Customer Churn Prediction for Telecommunications Media and Technologies (TMT) companies

By
LaShawn Gaines

The mobile telecommunications industry is highly competitive. According to Consumer Affairs as referenced by EY, it is estimated that 97% of all Americans own a smartphone. Because of this, maintaining a customer base cannot simply revolve around attracting new customers. In order to thrive in this competitive market, you must gain customers from other providers. Not only do you have to be able to woo customers from other providers, you must strategize ways to reduce the churn rate thus minimizing the loss of your customer base to other service providers.

In 2019, TechSee conducted a survey of TMT Customer Churn. They found that most telecom companies leveraged reactive methods of retaining their customers. TechSee's survey showed that 61% of the telecom companies attempted to retain their customers by offering discounts or apologies, but only after the customers canceled their contracts. Telecommunications companies have enormous amounts of data on their customers. Can this data somehow be used to build a customer churn prediction model in order to identify and contact customers who are most likely to churn before they cancel their contracts or service?

About the Data

This and IBM sample dataset is located here, on kaggle. It has a usability score of 8.82. It includes 7,043 records with 21 columns. The data captured is a sample of customers:

- that left the company within the last month
- the services the customer signed up for (phone, internet, online security, online backups, device protection, tech support, streaming content)
- customer account information (tenure, type of contract, payment methods, paperless billing, monthly charges and total charges)
- Customer demographic data (age, gender, if they are single/partnered, if they have dependents)

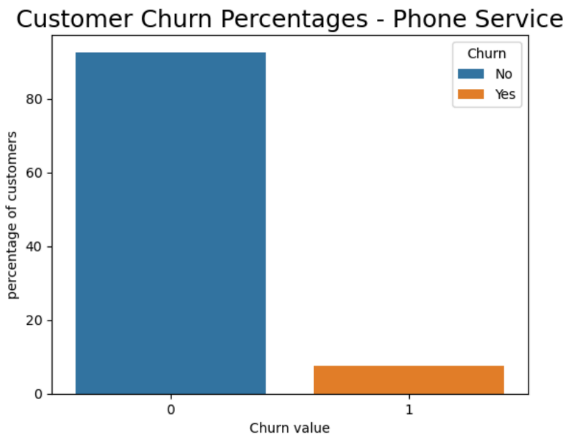
Data wrangling

With a usability score of 8.82 on kaggle, this dataset didn't require too much in the way of scrubbing. Below are some of the updates I made to the data set:

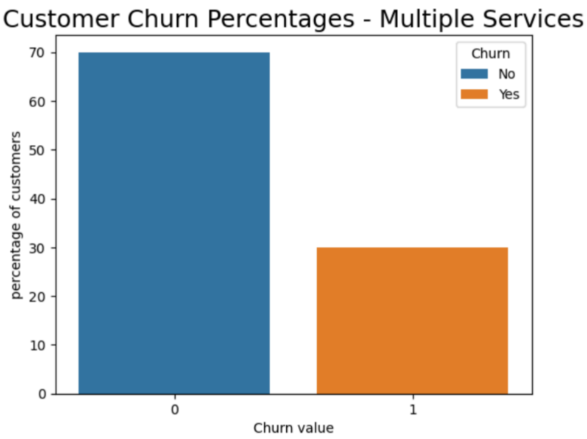
- total charges for the customers had an object data type, which I converted to float
- dropped 11 rows of data with missing values in the total charges column
- converted some of the descriptive features to numeric in order group them into bundles
    - creating phone_internet feature to count number of customers with bundled service
    - creating a number_of_services feature (count of all services subscribed)
- Dropped the customer ID column as it provides no insights

My preliminary findings from the data wrangling showed that of the final population of 7,032 customers in the data set, 1,869 churned during the month, giving us a churn rate of 26.58%.

When grouped by type of services the customer received, the lowest churn was for phone service customers and the highest was for customers leveraging multiple service offerings.



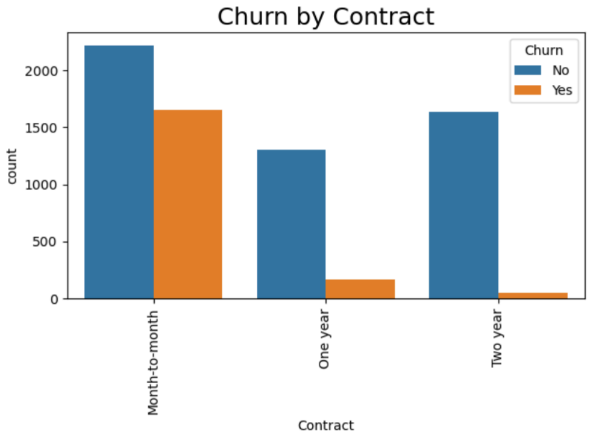
```
Churn Percentages - Phone Services
No: 92.57%
Yes: 7.43%
```

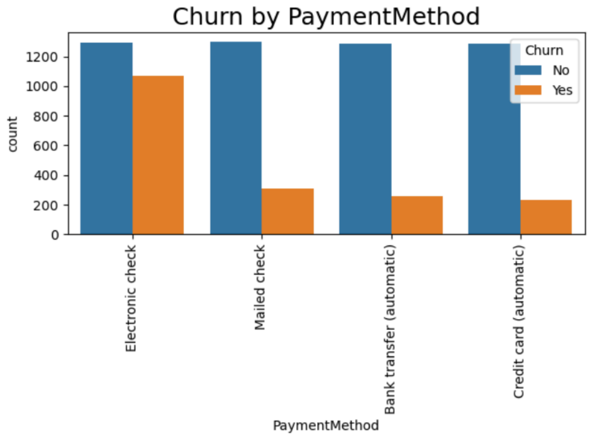


```
Churn Percentages - Multiple Services
No: 70.01%
Yes: 29.99%
```

EDA

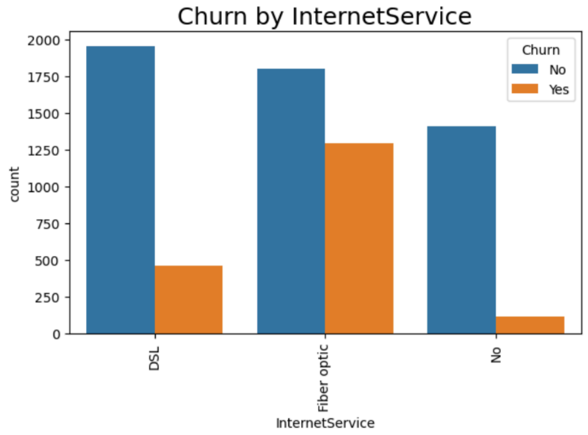
I began my EDA with a couple of bivariate analytical charts displaying churn by each of the other categorical features. I observed some interesting relationships. As it relates to customer account features, customers on month-to-month contracts are more likely to churn than those who are on one year or two year contract. This is logical as it is easier for the customer to leave with very few hurdles. In reviewing customer payment methods, I also noted those who pay via electronic check are also more likely to churn versus any other payment method.

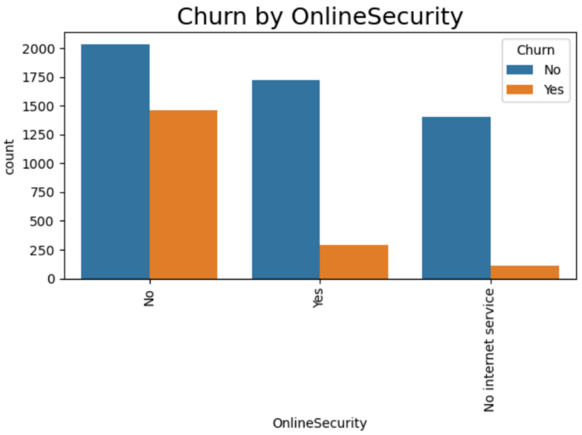

Churn by Contract


Churn by PaymentMethod

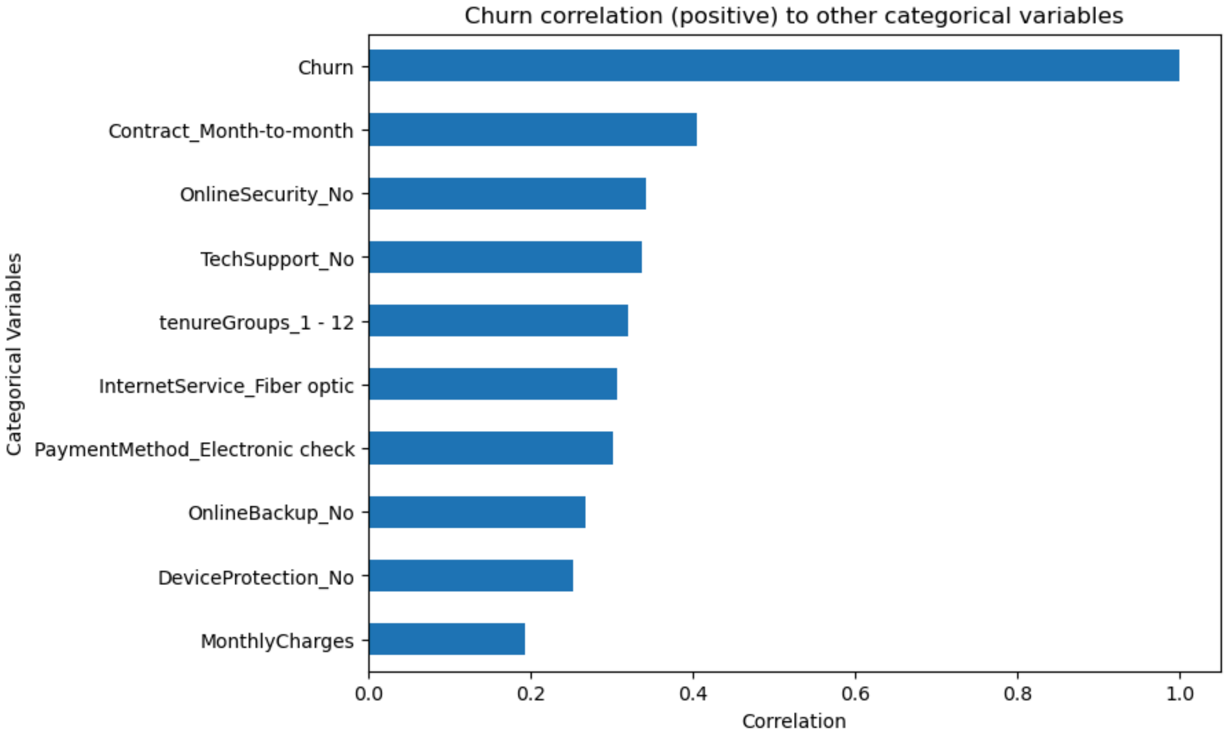
One can also observe some interesting relationships as it relates to the various services offered.

For the consumers subscribing to internet services, those using the fiber optic service are more likely to churn vs shoe on DSL. I also observed that customers who do not subscribe to the online security service are more likely to churn than those who do use it.

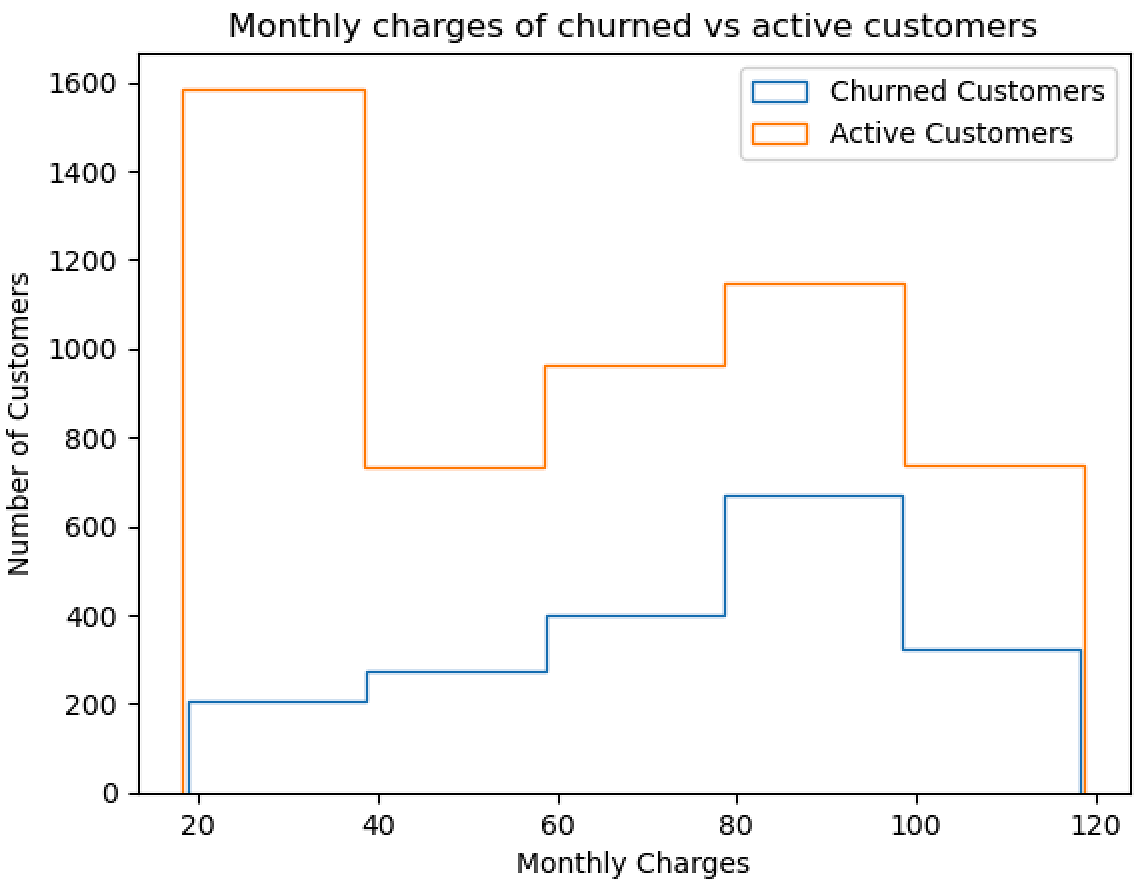

Churn by InternetService


Churn by OnlineSecurity

Next, I decided to see just how much these items may or may not be correlated to churn. I used the get_dummies method to encode the categorical features. This took my dataset from 24 columns to 59 columns. After charting the correlation between churn and the other categories, we can see some corroboration to our findings in our bivariate charts.



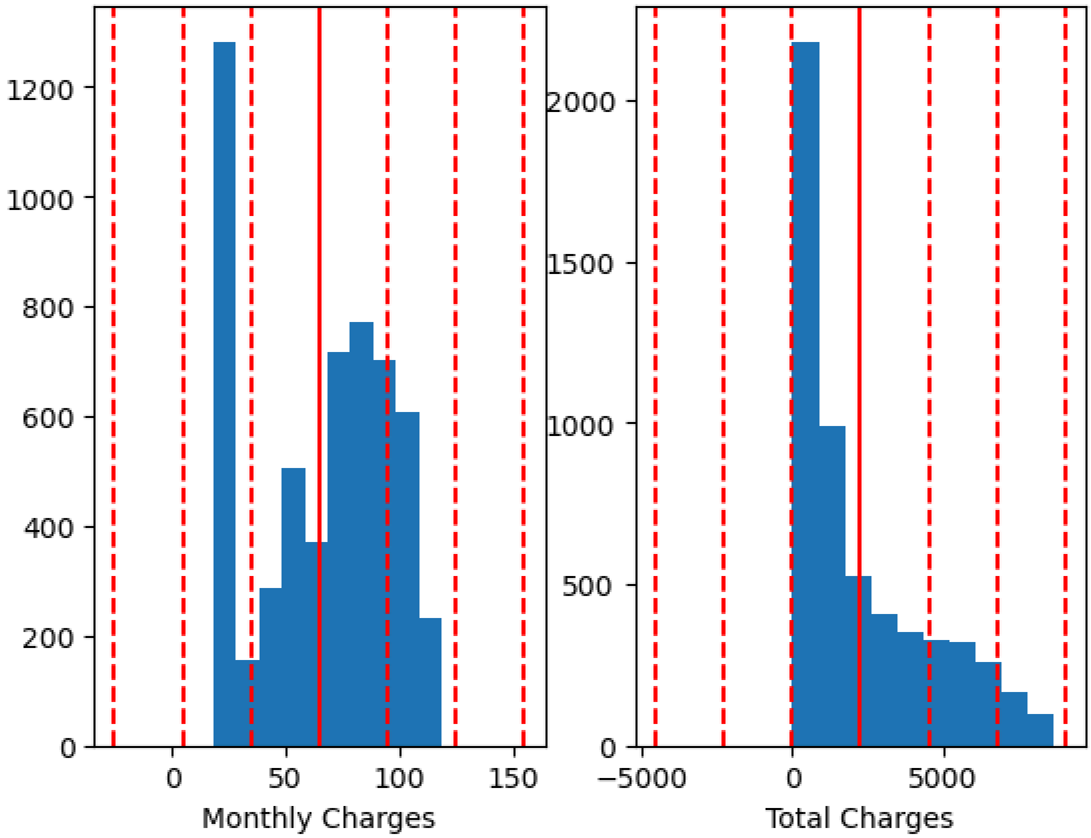
Although the correlation is somewhat weak between churn and monthly charges, we have a significant portion of our customer base with monthly charges between $20-40 per month. There also seems to be a spike in churn for those customers paying between $80-$100 per month.



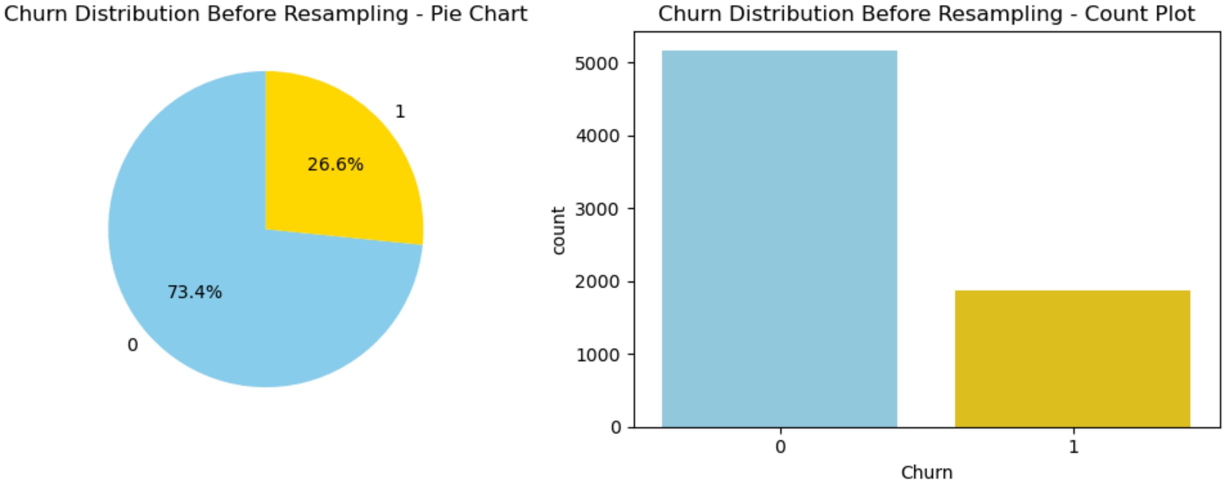
Preprocessing

After importing my scrubbed data files, I used dummy encoding (dropping the first feature) to prepare my categorical variables for machine learning. My final result was a dataframe with 7,032 records and 40 columns. My next step was to split my data into training and test sets.

While reviewing the monthly charges and total charges histograms, it was clear that the data was somewhat skewed. They are also of much higher magnitude than all the other fields. Therefore I chose to scale my data using StandardScaler.

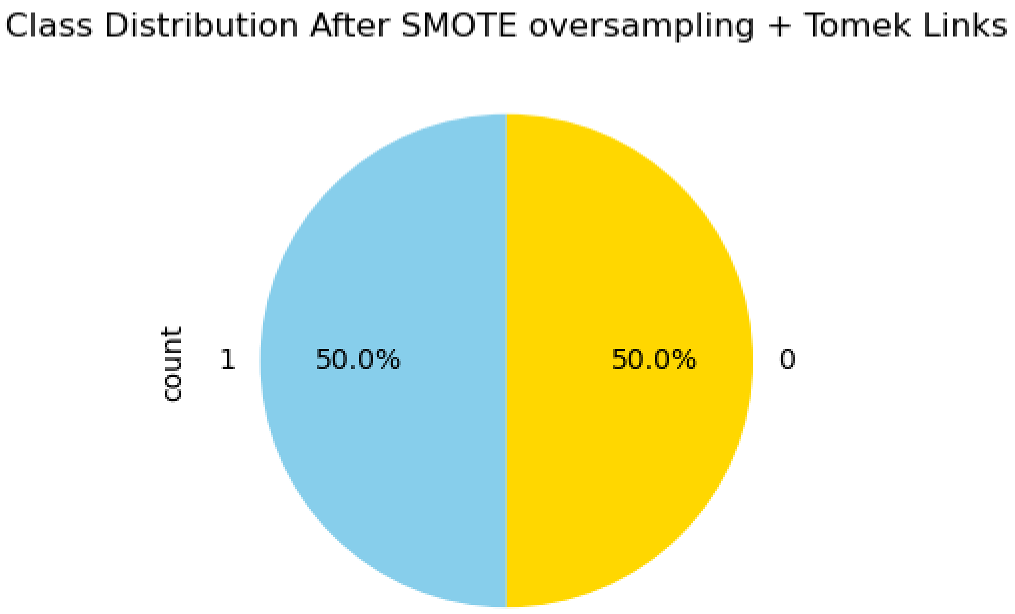


Next, given the nature of the dataset, an imbalance between the majority (active customers) and minority (churned) customers is to be expected. We found this to be true.


Churn Distribution Before Resampling - Pie Chart

Churn Distribution Before Resampling - Count Plot

As this dataset is relatively small, we determined that it was best to use an over-sampling technique vs undersampling. We did not want to lose any of the predictive power for classifying the majority class in our modeling. The SMOTE + Tomek links sampling method was applied.

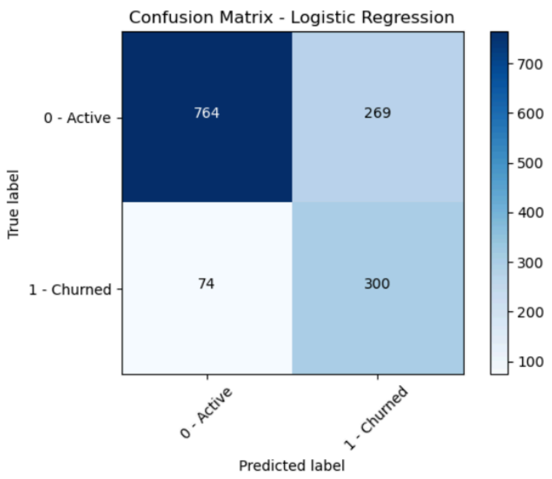

Class Distribution After SMOTE oversampling + Tomek Links

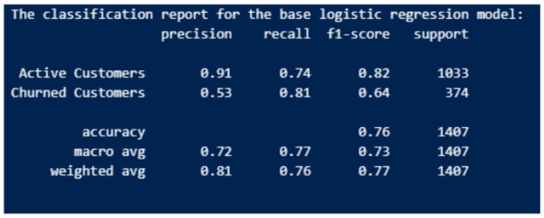
Modeling

Given that this is a classification model, the following algorithms were selected to predict customer churn:

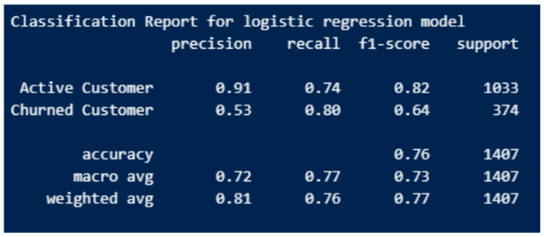
- Logistic regression
- RandomForest
- GradientBoosting
- KnearestNeighbors
- Support Vector Classifier

Each algorithm was first run with the default parameters to establish a baseline. This baseline was then compared to the results after hypertuning the model using a grid search with 5-fold cross validation. Given that this is a classification model, I have chosen to measure them based on recall. Assuming that it is better to have a few false positives vs false negatives. With that said, the logistic regression model produced the highest recall score for churned customers. The recall scores for the base model and the hyperparameter-tuned model were nearly identical.





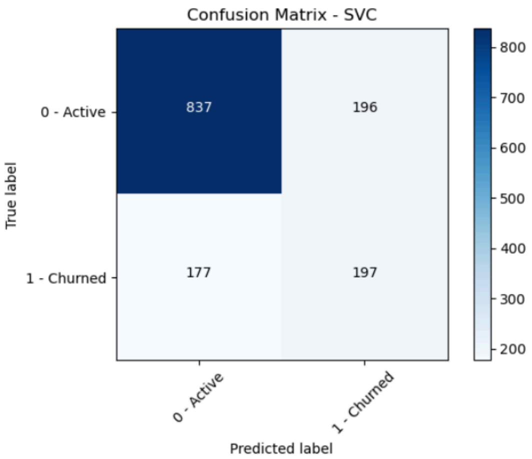
```
The classification report for the base logistic regression model:
                   precision    recall  f1-score   support

 Active Customers       0.91      0.74      0.82      1033
Churned Customers       0.53      0.81      0.64       374

         accuracy                           0.76      1407
        macro avg       0.72      0.77      0.73      1407
     weighted avg       0.81      0.76      0.77      1407
```

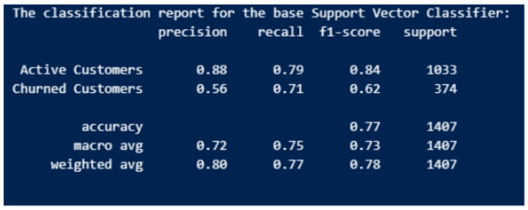


```
Classification Report for logistic regression model
                  precision    recall  f1-score   support

 Active Customer       0.91      0.74      0.82      1033
Churned Customer       0.53      0.80      0.64       374

        accuracy                           0.76      1407
       macro avg       0.72      0.77      0.73      1407
    weighted avg       0.81      0.76      0.77      1407
```

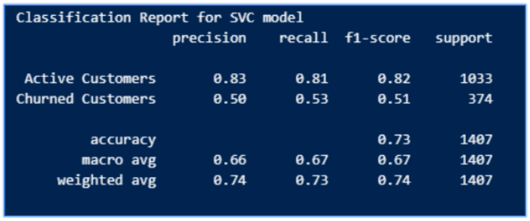
It is also worth noting that there is the potential for a high proportion of false positives. If this model is used for determining which customers to contact, we may have quite a bit of wasted resources to make those additional contacts.

Coming in just behind logistic regression, was the Support vector classifier model. Although it performed relatively well compared to Random Forest, Gradient Boosting and KNN, it is computationally very expensive.



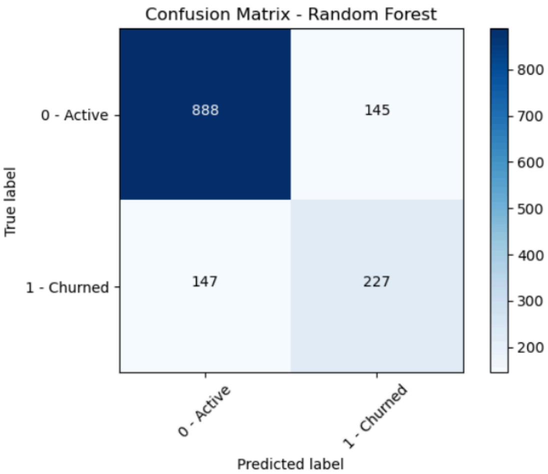


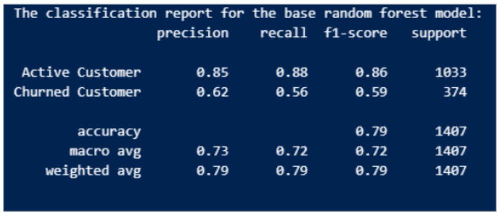
```
The classification report for the base Support Vector Classifier:
                   precision    recall  f1-score   support

 Active Customers       0.88      0.79      0.84      1033
Churned Customers       0.56      0.71      0.62       374

         accuracy                           0.77      1407
        macro avg       0.72      0.75      0.73      1407
     weighted avg       0.80      0.77      0.78      1407
```



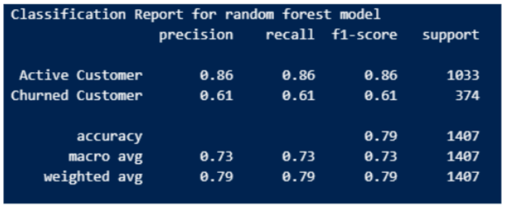
```
Classification Report for SVC model
                   precision    recall  f1-score   support

 Active Customers       0.83      0.81      0.82      1033
Churned Customers       0.50      0.53      0.51       374

         accuracy                           0.73      1407
        macro avg       0.66      0.67      0.67      1407
     weighted avg       0.74      0.73      0.74      1407
```

Random Forest appeared to reduce the number of false positives, but with the caveat of producing more false negatives. Gradient boosting produced nearly identical results. K-Nearest Neighbors produced the least reliable results.



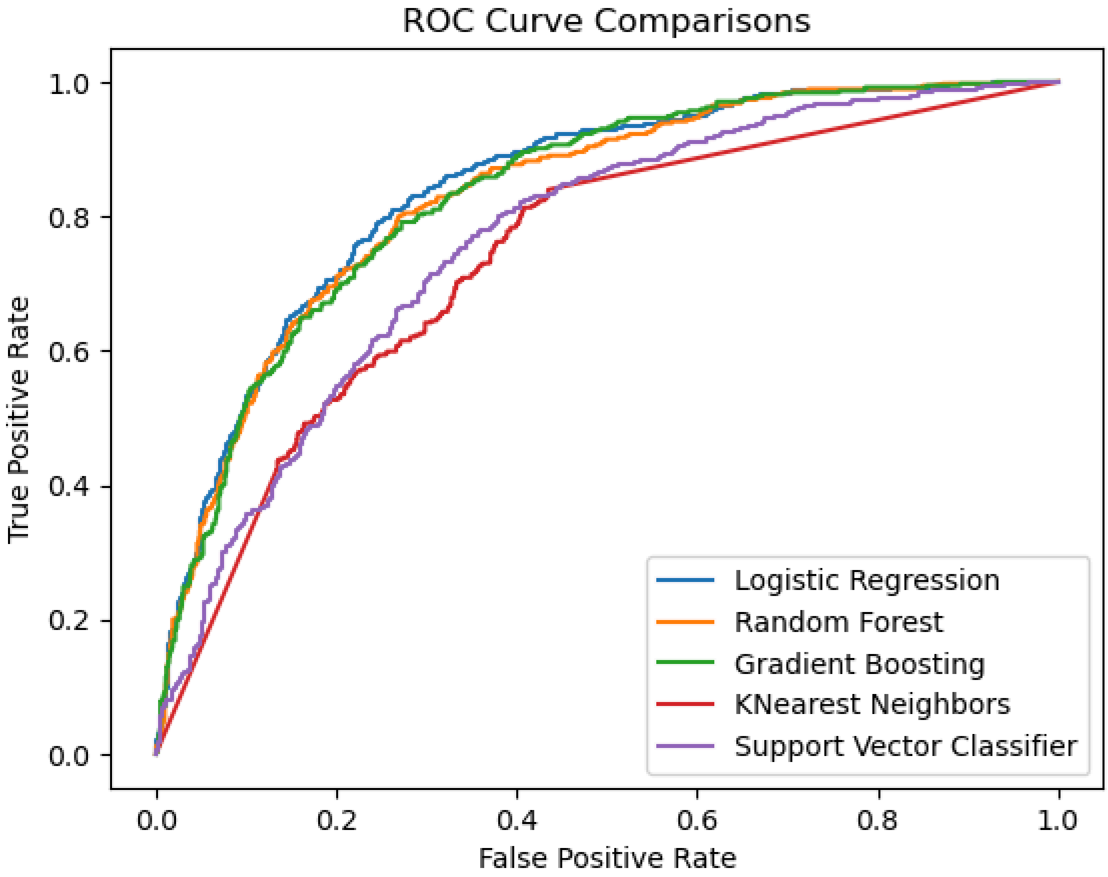


```
The classification report for the base random forest model:
                  precision    recall  f1-score   support

 Active Customer       0.85      0.88      0.86      1033
Churned Customer       0.62      0.56      0.59       374

        accuracy                           0.79      1407
       macro avg       0.73      0.72      0.72      1407
    weighted avg       0.79      0.79      0.79      1407
```



```
Classification Report for random forest model
                  precision    recall  f1-score   support

 Active Customer       0.86      0.86      0.86      1033
Churned Customer       0.61      0.61      0.61       374

        accuracy                           0.79      1407
       macro avg       0.73      0.73      0.73      1407
    weighted avg       0.79      0.79      0.79      1407
```

When reviewing the respective ROC curves, all the model performances were relatively comparable. However, given the best recall performance, my recommendation would be the logistic regression model.



Potential future prediction model enhancements

Enhancements could include additional preprocessing steps, such as filtering the features down to the most influential. This could reduce the computational resources needed to run all of the algorithms. I leveraged the StandardScaler in this model to scale my data. Perhaps leveraging MinMaxScaler or a combination of the two scalers could improve performance. Lastly, I would also potentially try different resampling methods, such as Random over sampling or ADASYN.

Lastly, In each of the algorithms, the results after hyperparameter tuning were close to the base measurement or in some cases worse. This could perhaps be the result of paring down the parameters in the param_grids in order to avoid overloading computational resources. With a filtered list of features by importance, may allow additional resources to achieve better results.

Business recommendations for customer retention

Historically speaking, the mobile telecom industry required one to two year contracts with customers as a part of doing business. Often-times they would include a mobile phone in the price of your service. As the devices are quite expensive, ensuring the customer was locked into a contract was necessary to cover equipment costs in addition to the service provided.

In 2013, T-mobile launched their "un-carrier" strategy, which eliminated customer contracts, thus beginning the now-normal month-to-month service agreements with customers. This just reiterates the importance of reducing churn.

During the EDA process, we noted several services that were negatively correlated to churn, thus making it less likely the customer would churn if they used that service. Perhaps bundling these services into Phone or Internet could reduce the customers likelihood of churning. Those services are:

- Tech Support
- Online Security
- Online Back-up service

We also noted that consumers on Fiber Optic internet were more likely to churn. Further research should be considered to ensure there isn't a performance issue. It was also noted that customers paying between $80 - $100 a month are more likely to churn also. Therefore proactively contacting these customers with potential discounts or additional service bundles may help stifle the urge for the customer to leave for a competitor.

Finally, there were some billing and payment methods that were correlated with churn. Customers leveraging paperless billing were more likely to churn than those with paper billing. Additional research is needed to ensure that the bills are actually being delivered to the customer. They could potentially be getting caught in a spam filter. Likewise, those paying with electronic checks were more likely to churn. There may be systemic issues that need to be resolved to enhance the customer's experience with electronic checks.