Springboard Capstone Project 1:
Predicting Telecom Churn Customers

Final Report

Venkatramaiah chadalavada

Problem:

For most businesses, new customer acquisition is an expensive and labor-intensive effort. While this is critical for the early stages of a business, once a customer base is built, there should be increased effort in customer retention. Satisfied, existing customers are vital: they reduce the expenditure on marketing, they provide free word of mouth advertising, they are more likely to provide valued feedback and are more likely to pay for premium features/products.

Churn, one of the main metrics used to measure customer retention, is defined as the rate at which customers stop subscribing to a service.

For my Capstone Project 1, I have built models that predict the probability of whether customers of a telecom company will churn or not. I have also conducted analyses to identify the profile of a churn customer. This project will help subscription-based businesses identify customers that are prone to churn and make more informed decisions on how they may retain these individuals.

Data Set:

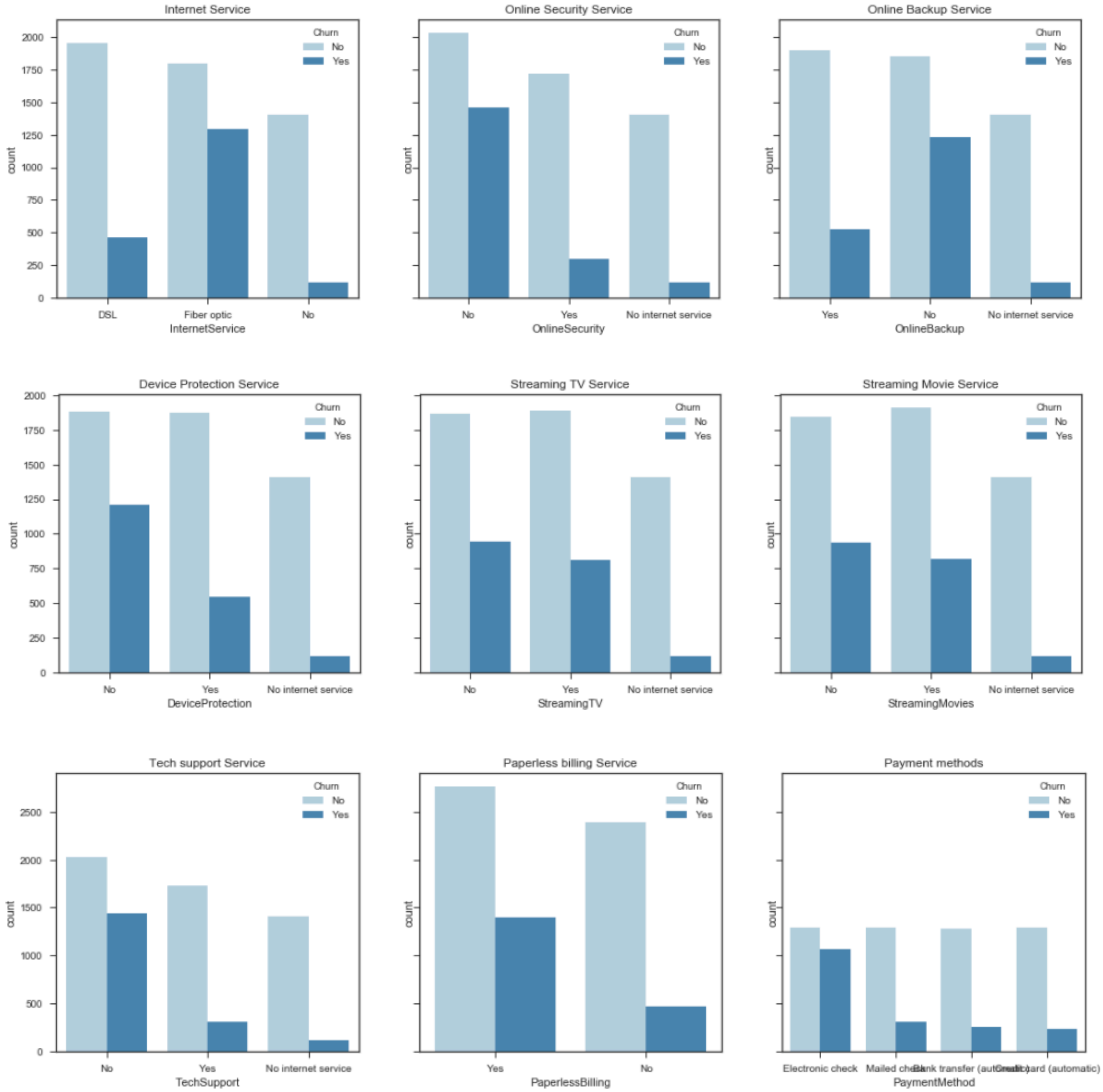
The data set used for this project is from Kaggle and can be found here:
https://www.kaggle.com/blastchar/telco-customer-churn

EDA and Data Wrangling:

As a first step, data wrangling and cleaning was performed on the data set, which consisted of 7043 observations with 21 features (4 are numerical, 17 are categorical). The features are wide ranging and include customer demographics, types of subscribed services, payment methods, monthly charges and length of customer tenure. The data set was fairly clean to begin with but did require the removal of 11 observations with null entries and updating of one numerical feature, TotalCharges, from object to float type.

Data Storytelling:

There were several questions I wanted to answer from the data set, including the demographics of churn customers and the types of services/payment methods preferred by churn customers. After analyzing the results, we see that there are some noticeable differences between churn customers and non-churn customers. One notable difference is in the services that these two customers types subscribe to, as depicted in the charts shown below.



Churn customers primarily subscribe to fiber internet service. They overwhelmingly prefer to pay by electronic check compared to all other payment methods. Non-churn customers are more likely to sign up for optional services such as online security, online backup, device protection and tech support than churn customers. Additionally, churn customers use month-to-month contracts at a much higher rate than non-churn customers (88.5% for churn compared to 43% for non-churn).

When examining the demographics of the customers in the data set, it appears that gender does not play a significant factor in determining whether a customer will churn (both churn at approx. 26%). However, customers without dependents are twice as likely to churn as those with dependents and senior citizens churn at a higher rate than non-senior citizens (41.7% for senior citizens compared to 23.7% for non-senior citizens).

Inferential Statistics:

To determine if there are statistically significant differences between churn and non-churn customers, hypothesis testing was applied to different features in the data set. Both bootstrap and frequentist methods were used. Using a threshold for significance of 0.05, it was observed that there are statistically significant differences between churn and non-churn customers for:

- monthly charges
- customer tenure
- percentage that have fiber optic internet service
- percentage that are on month-to-month contracts

It was also determined for the chosen threshold, statistical significance cannot be ascertained for the difference between churn and non-churn customers based on gender.

In Depth Analysis with Machine Learning:

Machine learning models for this project would be making a categorical prediction (whether customers churn or not) and the dataset already has labels that indicate whether each observation was a churn/non-churn customer; because of this, we would be using supervised classification methods.

As a first step, the dataset was split into training and test sets (80:20 split) and several classification algorithms (k-NN, Lasso Logistic Regression, Ridge Logistic Regression and Random Forest) were fitted to the training set. Hyperparameter tuning was applied to each of these classifiers to optimize results.

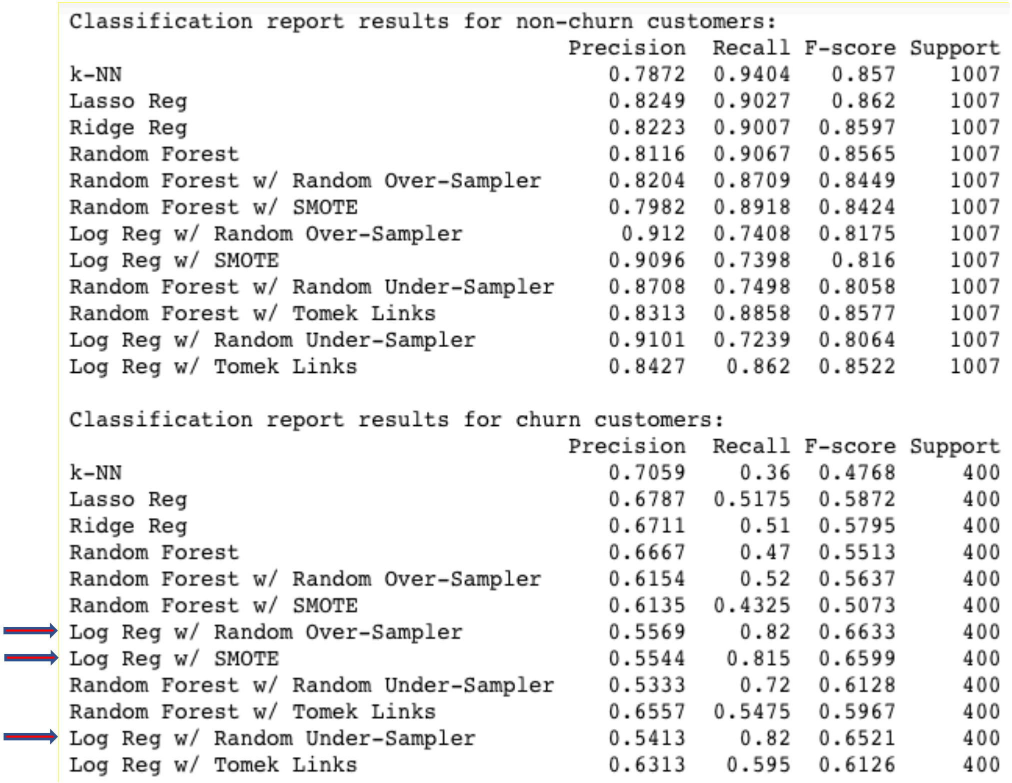
The classification reports for each algorithm was captured, containing the precision (percentage of the predicted positive observations that are actually positive), recall (percentage of all actual positive observations that were predicted correctly as positive) and F1-score (the harmonic average of precision and recall). From the initial exploratory data analysis, it was determined that there is a data imbalance in the dataset; approx. 74% of the customers are classified as non-churn. Because of this, we cannot use a simple accuracy score to measure the performance of the classifiers.

For this problem, we are most concerned with predicting as many of the actual churn customers as possible. Therefore, the primary metric that is used to judge algorithm performance is recall of churn observations. A business would be most concerned with correctly identifying as many of actual churn customers as it can so that it can take the proper actions to retain these customers. While having some false positive churn classifications is not a big concern, we do want to minimize the number of false negative churn classifications.

When examining the classification report results for these classifiers, they performed moderately well (the classifiers had a churn recall score of approx. 50%, except k-NN, which had a 36% churn recall score) but there was room for improvement.

To directly address the dataset imbalance, resampling techniques were applied. Oversampling techniques (SMOTE and random oversampling) and undersampling techniques (Tomek links and random undersampling) were combined with Random Forest and Logistic Regression classifiers. While a slight performance improvement was seen on Random Forest when it was paired with random undersampling, the biggest performance gains were seen in Logistic Regression pairing with SMOTE, random oversampling and random undersampling.

Below are the classification report results with the test set.



```
Classification report results for non-churn customers:
                                      Precision  Recall F-score Support
k-NN                                     0.7872  0.9404   0.857    1007
Lasso Reg                                0.8249  0.9027   0.862    1007
Ridge Reg                                0.8223  0.9007  0.8597    1007
Random Forest                            0.8116  0.9067  0.8565    1007
Random Forest w/ Random Over-Sampler     0.8204  0.8709  0.8449    1007
Random Forest w/ SMOTE                   0.7982  0.8918  0.8424    1007
Log Reg w/ Random Over-Sampler            0.912  0.7408  0.8175    1007
Log Reg w/ SMOTE                         0.9096  0.7398   0.816    1007
Random Forest w/ Random Under-Sampler    0.8708  0.7498  0.8058    1007
Random Forest w/ Tomek Links             0.8313  0.8858  0.8577    1007
Log Reg w/ Random Under-Sampler          0.9101  0.7239  0.8064    1007
Log Reg w/ Tomek Links                   0.8427   0.862  0.8522    1007

Classification report results for churn customers:
                                      Precision  Recall F-score Support
k-NN                                     0.7059    0.36  0.4768     400
Lasso Reg                                0.6787  0.5175  0.5872     400
Ridge Reg                                0.6711    0.51  0.5795     400
Random Forest                            0.6667    0.47  0.5513     400
Random Forest w/ Random Over-Sampler     0.6154    0.52  0.5637     400
Random Forest w/ SMOTE                   0.6135  0.4325  0.5073     400
Log Reg w/ Random Over-Sampler           0.5569    0.82  0.6633     400
Log Reg w/ SMOTE                         0.5544   0.815  0.6599     400
Random Forest w/ Random Under-Sampler    0.5333    0.72  0.6128     400
Random Forest w/ Tomek Links             0.6557  0.5475  0.5967     400
Log Reg w/ Random Under-Sampler          0.5413    0.82  0.6521     400
Log Reg w/ Tomek Links                   0.6313   0.595  0.6126     400
```

There appears to be an inverse correlation between churn customer recall and non-churn customer recall as well as churn customer recall and churn customer precision. While the top performing classifiers have a lower churn customer precision value, it is an acceptable trade off. It would be better to identify more of the actual churning customers at the expense of slightly more incorrect churn predictions rather than failing to identify a large percentage of churning customers.

Results and Recommendations:

Based on the findings from this project, Logistic Regression paired with random oversampling or random undersampling is the most effective classifier for identifying churn customers. The performance metric that is of the greatest importance is churn recall; identifying as many churn customers gives the business an opportunity to take proactive steps to retain a larger portion of customers that are potentially stopping service.

From exploring the data, the profile of the churn customers becomes clearer. They tend to subscribe primarily to fiber optic internet service with few other services besides streaming TV and movies. They are on month-to-month contracts and choose paperless billing at a higher rate than their non-churn counterparts. Additionally, while senior citizens account for a smaller percentage of the customer base, they are almost twice as likely to churn as non-senior citizens.

Given these insights, the company can take some steps to help reduce churn rate:

1. Conduct interviews and surveys to identify services/promotions that are preferred by senior citizens.
2. For high risk churn customers, offer a price promotion for extended service prior to a year and a half of tenure (the average churn customer stops service at approximately 18 months).
3. Collect age of customers. It is well known that different age groups have different consumption and expenditure habits. With this additional information, it will be possible to gain more granular insights of churn customers by age range.

Future Work:

Although it is not included in this report, I plan on performing feature importance analysis using Logistic Regression in future work. By selecting only the most important features, we can reduce the number of features and potentially improve model performance.