# Predicting Telco Churn with Supervised Learning Models

Joey Chao

# Introduction

- Churn is the rate of customers are leaving a group.
- Customer attrition is an important factor to analyze for businesses that rely on a subscription based model.
- Companies often focus on customer retention programs such as discounts or targeted offers because the cost of retaining a customer is lower than obtaining a new one.
- Modelling customer churn allows businesses to understand the factors that lead to churn as well as where to focus their efforts.

# Data

**Source:** *https://www.kaggle.com/blastchar/telco-customer-churn*

**Dataset:** Contains 7044 unique customers as well as demographic, technical,  and billing/payment informations

**Demographic Data:** Gender, dependents, senior, married

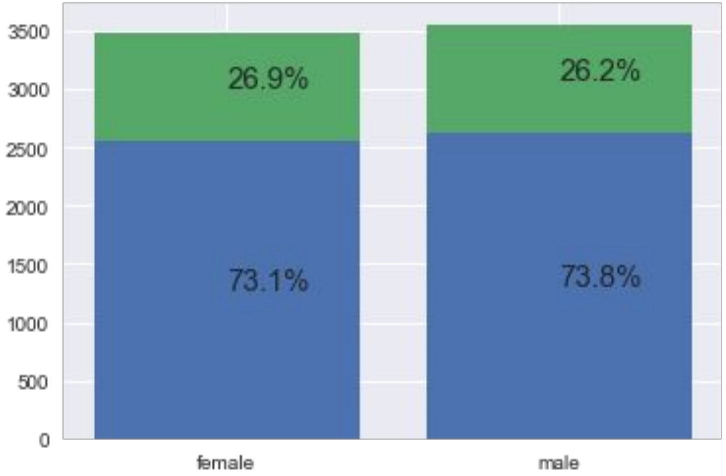
**Technical Data:** # of lines, internet service, tech support, device protection/backup and streaming.

**Billing/Payment:** Billing method, payment method, tenure

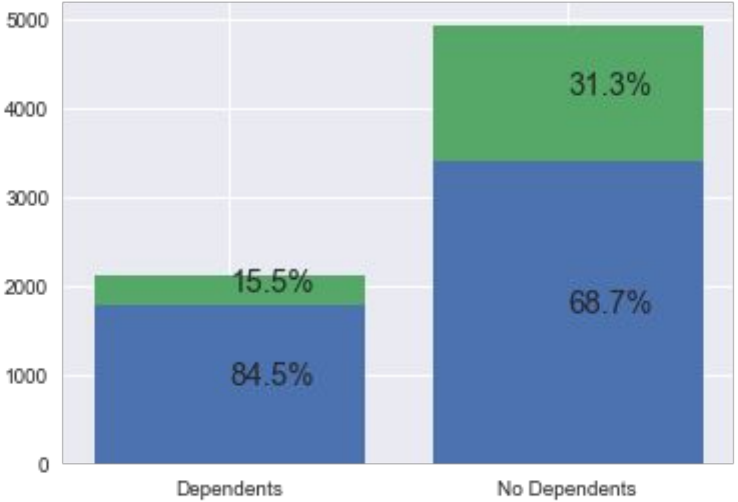
**Target of dataset:** Churn

# Demographics

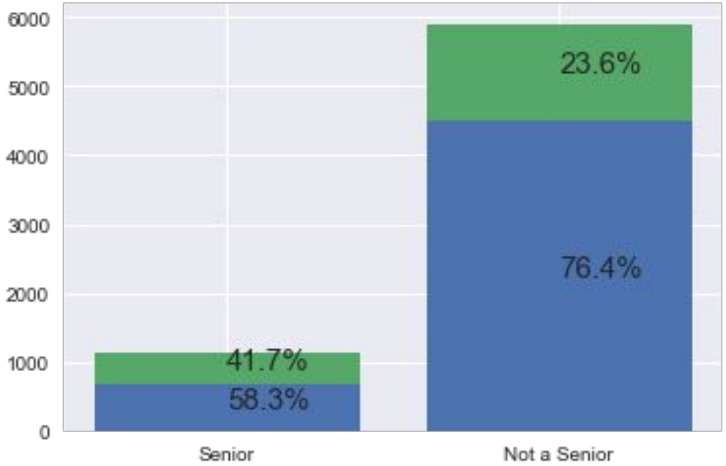
- Of the 7044 entries, 26.5% were classified Churn.
- Some variables had a greater correlation with Churn than others.




Churn by Gender



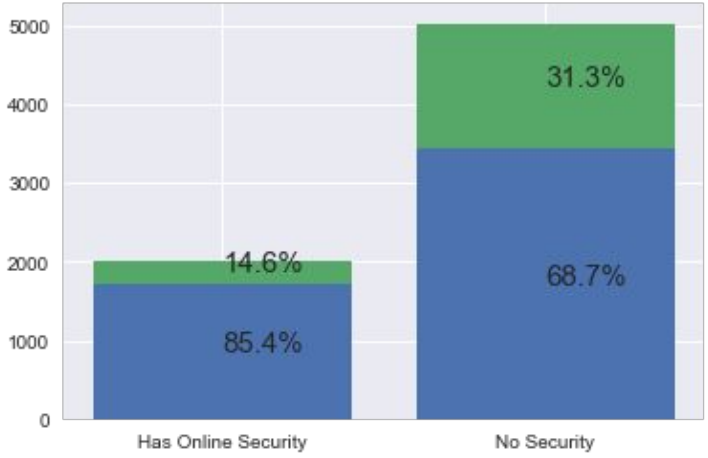

Dependents Churn

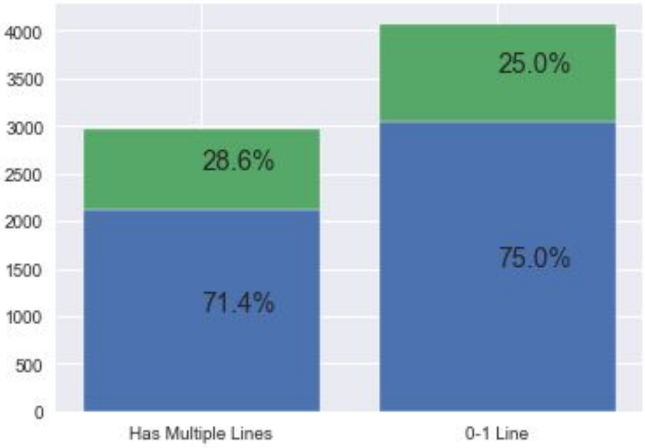


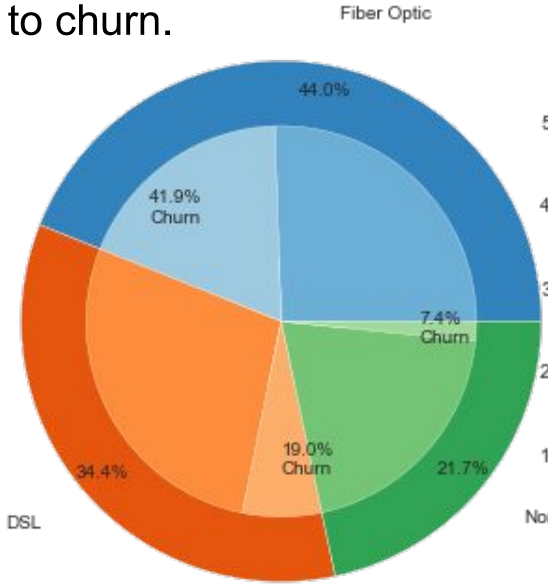

Churn in Seniors

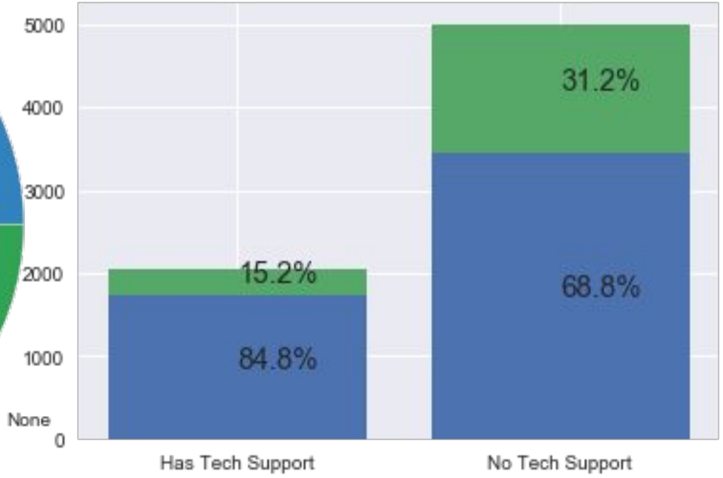
# Other Services

- Of the internet services, fiber-optic internet was the most popular option, but was also correlated with higher churn rate.
- Customers without online security and tech support were also more likely to churn.



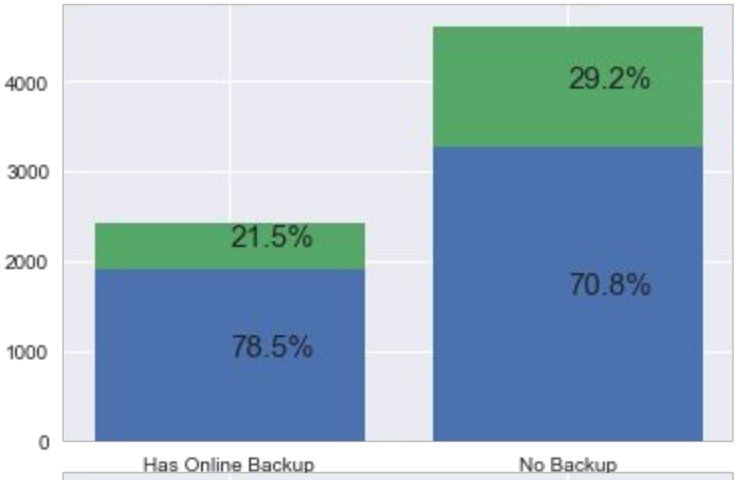



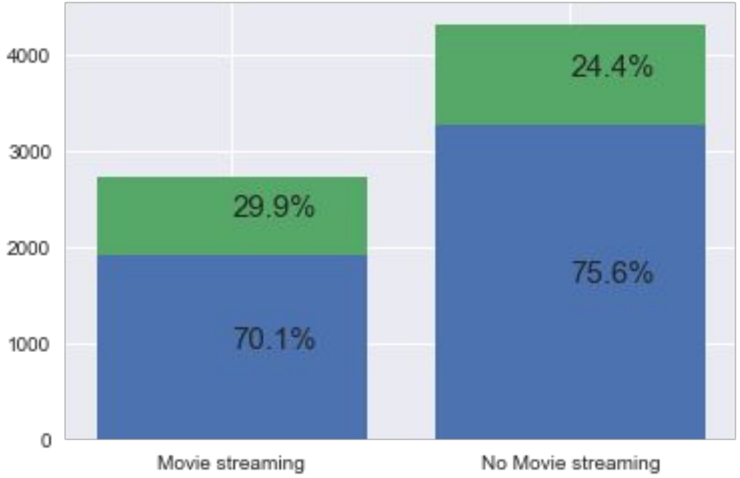



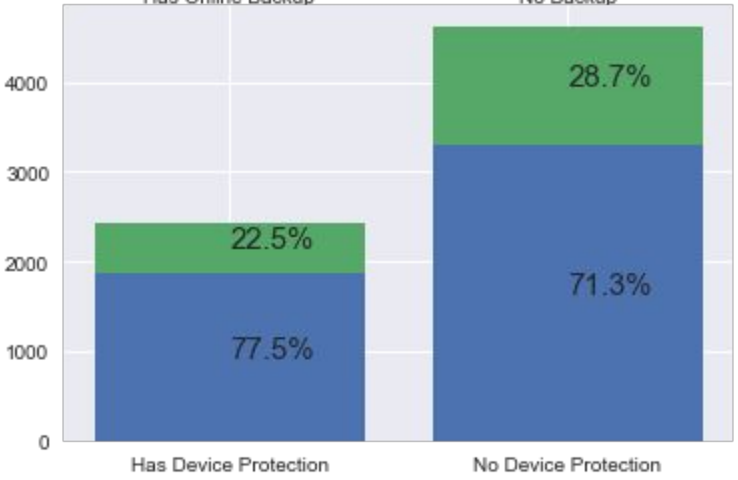



# Other Services

- Not having online backup and device protection was slightly correlated with increased churn.
- Movie streaming correlated with increased churn.

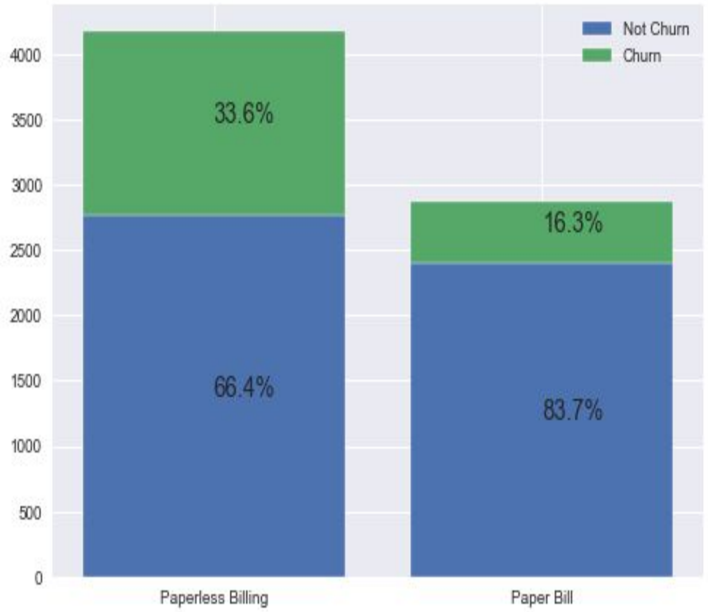


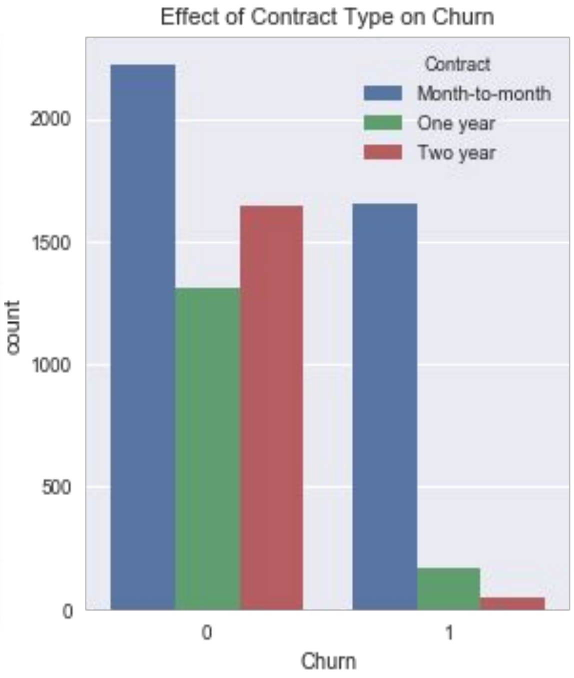



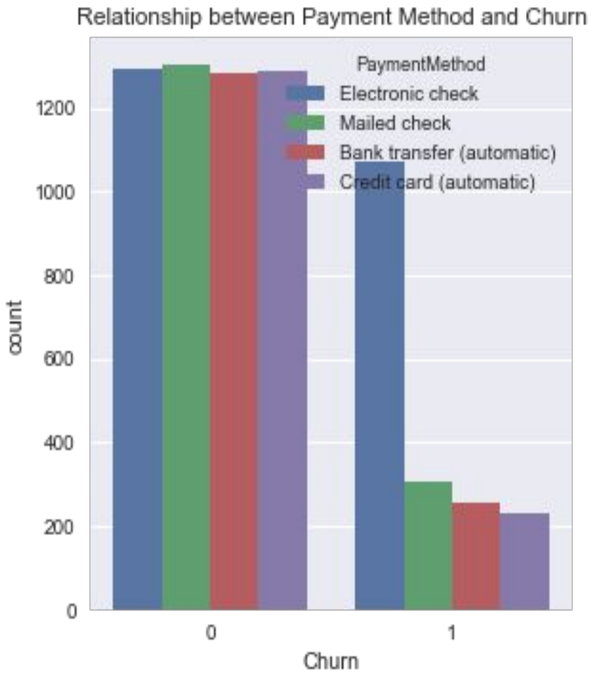



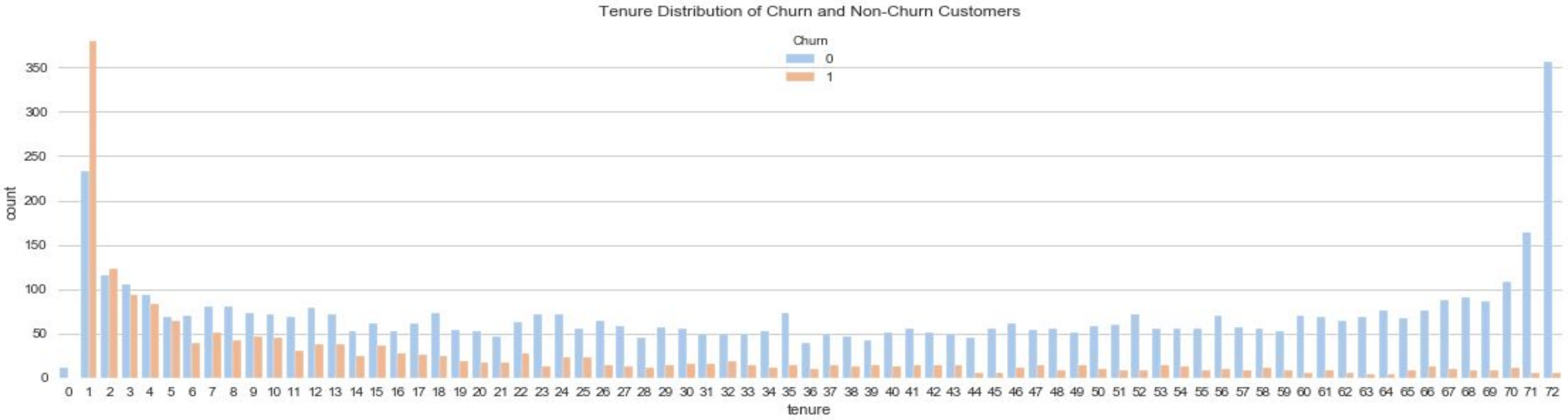
# Billing and Payment

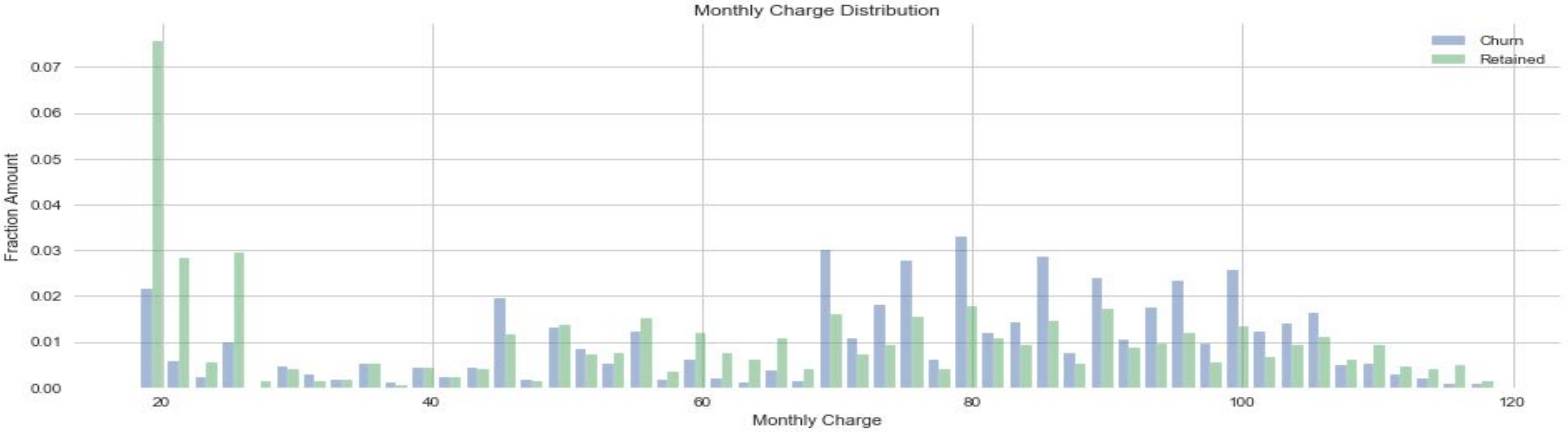
Billing Method




Tenure Distribution of Churn and Non-Churn Customers
Churn
0
1
count
tenure


Monthly Charge Distribution
Churn
Retained
Fraction Amount
Monthly Charge

# Modelling

- We are primarily interested in predicting positive churn. I chose model parameters that best improved performance predicting positive churn even if it was detrimental to predicting negative churn.
- Scaled the Monthly Charges and tenure columns with standard scalar.
- Deleted the Total Charges column because it could be derived from tenure*Monthly Charges.
- Models used: Logistic Regression, RFC, XGBoost, KNN, SVC

# Trial 1 Models

- I split data into 70/30 training and testing data.
- The models in this trial did not predict churn customers effectively, likely due to the large class imbalance.
- Of the models, logistic regression performed the best, and KNN the worst.

**Logistic Regression**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.90 | 0.87 | 1560 |
| 1 | 0.65 | 0.54 | 0.59 | 553 |
| weighted avg | 0.79 | 0.80 | 0.79 | 2113 |

**Random Forest Classifier**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.90 | 0.87 | 1560 |
| 1 | 0.65 | 0.52 | 0.57 | 553 |
| weighted avg | 0.79 | 0.80 | 0.79 | 2113 |

**XGBoost**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.89 | 0.87 | 1560 |
| 1 | 0.64 | 0.54 | 0.58 | 553 |
| weighted avg | 0.79 | 0.80 | 0.79 | 2113 |

**SVC (linear)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.91 | 0.86 | 1560 |
| 1 | 0.63 | 0.44 | 0.52 | 553 |
| weighted avg | 0.77 | 0.79 | 0.77 | 2113 |

**KNN**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.86 | 0.84 | 1560 |
| 1 | 0.54 | 0.48 | 0.51 | 553 |
| weighted avg | 0.75 | 0.76 | 0.75 | 2113 |

# Trial 2 Models - SMOTE

- In order to deal with the class imbalance, I used SMOTE to oversample the minority class (Churn).
- I split the data first and only applied SMOTE to the training set to prevent bias in the testing set.
- One interesting behavior to note was the accuracy score discrepancies between the training and testing sets in Log model and the RFC which suggests some overfitting.
- Results in this trial were almost the same for most models. SVC improved greatly.

**Logistic Regression (SMOTE)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.90 | 0.87 | 1560 |
| 1 | 0.65 | 0.54 | 0.59 | 553 |
| weighted avg | 0.80 | 0.75 | 0.76 | 2113 |

**Random Forest (SMOTE)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.91 | 0.84 | 1560 |
| 1 | 0.66 | 0.47 | 0.55 | 553 |
| weighted avg | 0.80 | 0.78 | 0.79 | 2113 |

**XGBoost (SMOTE)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.89 | 0.87 | 1560 |
| 1 | 0.63 | 0.54 | 0.58 | 553 |
| weighted avg | 0.79 | 0.79 | 0.79 | 2113 |

**SVC (SMOTE)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.68 | 0.78 | 1560 |
| 1 | 0.47 | 0.80 | 0.59 | 553 |
| weighted avg | 0.79 | 0.71 | 0.73 | 2113 |

**KNN (SMOTE)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.87 | 0.85 | 1560 |
| 1 | 0.59 | 0.53 | 0.56 | 553 |
| weighted avg | 0.76 | 0.70 | 0.72 | 2113 |

# Trial 3 Models

- I wanted to see if I could get better or comparable results by using PCA to limit the number of variables.
- 16 features explains 91% of the variance.
- In addition, I attempted to standardize the data by winsorizing Monthly Charges and tenure to deal with outliers.

**Logistic Regression (PCA)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.75 | 0.82 | 1560 |
| 1 | 0.52 | 0.77 | 0.62 | 553 |
| weighted avg | 0.80 | 0.75 | 0.76 | 2113 |

**Random Forest (PCA)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.78 | 0.83 | 1560 |
| 1 | 0.54 | 0.73 | 0.62 | 553 |
| weighted avg | 0.79 | 0.76 | 0.77 | 2113 |

**XGBoost (PCA)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.78 | 0.83 | 1560 |
| 1 | 0.53 | 0.75 | 0.62 | 553 |
| weighted avg | 0.79 | 0.76 | 0.77 | 2113 |

**SVC (PCA)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.77 | 0.82 | 1560 |
| 1 | 0.52 | 0.73 | 0.61 | 553 |
| weighted avg | 0.80 | 0.72 | 0.73 | 2113 |

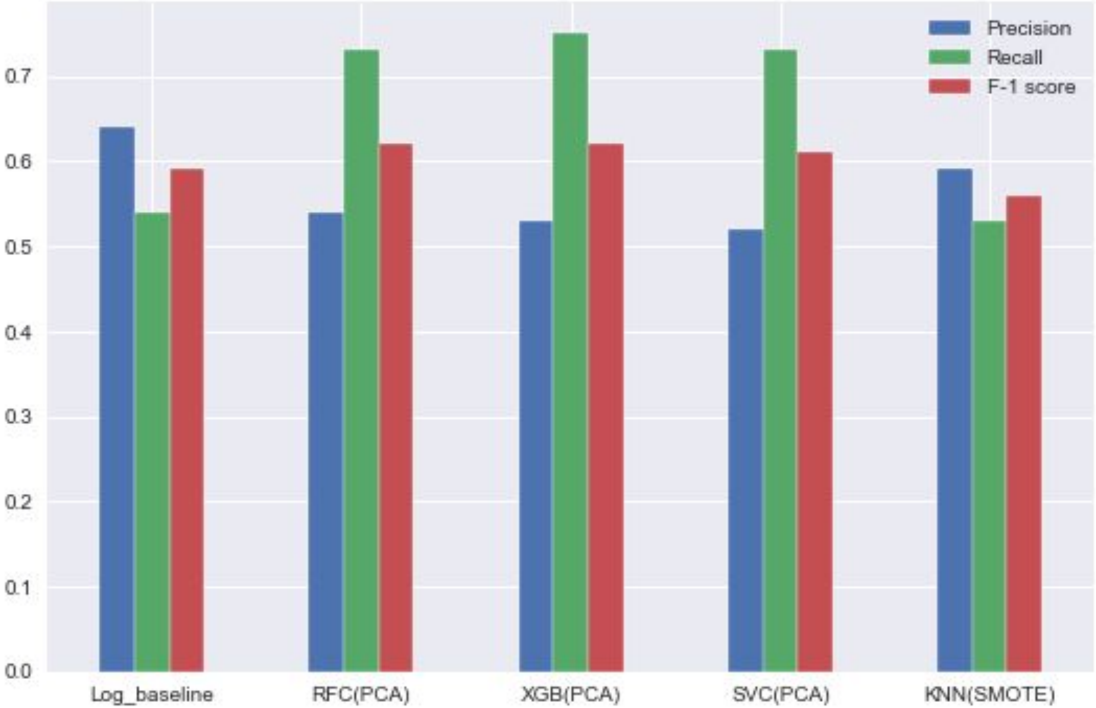
**KNN (PCA)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.74 | 0.73 | 1560 |
| 1 | 0.38 | 0.60 | 0.47 | 553 |
| weighted avg | 0.75 | 0.71 | 0.72 | 2113 |

# Model Performances

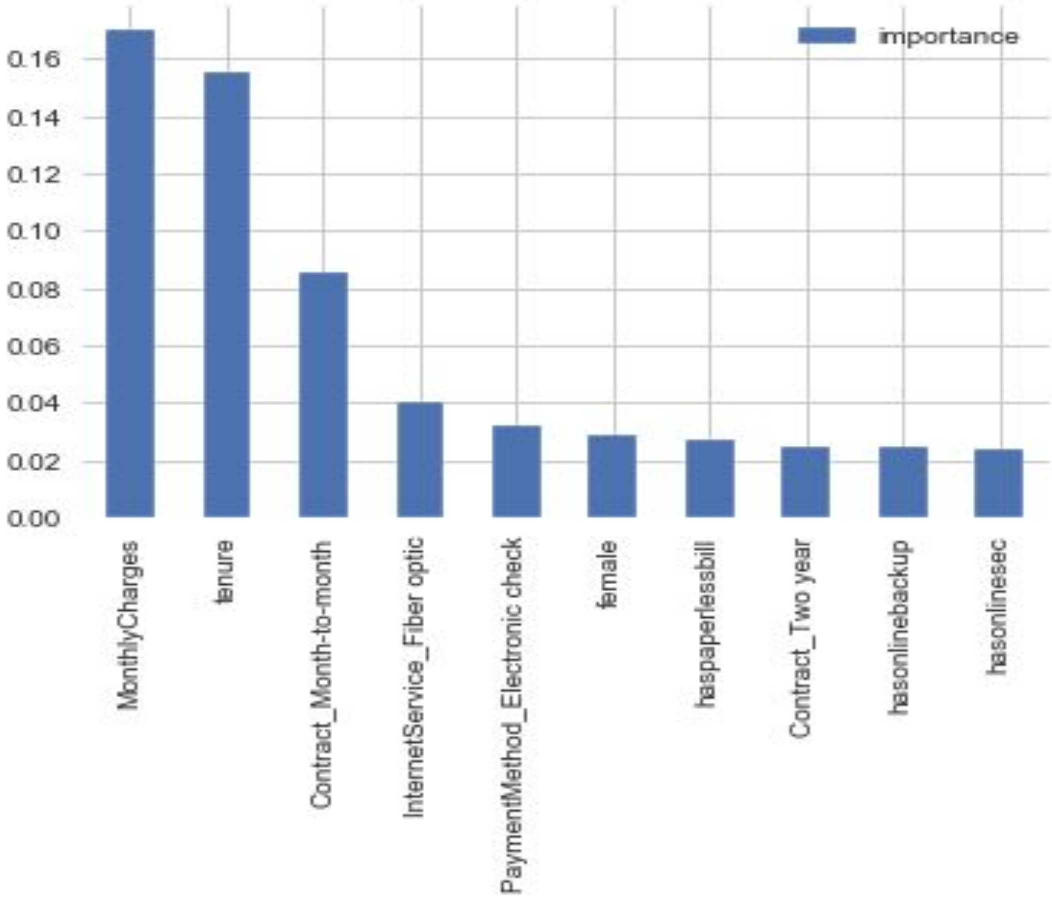
| | precision | recall | f1-score |
|---|---|---|---|
| Log (Baseline) | 0.64 | 0.54 | 0.59 |
| RFC (PCA) | 0.54 | 0.73 | 0.62 |
| XGB (PCA) | 0.53 | 0.75 | 0.62 |
| SVC (PCA) | 0.52 | 0.73 | 0.61 |
| KNN (SMOTE) | 0.59 | 0.53 | 0.56 |

- XGB on PCA reduced features was the most sensitive model.
- The baseline log model was trained on the original feature set.
- Improving recall by tuning parameters often resulted in lower precision.



# Feature Importances

- Based on the Random Forest Classifier, these are the top 10 most important features.
- Monthly Charge has the largest effect on customer churn.



## Challenges

- All of the models had a much lower ability to predict positive churn.
- The  size of the dataset is relatively small. In order to get better results, more data may be needed in addition to more entries. Time frame of collected data will also be helpful.
- The information given is relatively limited. Overall, positive churn is harder to predict because customers may leave for a variety of reasons not captured by the data collected.
- Computing power limited my ability to tune the parameters effectively.
- In addition to synthetic oversampling, I also tried undersampling the majority class as well as arbitrarily choosing equal numbers of samples from each class.

# Conclusion

- 
- Modelling customer churn can help focus customer retention programs.
- Predicting how many customers will leave the business also helps to forecast company revenue.
- In the future, gathering more detailed information about customers can help increase prediction strength.