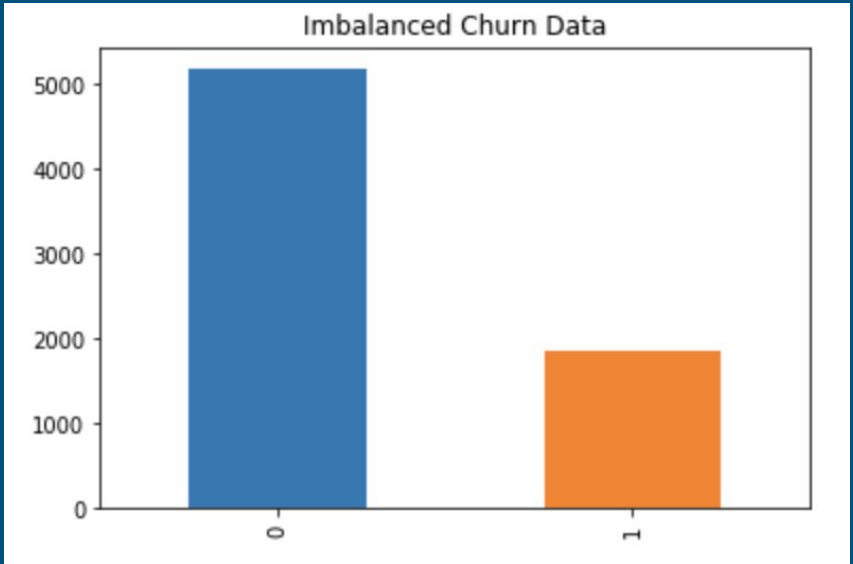
To Churn or Not to Churn?

By Daniel Nissani and Wei Duan

What data do we have?

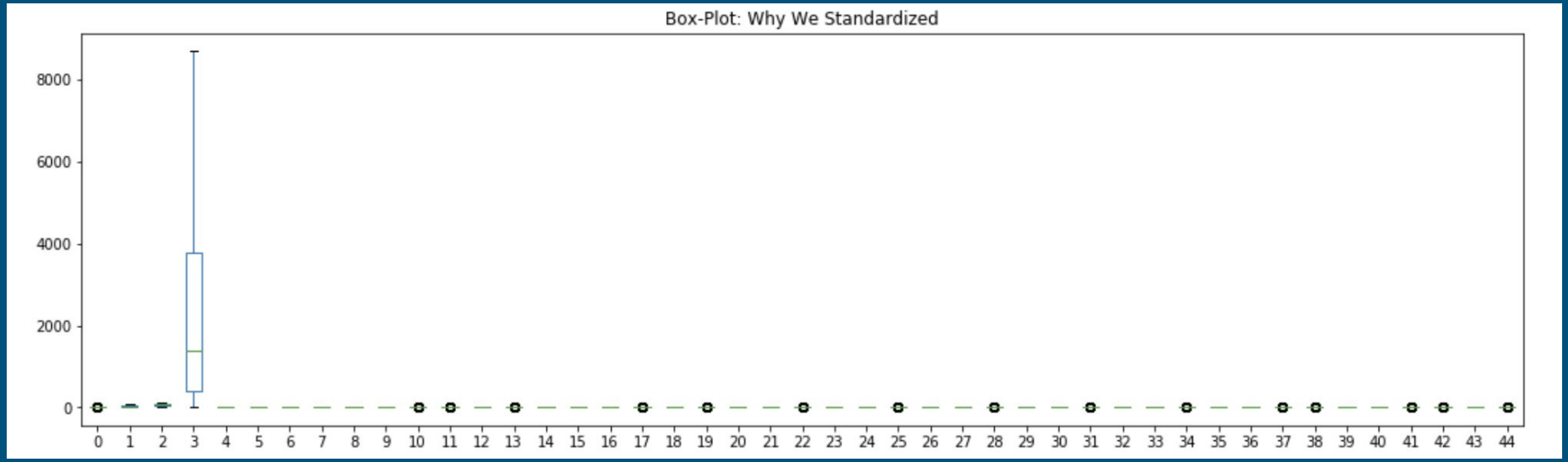
1) Demographic information (gender, age, dependents, etc.)
    a) Are they senior citizens or not?
2) Services (internet, phone, TV, security, backup, tech support, etc.)
    a) Do they have multiple lines?
3) Payment options (contract, month by month, total cost, etc.)
    a) Are they paperless or not?

Imbalanced Data



Most of the churn data had people that did not churn

Necessary to Standardize



Total cost has much greater variance and range.

Model

We used 10-fold cross validation to train models, AUC scores and accuracies are averaged on all 10 folds.

Logistic Regression

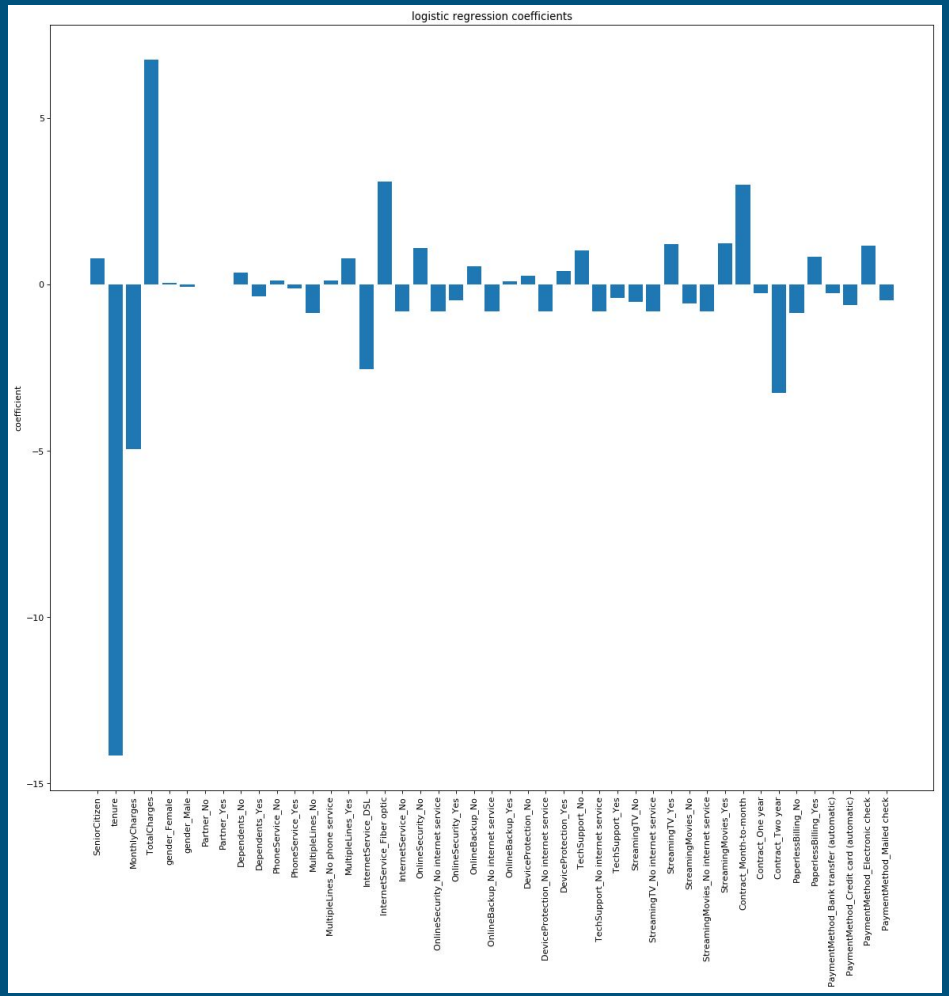
In balanced data, average test accuracy is 0.7531, average AUC is 0.7685

In original data, average test accuracy is 0.8030, average AUC is 0.7214

Random Forest

In balanced data, average test accuracy is 0.7732, average AUC is 0.6979

In original data, average test accuracy is 0.7808, average AUC is 0.6747



Analysis on correlations

From the bar chart of coefficients of logistic regression, we find that for customers who have larger “tenure” value, probability of churning is very low. Those who are charged a lot are more likely to churn, which are reasonable from the perspective of business.

Recommendation on business strategies

According to the coefficients of logistic regression, we recommend that:

customers who use fiber optic internet service have significant churning probability than those who use DSL. Considering that this is a telecom provider, a favorable strategy might be cooperating with fiber optic providers so that customers can get discount in installing fiber optics for