Report on Predicting Telecom Customer Churn

Problem Identification:

What is Customer Churn?

Churn in a business setting refers to losing an acquired, potentially profitable customer. The definition of churn can vary by industry. Customer churn means a customer's ending their relationship with a business for any reason. Although churn is inevitable at a certain level, a high customer churn rate is a reason for failing to reach business goals. So identifying customers who would churn is very important for business, as it can be difficult for established businesses with a broad customer base to identify customers at risk. Acquiring a new customer is always more expensive than retaining an existing one. Hence, not letting them churn is the key to a sustained revenue stream.

Why does it matter?

We could put this model into production and deliver results to a "call center" to offer good deals to "in-danger" customers. Calling people and offering deals cost money so you have a limited amount of people you can say they have to call. How do you decide who? What is the best metric you can use/optimize? Acquiring a new customer is always more expensive than retaining an existing one. Hence, not letting them churn is the key to a sustained revenue stream.

Problem Statement: Predict the telecom customer attrition rate and find the potential churning customers by analyzing data for a specific period of

six years time to minimize the churn rate by providing better service for the company's growth and financial stability.

The Data:

The dataset is one provided by IBM. The Data consists of information of around 7043 customers' service contracts, the type of services offered, payment details etc. Data was extracted from IBM Cognos Analytics Data Collection.
https://community.ibm.com/accelerators/catalog/content/Customer-churn

Approach:

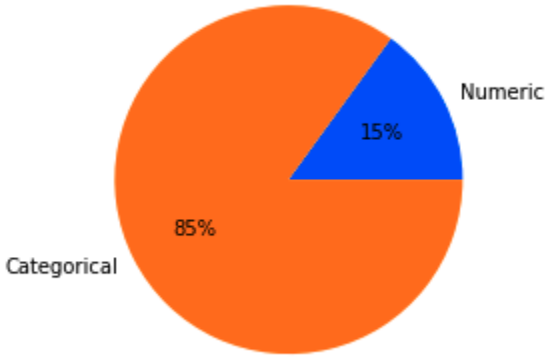
Customer churn models address a classification issue. We examine past user activity and examine their related features and use models to predict the probability that a customer will churn or be retained.

If we can predict customer segments in danger of churning then we can take proactive measures to try and retain customer segments with a high probability of churn.

Insights could assist in designing an intervention model to consider how the level of intervention could affect the churn percentages and customer lifetime value. Additionally, implementing effective experimentation across multiple customer segments for reducing churn and promoting retention.

Data Wrangling:

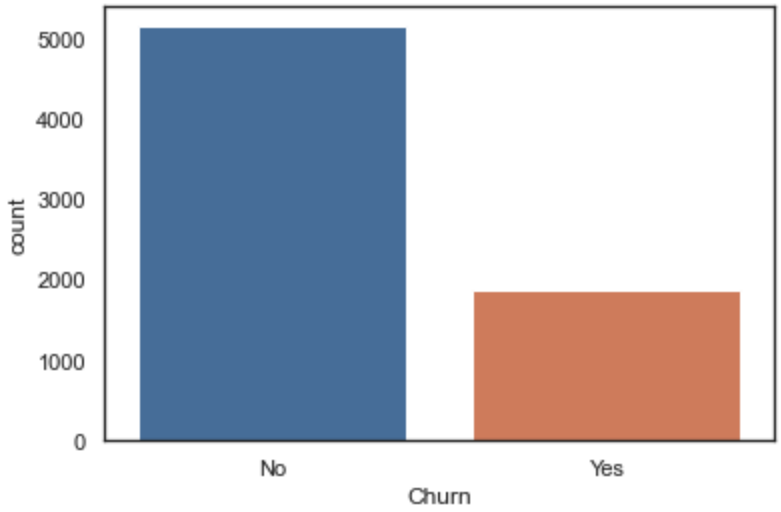
There was not much data wrangling necessary with this dataset. There was no null values, duplicated data or ambiguous feature labels. A few data type needed to be changed. The dataset initially consisted of 20 columns including the target variable “Churn” which was a binary feature consisting of “yes” and “no”. Most of the other features were categorical except for the variables “tenure”, “MonthsCharges”, and “TotalCharges”.



EDA:

After examining the features carefully, it made sense to drop the customerID column after no duplicates were detected. Then a count plot was made to examine the imbalance of the “Churn” feature. According to the data, 26.578% of the telecom customers churned over the six year period. How do we know it’s a six-year span? The churn attribute lists customer churn status for the month on which the data is collected, therefore from the 'tenure' column you can make an estimate. From df['tenure'].max() we get that maximum value is 72 months or 6 years. This is a very large percentage, even for the telecom industry which was 21% in 2020 according to Statista. To deal with this imbalance, we should use a method from the imbalance-learn library in the preprocessing stage.

```
The percentage of customer churn:
 No      73.42150170648463%
Yes     26.578498293515356%
Name: Churn, dtype: object
Total count of churners:
 No      5163
Yes     1869
Name: Churn, dtype: int64
```



Data Segmentation:

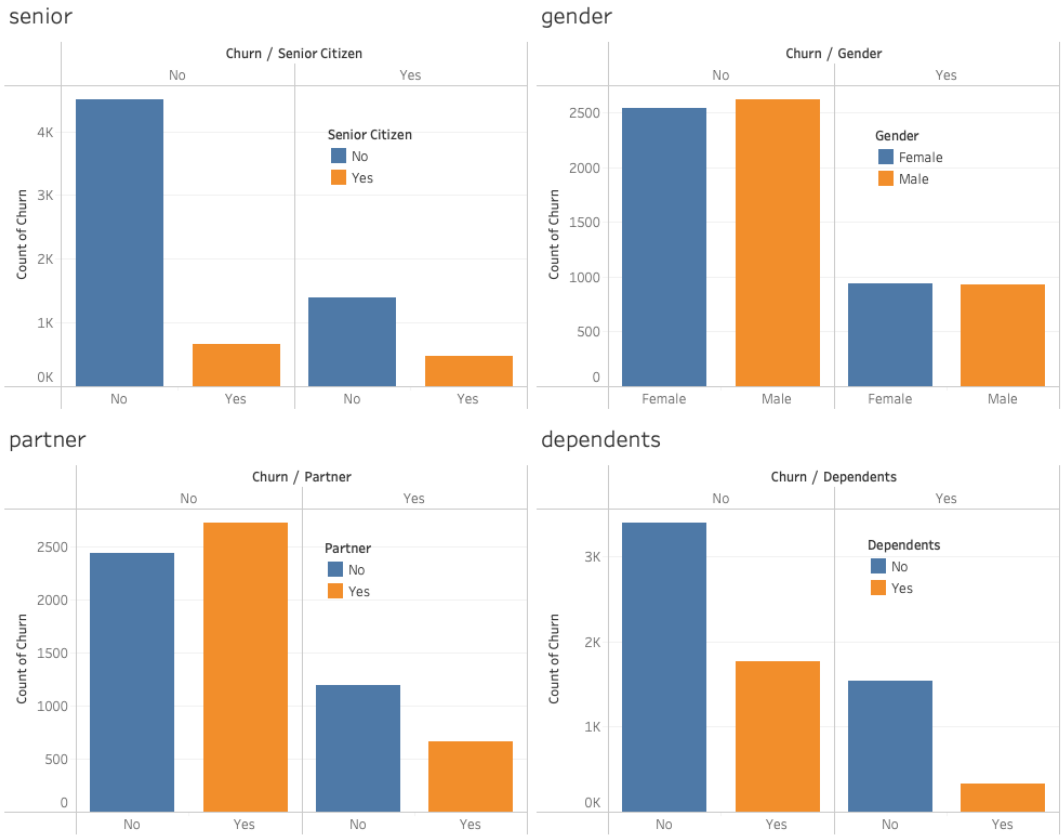
The data can be further segmented, beyond that just numerical and categorical. Features share different themes that can be compartmentalized into three bins regardless of their data type.

User Information - Senior Citizen, Gender, Partner, Dependents
These are all binary classifications consisting of “yes” or “no”.

Service Information - Streaming TV, Streaming Movies, Phone Service, Internet Service, Multiple Lines, Device Protection, Tech Support, Online Backup, Online Security

Account Information - Contract, Payment Method, Paperless Billing, Monthly Charges, Total Charges, Tenure

User Information

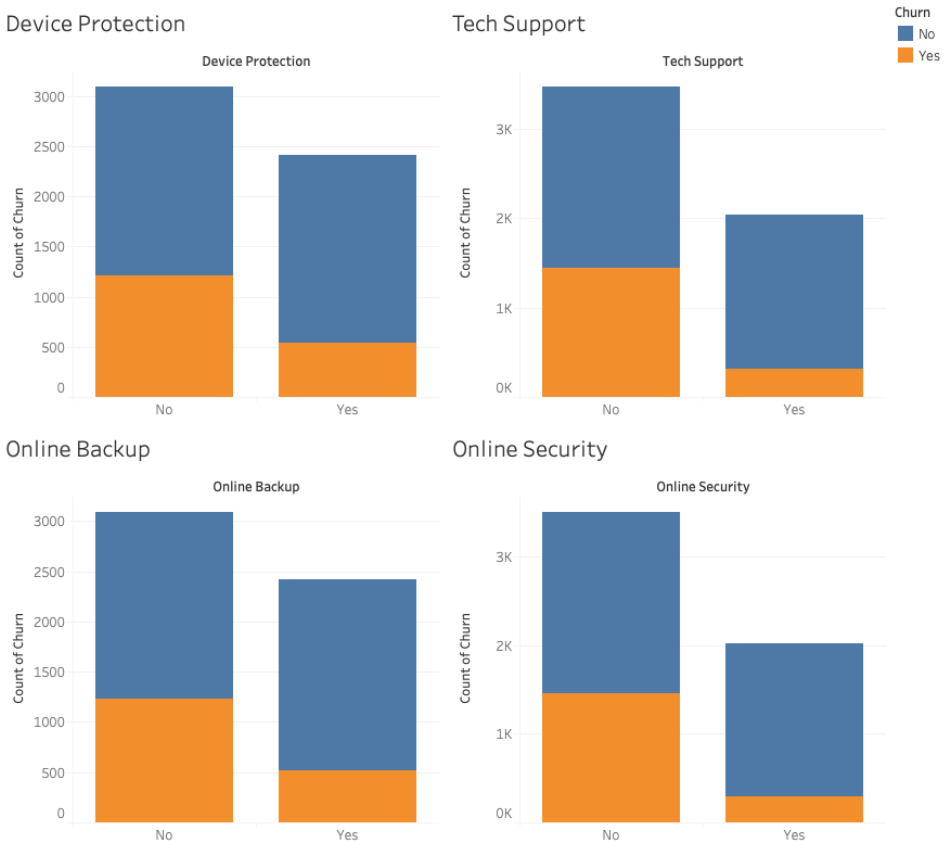


Takeaways:

- Senior citizens are less likely to churn
- Customer with no partner less likely to churn
- Customers with dependents are less likely to churn

The demographic features do not give us much statistical logic to go on other than young and single customers with no children are more likely to churn. This could be due to living more flexible lives which are more conducive to frequent change and flexibility.

Service Information



Takeaways:

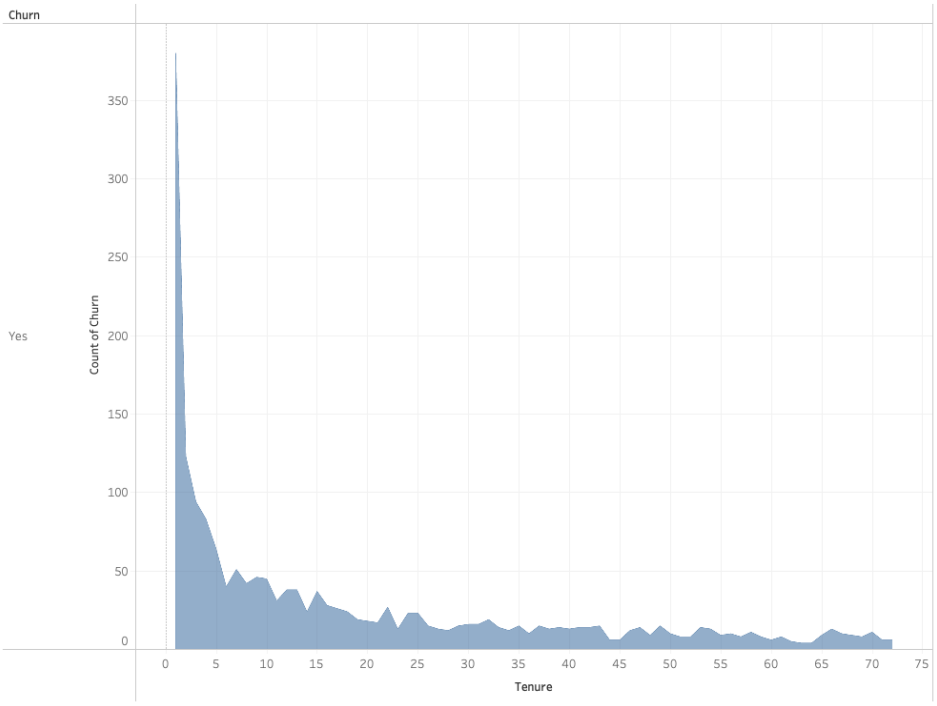
- Customers that do not subscribe to device protection, tech support, online backup, or online security are more likely to churn.

- Customer churn rate when not subscribed to
  Device protection - 39.14%
  Tech support - 41.65%
  Online backup - 39.94%
  Online security - 41.78%

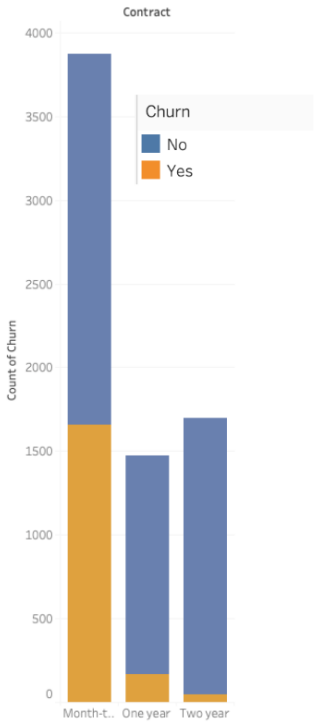
Customers using just the phone service can churn more easily than those customers that rely on a telecom business for many services.

Account Information

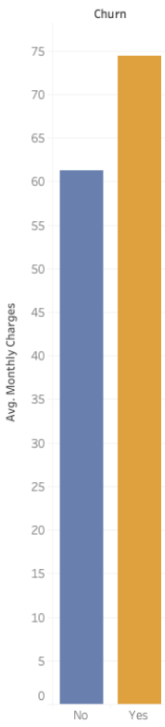
Tenure



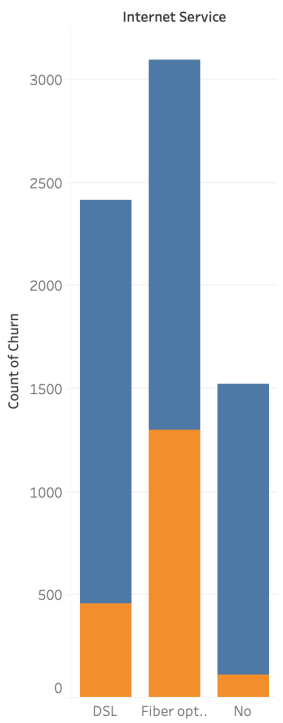
Contract



monthly



Internet Service



Takeaways:

- Customers with month-to-month contracts are more likely to churn.
- Customers with high monthly charges are more likely to churn.
- Customers who churn are likely to do so within two years.
- Customers that subscribed to internet services using fiber optic cable were more likely to churn.
    Fiber optic - 41.89% churn rate
    DSL - 19% churn rate
    No internet - 7.43% churn rate

There is less of a theme involved with respect to the account features. The takeaways are all logical individually. Customers not on long terms contracts are able to churn more easily. Customers with high bills are motivated to switch services. Customers who have not setup automatic payments are more likely to churn. Finally, customers are likely to churn within two years of service as it doesn't take long to make this decision.

Preprocessing:

As mentioned before, all customer churn problems have to deal with imbalanced data, and this dataset has a 27% churn rate. There were three options to use from RandomUnderSampler, RandomOverSampler, and SMOTE. All three are from the imbalanced learn library. I ran models with all three and the RandomUnderSampler had the best outcomes.

The following preprocessing steps were taken with the data:

- Separating the target variable and assigning it to y
- Fitting the target variable using LabelEncoder
- Standardization of numeric data fitting StandardScaler
- Creating dummy variables for categorical data

- Splitting the training and testing data (test size = 0.25)
- Fit the training data with RandomUderSampler

Modeling:

Before running a cross-validation score on a variety of default models, it is important to know what metrics define success with this particular problem.

What Measures be focused on?

With imbalanced data, it is best to avoid accuracy as the metric to judge a model. It is more important that we identify the churned customers than the retained ones. Recall is a good metric to score models if wanting to avoid False Negatives, or in other words, people that will probably be leaving but you fail to detect them. At the same time, the model needs to be decent in precision. Since it is a trade-off, we will plot the precision-recall curve and F1 Score. It is possible to adjust the threshold in the Precision-Recall curve and set a high recall by decreasing the threshold. When it comes to binary classification, many use ROC AUC, but recall is the best option since the important thing is to target customers that are leaving the company with marketing campaigns. Much like the ROC curve, The precision-recall curve is used for evaluating the performance of binary classification algorithms. The difference is it is often used in situations where classes are heavily imbalanced. The focus of this project is the attrition rate.

Attrition rate=number of customers who left/(total left + total stayed)
Retention rate=number of customers who stayed/(total left + total stayed)

Default Model F1 Scores:

Algorithm	F1 Score (Mean)	F1 Score (Standard Deviation)
Random Forest	**0.485**	**0.344**
Logistic Regression	**0.483**	**0.346**
Adaboost	**0.484**	**0.347**
Decision Tree	**0.472**	**0.334**
Gradient Boosting	**0.483**	**0.345**
SVC	**0.470**	**0.338**

Tuning the Random Forest Model:

Using RandomSearch to find the best parameters provide the following:

RandomForestClassifier(bootstrap=False, max_depth=30, max_features='sqrt', min_samples_leaf=2, min_samples_split=15, n_estimators=500)

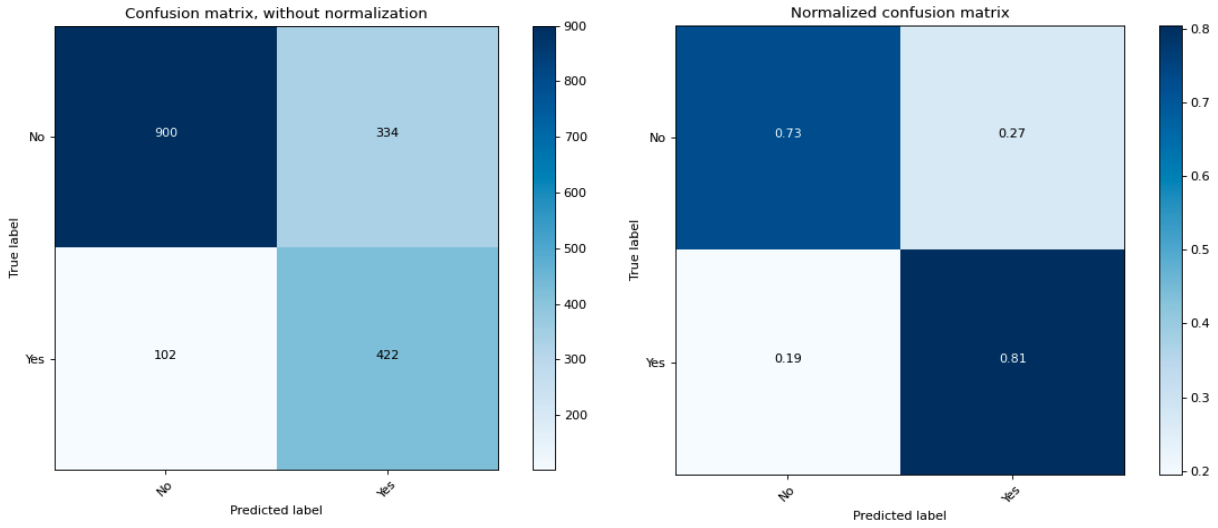
Classification Report:

Labels	Precision	Recall	F1
Retained Customers	**0.90**	**0.73**	**0.81**

Churned Customers	**0.56**	**0.81**	**0.66**
Accuracy			**0.75**
Macro Average	**0.73**	**0.77**	**0.73**
Weighted Average	**0.80**	**0.75**	**0.76**

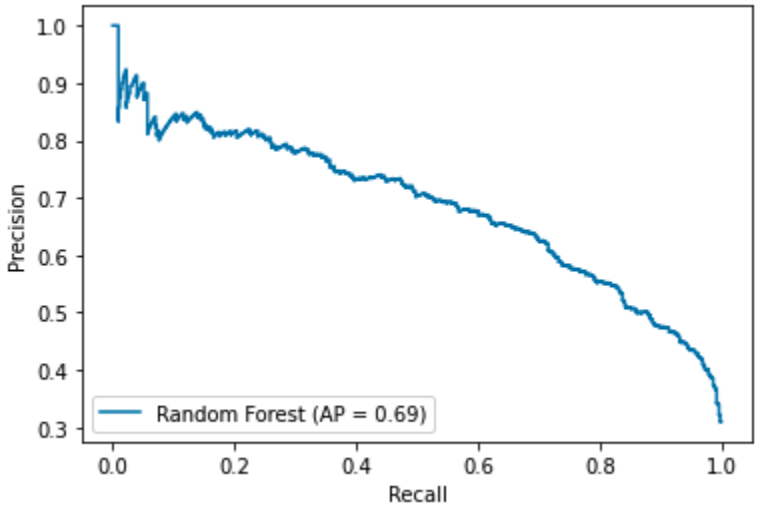
This is quite an improvement from the default Random Forest model. The F1 score has gone from 0.48 to 0.75 and the recall scores are promising.

Confusion Matrix:

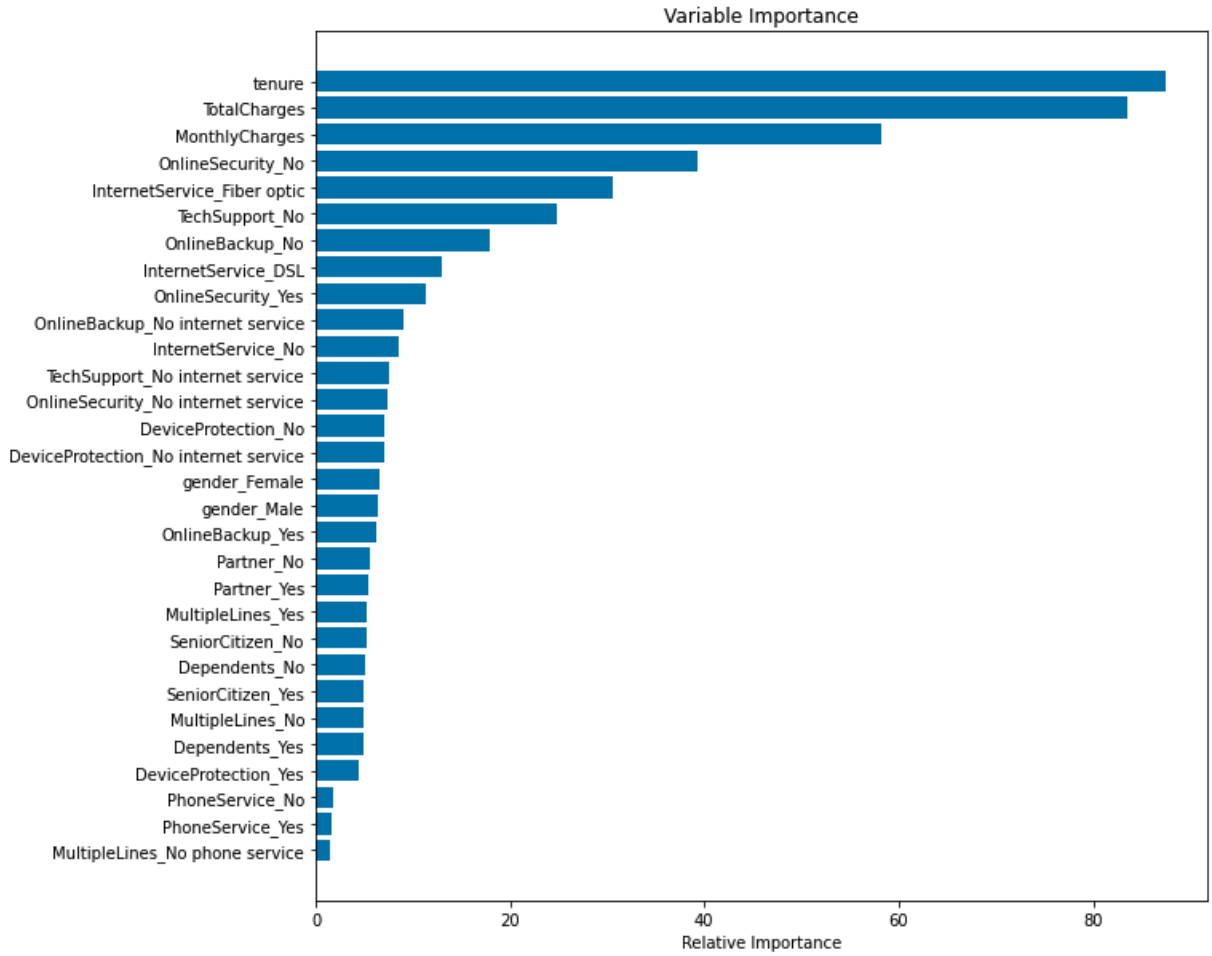


The false negative score is 19%, and the model identified 422 churn customers out of 524 actual churn customers.

Precision-Recall Curve:



Variable Importance:



Recommendations

- Filter customers that have subscribed for less than two years, who have high monthly billing statements, that use fiber optic internet services, and that do not subscribe to online security, tech support, and online backup.
- Reach out to prospective churners through email with surveys regarding customer satisfaction, the likelihood of churn, and possible competitors, follow up with incentives
- Target these customers by building customer loyalty through discounts on longer contracts.
- Promotions should be centered around deals on longer contracts, partial credits for monthly billing statements, and discounts on bundled subscription services.
- Reevaluate price optimization through the use of mathematical tools to determine how customers will respond to different prices for its products and services through different channels
- Reevaluate the quality of fiber optic internet service, or customer service regarding fiber optic internet service, as it could be causing churn

Constraints

A customer's likelihood of churning will be affected by other competitive factors such as the number of companies in the area that offer the same services, their pricing strategies, and promotions. This dataset doesn't have access to those features. An interesting aspect to explore would be a customer's personal experience/satisfaction with the company, particularly details regarding their customer service experience.