Part I: Research Question

1. Which customers are more likely to terminate services (Churn) with the telecommunication company?

2. The goal is to determine which variables indicate that a customer will terminate their service. This information is critical in reducing customer churn. Churn is the percentage of customers who terminated service over a given period. (Data Files and Associated Dictionary Files) The ability to predict which customers are at risk of canceling their service can allow the company to offer incentives for the customer to stay.

Part II: Method Justification

1. When using logistic regression, there are a few assumptions that analysis must make. The first is that the target variable outcome is binary. The target variable has two possible outcomes. Next, the observations are independent of each other. There are no duplicates. Third, multicollinearity does not exist between predictive variables. The predictive variables are not highly correlated. There are no extreme outliers. The sample size is sufficiently large. Finally, there should be a linear relationship between predictive variables and the logit of the target variable. (Zach, 2020)

2. Python was the language of choice for many reasons. It offers the versatility to do simple calculations as well as complex ones. It works across all platforms – Macs, Windows, Linux. Python runs faster than R. Most importantly, Python has an extensive library of packages like Numpy, Scikit-learn, and Mathplotlib. These packages allow Python to handle data analysis and machine learning efficiently. (Massaron & Boschetti, 2016)

3. Logistic regression is an appropriate technique for this question because the target variable is binary, with the outcomes of Churn being yes or no. Logistic regression is often used for this type of classification.

Part III: Data Preparation

1. There are several steps to preparing the data for analysis. First, unnecessary columns will be removed. The raw data contains 50 columns, many that are not relevant to answering the question. Next, look for missing, duplicate, or outlier values. Finally, enumerate all the categorical data using dummy variables.

2. The target variable for this question is Churn. No makes up about 70% of the churn values, with the other 30% being Yes.

The continuous numeric predictive variables include Income, Outage_sec_perweek, MonthlyCharge, Bandwidth_GB_Year. Income has a mean of about $39806 annually with a minimum of $348 and a maximum of $258900. Outage_sec_perweek is the average number of seconds per week that the customer's neighborhood experiences a service outage. Outages have a mean of 10 seconds with a minimum and maximum of 0.09 and 21 seconds. MonthlyCharge is the average monthly charge for the customer. MonthlyCharge ranges from $79.97 to $290.16, with an average of $172.62. Bandwidth_GB_Year is the total amount of data, in gigabytes, used by the customer in a year. The amount of bandwidth used per year has a minimum of 155 gigs and a maximum of 7158 gigs. The average usage is 3392 gigs per year.

The discrete numeric variables are Population, Children, Age, Email, Contacts, Yearly_equip_failure. Population is the number of people within a mile radius of a customer. The average is 9756 with a minimum of 0 and a maximum of 111850. Children are the number of children in the household at the time of signup. The maximum is 10 with a minimum of 0 and a mean of 2. Age is at the time of enrollment. The average age of customers is 53 years, with a minimum of 18 and a maximum of 89. Email counts the number of emails sent to the company in the last year. The count ranges from 1 to 23, with a mean of 12. Contacts are the number of times the customer contacted technical support. Contacts range from 0 to 7, with an average of less than 1. Yearly_equip_failure counts the number of times customer’s equipment has failed and replaced in the past year. The maximum is 6 with both mean and minimum of 0.

The categorical predictive variables are Area (rural, urban, suburban), Marital, Gender, Churn (ended service in the last month), Techie, Contract (month-to-month, one year, two year), Port_modem(portable modem), Tablet, InternetService(DSL, fiber optic, none), Phone, Multiple(multiple lines), OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod.

Items 1 through 8 are ordinal values. Customers’ responses on a survey of which factors are most important to them, with one being most important and eight being least important. The features included timely response, timely fixes, timely replacements, reliability, options, respectful response, courteous exchange, and evidence of active listening.

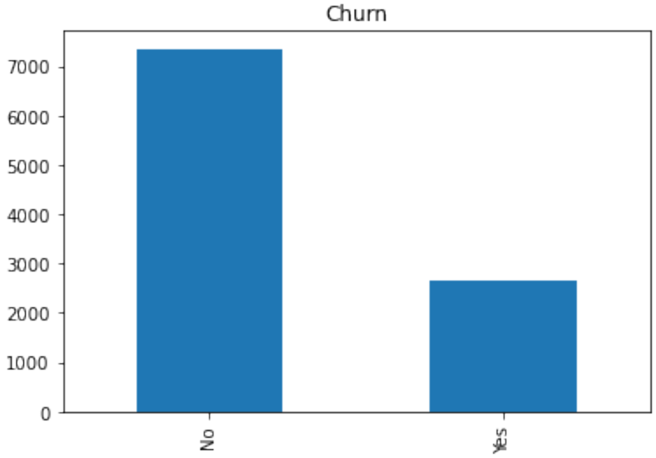
3. After loading the file, these steps used to prepare for analysis were:

   a. Use .head() to view the first five rows to ensure the data set loaded and get a glance at the information it holds.
   b. Use .shape() to see the number of columns and rows.
   c. Use .info() to see all the column names and the data types.
   d. Use .drop to remove CaseOrder, Customer_id, Interaction, UID, Lat, Lng, TimeZone, Job, City, County
   e. Look at zip code and state to decide if they should be dropped. Use .nunique() to see the total number of zip codes and states included in the data set. There are 8583 unique zip codes and 52 states. Including these variables will make the regression equation too unwieldy to glean any information. State and Zip were dropped.
   f. Check for missing values using .isnull().count().
   g. Check for duplications using .duplicated()
   h. Use .describe() to look for outliers in the numeric variables.
   i. Create dummy variable using .get_dummies on Area, Marital, Gender, Churn, Techie, Contract, Port_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV,

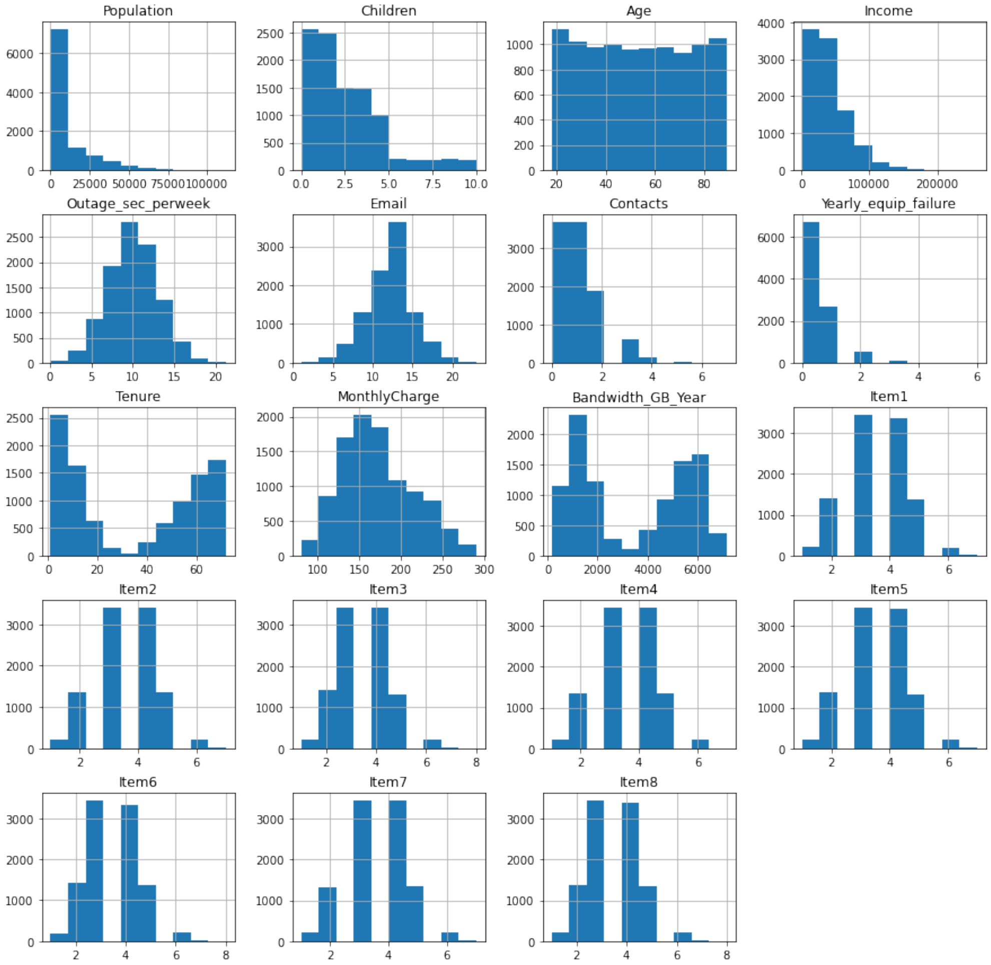
StreamingMovies, PaperlessBilling, PaymentMethod. Remove the No columns. (Langford, 2017)

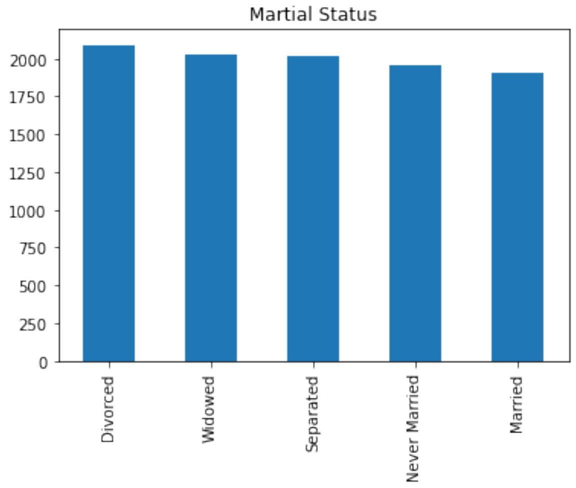
4.

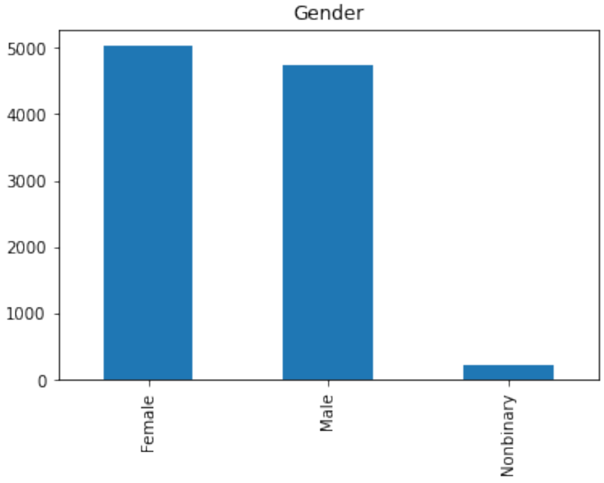
Target Variable: Churn

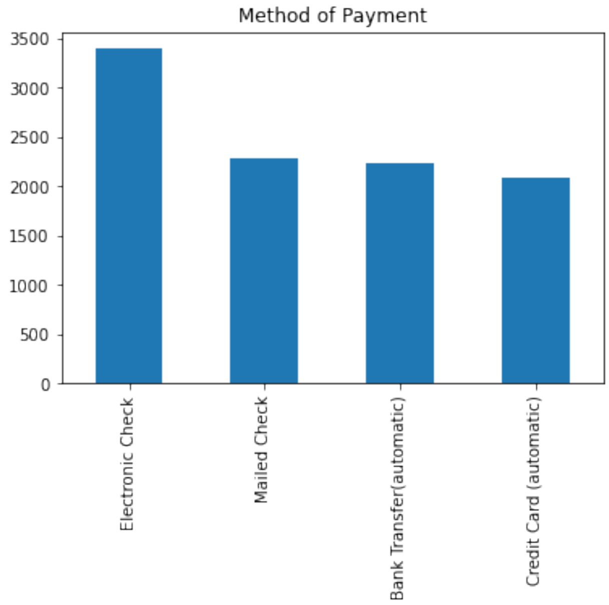


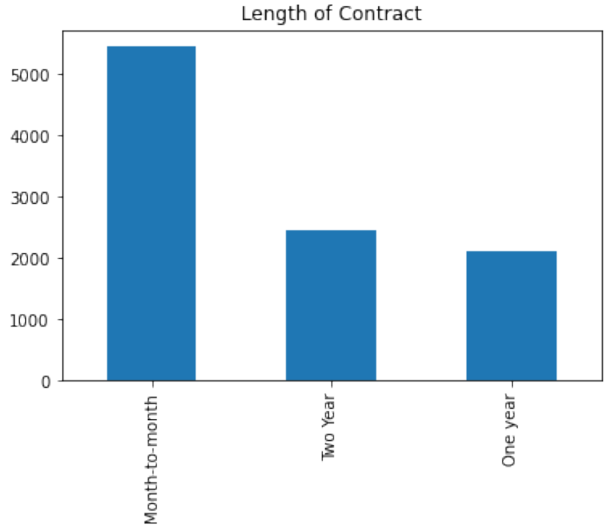
Univariate Visualizations

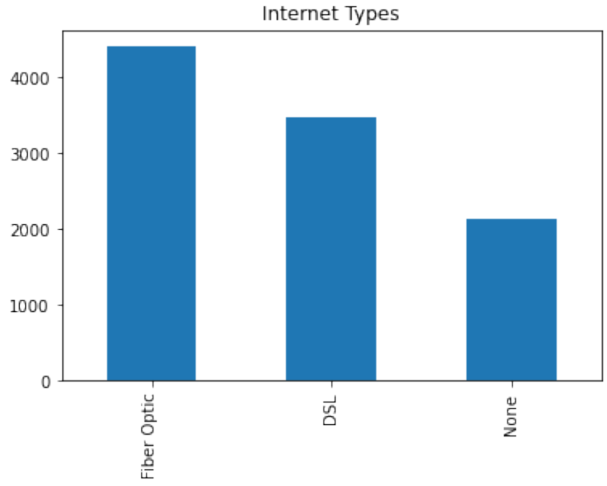


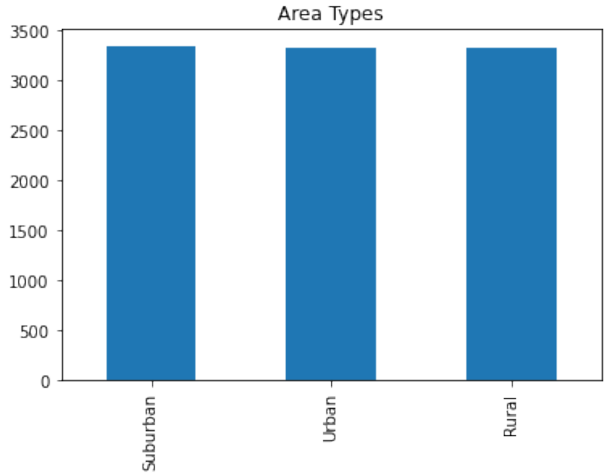

Martial Status


Gender

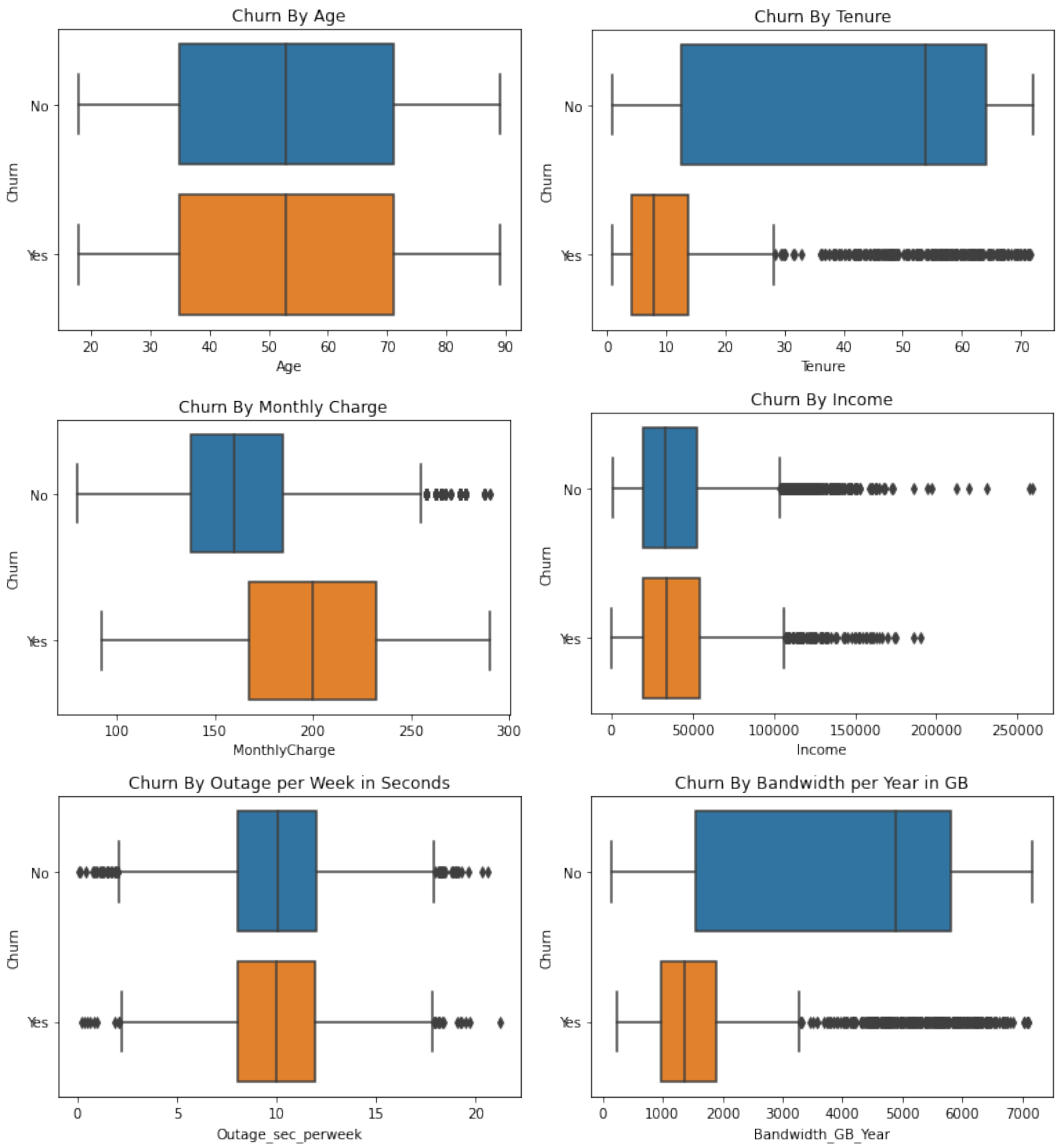

Method of Payment


Length of Contract


Internet Types


Area Types

Bivariate Visualizations



Churn By Age

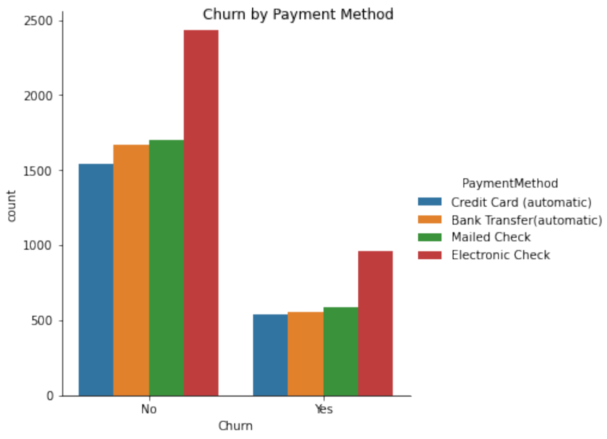
Churn By Tenure

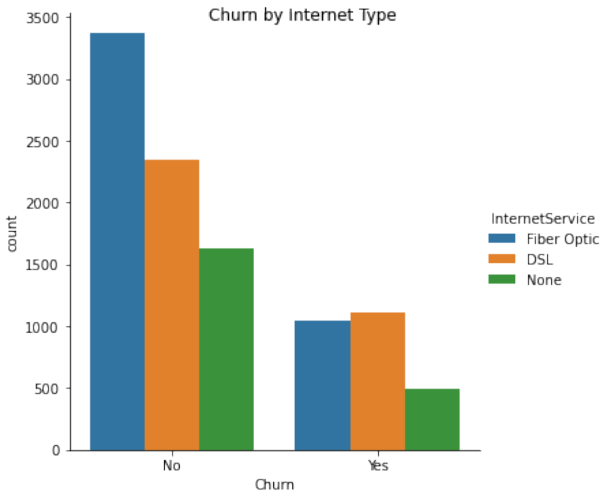
Churn By Monthly Charge

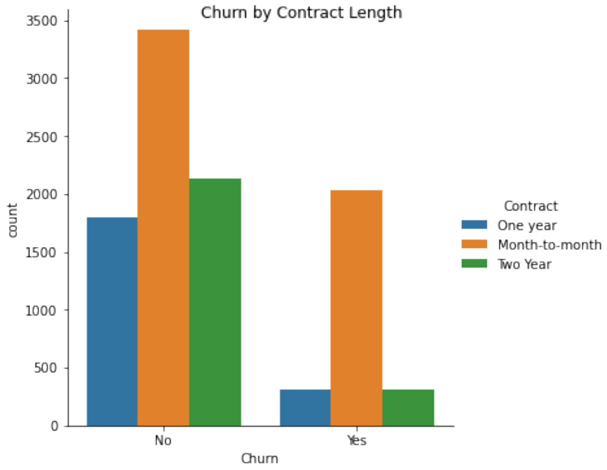
Churn By Income

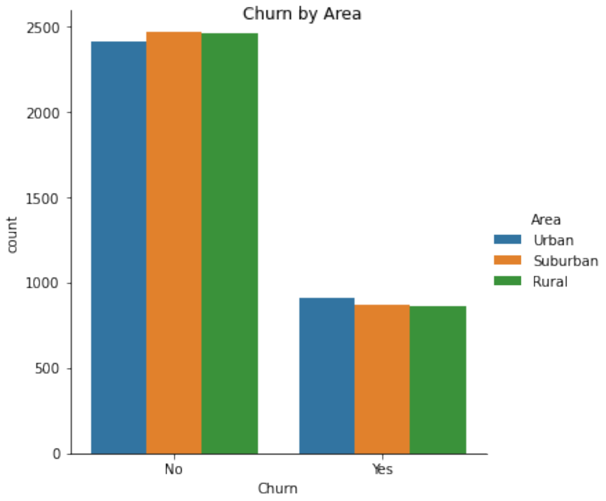
Churn By Outage per Week in Seconds

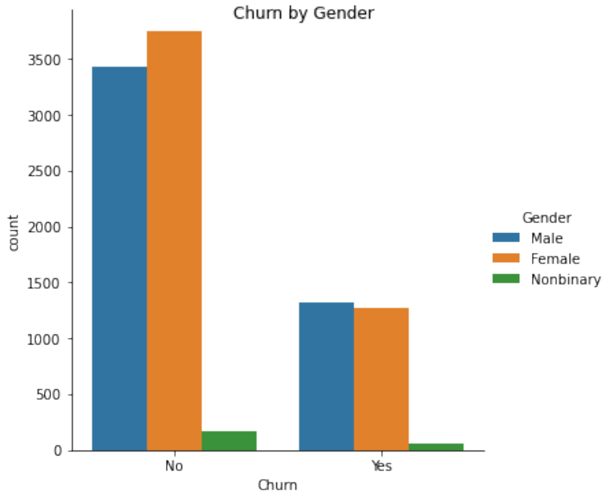
Churn By Bandwidth per Year in GB

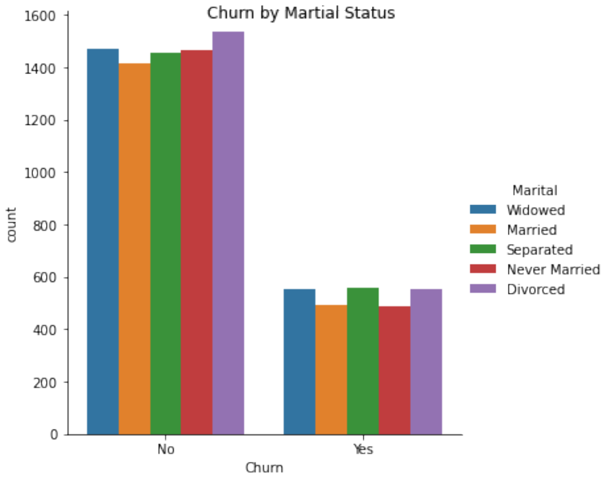










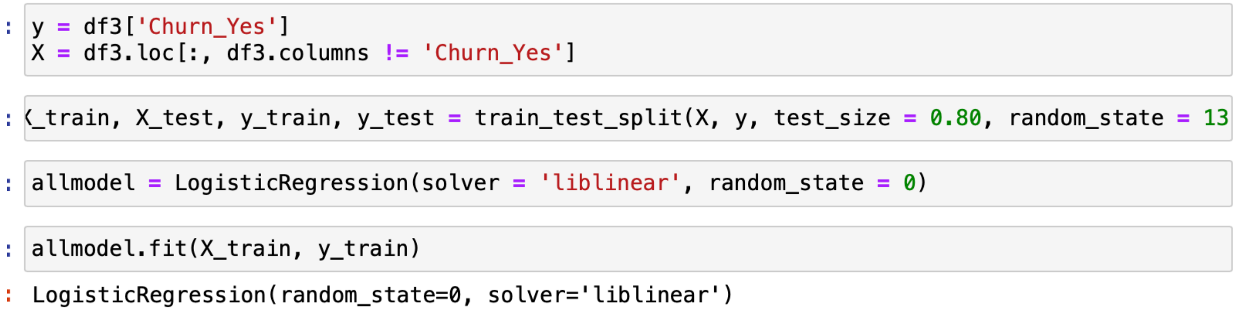


5. Copy of prepared data attached as logistic_prepared_churn.csv

Part IV: Model Comparison and Analysis

Model Comparison

1. The initial model was built with the following code:



```
y = df3['Churn_Yes']
X = df3.loc[:, df3.columns != 'Churn_Yes']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.80, random_state = 13
```

```
allmodel = LogisticRegression(solver = 'liblinear', random_state = 0)
```

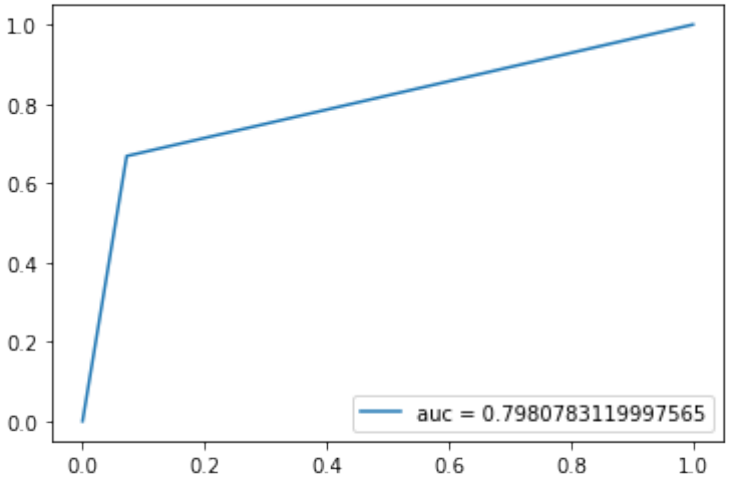
```
allmodel.fit(X_train, y_train)
```

```
LogisticRegression(random_state=0, solver='liblinear')
```

The model was scored using accuracy, precision, and recall. Accuracy measures the percent correct out of all the data points. Precision measures the percent of Churn's that are true Churn's. Recall is the percent of true Churn's to all Churn's. They are as follows:

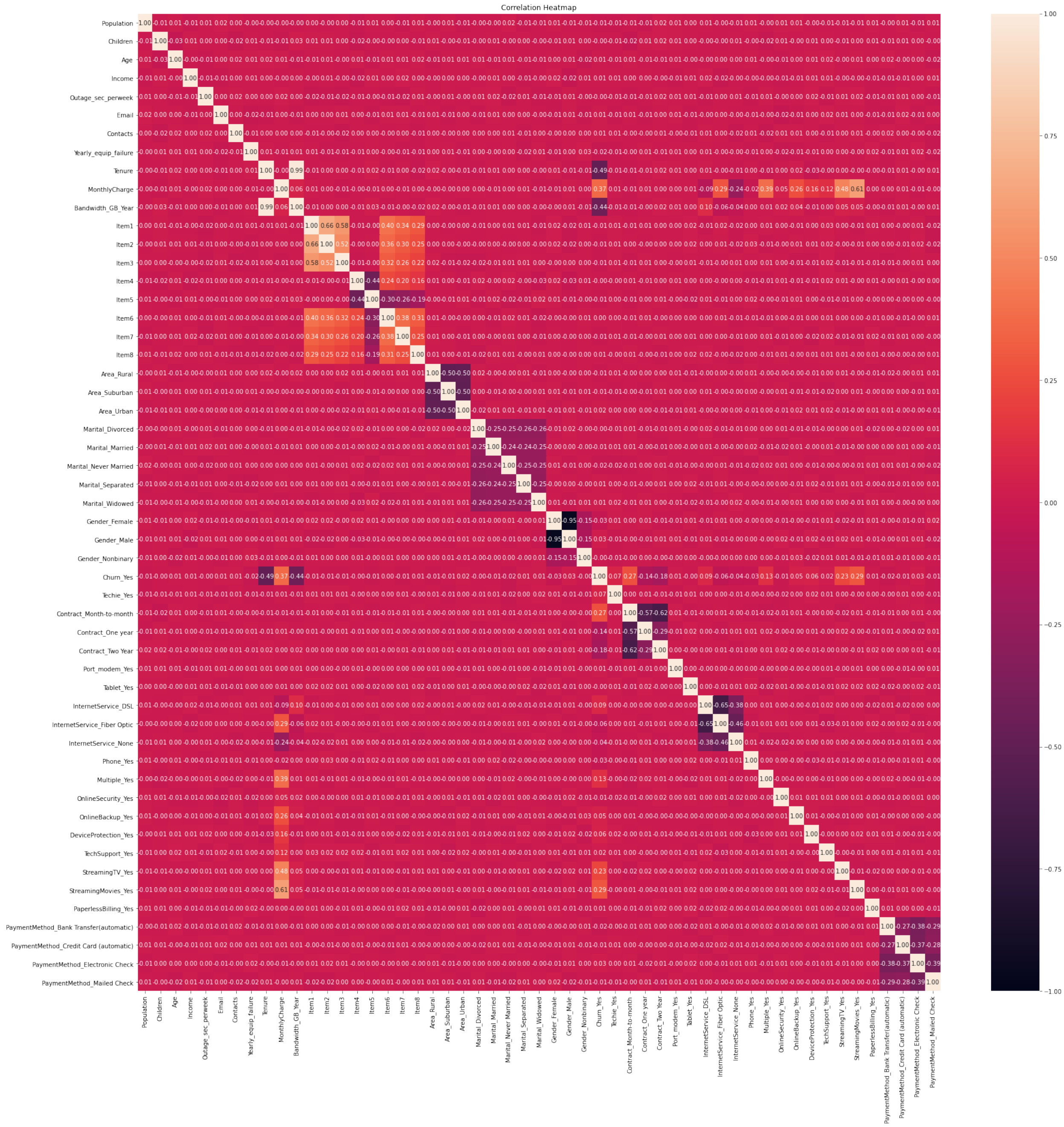
Accuracy: 0.859
Precision: 0.769
Recall: 0.668

A ROC (Receiver Operation Characteristic) curve was also created to measure the true positive rate versus the false positive rate. The score for this model was .798. (Navlanu, 2019)

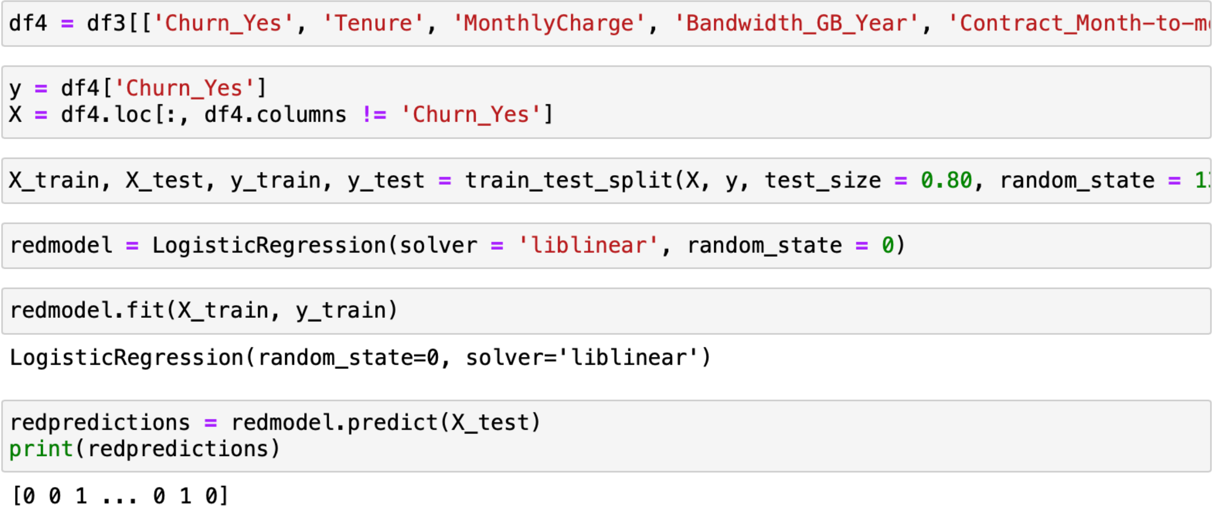


2. To create the reduced model, correlation was used. A heat map was created to find which predictive variables have the highest correlation values with Churn_Yes. It was challenging to read all the values in the Churn_Yes row. A list was created of the correlation values, and the most correlated variables were chosen. The correlation fell between the absolute value of 0.23 and 0.48. These variables were: Tenure, MonthlyCharge, Bandwidth_GB_Year, Contract_Month-to-month, StreamingMovies_Yes, StreamingTV_Yes.

```
Tenure                                      -0.485475
MonthlyCharge                                0.372938
Bandwidth_GB_Year                           -0.441669
Contract_Month-to-month                      0.267653
StreamingTV_Yes                              0.230151
StreamingMovies_Yes                          0.289262
```



3. The reduced model was created using the following code:



```
df4 = df3[['Churn_Yes', 'Tenure', 'MonthlyCharge', 'Bandwidth_GB_Year', 'Contract_Month-to-m
```

```
y = df4['Churn_Yes']
X = df4.loc[:, df4.columns != 'Churn_Yes']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.80, random_state = 1
```

```
redmodel = LogisticRegression(solver = 'liblinear', random_state = 0)
```

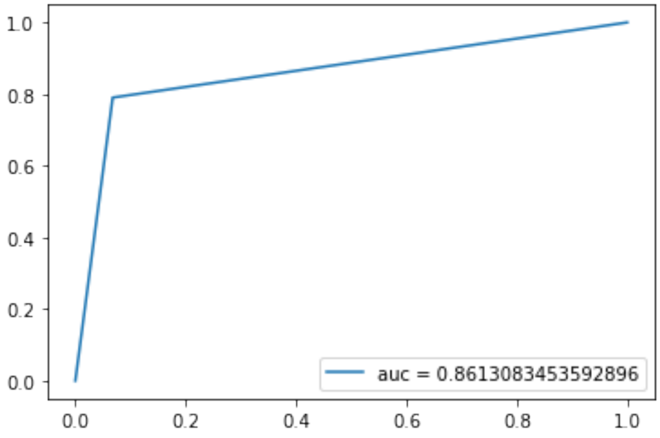
```
redmodel.fit(X_train, y_train)
```

```
LogisticRegression(random_state=0, solver='liblinear')
```

```
redpredictions = redmodel.predict(X_test)
print(redpredictions)
```

```
[0 0 1 ... 0 1 0]
```

The reduced model's scores are all an improvement over the previous model. The accuracy was 0.895, a gain of 3.6%. The precision is 0.808, an improvement of 3.9%. Recall had the most significant improvement going from 0.668 to 0.790, an increase of 12.2%. The ROC curve scored 0.861, an increase of 6.3%.



Analysis

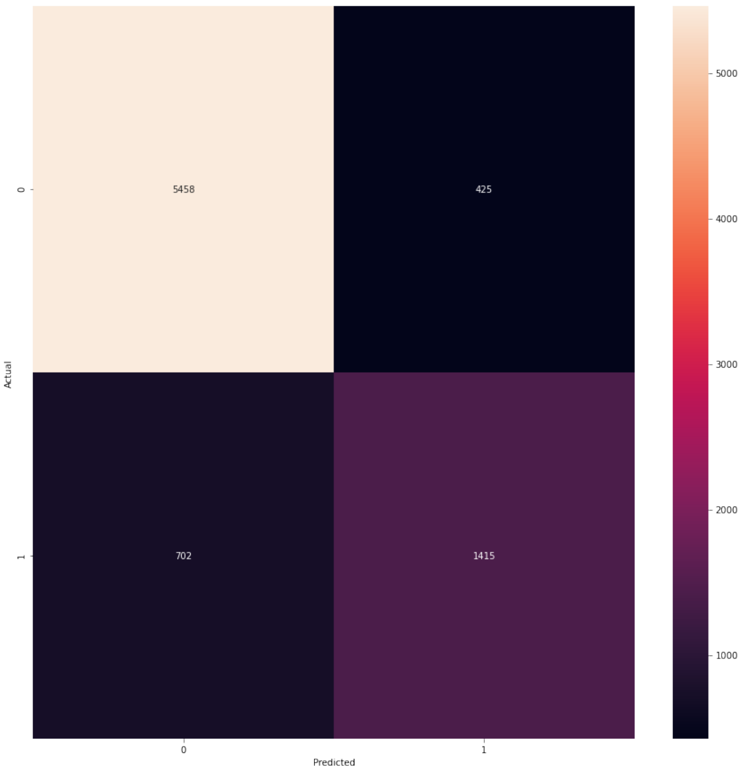
1. The focus of this logistic regression was Churn. Churn tracks if the customer discontinued their services within the last month. A model using all the given predictor variables was built using sklearn's LogisticRegression. The model had the following scores: Accuracy 0.859, Precision 0.769, Recall 0.668, and ROC 0.798.

Correlation was used to reduce the number of predictor variables. A heatmap and printed list of correlations were used to find the variables with the highest correlation values with Churn_Yes. These variables were Tenure, MonthlyCharge, Bandwidth_GB_Year, Contract_Month-to-month, StreamingMovies_Yes, and StreamingTV_Yes. This method was used because one of the assumptions of logistic regressions is that there is a linear relationship between the predictor and target variables. Using the variables with the strongest correlations helps to reinforce this assumption.

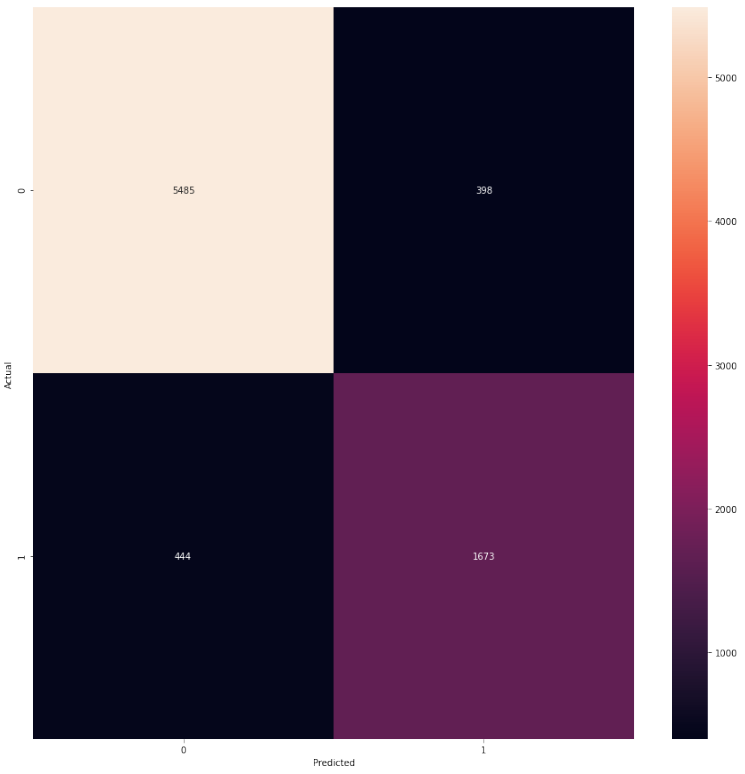
A second model was built using the variables mentioned above using the same coding. This model had the following scores: Accuracy 0.895 (an improvement of 3.6%), Precision 0.808 (an improvement of 3.9%), Recall 0.790 (an increase of 12.2%), ROC 0.861 (an increase of 6.3%).

2. Confusion matrices were crested for both models.

All Variable Model



Reduced Variable Model



When comparing the two models, there is an improvement in all areas. The largest is in type II errors. These are people who canceled service but were predicted to continue. The reduced model correctly classified 285 more people than the model using all variables.

3. Code see attached file.

Part V: Data Summary and Implications

1. The equation for the reduced model is:

y = 2.377(Contract_Month-to-month) + 1.663(StreamingMovies_Yes) + 1.454(StreamingTV_Yes) – 0.2719(Tenure) + 0.0078(MonthlyCharge) + 0.0022(Bandwidth_GB_Year)

There are both positive and negative coefficients in the equation for the reduced model. A positive coefficient indicates a variable that increases the likelihood that a customer will churn and stop services next month. Five variables fall into this category. The largest contributor to churn is having a month-to-month contract with a rate of 2.377. StreamingMovies_Yes is second with 1.663, followed closely with Streaming_TV with 1.454. MonthlyCharge and Bandwidth_GB_Year contributes a minimal amount to churn with 0.0078 and 0.0022. The only variable that negatively impacts Churn is tenure with 0.2719.

There are some limitations to this analysis. The provided data set represents a small portion of the total population. A quick Google search showed that three of the largest telecommunications companies have 20 million customers each. The data set with 10,000 records is not very large in comparison. The time frame is also limited. The needs of customers quickly change as technology changes. Looking at this narrow timeframe does not show an accurate picture of what is happening over time.

2. Based on this model, it is recommended that the company try to get customers to sign a contract rather than stay month-to-month. This change has the greatest impact on the churn rate. Also, the company should see why their streaming services are increasing the churn rate. Asking for customer feedback could lead to changes that reduce this influence on Churn.

Works Cited

Data Files and Associated Dictionary Files. (n.d.). Retrieved May 2021, from Western Governors University: https://access.wgu.edu/ASP3/aap/content/g9rke9s0rlc9ejd92md0.html

Langford, R. (2017, March 24). *The Dummy's Guide to Creating Dummy Variables*. Retrieved April 2021, from towards data science: https://towardsdatascience.com/the-dummys-guide-to-creating-dummy-variables-f21faddb1d40

Massaron, L., & Boschetti, A. (2016). *Regression analysis with Python.* Packt Publishing.

Navlanu, A. (2019, December 16). *datacamp*. Retrieved May 2021, from Understanding Logistic Regression in Python: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

Zach. (2020, October 13). *The 6 Assumptions of Logistic Regression*. Retrieved May 2021, from Statology: https://www.statology.org/assumptions-of-logistic-regression/