Telecom Churn Analysis

Javier Lopez

D208, Logistic Regression Modeling

Table of Contents

Part I. Research Question

A1. Research Question

What variables impact customer churn?

A2. Goals

Our goal is to understand the relationship between Churn and the explanatory variables. We seek to understand these variables deeply to create a model that will help us predict whether a customer is at risk of churn with some measure of confidence, using only a fraction of the explanatory variables available. The predictions will help stakeholders make business decisions regarding customer retention.

Part II. Method Justification

B1. Assumptions of a Logistic Regression Model

Logistic regression models make several assumptions. The first assumption is a binary outcome. The model assumes the target variable will have a binary outcome of zero or one. The model also assumes that observations are independent of each other. Furthermore, the multiple linear regression model assumes no high correlation between the explanatory variables, also known as multicollinearity. Lastly, there is the assumption of linearity of logit. The assumption is that the relationship between the

explanatory variables and the log-odds, the natural logarithm of the odds of an event occurring, of the dependent variable is linear.

B2. Benefits of Using Python

Python is a versatile programming language that offers many benefits through data analytics. Python offers benefits in data preprocessing through multiple libraries, including Pandas, Numpy, Statsmodels, Sklearn, and Scikit-learn. We were able to make use of Pandas and Numpy for data cleaning. Sklearn allowed us to encode categorical variables, essential for preparing data for logistic regression. Furthermore, Python plays a critical role in model development. Using the Scikit-learn library, we created stepwise regression models and quantified the models' performance using the F1 and accuracy scores.

B3. Justification: Logistic Regression

Logistic regression is an appropriate technique for answering the research question: "What variables impact customer churn?" because it allows us to examine and test the relationship between a dependent variable, customer churn, and multiple independent variables, factors that may impact customer churn. By using logistic regression, we can test the impact of the independent variables on the dependent variable while controlling the effects of other variables. This process helps us identify the most significant factors that affect bandwidth usage and how these factors relate to each other. Additionally, logistic regression aids us in the creation of predictive models that can help us estimate future risks of churn for customers based on their

characteristics. The most important factor in using logistic regression is the model's ability to handle categorical and continuous explanatory variables, which is useful, as a mix of continuous and categorical variables may impact churn.

Part III. Data Preparation

C1. Data Cleaning Goals and Steps

Our goal for data cleaning was to rename non-descriptive columns, impute missing values, remove duplicate values, and identify outliers. We began by implementing the info function and renaming items one through eight columns. We identified less meaningful columns that would not explain the target variable, Churn. We then created a copy of our data frame to preserve the original data and checked for missing values using Pandas' "isna" function.

We continued to utilize Pandas' "duplicated" function to check for duplicate values; there were none. Before checking for outliers, we calculated the z-scores for our data and, once identified, removed any observations with absolute values greater than three; those were considered outliers. We lost just 9.04% of our observations, leaving us with 90.96% of the data for training and testing. Lastly, we merged our original data frame, keeping only the observations from the index values in the z-scores data frame.

```
## View data types
df.info()
## Rename survey columns
```

```
df.rename({
    'Item1':'TimelyResponse',
    'Item2':'TimelyFixes',
    'Item3':'TimelyReplacements',
    'Item4':'Reliability',
    'Item5':'Options',
    'Item6':'RespectfulResponse',
    'Item7':'CourteousExchange',
    'Item8':'ActiveListening'
}, axis=1, inplace=True)
## View summary statistics
df.describe()
## Drop less meaningful columns
df = df.drop(['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng',
              'Area', 'TimeZone', 'Job', 'Marital', 'Gender', 'Email', 'Multiple', 'OnlineSecurity',
              'OnlineBackup', 'DeviceProtection', 'TechSupport', 'PaperlessBilling', 'PaymentMethod'
             ], axis=1)
## Create copy of dataframe
df1 = df.copy()
## Check for missing values
df1.isna().sum()
## Check for duplicate values
df1.duplicated().value_counts()
## Check for outliers
df1.describe()
## Separate object variables
df2 = pd.DataFrame([df1[col] for col in df1.columns if df1[col].dtype != 'object']).transpose()
df2.info()
```

```
## Normalize data and exclude outliers
df2 = df2[zscore(df2).abs() < 3]
## Count outliers
df2.isna().sum()
## Drop outlier values
df2.dropna(inplace=True)
df2.info()
memory usage: 1.3 MB
## Measure data loss
lost = ((len(df1) - len(df2))/len(df1))*100
remaining = 100 - lost
print('{}% of data lost\n{}% of data remains'.format(round(lost, 2), remaining))
## Combine dataframes
df = df.loc[df2.index]
df1 = df1.loc[df2.index]
## Reset index values
df = df.reset_index(drop=True)
df1 = df1.reset_index(drop=True)
df1.info()
```

C2. Summary Statistics

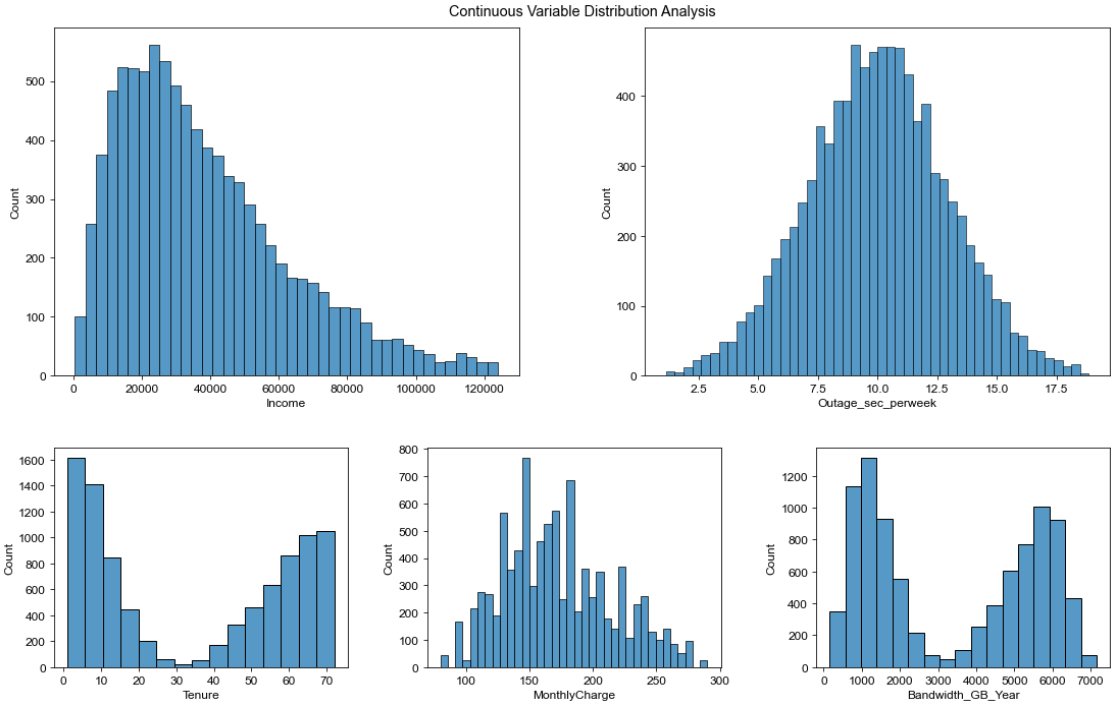
Our dependent variable, Churn, is a binary variable with 26.67% (2425 customers) churned and 73.33% (6671 customers) active. Our independent variables include a range of data points, including Population, Children, Age, Tenure, and many

others. The table below lists each independent variable's range, standard deviations, mean, median, and mode.

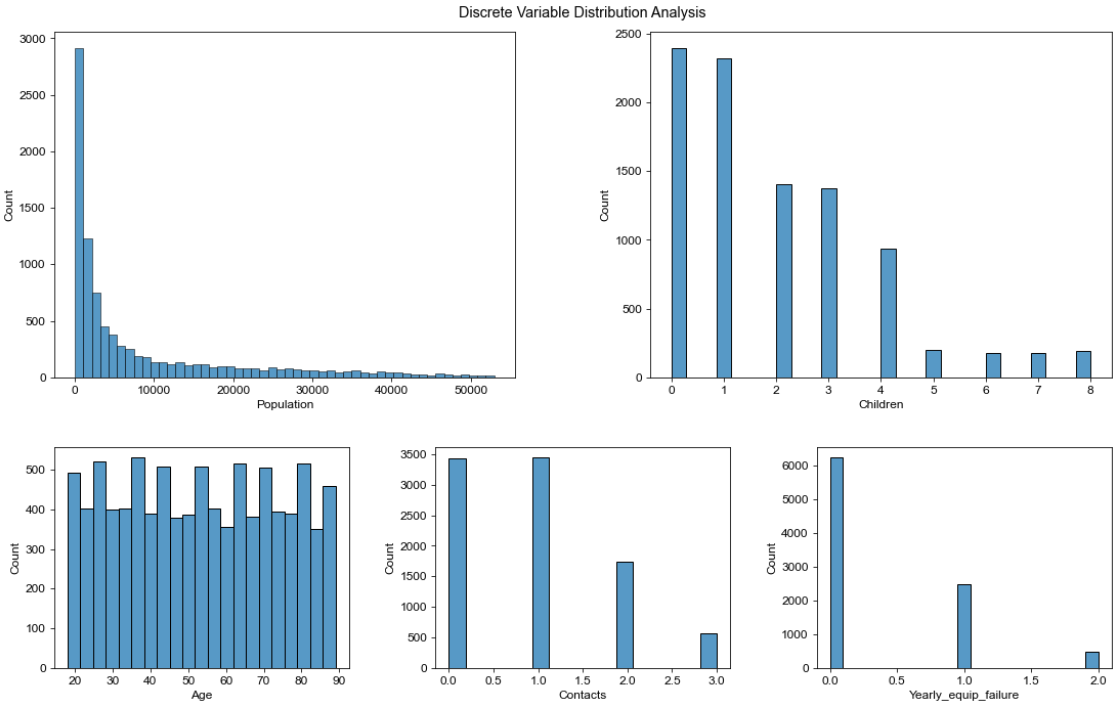
	min	max	std	mean	median	mode
Population	0.000000	52967.000000	11806.655958	8539.710642	2738.500000	0.000000
Children	0.000000	8.000000	1.895131	1.946020	1.000000	0.000000
Age	18.000000	89.000000	20.647938	53.092348	53.000000	75.000000
Income	348.670000	124025.100000	25097.606419	38315.115020	32759.635000	10530.090000
Outage_sec_perweek	1.144796	18.851730	2.924140	10.008946	10.019747	10.488750
Contacts	0.000000	3.000000	0.899321	0.941293	1.000000	1.000000
Yearly_equip_failure	0.000000	2.000000	0.582221	0.374780	0.000000	0.000000
Tenure	1.005104	71.999280	26.440792	34.435636	30.795855	55.449910
MonthlyCharge	79.978860	290.160419	43.002097	172.756644	169.937800	179.947600
Bandwidth_GB_Year	155.506715	7158.981530	2185.169628	3380.549302	3170.731427	155.506715
TimelyResponse	1.000000	6.000000	1.014185	3.474055	3.000000	3.000000
TimelyFixes	1.000000	6.000000	1.020008	3.493624	4.000000	4.000000
TimelyReplacements	1.000000	6.000000	1.015795	3.473505	3.000000	3.000000
Reliability	1.000000	6.000000	1.020320	3.491974	3.000000	3.000000
Options	1.000000	6.000000	1.015341	3.490765	3.000000	3.000000
RespectfulResponse	1.000000	6.000000	1.020559	3.484609	3.000000	3.000000
CourteousExchange	1.000000	6.000000	1.020243	3.499120	3.000000	3.000000
ActiveListening	1.000000	6.000000	1.018370	3.482630	3.000000	3.000000

C3. Distribution Visualizations

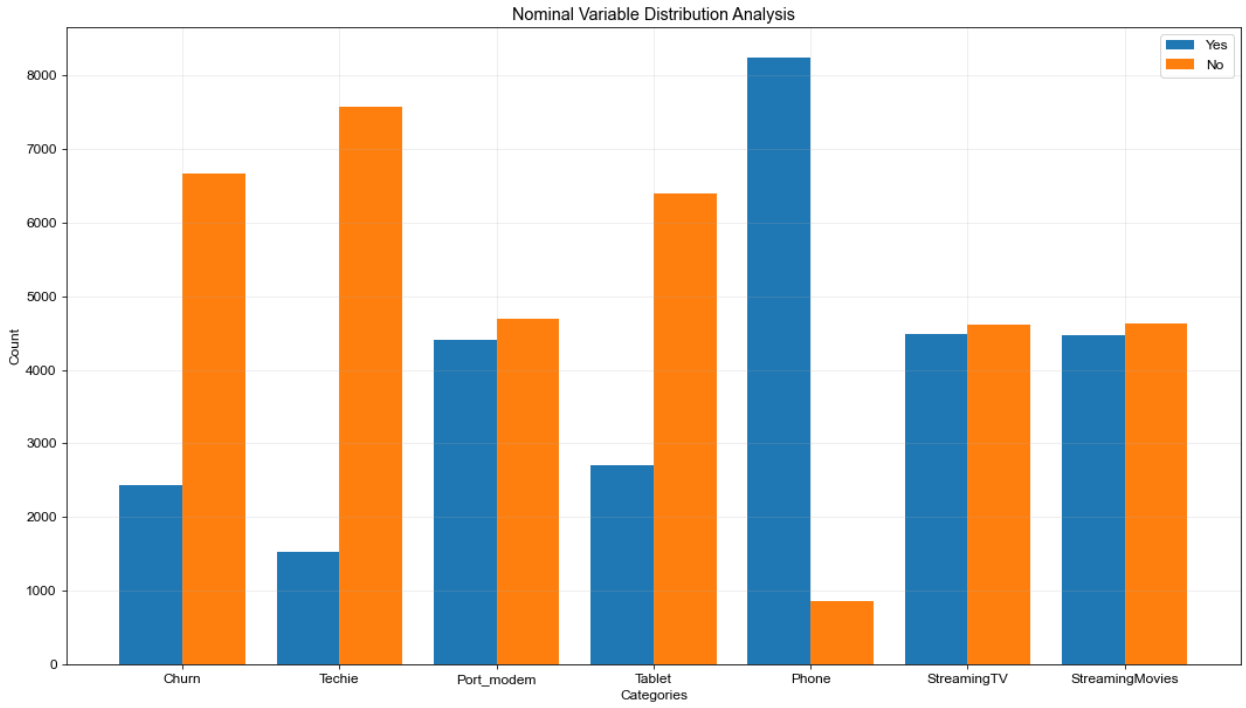
Univariate Visualizations of Continuous Variables

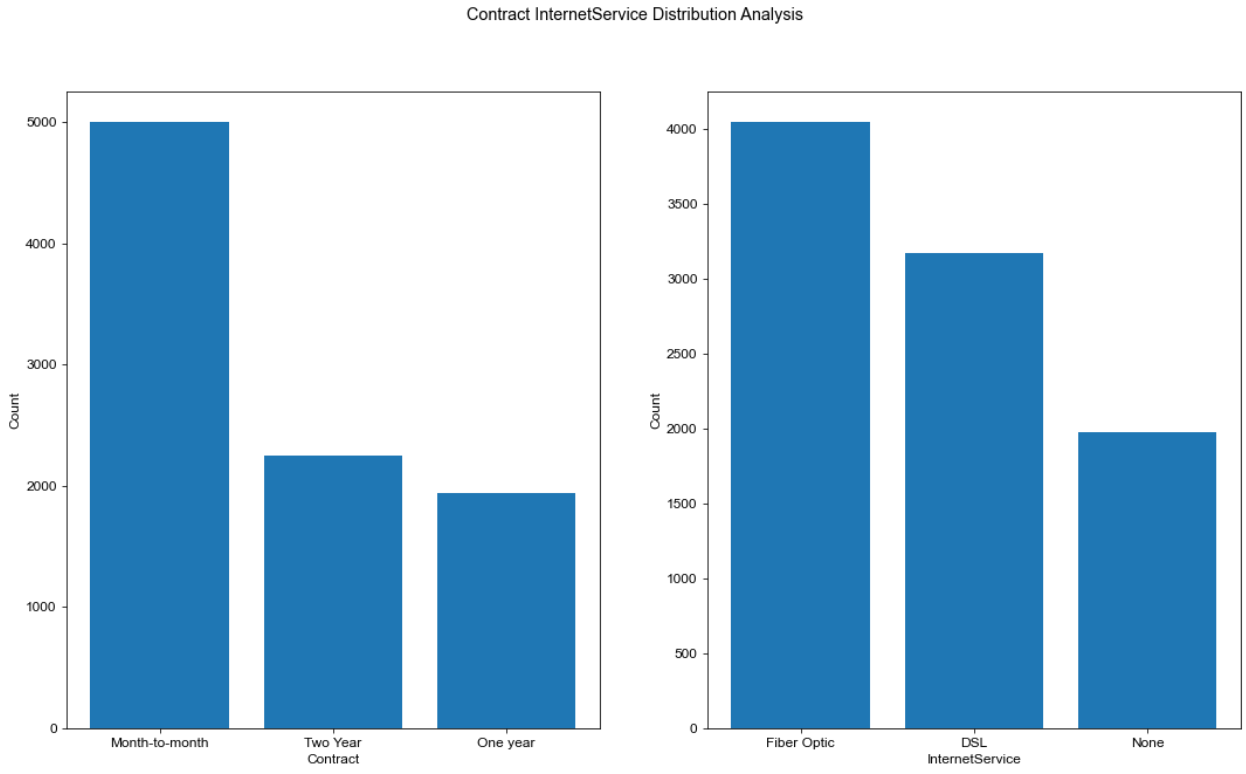


Univariate Visualizations of Discrete Variables

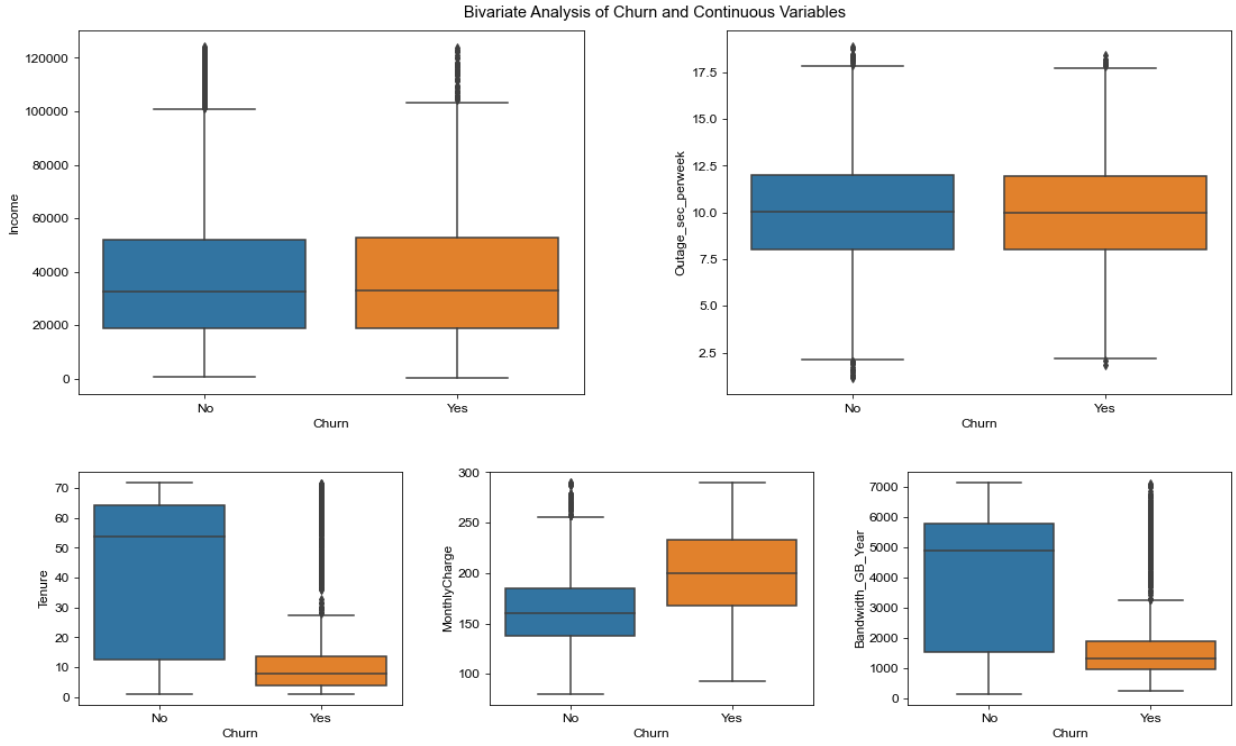


Univariate Visualizations of Nominal Variables

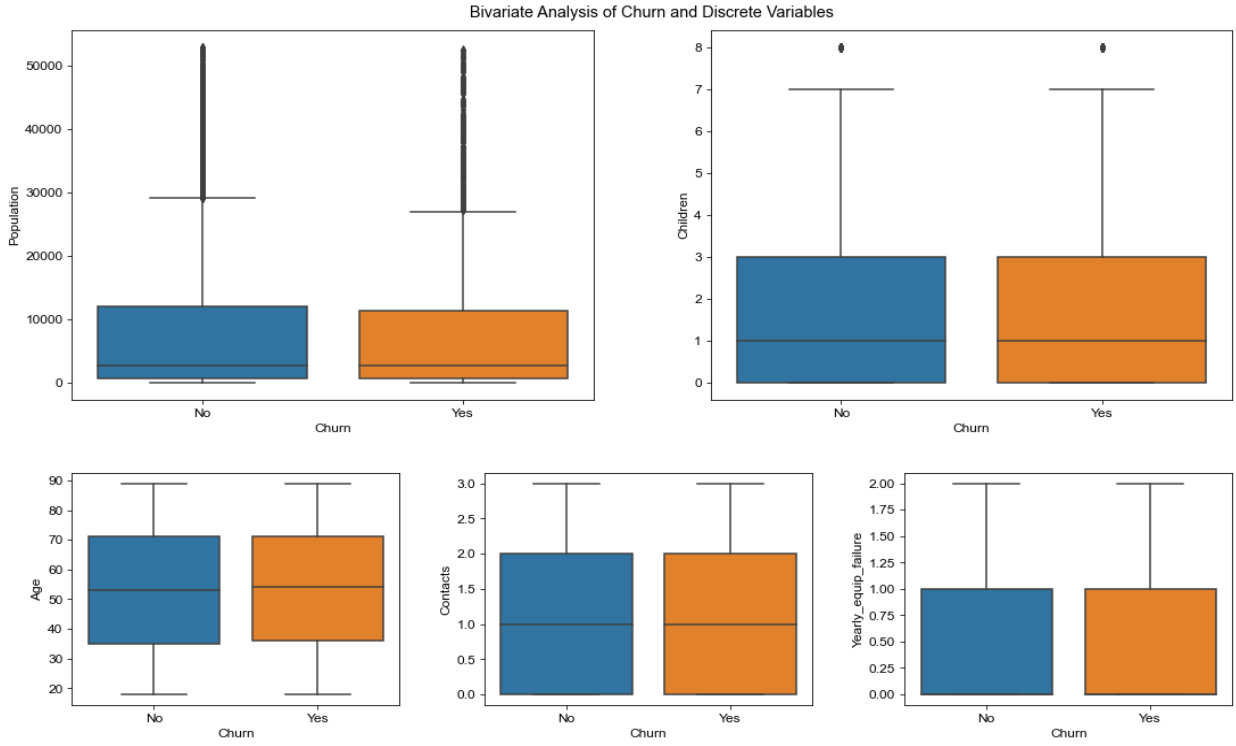




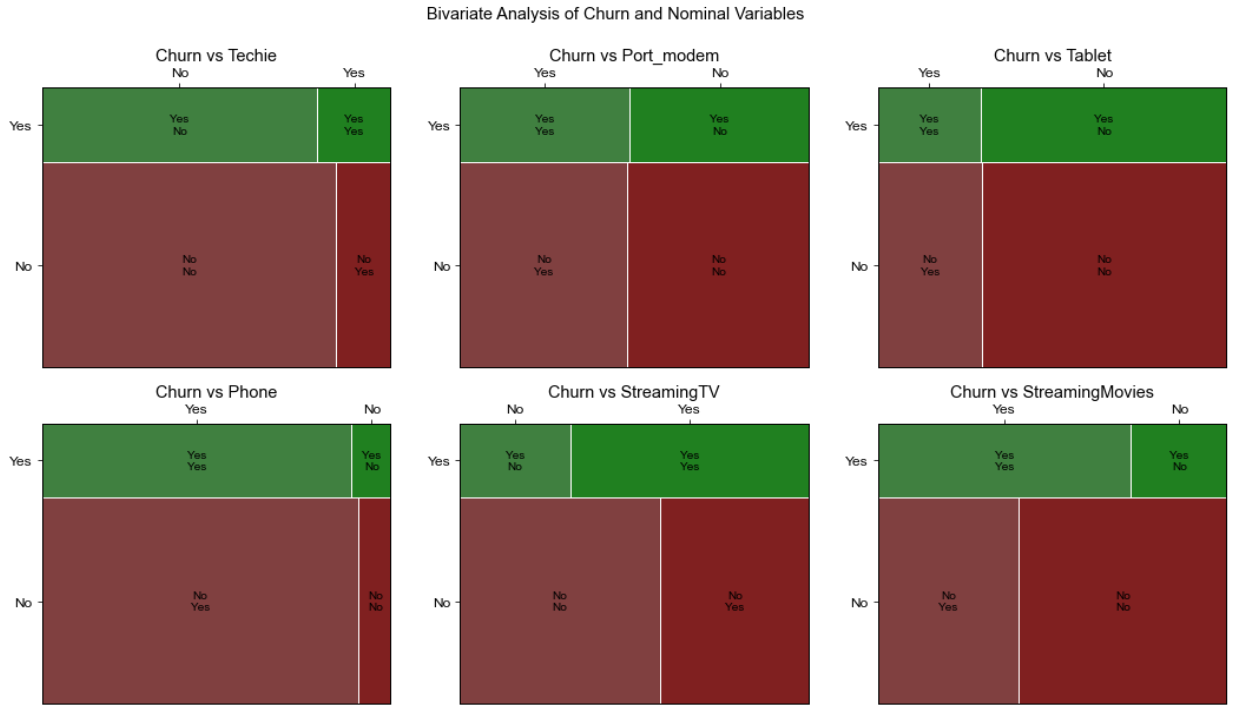
Bivariate Analysis of Continuous Variables



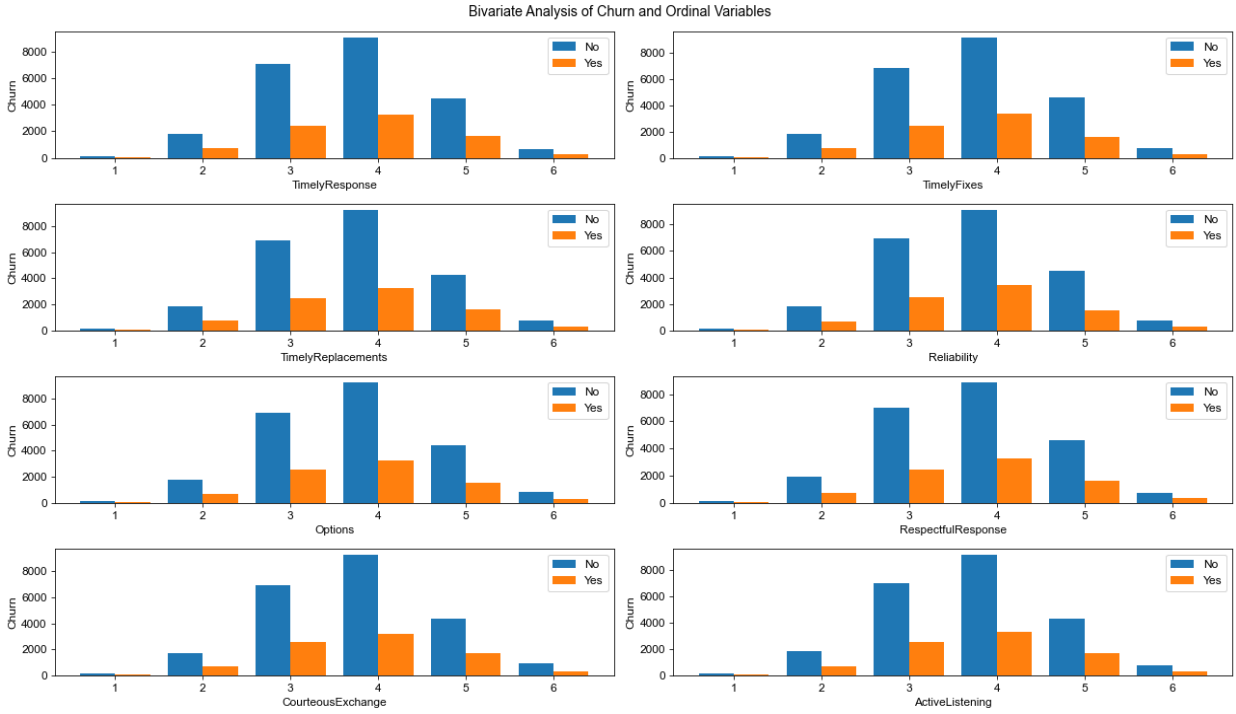
Bivariate Analysis of Discrete Variables



Bivariate Distribution of Nominal Variables



Bivariate Distribution of Ordinal Variables



C4. Data Transformation Goals and Steps

Our goals for the data transformation process were to encode all binary nominal variables with one and zero using the label encoder from Sklearn, encode ordinal variables with their respective rank, and create the necessary columns to explain the remainder of our data. We made two columns, DSL and FiberOptic, to describe the type of internet service customers were subscribed to. We used binary to encode the two columns and then transformed the InternetService columns to represent whether a customer subscribed to an Internet service. InternetService was encoded with one if the customer was signed up for DSL or fiber optic internet and zero if the customer was subscribed to neither internet service. We then encoded the contract variable with the respective years the customer signed.

```
## Create InternetDSL and InternetFiberOptic columns
dsl = []
fiber = []
for i in df1.InternetService:
    if i == 'DSL':
        dsl.append(1)
        fiber.append(0)
    elif i == 'Fiber Optic':
        dsl.append(0)
        fiber.append(1)
    else:
        dsl.append(0)
        fiber.append(0)
```

```
df1['InternetDSL'] = dsl

df1['InternetFiberOptic'] = fiber

df1.head()

## Encode InternetService column

internet_service = {'DSL':'Yes', 'Fiber Optic':'Yes', 'None':'No'}

df1.InternetService.replace(internet_service, inplace=True)

## Initiate label encoder

le = LabelEncoder()

## Encode variables

for col in df1.columns:

    if 'Yes' in df1[col].values:

        df1[col] = le.fit_transform(df1[col])

df1.info()

contract = {'Month-to-month':0, 'One year':1, 'Two Year':2}

df1.Contract.replace(contract, inplace=True)

df1.info()
```

C5. Prepared Data CSV

```
## Store clean data as CSV

df1.to_csv('churn_logistic_regression.csv')
```

Part IV. Model Comparison and Analysis

D1. Initial Logistic Regression Model

	coef
Population	-0.000004
Children	-0.026484
Age	0.001325
Income	-0.000003
Outage_sec_perweek	-0.195501
Contacts	-0.015508
Yearly_equip_failure	-0.008455
Techie	0.005448
Contract	-0.098930
Port_modem	-0.008253
Tablet	-0.006291
InternetService	-0.014986
Phone	-0.021932
StreamingTV	0.023062
StreamingMovies	0.029436
Tenure	-0.208926
MonthlyCharge	0.023637
Bandwidth_GB_Year	0.001626
TimelyResponse	-0.077138
TimelyFixes	-0.075699
TimelyReplacements	-0.080270
Reliability	-0.076059
Options	-0.076428
RespectfulResponse	-0.074785
CourteousExchange	-0.071998
ActiveListening	-0.074523
InternetDSL	0.024950
InternetFiberOptic	-0.039937

D2. Justification: Model Reduction Method

Our initial model fit our data well with a high accuracy score, but it contained an excessive amount of variables, twenty-eight variables, and a constant, to be exact. To reduce our model, we implemented the use of step-forward feature selection. We began by individually testing each feature in a logistic regression model and recording each model's F1 and accuracy scores. The F1 score is the harmonic mean of precision and

recall, two metrics that measure true positives and the sensitivity of our model's predictions.

We selected the independent variable with the highest correlation coefficient compared to churn and stored the variables, F1, and accuracy scores in a data frame. We then repeated the feature selection process but removed the previously selected variables from the data set. We kept the variables used in each model with their corresponding F1 and accuracy scores. Once we iterated through all possible independent variables, we sorted the results data frame and selected the model with the highest F1 and accuracy scores.

We then tested the features chosen for multicollinearity using the variance inflation factor formula from Sklearn. We removed the variable with the highest coefficient above five, and made our final model using the reduced features. The step-forward feature selection method and checking for multicollinearity would yield the best model possible using the least possible explanatory variables. Ultimately, our process proved successful, producing a model with a comparable F1 and accuracy score to the initial model and independent variables with little correlation.

D3. Reduced Logistic Regression Model

Step-Forward Feature Selection Scores and Variance Inflation Factors

Step 1

	f1	accuracy	features
0	0.689778	0.846526	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
1	0.685121	0.839930	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
2	0.683791	0.843008	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
3	0.678414	0.839490	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
4	0.676311	0.842568	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
5	0.675370	0.835972	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
6	0.675182	0.843448	[Tenure, Bandwidth_GB_Year, MonthlyCharge]
7	0.674868	0.837291	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
8	0.674545	0.842568	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
9	0.674256	0.836412	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
10	0.674216	0.835532	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
11	0.671341	0.835092	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
12	0.670742	0.834213	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
13	0.667838	0.833773	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
14	0.666085	0.831574	[Tenure, Bandwidth_GB_Year, MonthlyCharge, Opt...
15	0.633969	0.822779	[Tenure, Bandwidth_GB_Year]
16	0.455978	0.741865	[Tenure]

Step 2

	Variable	VIF
0	Tenure	1.000632
1	MonthlyCharge	1.000311
2	Options	1.000797
3	TimelyReplacements	1.000334
4	Age	1.000611
5	Contacts	1.000496

Reduced Model Summary

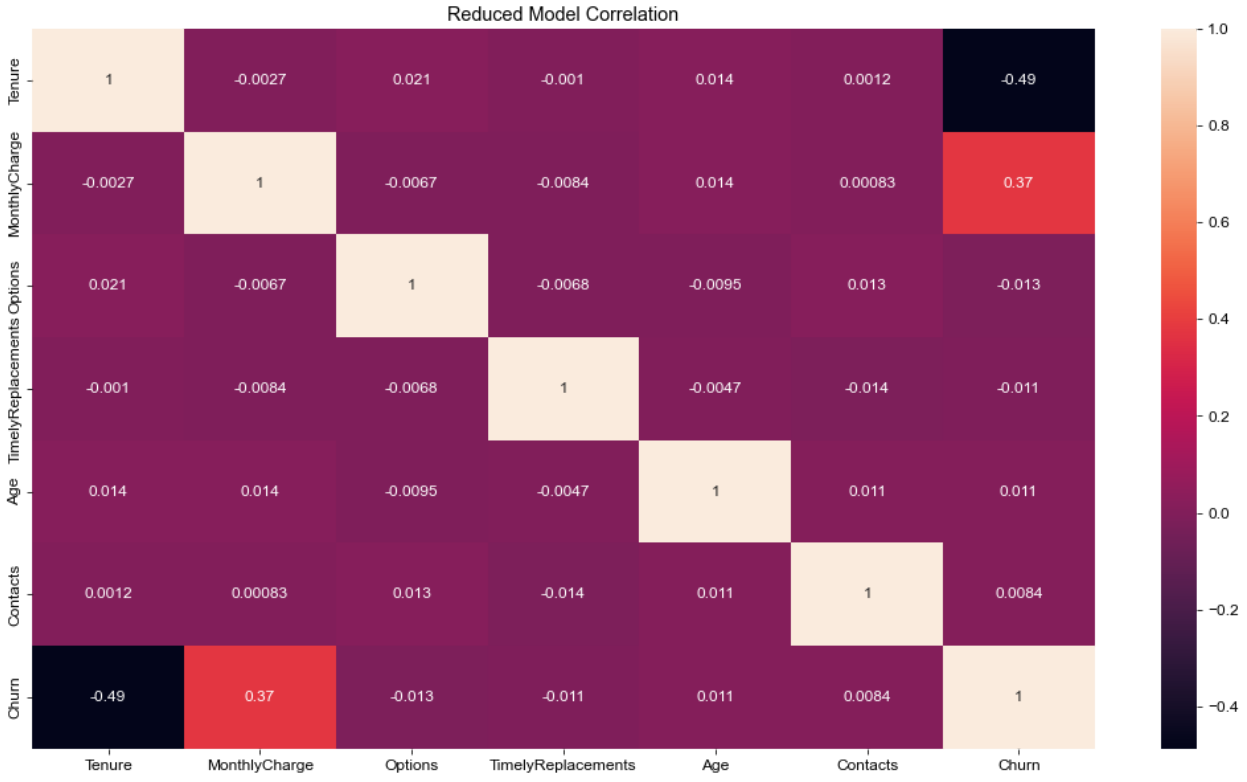
Step 3

```
                        coef
Tenure              -0.075068
MonthlyCharge        0.033517
Options             -0.011357
TimelyReplacements  -0.065911
Age                  0.003418
Contacts             0.035902


Final F1 Score: 0.6654804270462634
Final Accuracy Score 0.8346525945470537
Final Features: 6

Final features used:
Tenure
MonthlyCharge
Options
TimelyReplacements
Age
Contacts
```

E1. Data Analysis Process

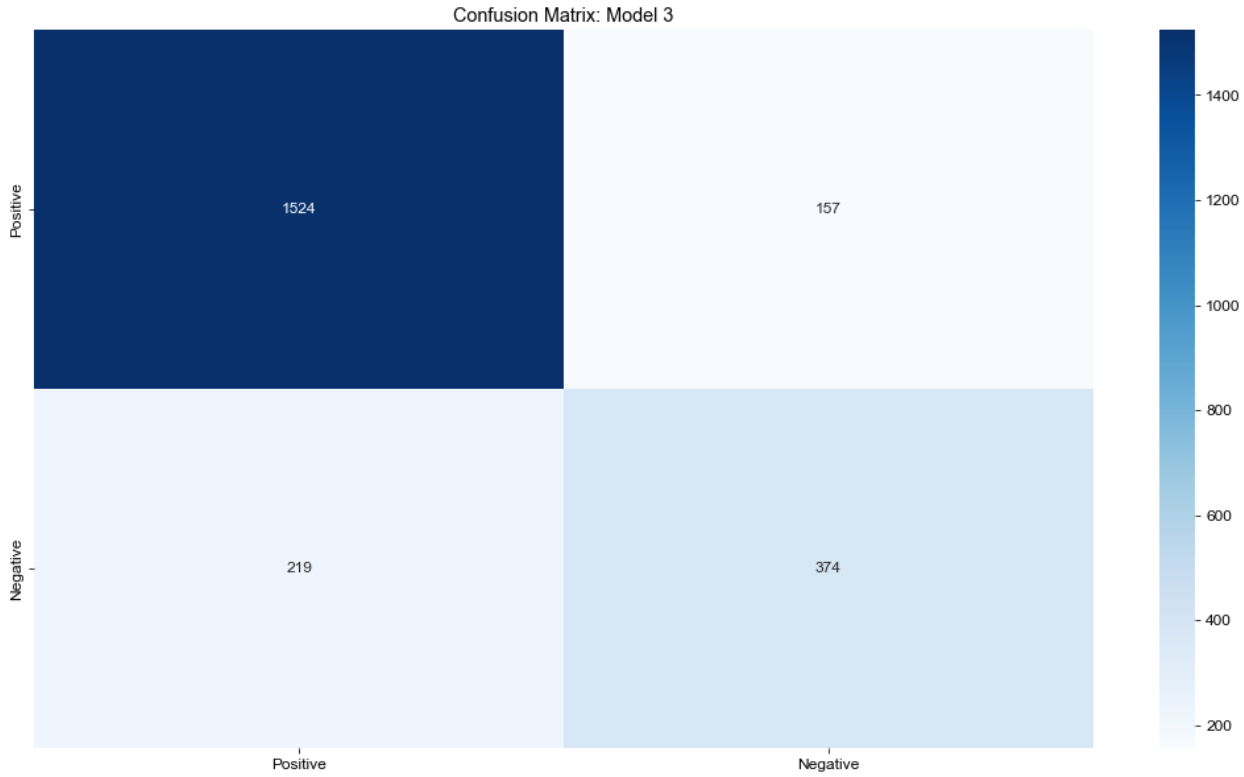
Our data analysis process consisted of cleaning and preparing our data, creating an initial model, step-forward feature selection, and feature variance reduction. We began with an initial model that took in data for twenty-eight independent variables and rendered an F1 score of 0.69 and an accuracy score of 0.84. After following the steps listed above, we produced a reduced model that took in data for just six independent variables and rendered an F1 score of 0.67 and an accuracy score of 0.83. Below is a detailed description of the F1 and accuracy scores and features used in our reduced model and a heatmap representing the correlation between the independent variable.

```
Initial F1 Score: 0.6899563318777293
Initial Accuracy Score 0.8438874230430958
Initial Features: 28
Final F1 Score: 0.6654804270462634
Final Accuracy Score 0.8346525945470537
Final Features: 6

Final features used:
Tenure
MonthlyCharge
Options
TimelyReplacements
Age
Contacts
```



E2. Results Analysis

Confusion Matrix



Accuracy Score

To get the accuracy score, we performed the following calculations:

accuracy = (number of correct predictions) / (total number of predictions)

Our model made 2,274 predictions. Out of those predictions, we can see from the confusion matrix above that 1,524 were accurate positives, 219 were false positives, 374 were accurate negatives, and 157 were false negatives. Using that information, we can sum the number of accurate positives and negatives, giving us 1,898 correct predictions. When divided by the total number of predictions, 2,274, we get an accuracy score of 0.835, or 83.5% accuracy.

E3. Executable Python File

Executable Python file is saved as: *logistic_regression_models.py*.

Part V. Data Summary and Implications

F1. Data Analysis Results

Reduced Model's Regression Equation

The equation for the reduced regression model is:

$$p = 1 / (1 + e^{-(intercept + b1x1 + \ldots + bnxn)})$$

Where p is the predicted probability of an event occurring for a given observation, intercept the coefficient of the constant term, b1 and bn are the coefficients for the independent variables, and x1 and xn are the values for the respective independent variables. The equation for our reduced regression would read as follows:

Churn = - 5.1215 - 0.0751(Tenure) + 0.0335(MonthlyCharge) - 0.0114 (Options) - 0.0659(TimelyReplacements) + 0.003418(Age) + 0.0359(Contacts)

Coefficient Interpretation

The coefficients of the reduced model tells us how each independent variable affects the probability of the dependent variable. For our reduced model, we can see that as tenure increases, the likelihood of customer churn decreases. We can also see that as the customer's monthly charges increase, so does their likelihood of churn. As

the customer's satisfaction with the options provided increases, there is a negligible decrease in the likelihood they will churn.

We can also point out that as a customer's satisfaction with timely replacements of devices increases, their likelihood to churn decreases. Interestingly, as customers' age increases, so does their likelihood to churn. Similarly, as the number of times a customer has to contact technical support increases, their likelihood to churn increases. We can use these variables from the regression equation to estimate how much customers value each independent variable as they impact their likelihood to churn. Alternatively, we could predict future churn or run specific scenarios if any independent variables were to change.

Statistical and Practical Significance

Regarding statistical significance, we can use the coefficient of each independent variable in our reduced logistic regression model and the exponential function (exp) to calculate the odds ratio of each independent variable as they relate to churn. An odds ratio greater than one indicates a higher probability of a customer not churning. In contrast, an odds ratio of less than one indicates that the odds of customers churning are higher. When calculating the odds ratio for each independent variable, we get the following:

Tenure	*0.9274*	*TimelyReplacements*	*0.9369*
MonthlyCharge	*1.034*	*Age*	*1.0034*
Options	*0.9887*	*Contacts*	*1.0369*

We also used the model's F1 and accuracy scores to determine the statistical significance of the model overall. With an F1 score of 0.67 compared to the initial model's F1 score of 0.69 and an accuracy score of 0.83 compared to the initial model's accuracy score of 0.84, we can feel confident that our model makes predictions nearly as accurate as the initial model.

The practical significance of our model is that a telecom company will be able to use the model for multiple prediction purposes. For example, a customer's likelihood of churn increases as their monthly charges increase. Those with a higher likelihood of churning also tend to contact technical support more often and are older. We can also use a customer's tenure to measure their likelihood to churn, as those who spend more time with the company are less likely to churn than newly subscribed customers. A telecom company can use insights like those mentioned above to decide on future marketing strategies and customer incentives.

Limitations

Despite having a fairly high accuracy score and using few data points to make those predictions, there are some limitations to a logistic regression model. One limitation is the logistic regression model's sensitivity to outliers. All outliers must be addressed in the data preparation process, as they can significantly impact the model's prediction power. Suppose the data has many outliers, and the analyst drops them altogether. The decision to drop too much of the data can impact the accuracy of the linear regression model, as it needs a sufficient sample size to estimate coefficients and make assumptions about the data's distribution.

F2. Recommendations

When interpreting the coefficients for our model, we saw that the coefficient for tenure, options and timely replacements are negative, indicating that as they increase, a customer's likelihood to churn decreases. Inversely, we saw that as a customer's monthly charges, age, and amount of times they contact technical support increase, so does their likelihood of churn. Using that information, we recommend that the telecom provider focus on providing more service options and ensuring that replacements are processed quickly and efficiently. I would also recommend the telecom provider review their pricing strategies to ensure they remain competitive in the market.

Based on the data, I recommend that the telecom company focus on customer retention, especially as customer age increases their likelihood of churn, but their tenure with the company decreases it. The telecom company can also use the model to predict churn for other customers with similar characteristics, which can inform future marketing, customer service training, and pricing strategies.

Part VI. Demonstration

G. Panopto Video

The Panopto video is linked in the project submission.

H. Code Sources

We did not use third-party code to create this project.

I. Citation Sources

We did not use in-text citations in this project.