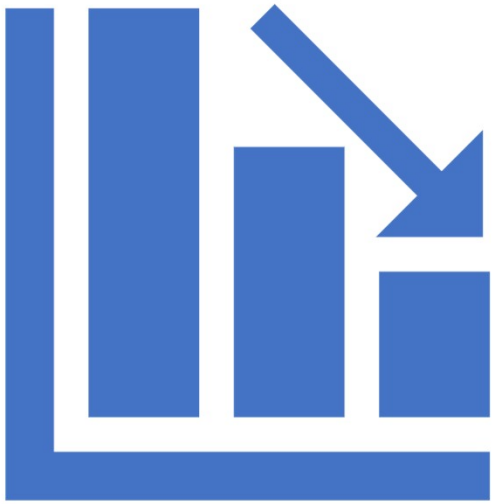
Customer Churn Prediction Analysis

Richard Jackson



Introduction

- Customer churn is a problem faced by every business, particularly those who are subscription-based services.
- CAC or the cost to acquire a new customer is sometimes considerably more to keep a current customer.
- Businesses need to know who is likely to cancel and take steps to try and keep those customers.
- Customer churn prediction is useful in helping businesses direct the appropriate efforts towards those customers most likely to churn.

Methodology

- Utilize features of customers both current and churned to determine which aspects of customer behavior are determinative of likelihood of cancellation of service.
- Customer churn analysis using
  - Logistic Regression
  - Random Forest model

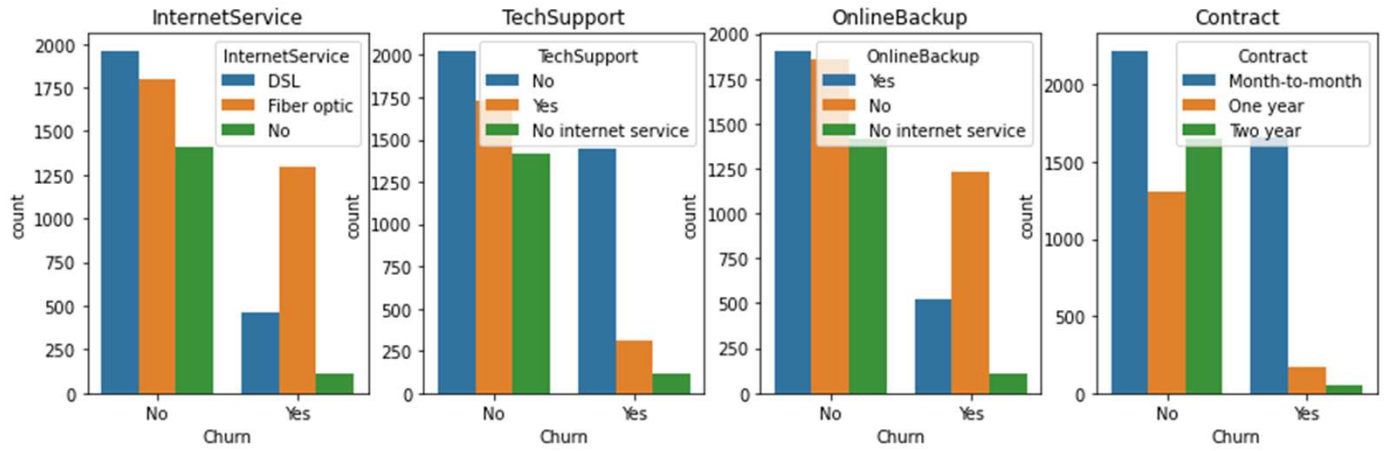
Data

- Retrieved from IBM
- Has 7043 customers – 5174 current and 1869 churned
- Has 21 features and over 7043 rows

```
df.head()
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No

Exploratory Data Analysis

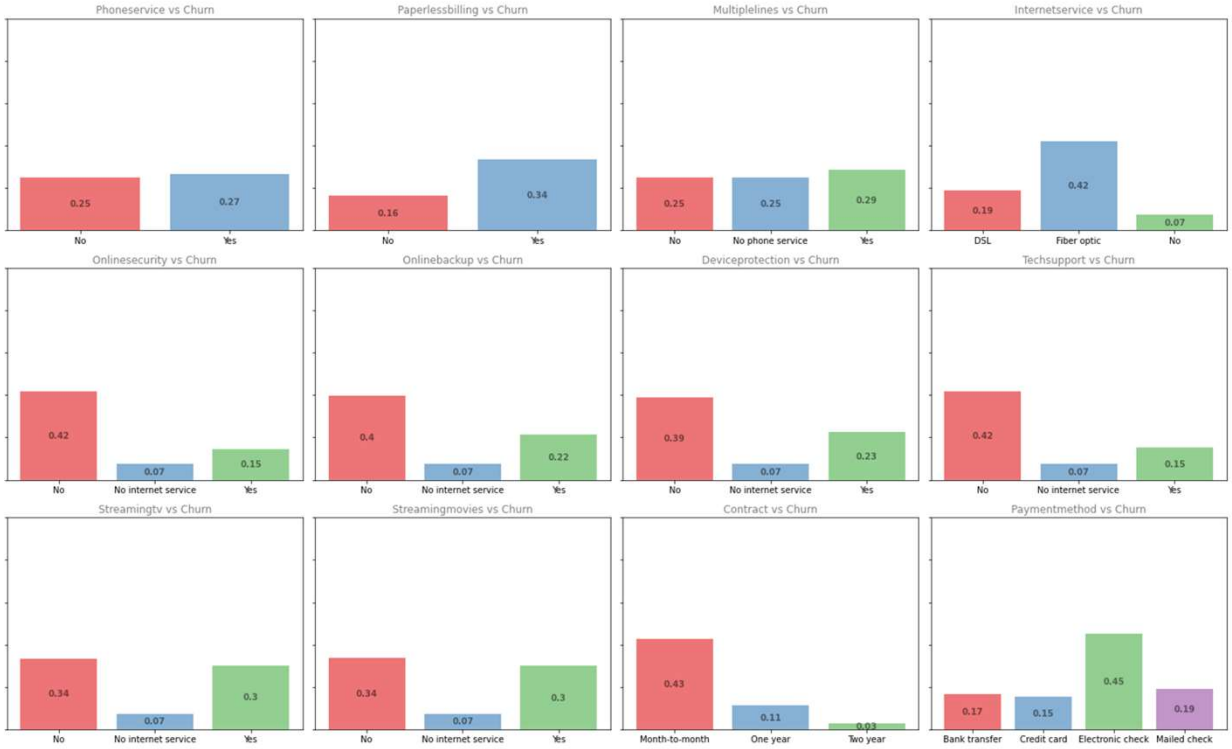


•Customers with no online security or backup, no device protection, and no tech support are from two to three times more likely to churn.

•Customers with no internet service are unlikely to churn.

•Customers with month-to-month contracts are almost four times more likely to churn than customers with yearly contracts. Two-year contractors are very unlikely to churn.

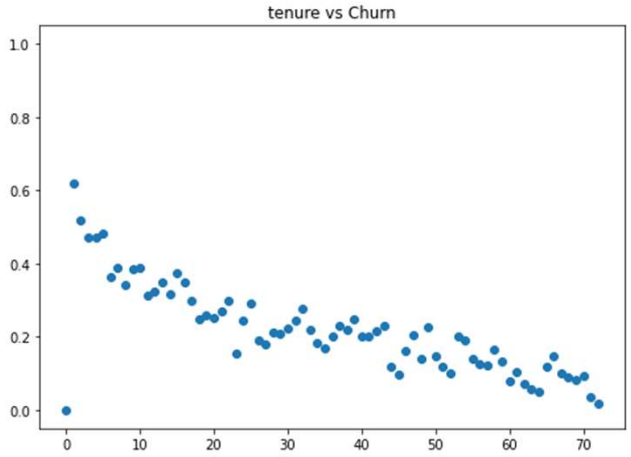
Exploratory Data Analysis

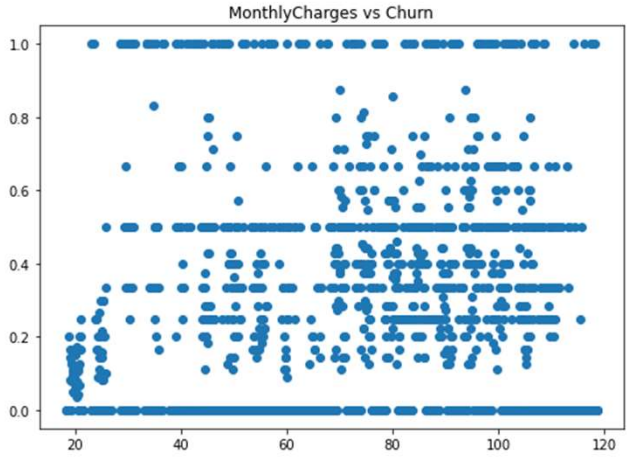


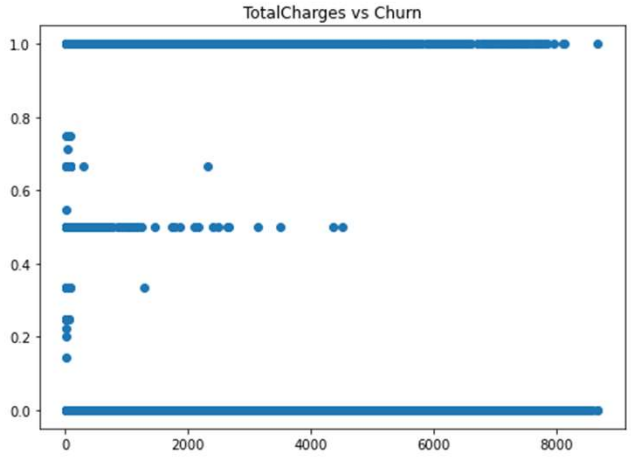
- Customers without dependents are two times more likely to churn.
- Customers that use paperless billing and optical fiber are more likely to churn.
- Customers that use electronic checks to pay their bills are more likely to churn.

Correlations

- There is a direct correlation between Tenure and Churn







Machine Learning Models

```
#Check precision, recall, f1-score
print( classification_report(y_test, predictions) )
```

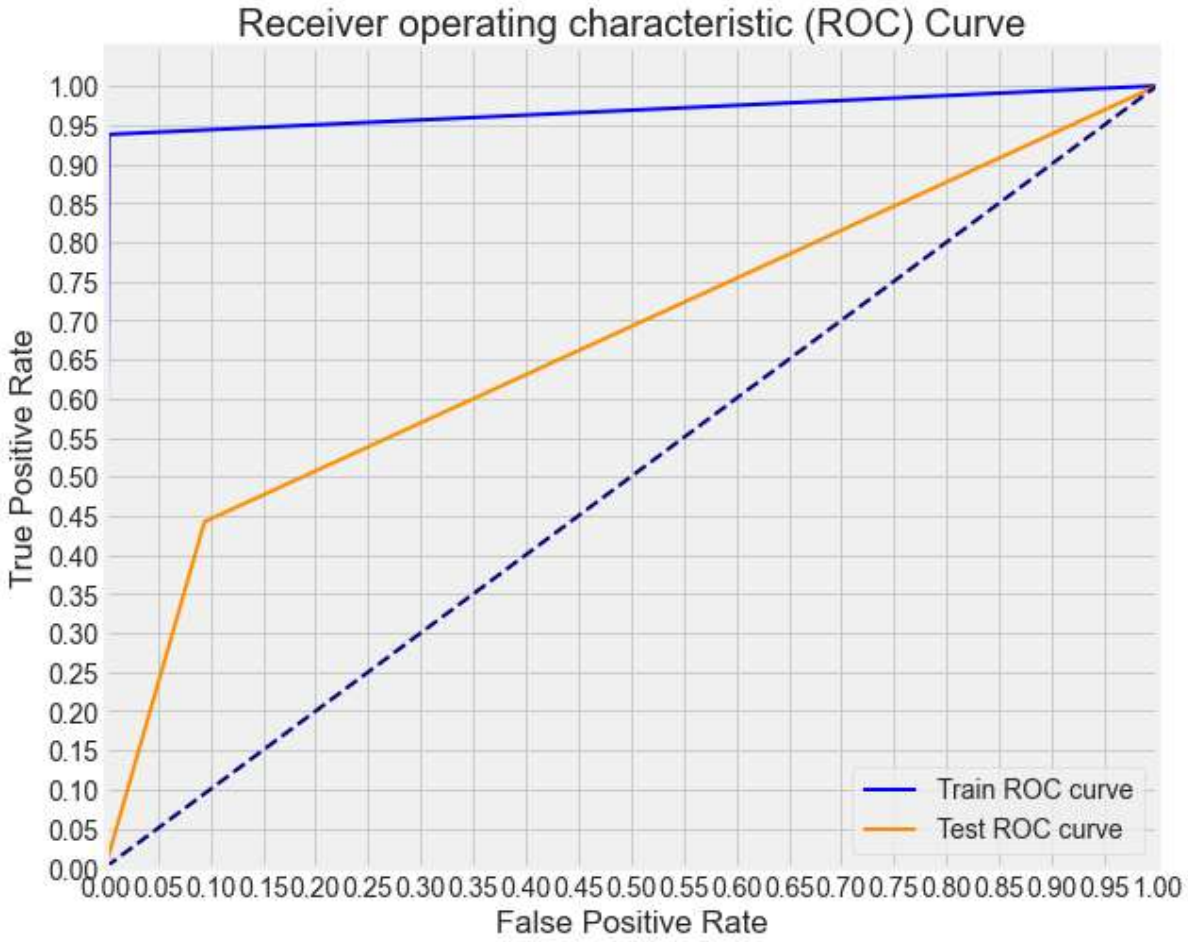
	precision	recall	f1-score	support
0	0.86	0.90	0.88	1036
1	0.68	0.58	0.62	373
accuracy			0.82	1409
macro avg	0.77	0.74	0.75	1409
weighted avg	0.81	0.82	0.81	1409

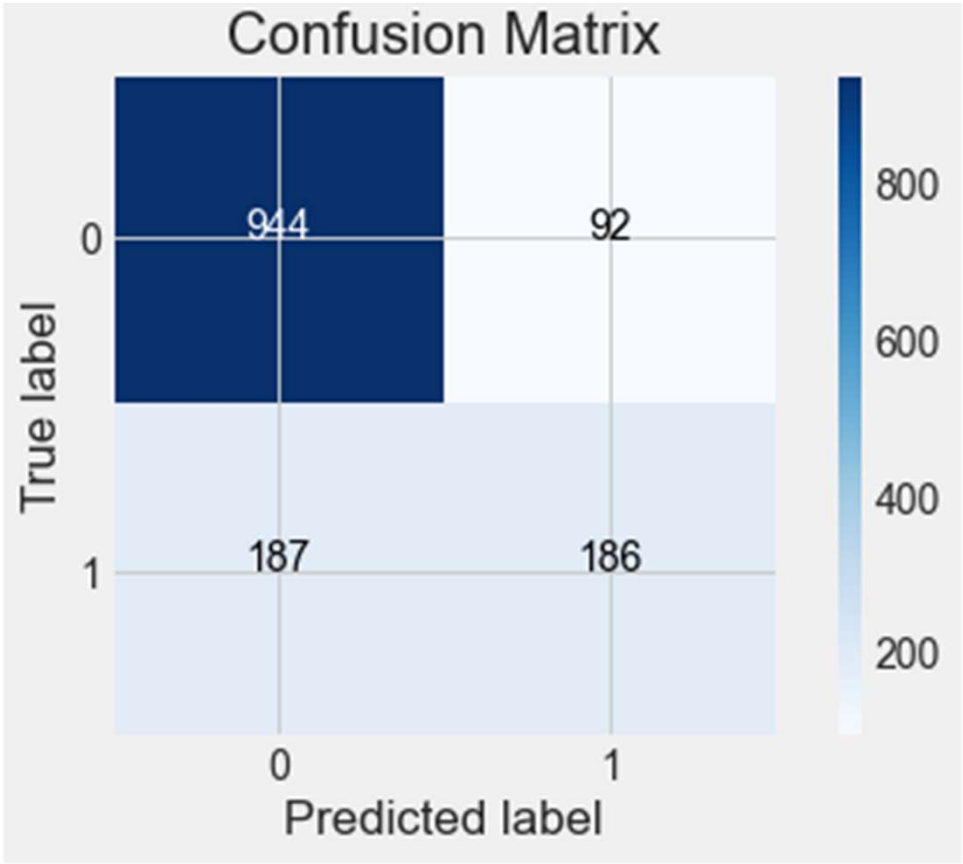
I first tried a Logistic Regression model using Scikit-Learn

The recall of this model is ≈ 90% which means it correctly identified 90% of the customers which were retained. The precision was ≈ 86% and the f1 score was 88%. The accuracy of this model is about 82%.

Random Forest model

- I then created a random forest model
- This model had slightly better recall than Logistic regression at 91% but had less precision and less accuracy than Logistic regression model.






```
test accuracy: 0.8019872249822569
train accuracy: 0.8168264110756124
```

Hyperparameter tuning the model

- I used Scikit-Learns Grid Search CV to find the optimum number of trees for the random forest model
- It did have a slight increase on the model's accuracy

Results

- **Even though hypertuning increased the Random Forest Model slighlty, the logistic regression model still perfomed better and is the model that I would use**

Future Improvements

- Add XGBoost Model
- Utilize techniques to combat the fact that this is an unbalanced data set.