Matthew Smith

# Final Report:
# Telecom Churn Analysis

## Problem Statement

A California telecom company is faced with customer churn of approximately 26% this quarter. The loss of customer revenue and the related cost of acquisition of new customers to keep income stable is likely a source of concern for management at the company. This analysis focuses on the source of customer churn and aims to provide insight into opportunities to boost customer retention.

## Data Wrangling

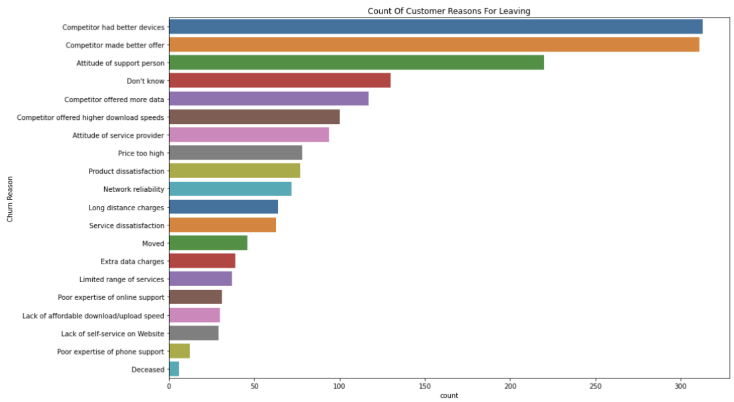
The raw dataset was found on Kaggle.com generated by Maven Analytics using a derivative of the IBM Dataset. The dataset included two three tables:

- data_dictionary - containing a map outlining the meaning of each field in the customer churn csv file
- customer_churn - our main table, containing 40 columns and 7043 rows
    - Key features of this dataset include: gender, age, marital status, number of dependents, number of referrals, tenure (in months), total revenue, & customer status
- zipcodes - a supporting table which contains population information for each respective zipcode

The data for this project was generally clean. Missing values had an identifiable pattern for their absence. One example of these missing fields would be “Avg Monthly GB Download” which was consistently missing on clients that did not have internet service. Imputation was done on missing values and formatting was done to update categorical values, such as marital status, to categorical variables (1 = Yes, 2 = No).

# Exploratory Data Analysis

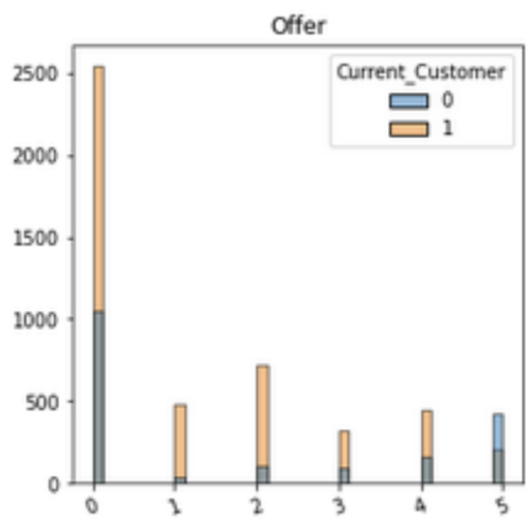
Once the data was prepared, simple analysis of the data was conducted to understand the relation of various features. Given that we were trying to identify sources of churn, customer reasons for leaving were charted and can be seen below.



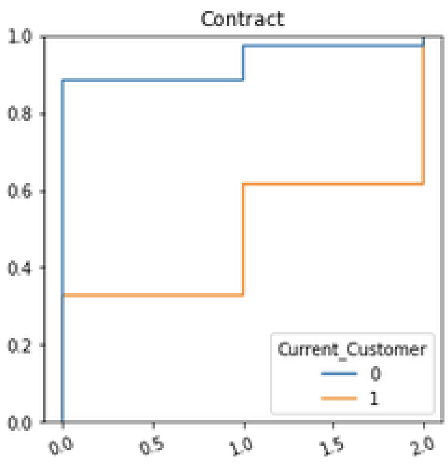

We learn from this chart that customers say they are leaving for many reasons but the top three reasons are:

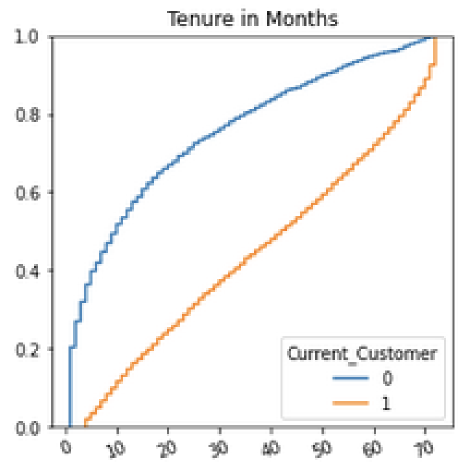
- Competitor had better devices
- Competitor made better offer
- Attitude of support person

Our data set doesn't provide insight into the devices provided by our telecom provider or the customer service conversations, but we do have information about the offer data.




Customers that churned (0) have a much higher count of Offer E (5) when compared to retained customers. Our data does not provide further insight into the contents of this offer or offers made by our competitors but could be the source of future analysis. We assume that an offer would take place at the beginning of a customer's term or at the time of renewal. The next two features we look at are, customer contract term & tenure (in months).

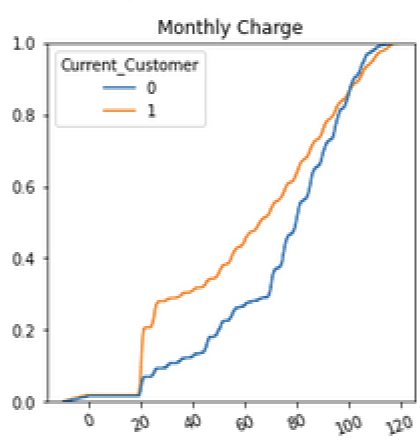


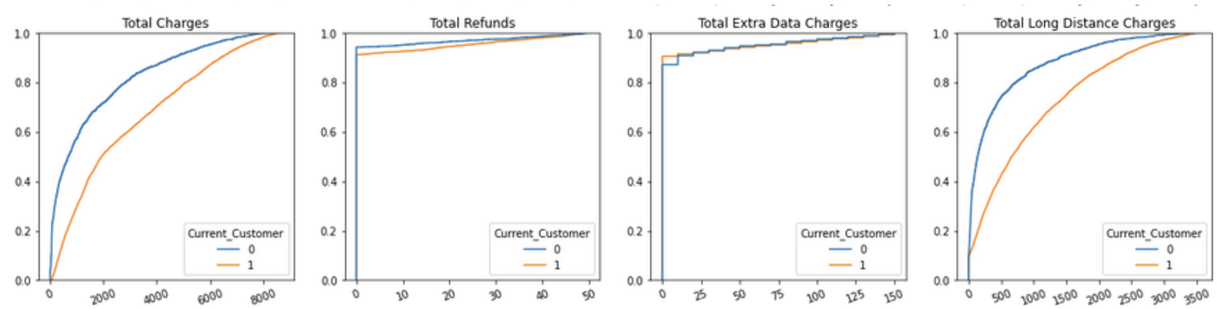

These charts show two key things:

1) Customers that churned (0) have much shorter contract terms compared to customers that were retained (1)
2) Customers that churned (0) have a shorter tenure in months compared to customers that were retained (1)

Given that we see churned customers have shorter contracts and shorter tenure, how does this impact revenue for these customers? A chart of the monthly charge shows that many of the churned customers pay around $70/month, which is higher than the average retained customer's monthly charge.

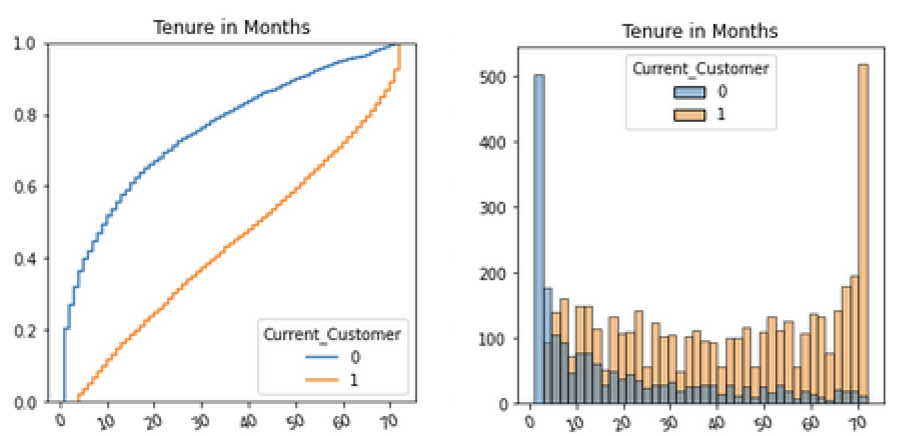



How does the monthly charge relate to the lifetime revenue for the customer?
Total Revenue for this data is calculated as (Total Charges - Total Refunds + Total Extra Data Charges + Total Long Distance Charges).





Although the monthly charges are higher for customers that churned, the total revenue is lower. This is likely to be a result of their shorter tenure.



# Modeling

Once the data was prepared, I used PyCaret's classification module to model the data using 13 classification models and ordered the results by accuracy.

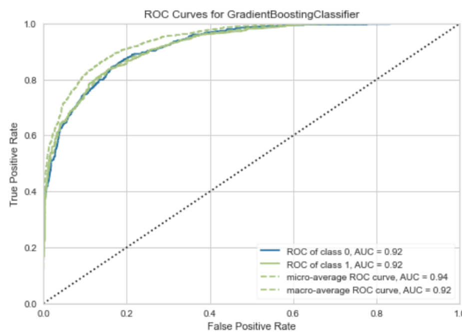
| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC | TT (Sec) |
|---|---|---|---|---|---|---|---|---|---|
| **gbc** | Gradient Boosting Classifier | 0.8774 | 0.9377 | 0.9464 | 0.8896 | 0.9171 | 0.6830 | 0.6885 | 0.2520 |
| **lightgbm** | Light Gradient Boosting Machine | 0.8742 | 0.9329 | 0.9356 | 0.8938 | 0.9141 | 0.6794 | 0.6823 | 0.0530 |
| **rf** | Random Forest Classifier | 0.8717 | 0.9252 | 0.9480 | 0.8819 | 0.9137 | 0.6652 | 0.6723 | 0.1220 |
| **ada** | Ada Boost Classifier | 0.8678 | 0.9265 | 0.9244 | 0.8946 | 0.9091 | 0.6668 | 0.6689 | 0.0800 |
| **et** | Extra Trees Classifier | 0.8637 | 0.9218 | 0.9298 | 0.8856 | 0.9071 | 0.6516 | 0.6547 | 0.1110 |
| **lda** | Linear Discriminant Analysis | 0.8439 | 0.9075 | 0.8900 | 0.8919 | 0.8908 | 0.6168 | 0.6176 | 0.0310 |
| **ridge** | Ridge Classifier | 0.8432 | 0.0000 | 0.9005 | 0.8830 | 0.8915 | 0.6085 | 0.6096 | 0.0220 |
| **dt** | Decision Tree Classifier | 0.8206 | 0.7800 | 0.8740 | 0.8752 | 0.8745 | 0.5597 | 0.5602 | 0.0230 |
| **knn** | K Neighbors Classifier | 0.7466 | 0.7117 | 0.8753 | 0.7927 | 0.8318 | 0.3229 | 0.3303 | 0.0320 |
| **nb** | Naive Bayes | 0.7188 | 0.8416 | 0.6735 | 0.9106 | 0.7738 | 0.4234 | 0.4584 | 0.0230 |
| **dummy** | Dummy Classifier | 0.7158 | 0.5000 | 1.0000 | 0.7158 | 0.8344 | 0.0000 | 0.0000 | 0.0170 |
| **svm** | SVM - Linear Kernel | 0.6743 | 0.0000 | 0.7467 | 0.8131 | 0.7435 | 0.2242 | 0.2674 | 0.0380 |
| **qda** | Quadratic Discriminant Analysis | 0.6423 | 0.6401 | 0.6456 | 0.8174 | 0.7054 | 0.2550 | 0.2713 | 0.0340 |

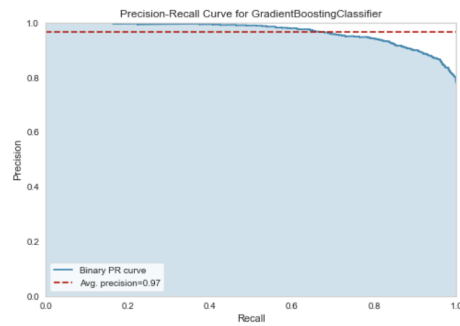
The best performing model was a gradient boosting classifier. Which was further tuned using the hyperparameters below.

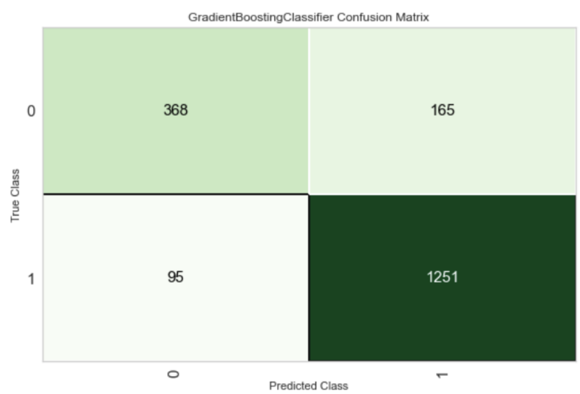
```
GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None,
                           learning_rate=0.4, loss='deviance', max_depth=7,
                           max_features='sqrt', max_leaf_nodes=None,
                           min_impurity_decrease=0.3, min_impurity_split=None,
                           min_samples_leaf=4, min_samples_split=10,
                           min_weight_fraction_leaf=0.0, n_estimators=190,
                           n_iter_no_change=None, presort='deprecated',
                           random_state=123, subsample=0.7, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

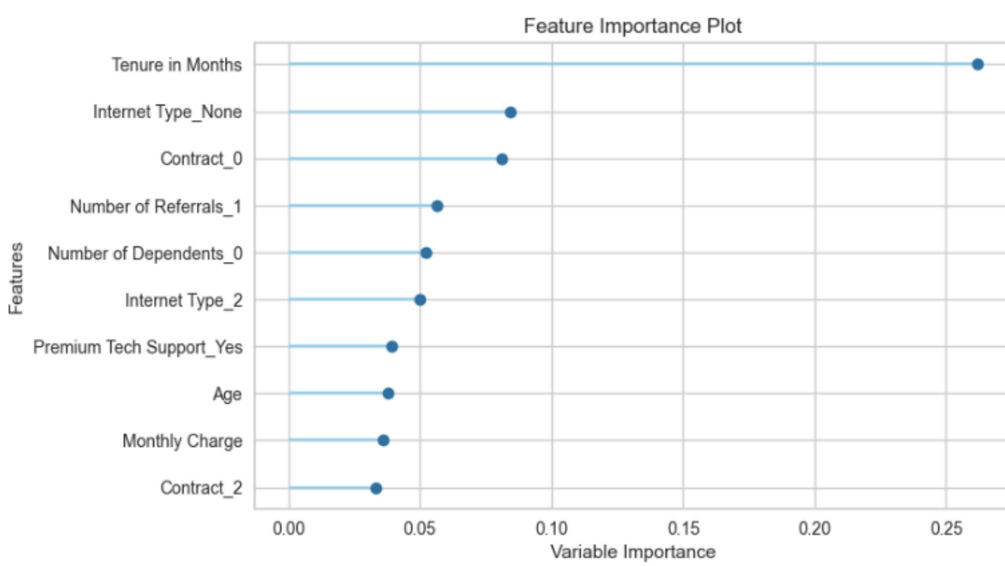
# Takeaways

The clear takeaway is that the most important feature is tenure in months. Customers are less likely to churn as their tenure with our telecom company increases.



# Recommendations

The customers that are most likely to churn are highly influenced by competitors' offers, pay a higher monthly rate compared to retained customers, and have minimal contract terms. Some options to consider for boosting customer retention are:

- Add contract terms to all plans
  - Reason: Customer tenure is a major factor in churn
    - Pros: Forces customer retention until end of contract, may make turnover more predictable
    - Cons: May be less attractive to potential customers
- Offer lower priced plans to beat competitors
  - Reason: Competitor made better offer is 2nd most common customer reason
  - Pros: More attractive to potential customers
  - Cons: Reduces profit per customer, may still result in churn
- Improve customer support quality
  - Reason: Customer Support attitude was the 3rd most common reason given for customer churn.
  - Pros: More satisfied customers, potentially higher retention rates of customers
  - Cons: Cost of training & cost of monitoring customer support

# Future Research

The focus of this analysis is to identify sources & indicators of customer churn. Additional analysis could be done on the profiles of existing customers, especially high revenue customers, or to identify the features of new customers. If additional data could be collected on competitor offerings, additional analysis regarding competitors “better devices” could be reviewed.