Truett Bloxsom, Gabriel James, Yuke Liu, Manas Rai
Group 20

1. Description of Project Goals

Our group investigated a Telecom Customer Churn for our project we found on Kaggle. Specifically, we wanted to identify key drivers of customer churn and develop a predictive model to identify customers who were likely to churn in the future. Customer churn is the tendency of customers to stop using a company's products and/or services over a certain window of time. An increase in customer churn directly and immediately translates to loss of revenue and, therefore, is an essential metric to measure and address in any industry. For instance, in property management, customer churn (i.e., failed resident retention) means loss of rent revenue and increased costs in repairs/renovations and marketing. A predictive model that can identify customers at risk of churning allows firms to better understand their customers and develop strategies to improve customer retention.

We found this business problem both important and compelling. We wanted to address a business challenge that was easily translatable from industry to industry and would be conducive to developing meaningful, actionable business solutions. Additionally, the size of the dataset (100 columns by 100,000 rows) meant not only more opportunity to understand the customer base and drivers of telecom customer churn but also more challenge in identifying the strongest predictors of customer churn and accounting for interactions among known and unknown variables.

2. Exploratory Analysis

This dataset was used in a churn modeling tournament at the Fuqua School of Business, Duke University. Each row represents a Telecom customer. The dataset contains 75 numeric columns, 24 non-numeric, and 1 customer primary key. 46 of the numeric columns are mean data of months 1, 2, and 3, and the churn field (binary categorical variable) represents whether a customer churned on month 5. This structuring of the data would enable a firm to identify customers at risk of churning and, over the course of a month, employ targeting strategies to reduce the likelihood that the customer churns.

Each value in the Customer_ID field was distinct, so no customers were being counted multiple times in our analysis. 49.56% of the customers in the dataset churned. Typically, telecom churn rates fall between 1.9 and 2.1%; this dataset was designed with high churn rate to provide a better training dataset for classification models[1]. The large number of continuous variables in this dataset proved particularly difficult to handle.

Most of the exploratory analysis we conducted centered on categorical variables. Figures 1-5 discuss marital status. Marital Status "M" was the most common status at 26,440 customers. However, married couples had the lowest mean number of total calls and lowest mean monthly minutes used. Additionally, all marital statuses had similar churn rates (Figure 6). Figure 6 shows that customers in dwelling type "M" had a higher average of total calls than customers in dwelling type "S".

We also looked at how the age of equipment impacted the minutes used and total calls of a customer. In Figure 7, Equipment Age is plotted against Mean Monthly Minutes Used. There is a general downward trend. The older the equipment, the fewer minutes a customer uses. Figure 8 corroborates this story: customers with newer equipment tend to have a higher number of Total Calls and a wider spread. Interestingly, some customers had nearly 100,000 calls over the span of their customer life.

3. Insights and Solutions

We first investigated the effect of length of service on churn. Figure 9 shows a skewed right distribution, with a mean length of service of 18.83 months, a median of 16 months, and a mode of 11 months. Plotting the churn distribution against the age of the customer's device in days yielded another skewed right distribution, with a mean of 391.93 days, a median of 342 days, and a mode of 310 days. We also plotted the age of the device in months against average monthly usage of the device (Figure 10). In general, as the age of the device increases, monthly usage of the device decreases, possibly suggesting that customers tend to use their devices more when they are new.

Figure 11 shows that customers with more expensive phones tend to have lower churn rates. We designated devices as cheap if their current price was below $60 and not cheap otherwise. Figure 12 adds more clarity to the behavior between cheap phones and churn rate. Though customers with more expensive phones were less likely to churn, of all customers who churned, customers with cheap phones had the longest length of service.

Figures 13, 14, and 15 provide more context about churned customers. Customers with non-dualband devices, WC devices, or refurbished phones were more likely to churn. Refurbished phones tended to have a shorter lifespan (Figure 16). Stratifying churn by length of residence (lor, measured in years) was negatively correlated with length of residence. We would expect customers living in multi-family and single-family rentals to be more likely to churn (Figure 17). Additionally, customers with more models issued were more likely to churn (Figure 18). Last, Figures 19 and 20 demonstrate that customers who later churn are, on average, decreasing their monthly usage more than non-churners are and that customers who later churn have higher mean roaming usage.

We then used a random forest model and XGBoost model to predict customer churn. The random forest had an out of sample AUC score of 0.683 (see Figure 21). Additionally, the model identified Age of Equipment (Days), Percentage Change in Monthly Minutes of Use vs Previous Three Month Average, Length of Service (Months), Mean Monthly Minutes Used, and Billing Adjusted Total Revenue Over the Life of the Customer as the most important features in the model (Figure 22). The XGBoost model had an out of sample AUC score of 0.693 (Figure 23). The model identified Price (Cheap/Not Cheap), Length of Service (Months), Handset Web Capability, Multiple Models (Y/N), and Age of Equipment (Days) as the most important features in the model (Figure 24).

We were surprised that Months of Service and Age of Equipment were such strong predictors of customer churn. It is possible that customers churn based on offers and contracts related to new models of phones, which would explain why customers churn most frequently between 10 and 12 months and are more active in the first few months of owning a device. It is also possible that competitors' first month discounts are high enough to motivate customers to switch providers.

4. Recommendations

Our first recommendation is to conduct a more thorough analysis with more data and to obtain documentation on categorical variables in order to better understand consumer behaviors. Assuming that we have correctly interpreted the categorical variables, it would be in the firm's best interest to investigate how new devices and contract signing deals change consumer preferences. Last, the firm should also investigate long term pricing strategies and develop stronger intuitions about customer retention incentives.

Truett Bloxsom, Gabriel James, Yuke Liu, Manas Rai
Group 20

Appendix

Figure 1:

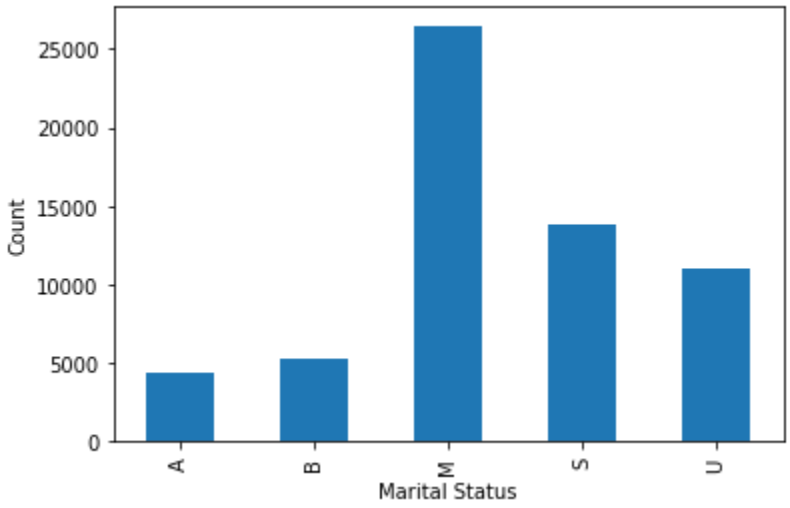


Figure 2:

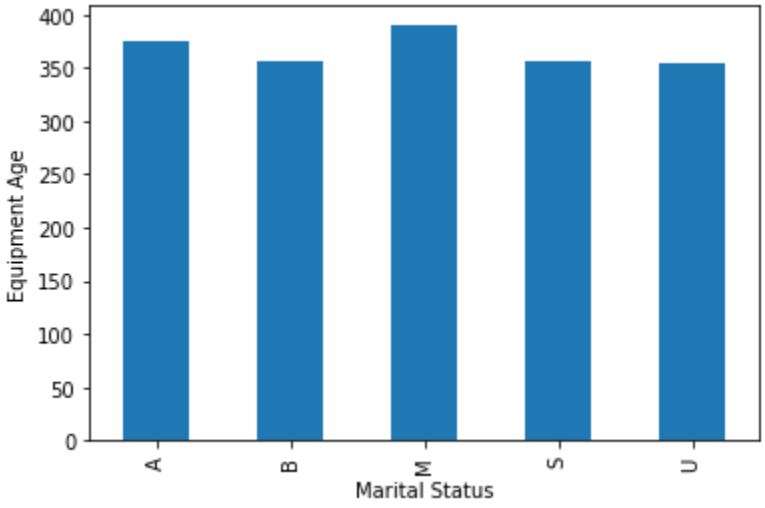


Figure 3:

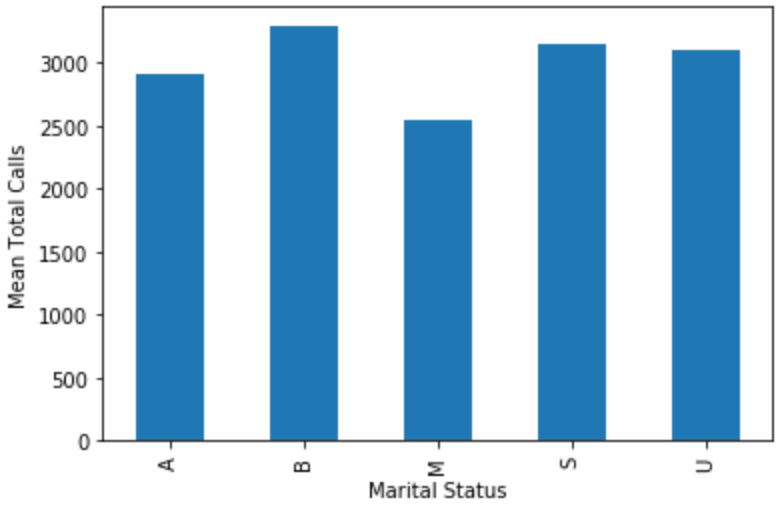


Figure 4:

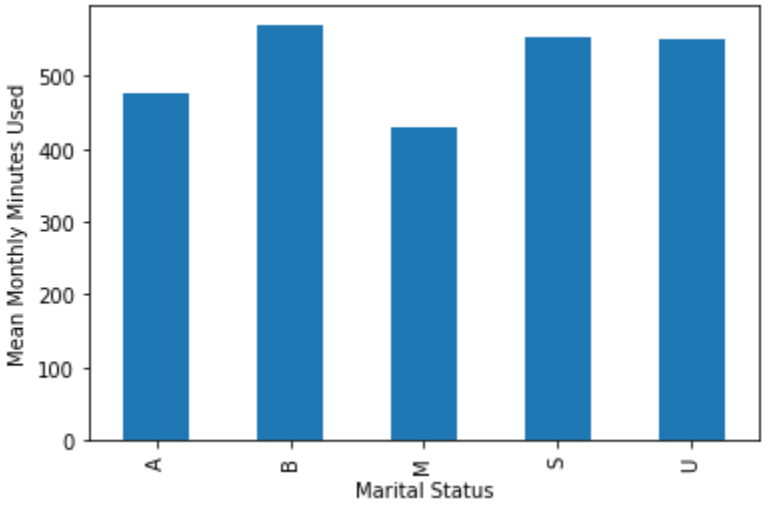


Figure 5:

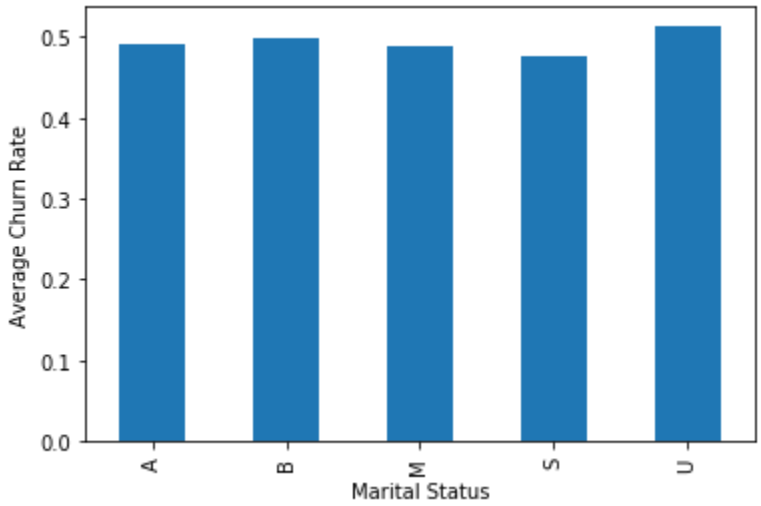


Figure 6:

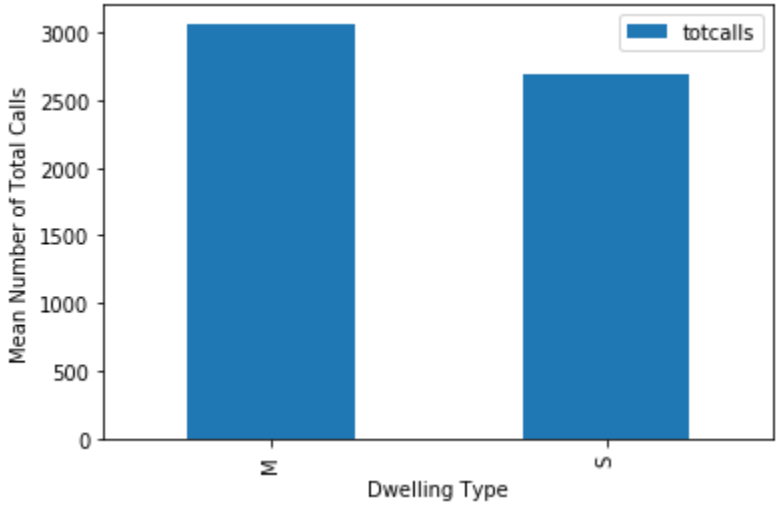


Figure 7:

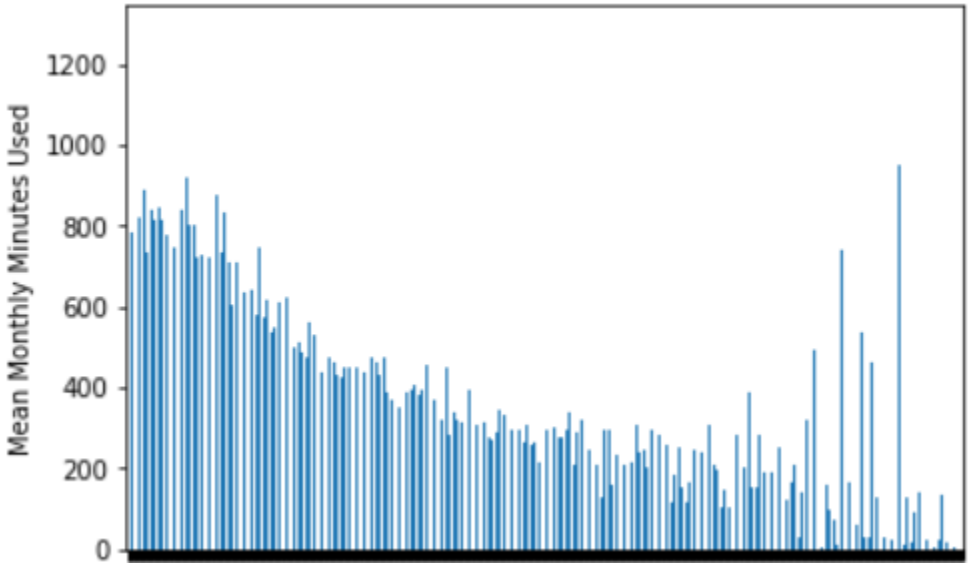


Figure 8:

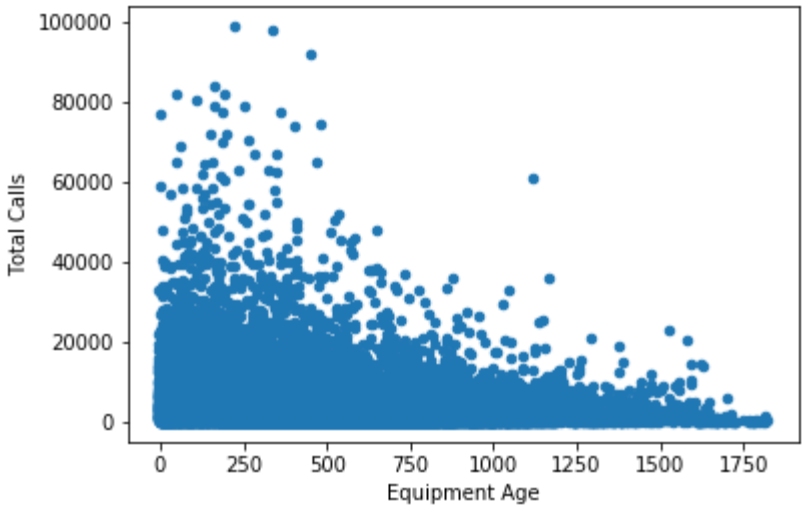


Figure 9:

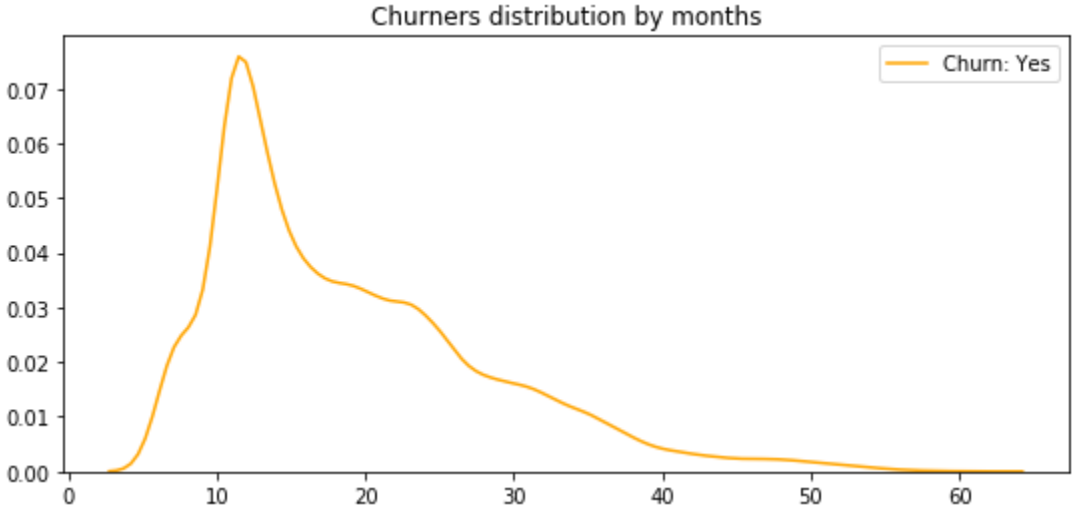


Figure 10:

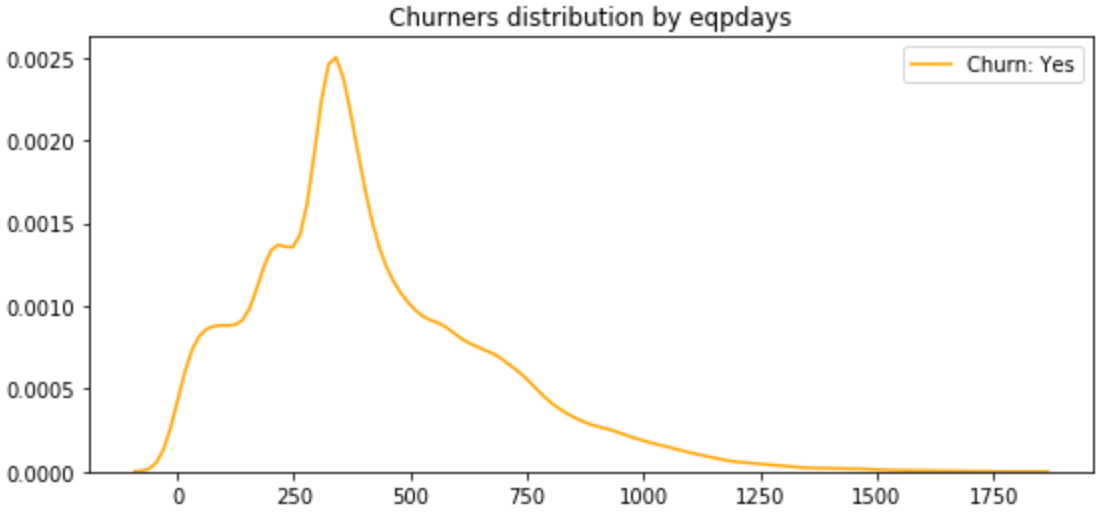


Figure 11:

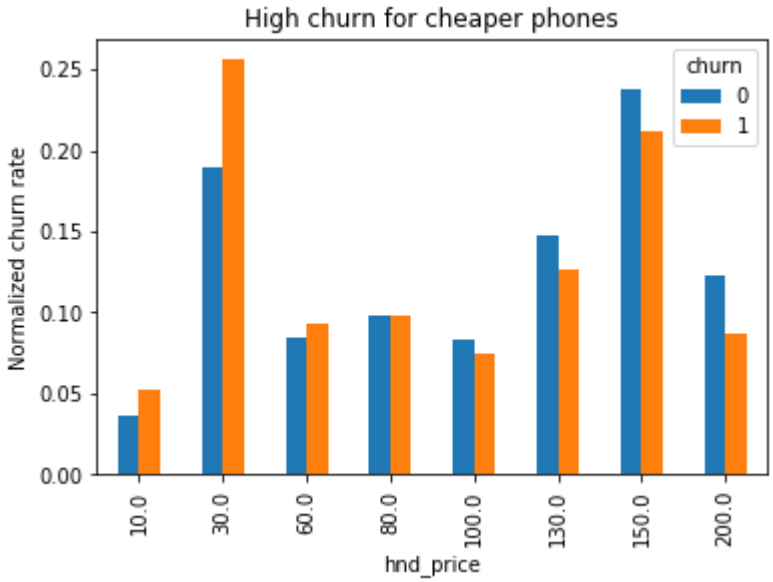


Figure 12:

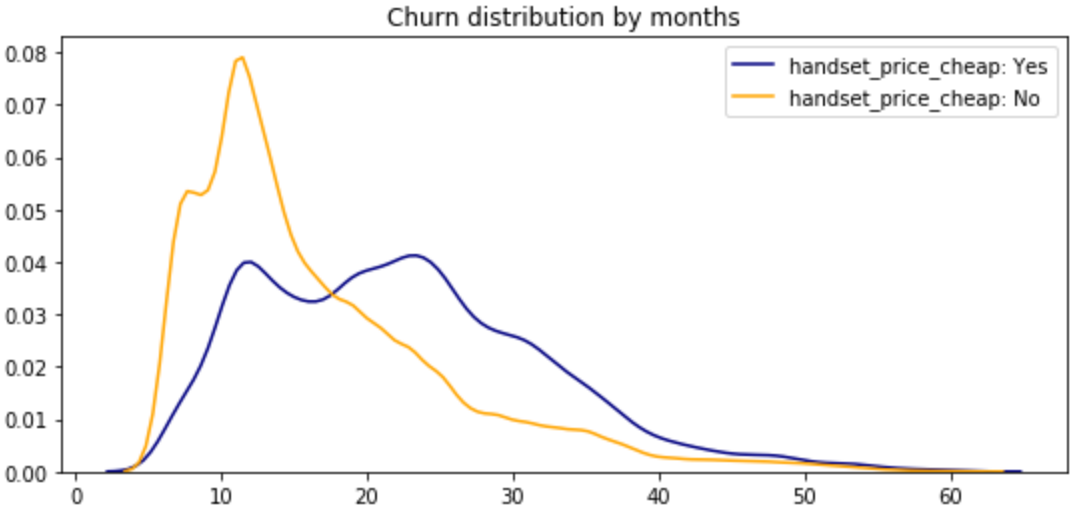


Figure 13:

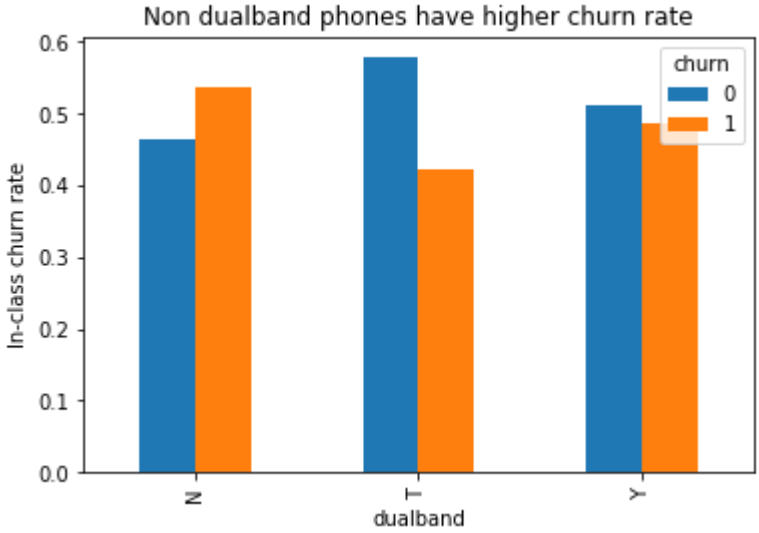


Figure 14:

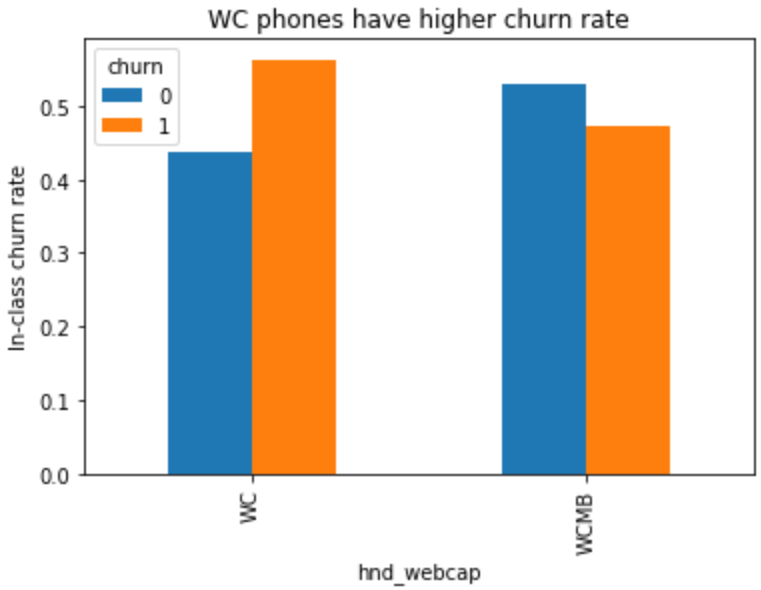


Figure 15:

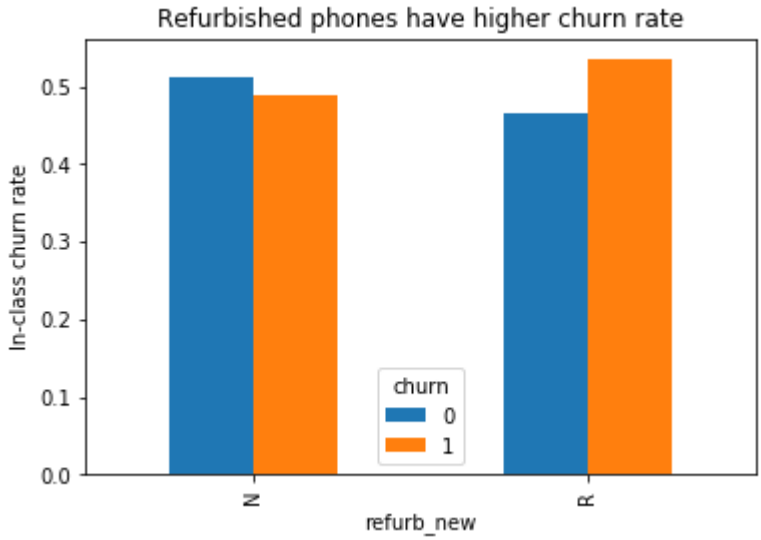


Figure 16:

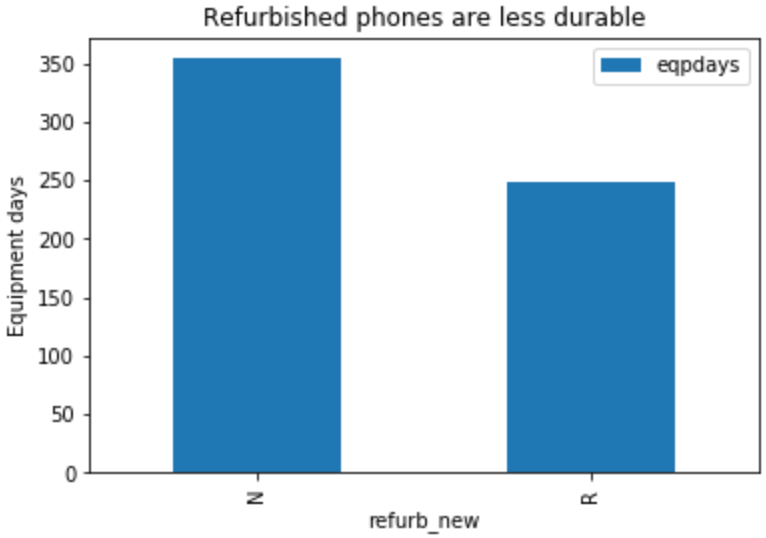


Figure 17:

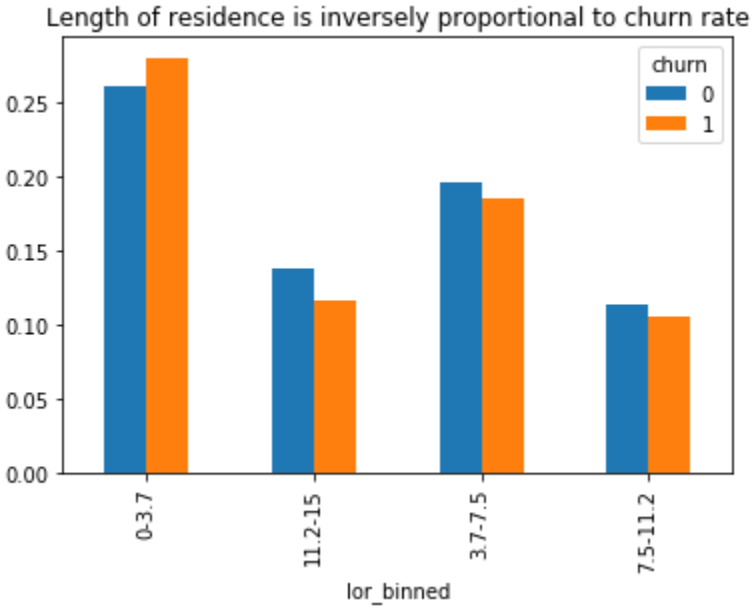


Figure 18:

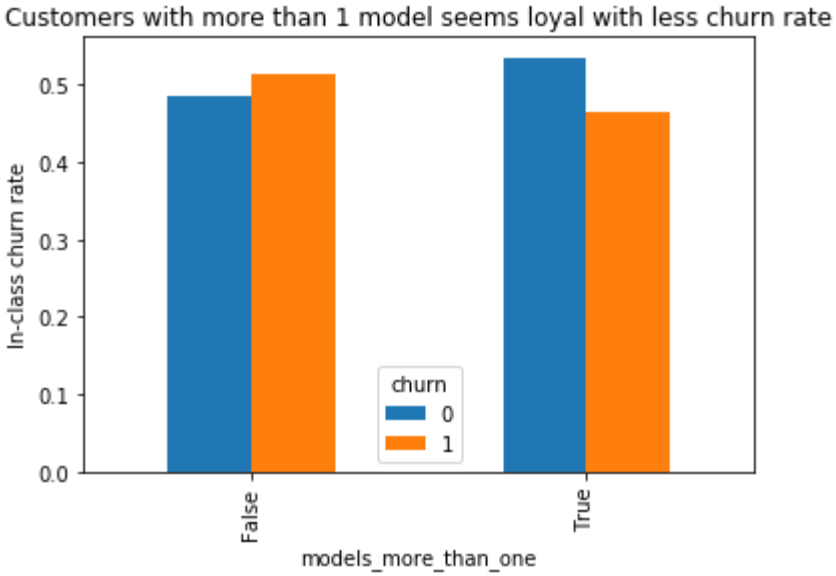


Figure 19:

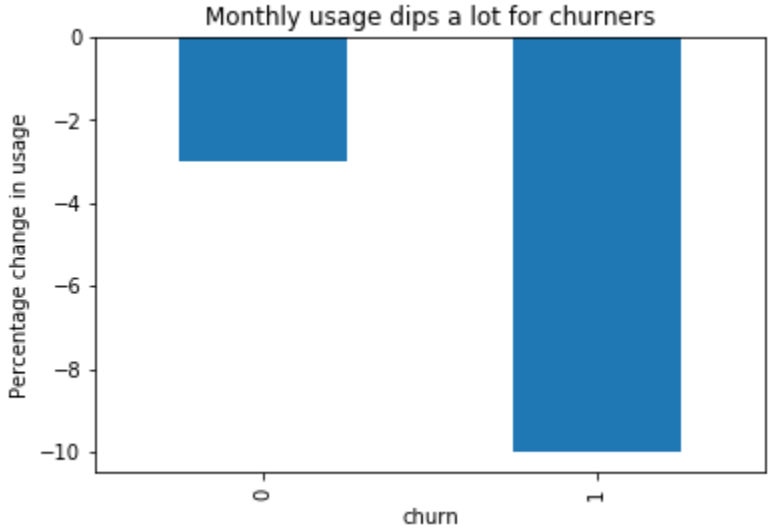


Figure 20:

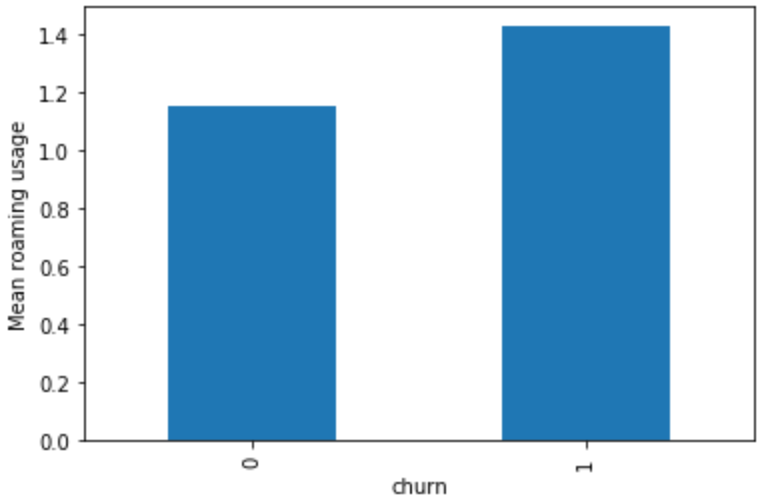


Figure 21:

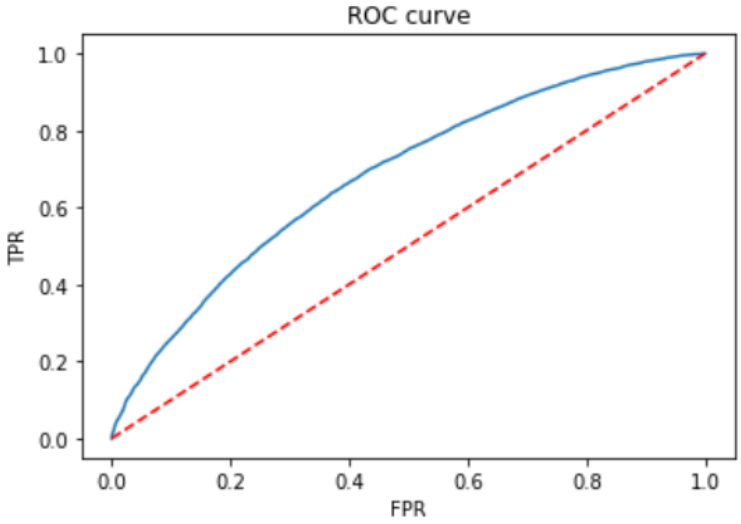


Figure 22:

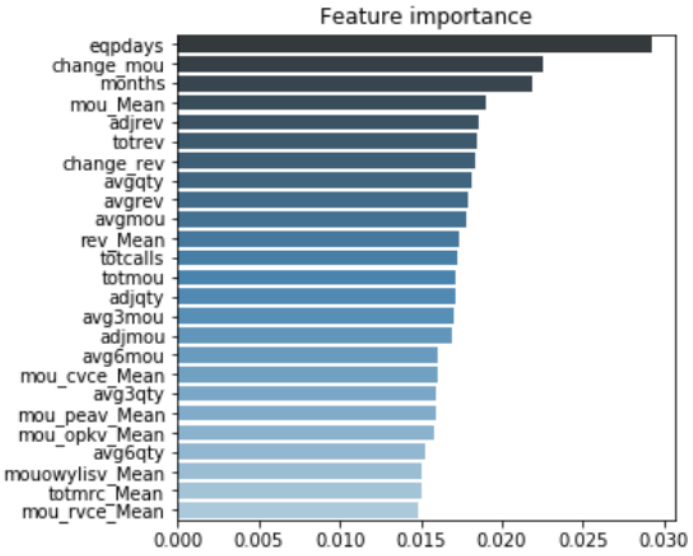


Figure 23:

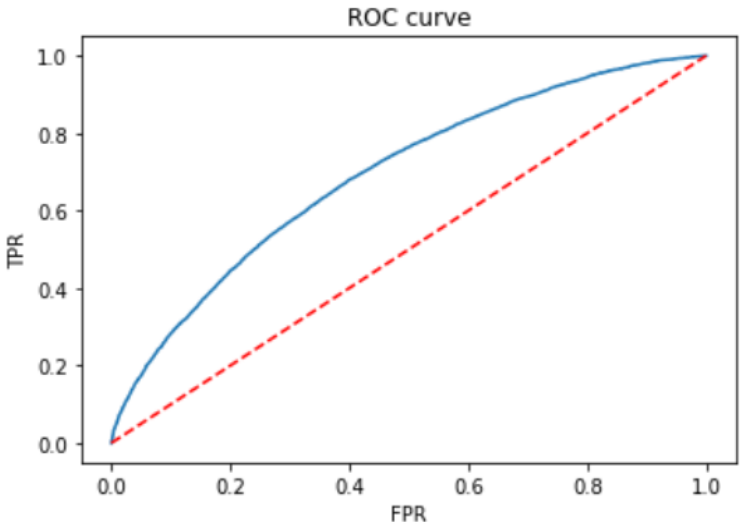
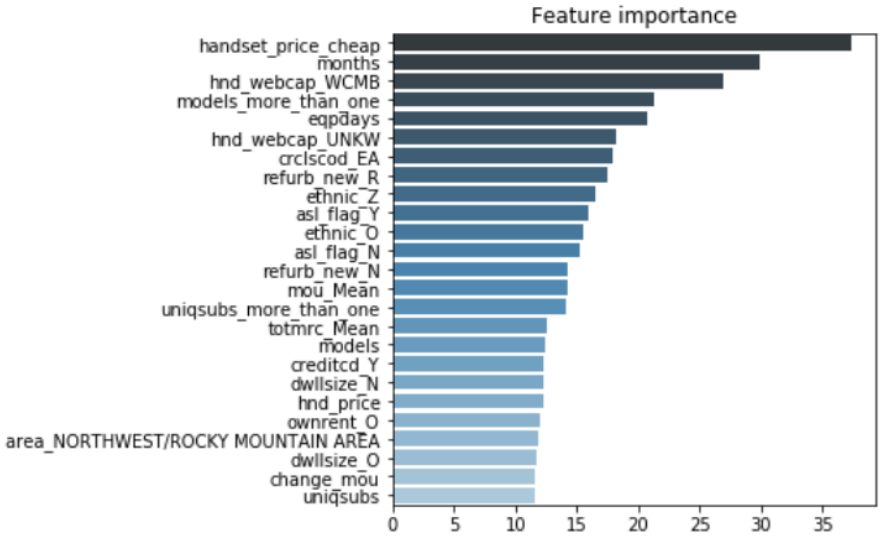


Figure 24:


Feature importance
handset_price_cheap
months
hnd_webcap_WCMB
models_more_than_one
eqpdays
hnd_webcap_UNKW
crclscod_EA
refurb_new_R
ethnic_Z
asl_flag_Y
ethnic_O
asl_flag_N
refurb_new_N
mou_Mean
uniqsubs_more_than_one
totmrc_Mean
models
creditcd_Y
dwllsize_N
hnd_price
ownrent_O
area_NORTHWEST/ROCKY MOUNTAIN AREA
dwllsize_O
change_mou
uniqsubs
0
5
10
15
20
25
30
35

Works Cited

Nelsin, S., Gupta, S., Kamakura, W., Lu, J., Mason, C. (2004). Defection Detection: Improving Predictive Accuracy of Customer Churn Models. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.489.5495&rep=rep1&type=pdf>.