Telecommunications Churn Analysis

Phone companies, among others, continuously face the challenge of maintaining a consistent customer base. Customers jump from company to company for various reasons. This can be costly because acquiring new customers is much more expensive than keeping current ones.

For this project, I will be looking at data provided by an anonymous telecommunications company on 7,000+ customers. The data includes mostly categorical information like: gender, if they are a senior citizen, if they have a spouse, or children, or if they use other products offered by the same company. Also included are some numerical fields like: tenure and monthly charges.

The end goal will be to predict if a particular person will discontinue service (churn). Having an accurate prediction if a customer will churn or not is useful to all companies, because this gives them a better prediction on future cash flow and also provides many insights into what they, as a company, are doing well for their customers and what they can improve on.

To achieve this, I plan on looking at multiple models: Random Forest, CatBoost, and XGBoost, and Neural Network to name a few, with the predictor variable being binomial, i.e. either the customer churned or didn't.

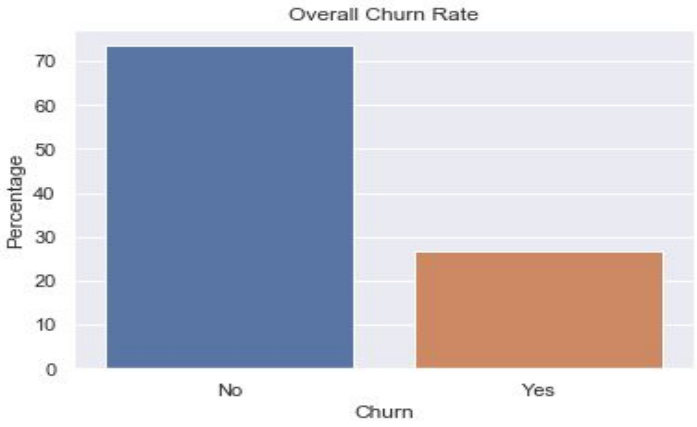
Exploratory Data Analysis

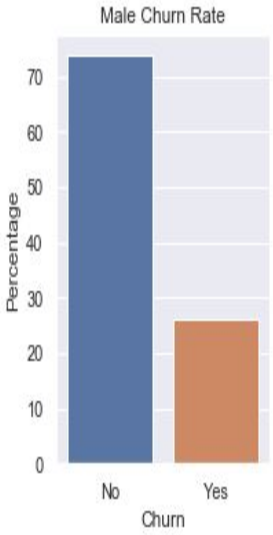
I acquired data from an anonymous telecommunications company on customers with multiple levels of information including if the customer ended dropping there service or not (churn). Below is a list of each variable with a short description:

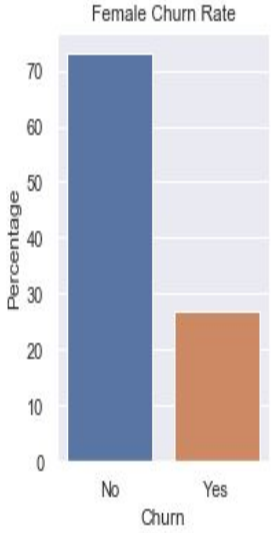
gender	**Male/Female**
SeniorCitizen	**Senior citizen or not(0,1)**
Partner	**Whether the person has a partner or not**
Dependents	**Whether the person has dependents or not**
tenure	**Number of months with the company**
PhoneService	**Whether the customer has phone service or not**
MultipleLines	**Whether the customer has multiple lines or not**
InternetService	**Customer's internet provider (DSL, Fiberoptic, No)**
OnlineSecurity	**Whether the customer has online security or not**
OnlineBackup	**Whether the customer has online backup or not**

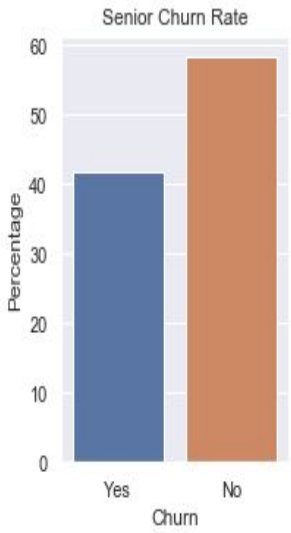
DeviceProtection	Whether the customer has device protection or not
TechSupport	Whether the customer has a tech support plan or not
StreamingTV	Whether the customer has streaming tv or not
StreamingMovies	Whether the customer has streaming movies or not
Contract	Contract term (Month-to-month, year, Two-year)
PaperlessBilling	Whether the customer has paperless billing or not
PaymentMethod	Customer's payment method (E-Check, Mailed Check,Bank,CC)
MonthlyCharges	Amount charged monthly
TotalCharges	Total amount charged
Churn	Whether customer churned or not

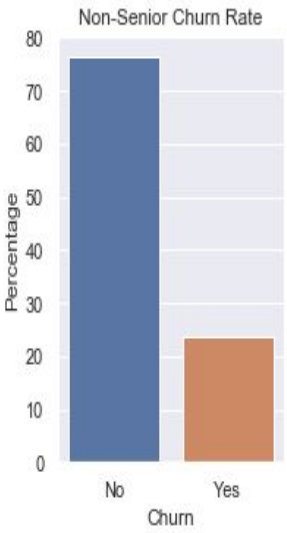
Next, I looked at some graphs to get a feel, visually, for the data starting with churn rate and then moving through each of the variables.

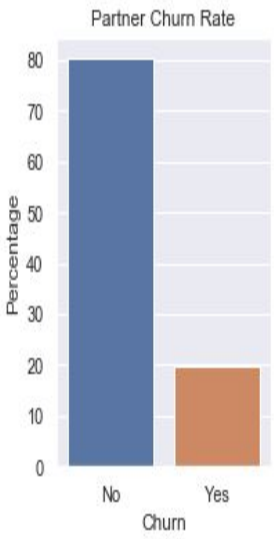


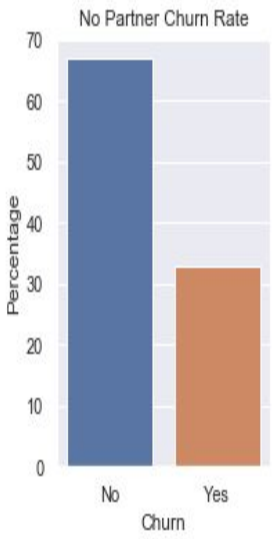


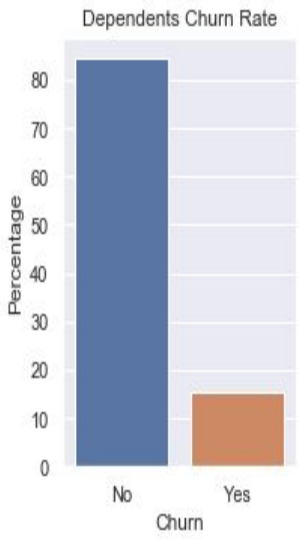


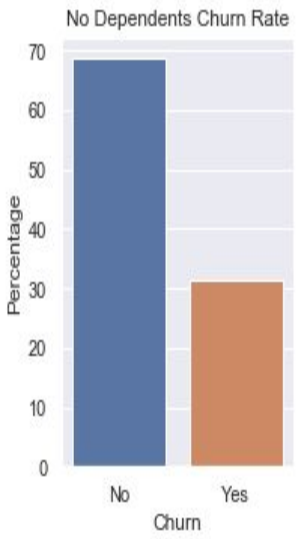


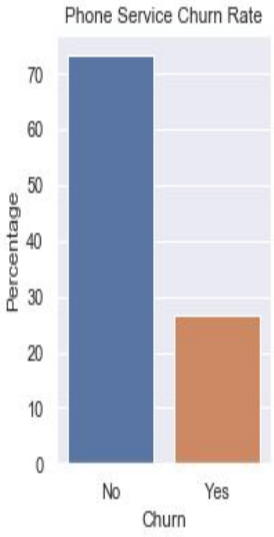


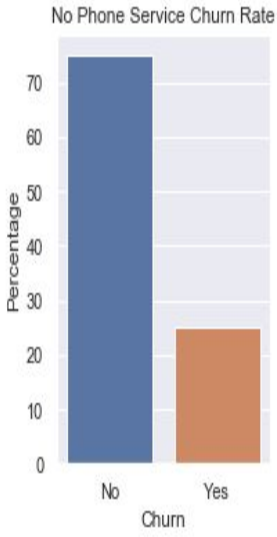

Partner Churn Rate
Percentage
Churn
No
Yes

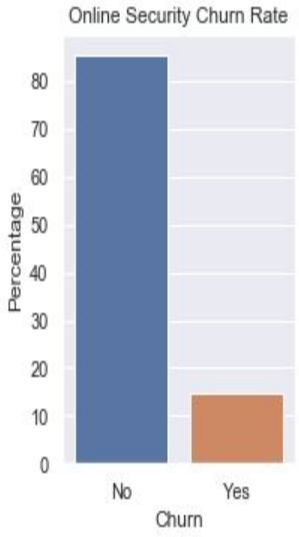

No Partner Churn Rate
Percentage
Churn
No
Yes

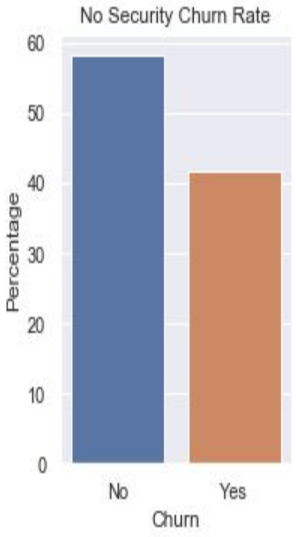

Dependents Churn Rate
Percentage
Churn
No
Yes

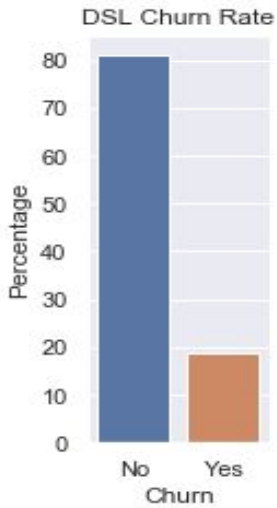

No Dependents Churn Rate
Percentage
Churn
No
Yes

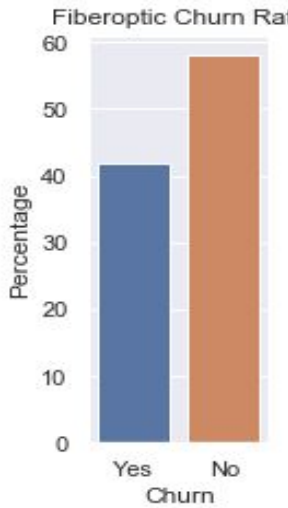

Phone Service Churn Rate
Percentage
Churn
No
Yes

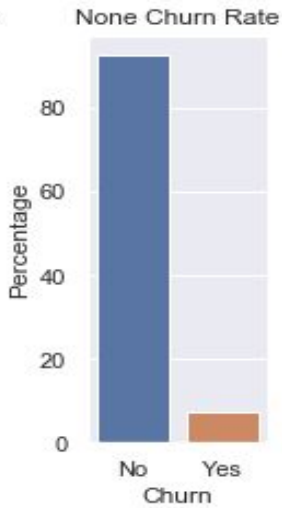

No Phone Service Churn Rate
Percentage
Churn
No
Yes


Online Security Churn Rate
Percentage
Churn
No
Yes


No Security Churn Rate
Percentage
Churn
No
Yes


DSL Churn Rate
Percentage
Churn
No
Yes


Fiberoptic Churn Rate
Percentage
Churn
Yes
No


None Churn Rate
Percentage
Churn
No
Yes

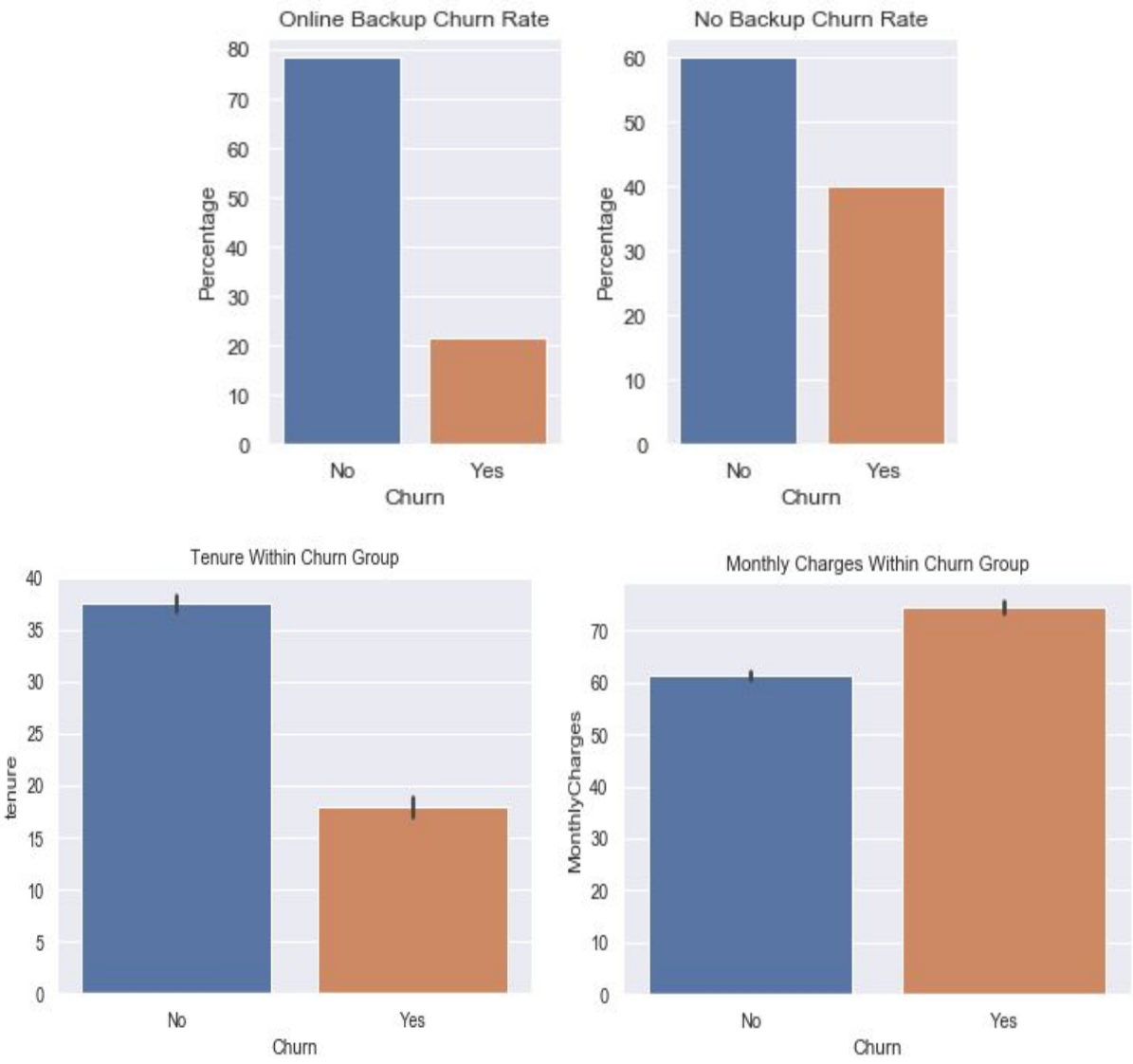


Figure 1.

In figure 1 above we can get a feel for what features might have predictive power regarding churn. For example, the churn rate of senior citizens is much higher than for non-senior citizens and the churn rate of people with a partner is lower than those without. The churn rate for people with dependents is lower than those without. And some variables are somewhat surprising in the fact that there isn't much of a difference in churn rate, for example, for people with and without phone service, the churn rate is very close to the same.

Inferential Statistical Analysis

The first thing I looked at was to see if there is any statistically strong relationships between any of the variables provided and Churn. Using a Chi-square analysis, which looks at comparing the expected amount of churn, given no relationship between churn and the chosen variable, and the observed amount of churn, I built a table showing any significant variables. First let's look at an example, let's look at testing a relationship between Partner status and Churn. Given the hypothesis below, we would expect the rate of Churn between the two groups to be about equal.

H_o: There is no relationship between partner status and churn.

H_a: There is a relationship between partner status and churn.

α=.05

Looking at a graph shows they don't appear to be equal, but we want to quantify if the difference is due to mere chance or if the observed difference is so large that chance is highly unlikely.

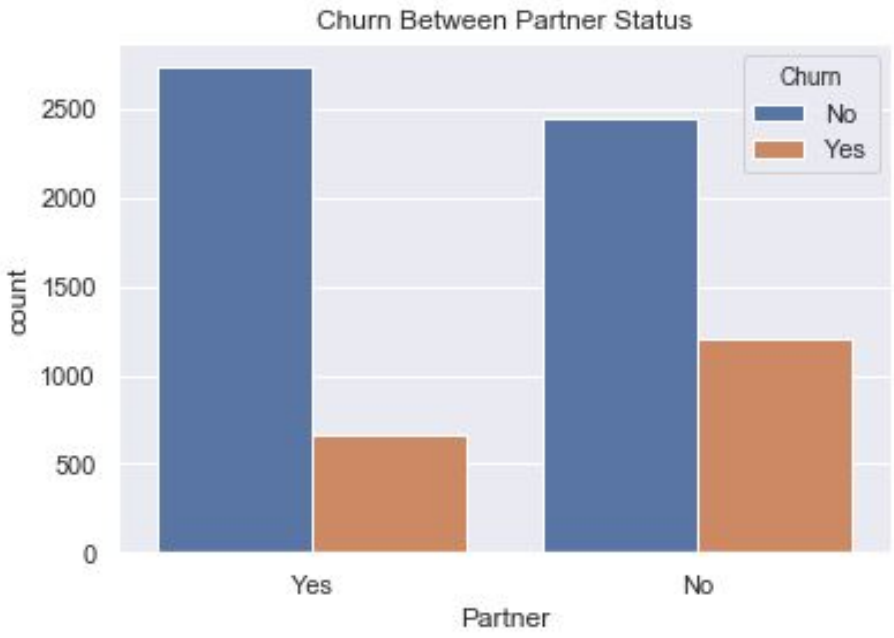


Figure 2.

Below is a table containing each variable and its respective Chi-square and p-value indicating if the probability that the observed difference is due to chance or randomness assuming the null hypothesis is true. In the example above, for the variable Partner status we get a p-value of

<.0001 and we can therefore conclude that there is a relationship between Partner status and Churn.

feature	chi2	pval
gender	0.52	0.9713
PhoneService	1.0	0.9091
MultipleLines	11.33	0.0787
Partner	159.41	<.0001
SeniorCitizen	160.35	<.0001
Dependents	189.94	<.0001
PaperlessBilling	259.16	<.0001
StreamingTV	374.2	<.0001
StreamingMovies	375.66	<.0001
DeviceProtection	558.42	<.0001
OnlineBackup	601.81	<.0001
PaymentMethod	648.14	<.0001
InternetService	732.31	<.0001
TechSupport	828.2	<.0001
OnlineSecurity	850.0	<.0001
Contract	1184.6	<.0001

Figure 3.

Now that we have found some variables that show a relationship to Churn, let's look at some relationships between variables. Understanding these relationships can help with other parts of business. Say for example, if a company wants to increase sales of longer term contracts. Well if we know what groups are more likely to purchase that then we can know where to invest in marketing. Let's start with looking at Partner status and Contract type.

H_o: There is no relationship between partner status and type of contract selected.

H_a: There is a relationship between partner status and type of contract selected.

α=.05

Figure 4.

Looking at the graph, we can see that single people are more likely to get into a month to month contract than people with a partner. And customers with a partner are more likely to sign one or two year contracts.

```
Chi-square Statistic : 617.2440921596583 ,p-value: 4.4441292473963605e-130
```

With p < .0001, we reject the null hypothesis in favor of the alternative and conclude that there is a relationship between partner status and the type of contract purchased.

H_o: There is no relationship between senior citizen status and type of contract selected.

H_a: There is a relationship between senior citizen status and type of contract selected.

α=.05

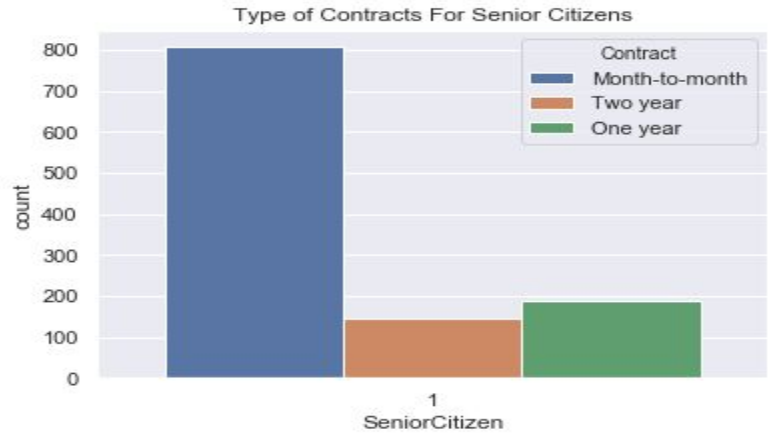


Figure 5.

The data suggests that Senior Citizens are more likely to be in month to month contracts.

```
Chi-square Statistic : 145.77676674829792 ,p-value: 6.041978569381217e-29
```

With p < .0001, we reject the null hypothesis in favor of the alternative and conclude that there is a relationship between senior citizen status and the type of contract purchased.

H_o: There is no relationship between partner status and selecting device protection.

H_a: There is a relationship between partner status and selecting device protection.

α=.05

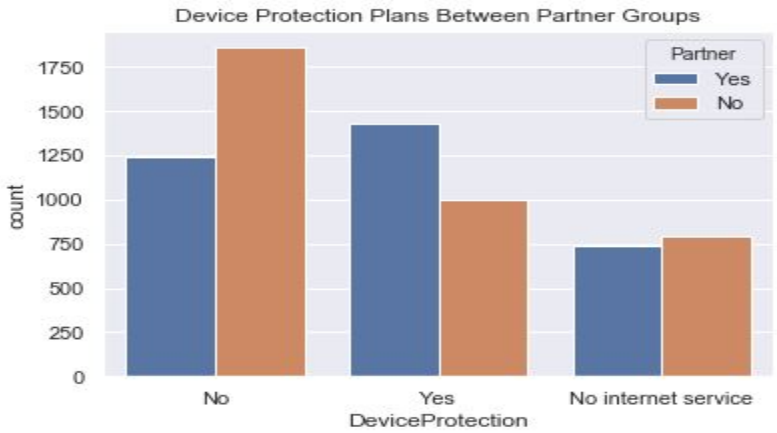


Figure 6.

The data suggests that of those who chose a device protection plan, most had a partner and of those who selected no online security, most didn't have a partner.

```
Chi-square Statistic : 195.40793551503089 ,p-value: 1.800485981859397e-39
```

With p < .0001, we reject the null hypothesis in favor of the alternative and conclude that there is a relationship between partner status and if they purchased device protection.

Machine Learning

In this section, I want to look at building a model to predict if an individual will churn based on the features provided. I will cover feature importance, model parameter tuning, and model evaluation. To begin, I fit a Random Forest Classifier using a probability prediction threshold of 0.5 and looked at the feature scores to get an idea of the most predictive features for the model, shown below. It looks like tenure and monthly charges are the most important features. However, as discussed earlier in this paper, many of the categorical features have a statistical

relationship with churn and therefore we want to keep each of those significant features in the model.

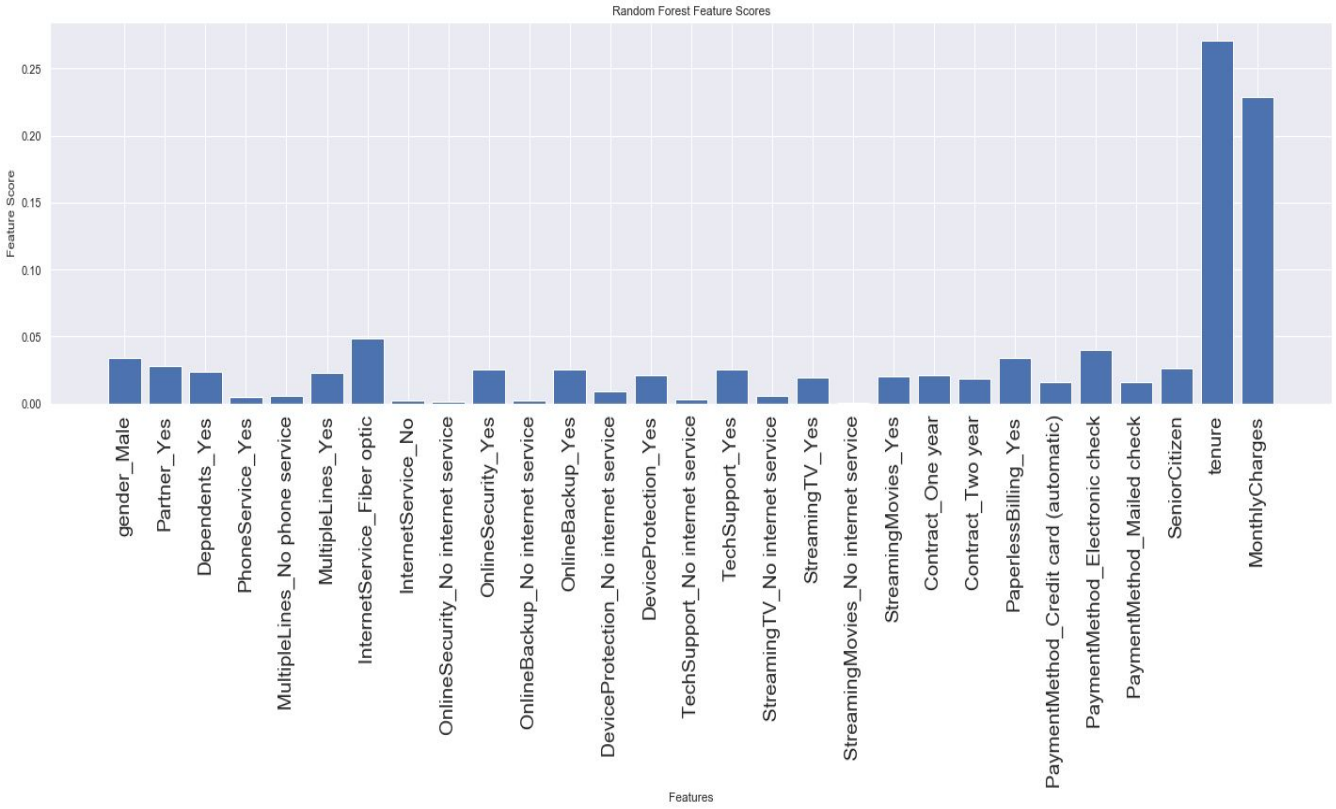


Figure 7.

Next, I looked at tuning the “n_estimators” parameter to get the best results. “N_estimators” in the sklearn library is the number of decision trees built out by the Random Forest Classifier. To determine the best results, I used the AUROC score (Area Under the Receiver Operating Characteristic) for comparisons. I also used a 5-fold cross validation process, which involves separating the training data into four training sets and one validation set. Then the model is trained on the training sets and scored on the validation set. This is done five times, rotating the validation set each time, to accumulate an average AUROC score for each parameter value.

```
{'n_estimators': 350}
```

The top score is for n_estimators = 350. Now that we have the features we want to keep and a tuned up Random Forest Classifier, I then looked at comparing different models to see if we can get better performance. For performance measurements, using a probability prediction threshold of 0.5, I looked at a confusion matrix, 5-fold cross validation ROC, and a Lift Graph for each of the models.

The Confusion Matrix looks at the model accuracy rates for true positive, true negative and the inaccuracy rates for false positive and false negative. In other words, the Confusion Matrix shows the rate at which the model predicted a customer would churn (1) and they did, as well as the rate at which the model predicted customer wouldn't churn (0) and they actually did churn.

In figure 8 below you will see the percentages and counts for each confusion matrix.

Random Forest Confusion Matrix

Actual	0	1
0.0	0.6769	0.0772
1.0	0.1227	0.1232

Random Forest Confusion Matrix

Actual	0	1
0	1192	136
1	216	217

XGBoost Confusion Matrix

Actual	0	1
0.0	0.674	0.0801
1.0	0.1221	0.1238

XGBoost Confusion Matrix

Actual	0	1
0	1187	141
1	215	218

CatBoost Confusion Matrix

Actual	0.0	1.0
0.0	0.6882	0.0659
1.0	0.1175	0.1283

CatBoost Confusion Matrix

Actual	0.0	1.0
0	1212	116
1	207	226

Log Regression Confusion Matrix

Actual	0	1
0.0	0.6786	0.0755
1.0	0.1153	0.1306

Log Regression Confusion Matrix

Actual	0	1
0	1195	133
1	203	230

Neural Networks Confusion Matrix

Actual	0	1
0.0	0.6854	0.0687
1.0	0.1175	0.1283

Neural Networks Confusion Matrix

Actual	0	1
0	1207	121
1	207	226

Stacked Confusion Matrix

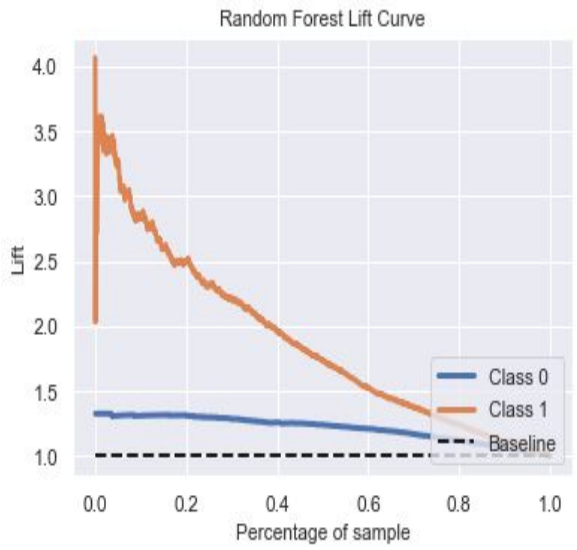
Actual	0	1
0.0	0.674	0.0801
1.0	0.1221	0.1238

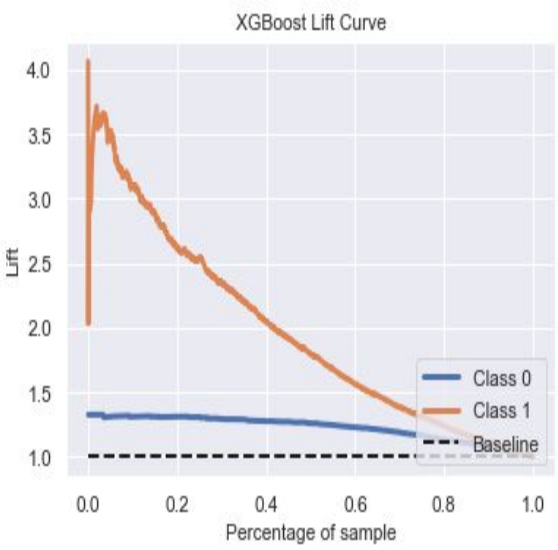
Stacked Confusion Matrix

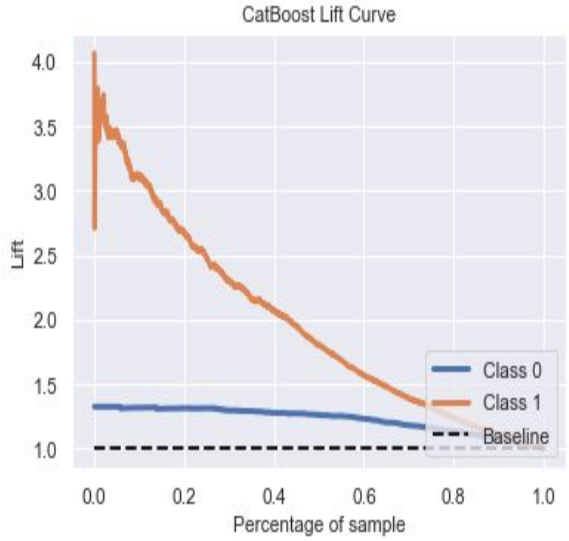
Actual	0	1
0	1187	141
1	215	218

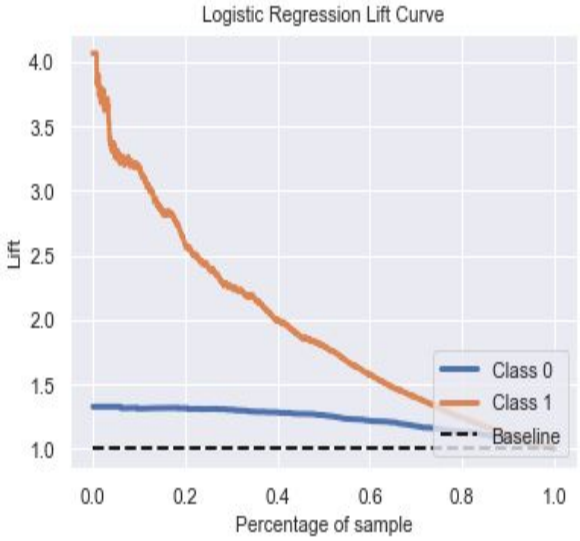
Figure 8.

The Lift Graph shows how much better the model predictions are above randomly guessing at each of the probability percentiles. The Y axis is the lift measurement, and means if you used predictions at X percentile you would achieve Y times better accuracy than random guessing.


Random Forest Lift Curve
Lift
Percentage of sample
Class 0
Class 1
Baseline
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost Lift Curve
Lift
Percentage of sample
Class 0
Class 1
Baseline
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.0
0.2
0.4
0.6
0.8
1.0


CatBoost Lift Curve
Lift
Percentage of sample
Class 0
Class 1
Baseline
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Logistic Regression Lift Curve
Lift
Percentage of sample
Class 0
Class 1
Baseline
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.0
0.2
0.4
0.6
0.8
1.0

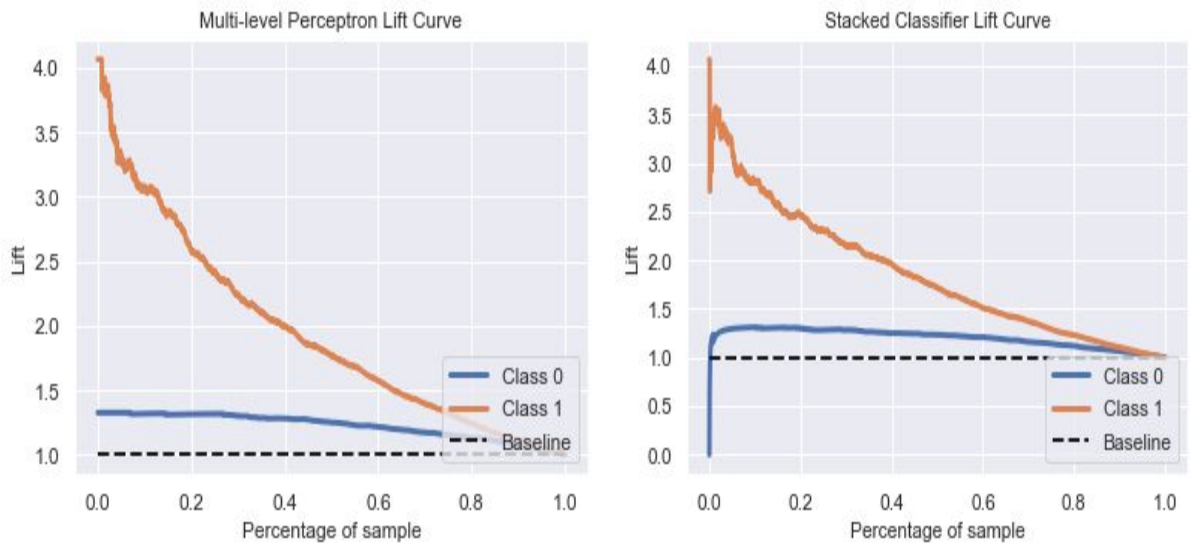


Figure 9.

ROC stands for Receiver Operating Characteristic, the ROC curve contains true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the graph indicates a true positive rate of 1 and a false positive rate of 0. Therefore a curve with a larger area under the curve (AUC) is better. For evaluating multiples models, I used AUROC with 5-fold cross validation.

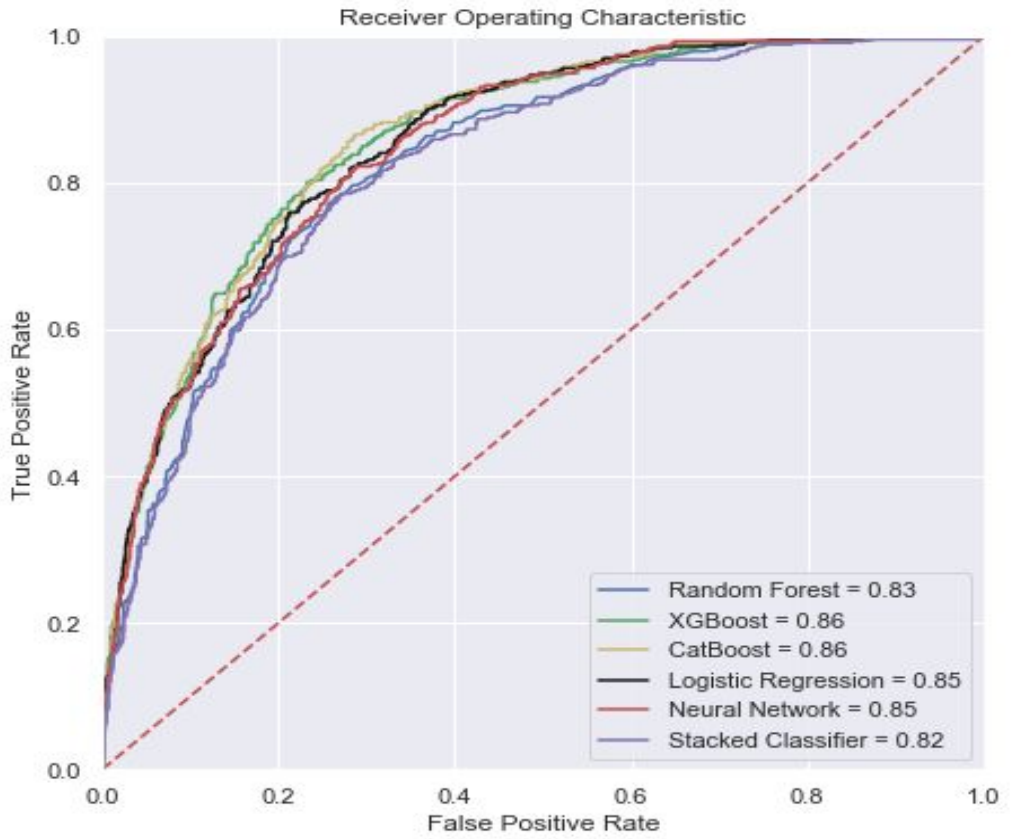


Figure 10.

The CatBoost and XGBoost models have the best AUROC scores, both at .86, however, the CatBoost model has better prediction accuracy on the test data.

Conclusion

When predicting if a customer will churn the most important features are: Senior Citizen, Partner, Dependents, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, and Payment Method. The data shows there are significant differences in churn rate within these categories. Using a CatBoost model we are able to achieve 82% prediction accuracy on individual churn.

Appendix

The data set can be found at: https://www.kaggle.com/puja19/telcom-customer-churn. The Python libraries used in this analysis are:

- Pandas v0.24.2
- Numpy v1.16.2
- Seaborn v0.9.0
- Matplotlib v3.0.3
- Scipy v1.2.1
- Sklearn v0.20.0
- Mlxtend v0.15.0.0

When setting up a model for machine learning, most models require the categorical features be converted to dummy variables. This means that instead of a feature containing strings with each of the categories, the feature contains 1 or 0, for example with a binary feature. Features with more than two categories, need to be split into multiple columns containing 0 and 1. However, the models will accept any numeric data, so it is possible to encode categorical features numerically (0,1,2,....n_categories). When done this way, the model treats the numbers ordinally, so three is greater than two and four is greater than three, etc. This is called label encoding and below is a comparison of the two types of categorical encoding.

Random Forest Label Encoding

Compared with the feature scores using dummy variables, in the graph below, the features are not split apart. The results are only slightly different, but Tenure and Monthly Charges are still the most predictive.

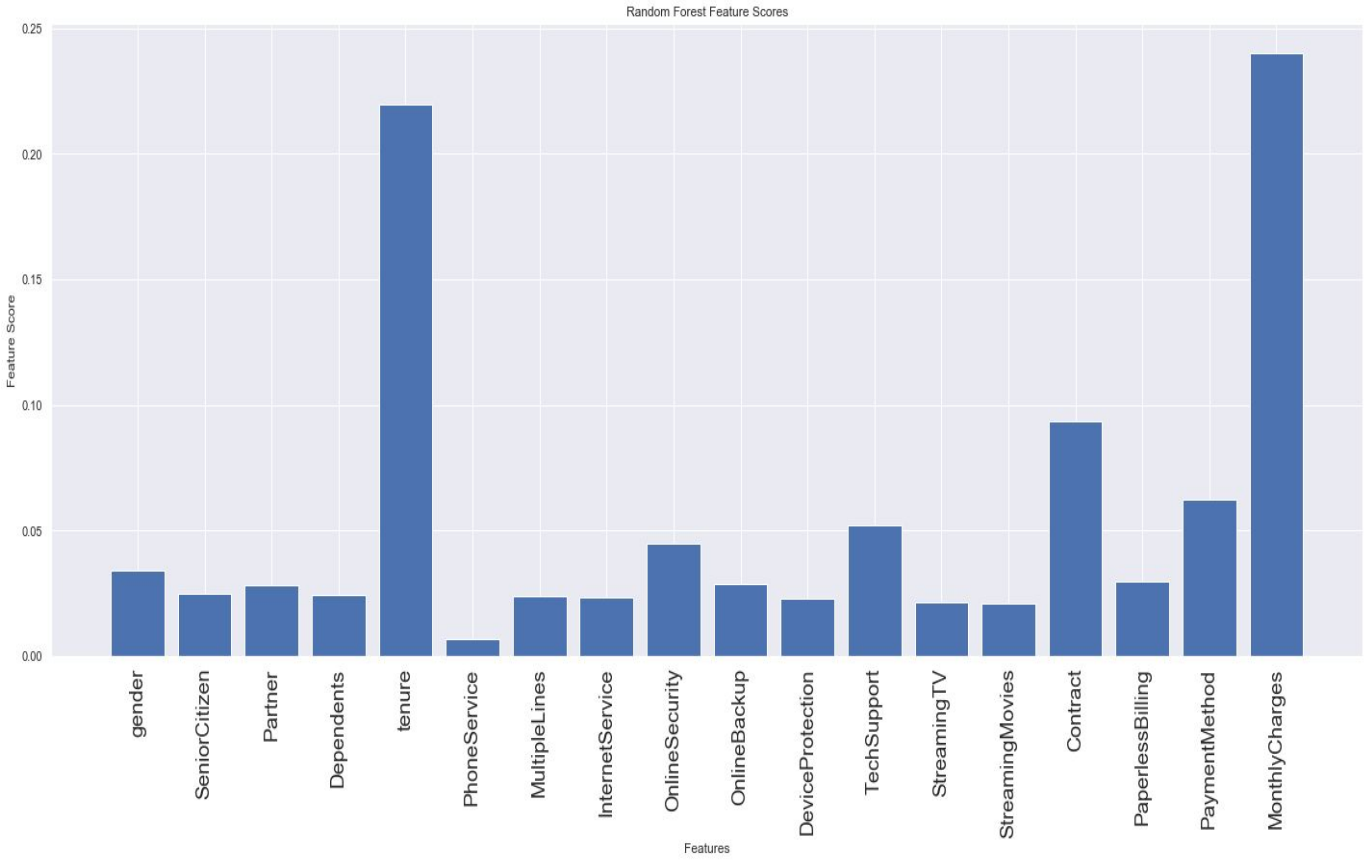


Figure 11.

The accuracy of each model came in very close to each other, with the label encoded model slightly higher at 81% and the dummy encoded model at 80.9% prediction accuracy.

```
[0.8103350369108461, 0.8097671777399205]
```

Looking at the Lift Curves below, only slight differences can be seen. The label encoded is on the left and the dummy encoded is on the right.

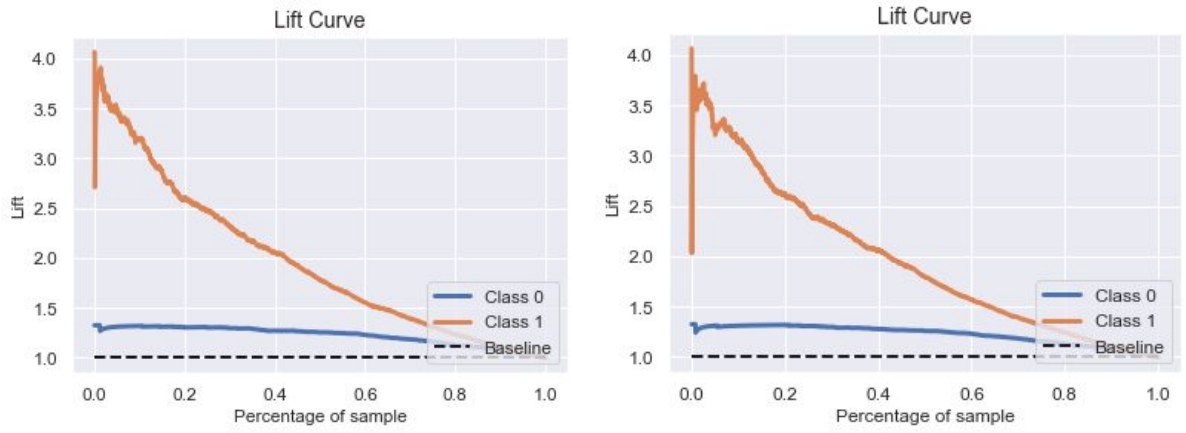


Figure 12.

This was the case for each of the models except the neural networks model. The label encoded data produced a model that predicted zero people would churn. As shown in the Lift Curves below, for the label encoded model (left) there is a significant drop off in “Class 0” (blue line) at the highest probabilities.

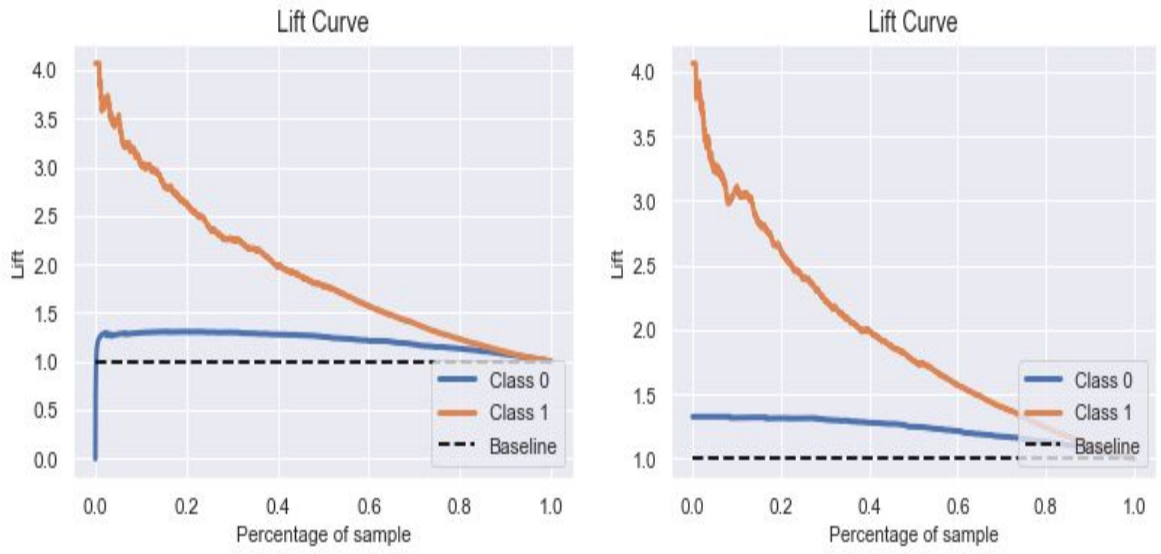


Figure 13.

Below are the ROC curves of each of the label encoded models. No significant changes can be seen compared with the ROC curve of the dummy encoded models shown earlier in this paper.

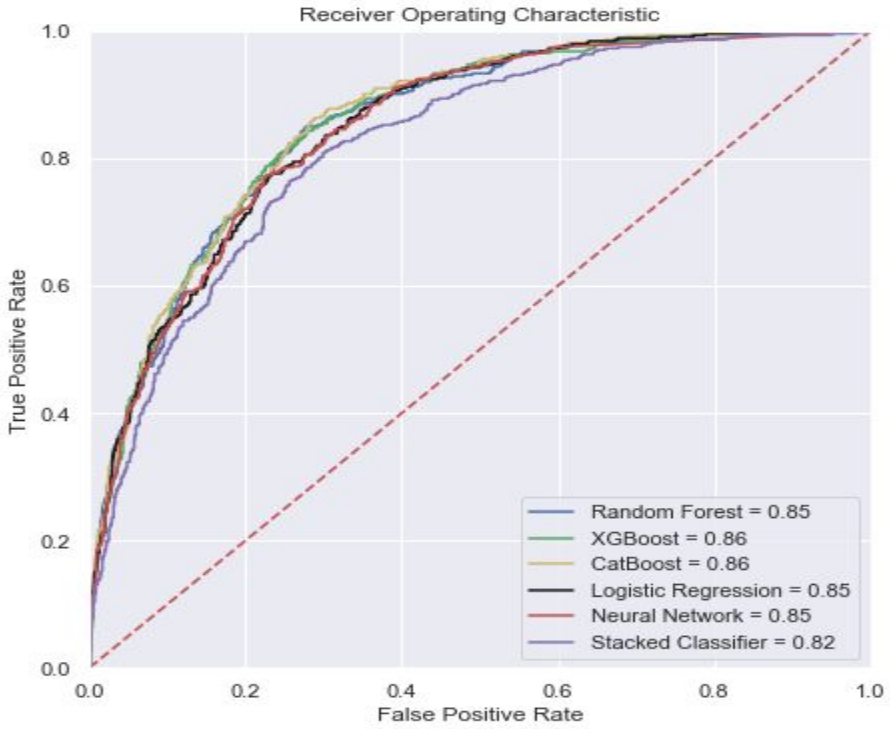

Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest = 0.85
XGBoost = 0.86
CatBoost = 0.86
Logistic Regression = 0.85
Neural Network = 0.85
Stacked Classifier = 0.82

Figure 14.