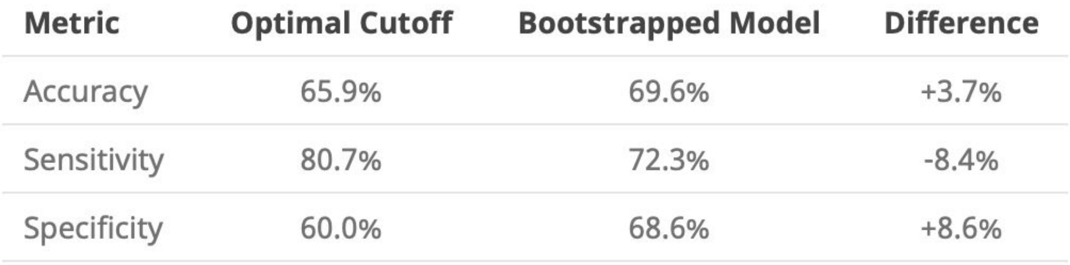

Methodology
Training Set
Fit Logistic Regression Model
Validation Set
Set Probability Threshold $p \geq 0.219$
Testing Set
Bootstrap Sample
Fit Logistic Regression Model
Validation Set
Testing Set

Results

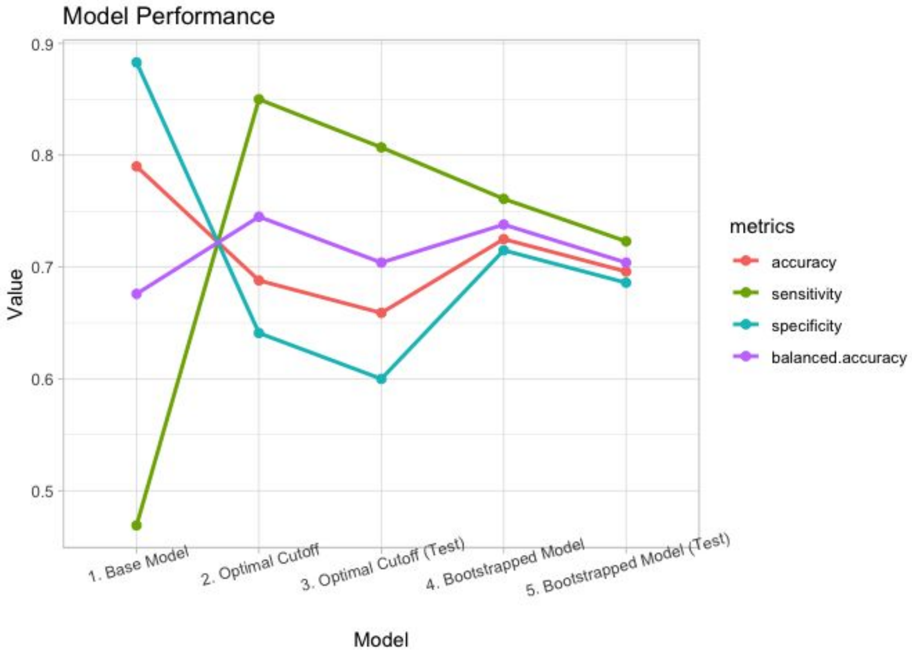
The performance of the models can be summarized in the table to the right.

Both models have a lower accuracy than the initial model, but they both have higher sensitivity. Since the bootstrapped model has a higher accuracy (+3.7%) and specificity (+8.6%), I decided to select that model as the final model. Compared to the original model, the bootstrapped model has a **54.2% greater performance** in correctly identifying customers who will churn.

All predictors included in this model were determined to be important in predicting churning, with monthly contract being the most influential predictor. Customers who have this type of contract have a **622% greater** odds of churning.



Metric	Optimal Cutoff	Bootstrapped Model	Difference
Accuracy	65.9%	69.6%	+3.7%
Sensitivity	80.7%	72.3%	-8.4%
Specificity	60.0%	68.6%	+8.6%



Recommendations

While the bootstrapped model is my recommended final model, the other model might be better if the cost of losing a customer is greater than the cost of implementing customer retention strategies due to its higher sensitivity (i.e., greater performance in identifying churning customers).

Overall, I recommend the following actions to be taken:

- Focus on customers who have a monthly contract. Offer incentives when signing up for a 1-year or 2-year contract.
- Overhaul the *Fiber Optic* internet service. It has a higher overall monthly cost, but there does not appear to be a difference in the quality of the services provided when compared to *DSL*.
- Examine the electronic check payment channel. If extra costs are associated with this kind of payment, find a way to decrease it. Another option is to make other methods more convenient.
- Properly inform customers about the offered services such as online protection, online backup, etc. and their benefits.