Predicting Customer Retention Outcomes

Gerald Burke

Western Governors University

D208: Predictive Modeling

Dr. Taylor Jensen

November 6, 2024

[Part I: Research Question 4](#_Toc182042808)

[A1. Research Question 4](#_Toc182042809)

[A2. Goals of the analysis 4](#_Toc182042810)

[Part II: Method Justification 5](#_Toc182042811)

[B1. Assumptions of Logistic Regression Modelling 5](#_Toc182042812)

[B2. Benefits of the Python Environment 6](#_Toc182042813)

[B3. Application of Logistic Regression 6](#_Toc182042814)

[Part III: Data Preparation 6](#_Toc182042815)

[C1. Data Cleaning Goals 6](#_Toc182042816)

[Duplicate Values 6](#_Toc182042817)

[Missing Values 7](#_Toc182042818)

[Outliers 7](#_Toc182042819)

[Evaluation 7](#_Toc182042820)

[C2. Variables for Analysis 11](#_Toc182042821)

[Dependent Variable 11](#_Toc182042822)

[Independent Variables 12](#_Toc182042823)

[C3. Visual Analysis 14](#_Toc182042824)

[Univariate Analysis 14](#_Toc182042825)

[Bivariate Analysis 16](#_Toc182042826)

[C4. Data Transformation 17](#_Toc182042827)

[C5. Prepared Data 19](#_Toc182042828)

[Part IV: Model Comparison and Analysis 19](#_Toc182042829)

[D1. Initial Model 19](#_Toc182042830)

[D3. Reduced Model 34](#_Toc182042831)

[E1. Comparison of Models 35](#_Toc182042832)

[E2. Analysis Process and Output 35](#_Toc182042833)

[E3. Code 35](#_Toc182042834)

[Part V: Data Summary and Implications 36](#_Toc182042835)

[F1. Results of the Analysis 36](#_Toc182042836)

[F2. Recommended Action 36](#_Toc182042837)

[Part VI: Demonstration 36](#_Toc182042838)

[G. Panopto Video 36](#_Toc182042839)

[H. Code Sources 37](#_Toc182042840)

[I. Sources 37](#_Toc182042841)

# Part I: Research Question

## A1. Research Question

What factors, if any, are the most impactful on predicting rates of customer churn?

## A2. Goals of the analysis

Customer churn is a massive issue in telecommunications. Customer churn is defined as whether or not a customer discontinued service in the last month. My analysis will focus on a mix of features, both those within the control of the organization, and those that are not, but could point to potential user categories for further analysis.

Those features being examined that fall into the former category are:

Item1:

* Type: Ordinal
* Example: 5
* Description: Timely response

Item2:

* Type: Ordinal
* Example: 5
* Description: Timely fixes

Item4:

* Type: Ordinal
* Example: 3
* Description: Reliability

Item5:

* Type: Ordinal
* Example: 4
* Description: Options

Item6:

* Type: Ordinal
* Example: 4
* Description: Respectful response

These features are characterized by the importance customers place on certain aspects of the services rendered. If there is a relationship between churn and one of this categories, it could point to a potential failure in meeting customer expectations. Problem areas in these categories can be addressed with training and institutional policies.

The features that fall into the latter category are:

Gender:

* Type: Nominal
* Example: Male
* Description: Customer self-identification as male, female, or nonbinary

Age:

* Type: Continuous/Discrete (based on treatment)
* Example: 68
* Description: Age of customer as reported in sign-up information

Income:

* Type: Continuous
* Example: 28561.99
* Description: Annual income of customer as reported at time of sign-up

Marital:

* Type: Nominal
* Example: Widowed
* Description: Marital status of customer as reported in sign-up information

Children:

* Type: Discrete
* Example: 0
* Description: Number of children in customer’s household as reported in sign-up information

These features are largely demographic data that create a profile of the user. If variables in this category can be found to accurately predict customer churn, further research into the treatment of customers that fall within certain demographics could potentially yield better outcomes in the future.

# Part II: Method Justification

## B1. Assumptions of Logistic Regression Modelling

Logistic regression requires a number of assumptions to be met to assure the accuracy of the analysis. These are:

1. Logistic regression requires the dependent variable to be categorical
2. Logistic regression requires observations to be independent of one another
3. Logistic regression requires little to no multicollinearity among independent variables
4. Logistic regression requires a large sample size. i.e. a minimum of 10 cases with the least frequent outcome.

## B2. Benefits of the Python Environment

I chose to approach this project using python. My two primary reasons are:

1. Python is a language I am both professionally and academically experienced in.
2. Incredibly robust and powerful tools and libraries are available to assist in performing data analysis in Python.

The libraries I primarily used for this task are:

* Pandas – A framework for managing data utilizing powerful and robust objects called data frames.
* NumPy – A library of mathematical functions that expand the mathematical analysis available in vanilla Python.
* Statsmodels – A library full of tools for performing statistical tests and analyses, includes the Logit function to enable my logistic regression analysis.
* SciKit Learn – A library of tools for scientific analysis. Used primarily in this project for data scaling and variable re-expression.
* MatPlotLib – A library of tools for plotting various graphs and charts.
* Seaborn – An extension of MatPlotLib designed to provide more specialized graphs and charts for MatPlotLib users.

## B3. Application of Logistic Regression

Logistic regression was the most appropriate tool for this analysis as it is designed to compare a categorical with one or more continuous or categorical variables. Given I was testing what independent variables impacted the dependent variable, churn, the approach fit the hypothesis being tested. Another benefit was my ability to cast a wide net and test a number of independent variables for the sake of developing a model that could provide insight into action for the project stakeholders.

# Part III: Data Preparation

## C1. Data Cleaning Goals

My cleaning plan primarily focused on rooting out duplicate records, missing values, and outliers.

### Duplicate Values

I performed the following action to detect duplicates:

#Check for duplicates - None Found

print(df.duplicated().value\_counts())

The result was that no duplicate values were found in the set.

False 10000

### Missing Values

I performed the following action to detect missing values:

#Check for missing values - None Found

print(df.isna().any())

The result was that no missing values were detected in the set.

Name: count, dtype: int64

Churn False

Item1 False

Item2 False

Item4 False

Item5 False

Item6 False

Gender False

Age False

Income False

Marital False

Children False

dtype: bool

### Outliers

My method for detecting outliers was a combination of three techniques. I printed a report that featured minimum, maximum, range, and mean of all the numeric variables. This report also features an evaluation of values with z-scores above 3 and below -3. It also examines values above Q4 and below Q1, with a minimum, maximum, and range of outlier values outside of the Q1-Q4 range.

The other two techniques involve plotting the variables for visual evaluation of outliers. The plots chosen for the analysis were histograms and box plots. Below is my evaluation of each independent numeric variable.

### Evaluation

I have organized my evaluation into three sections per variable.

1. The Python generated report of values and ranges
2. The generated plots
3. Conclusions from the analysis and justification for chose mitigation method

Below is the code I used to generate the reports:

#Loop through the numeric variables and generate a report for each

for column in numeric\_columns:

#Generate name for new column based on original column name

name = 'Z\_Score\_' + column

#Calculate z-score and assign its value to fields in new column

df[name] = stats.zscore(df[column])

#Print relevant information about values in the set

print(f'++++========={column}========++++')

print(f'Minimum value: {df[column].min()}')

print(f'Maximum value: {df[column].max()}')

print(f'Value range: {df[column].max() - df[column].min()}')

print(f'Mean Value: {df[column].mean()}')

print(f'Values with a z-score over 3: {sum(stats.zscore(df[column]) > 3)}')

print(f'Values with a z-score under -3: {sum(stats.zscore(df[column]) < -3)}')

#Calculate quantiles

q1 = np.quantile(df[column], 0.25)

q3 = np.quantile(df[column], 0.75)

lower\_threshold = q1 - 1.5 \* stats.iqr(df[column])

upper\_threshold = q3 + 1.5 \* stats.iqr(df[column])

print(f'Values below IQR lower threshold: {(df[column] < lower\_threshold).sum()}')

print(f'Values above IQR upper threshold: {(df[column] > upper\_threshold).sum()}')

outliers = df.query(column + ' < ' + str(lower\_threshold) + ' | ' + column + ' > ' + str(upper\_threshold))

print(f'Minimum Outlier Value: {outliers[column].min()}')

print(f'Maximum Outlier Value: {outliers[column].max()}')

print(f'Range of Outlier Values: {outliers[column].max() - outliers[column].min()}')

#Plot the histogram

plt.hist(df[name])

plt.title(column)

plt.show()

#Plot the boxplots

boxplot=sns.boxplot(x=column, data=df)

plt.show()

#### Children

##### Report

++++=========Children========++++

Minimum value: 0

Maximum value: 10

Value range: 10

Mean Value: 2.0877

Values with a z-score over 3: 191

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

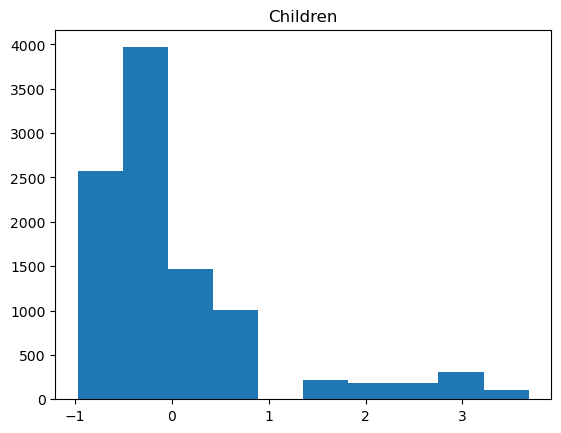
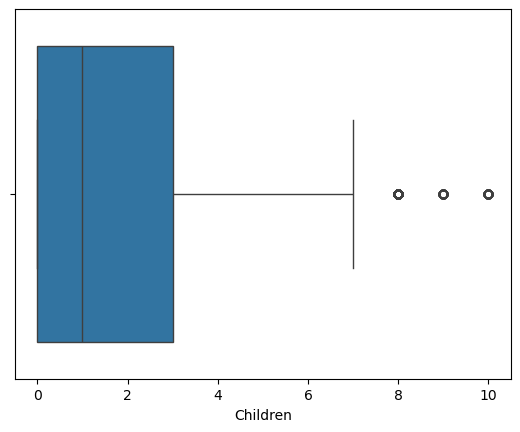
Values above IQR upper threshold: 401

Minimum Outlier Value: 8

Maximum Outlier Value: 10

Range of Outlier Values: 2

##### Graphs

##### Conclusion and justification

The histogram featured shows the isolated columns generally associated with outliers. The count of values above Q4 is 401, or roughly 0.04% of the set. The rang of values is small with a minimum of 0 and a maximum of 10. The values that are above the Q4 threshold represent ~0.04% of the set. With these considerations, I decided to retain the found outliers.

#### Age

##### Report

++++=========Age========++++

Minimum value: 18

Maximum value: 89

Value range: 71

Mean Value: 53.0784

Values with a z-score over 3: 0

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

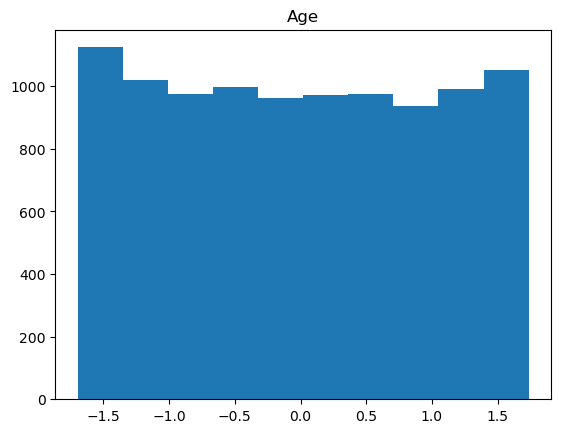
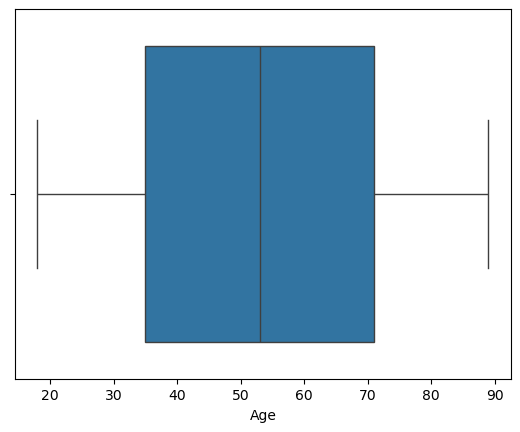
Values above IQR upper threshold: 0

Minimum Outlier Value: nan

Maximum Outlier Value: nan

Range of Outlier Values: nan

##### Graphs

##### Conclusion and justification

No outliers were found in the Age column, so no further mitigation or analysis was necessary.

#### Income

##### Report

++++=========Income========++++

Minimum value: 348.67

Maximum value: 258900.7

Value range: 258552.03

Mean Value: 39806.926771

Values with a z-score over 3: 145

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

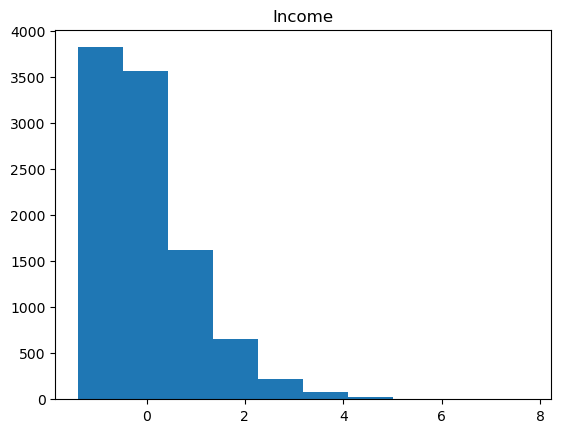
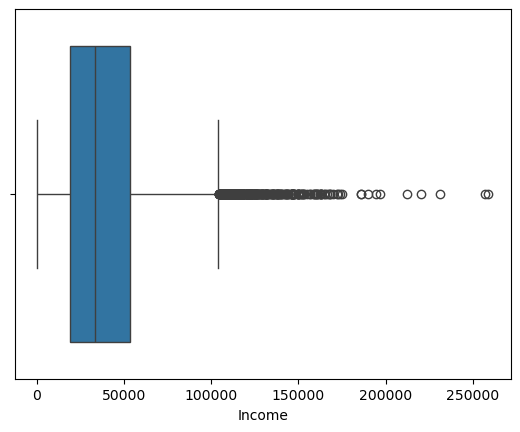
Values above IQR upper threshold: 336

Minimum Outlier Value: 104362.5

Maximum Outlier Value: 258900.7

Range of Outlier Values: 154538.2

##### Graph

##### Conclusion and justification

Outliers were found within the column. These were expected within income and represent reasonable values. None of the values appear to be erroneous. For those reasons my chosen course of treatment was to retain the values.

## C2. Variables for Analysis

For each variable in the set, I generated a report of relevant statistical information using the below formula:

#Print reports for associated variables

for variable in df.columns:

print(f'========Begin Report========')

print(f'========{variable}========')

print(df[variable].describe())

if pd.api.types.is\_numeric\_dtype(df[variable]):

print(f'Median: {df[variable].median()}')

else:

print(f'Unique Values: {df[variable].unique()}')

print(f'Unique Value Counts: {df[variable].value\_counts()}')

print(f'========End Report========')

Each report leverages the *describe()* function to get some basic statistics. Further statistics were generated based on whether the values in the column were numeric or categorical. For this I use *pd.api.types.is\_numeric\_dtype()* to distinguish between the two. For numeric values I also included the median in the report. For the categorical values, I included a list of the unique values present in the column and a count of each unique value.

### Dependent Variable

Churn:

========Begin Report========

========Churn========

count 10000

unique 2

top No

freq 7350

Name: Churn, dtype: object

Unique Values: ['No' 'Yes']

Unique Value Counts: Churn

No 7350

Yes 2650

Name: count, dtype: int64

========End Report========

### Independent Variables

Item1:

========Begin Report========

========Item1========

count 10000.000000

mean 3.490800

std 1.037797

min 1.000000

25% 3.000000

50% 3.000000

75% 4.000000

max 7.000000

Name: Item1, dtype: float64

Median: 3.0

========End Report========

Item2

========Begin Report========

========Item2========

count 10000.000000

mean 3.505100

std 1.034641

min 1.000000

25% 3.000000

50% 4.000000

75% 4.000000

max 7.000000

Name: Item2, dtype: float64

Median: 4.0

========End Report========

Item4:

========Begin Report========

========Item4========

count 10000.000000

mean 3.497500

std 1.025816

min 1.000000

25% 3.000000

50% 3.000000

75% 4.000000

max 7.000000

Name: Item4, dtype: float64

Median: 3.0

========End Report========

Item5:

========Begin Report========

========Item5========

count 10000.000000

mean 3.492900

std 1.024819

min 1.000000

25% 3.000000

50% 3.000000

75% 4.000000

max 7.000000

Name: Item5, dtype: float64

Median: 3.0

========End Report========

Item6:

========Begin Report========

========Item6========

count 10000.000000

mean 3.497300

std 1.033586

min 1.000000

25% 3.000000

50% 3.000000

75% 4.000000

max 8.000000

Name: Item6, dtype: float64

Median: 3.0

========End Report========

Gender:

========Begin Report========

========Gender========

count 10000

unique 3

top Female

freq 5025

Name: Gender, dtype: object

Unique Values: ['Male' 'Female' 'Nonbinary']

Unique Value Counts: Gender

Female 5025

Male 4744

Nonbinary 231

Name: count, dtype: int64

========End Report========

Age:

========Begin Report========

========Age========

count 10000.000000

mean 53.078400

std 20.698882

min 18.000000

25% 35.000000

50% 53.000000

75% 71.000000

max 89.000000

Name: Age, dtype: float64

Median: 53.0

========End Report========

Income:

========Begin Report========

========Income========

count 10000.000000

mean 39806.926771

std 28199.916702

min 348.670000

25% 19224.717500

50% 33170.605000

75% 53246.170000

max 258900.700000

Name: Income, dtype: float64

Median: 33170.604999999996

========End Report========

Marital:

========Begin Report========

========Marital========

count 10000

unique 5

top Divorced

freq 2092

Name: Marital, dtype: object

Unique Values: ['Widowed' 'Married' 'Separated' 'Never Married' 'Divorced']

Unique Value Counts: Marital

Divorced 2092

Widowed 2027

Separated 2014

Never Married 1956

Married 1911

Name: count, dtype: int64

========End Report========

Children:

========Begin Report========

========Children========

count 10000.0000

mean 2.0877

std 2.1472

min 0.0000

25% 0.0000

50% 1.0000

75% 3.0000

max 10.0000

Name: Children, dtype: float64

Median: 1.0

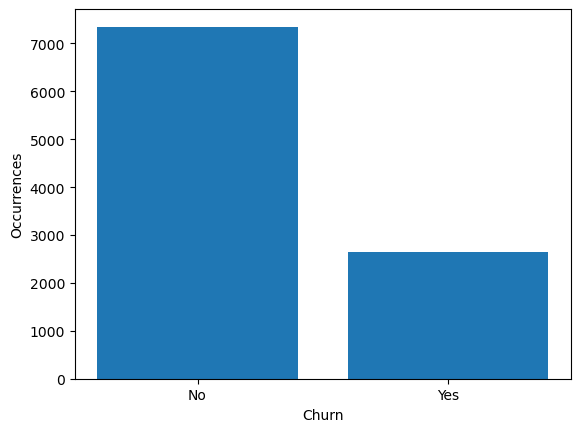
========End Report========

## C3. Visual Analysis

### Univariate Analysis

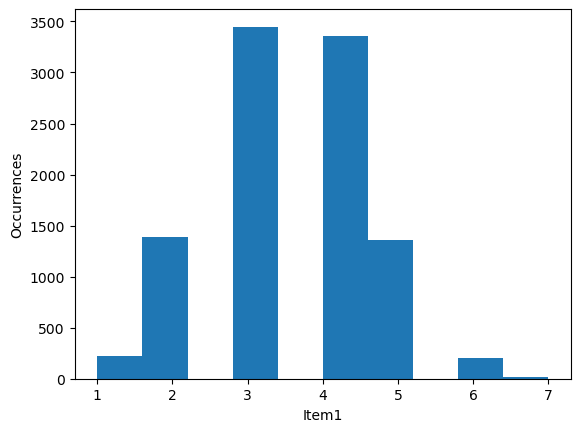
#### Dependent Variable

##### Churn

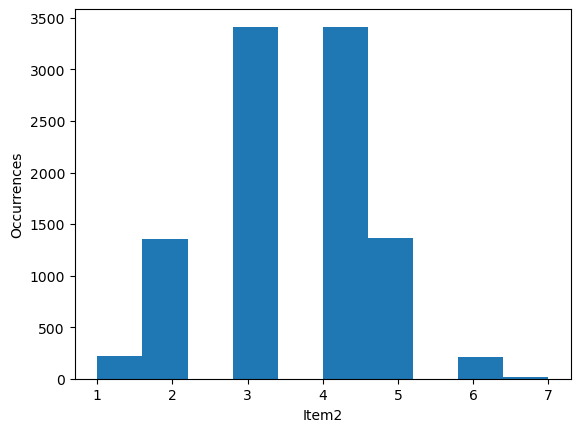


#### Independent Variables

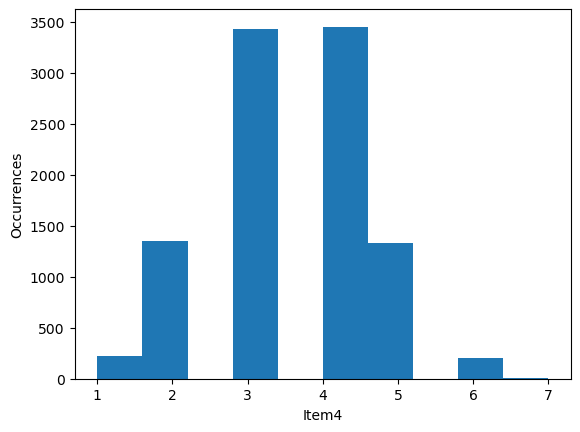
##### Item1



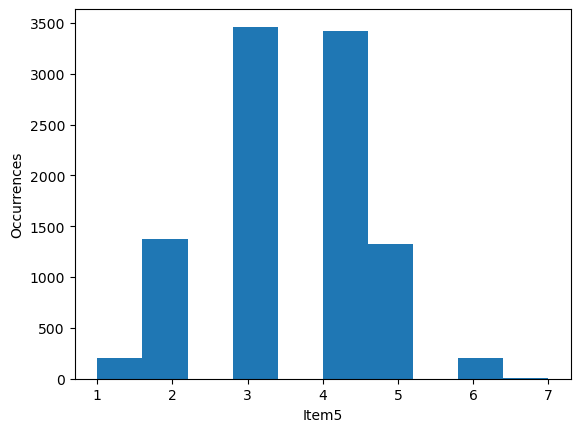
##### Item2



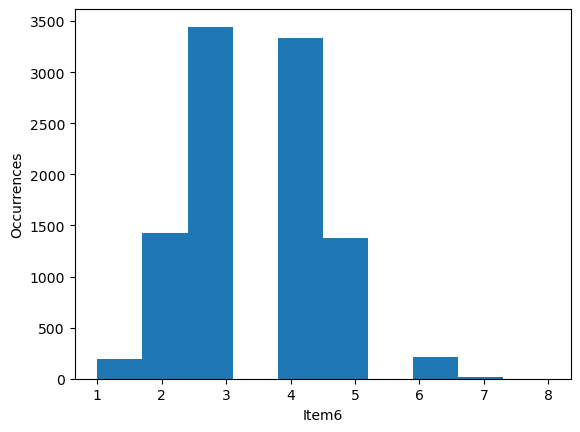
##### Item4



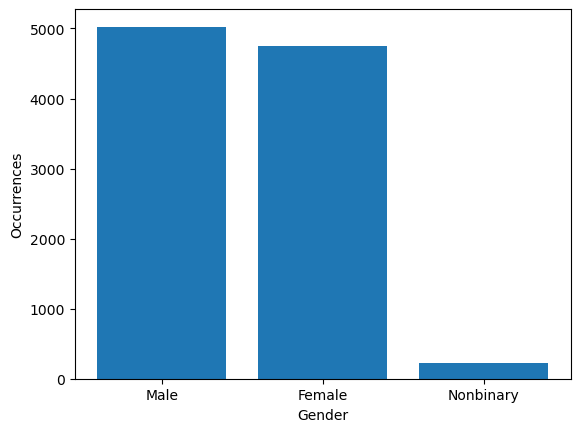
##### Item5



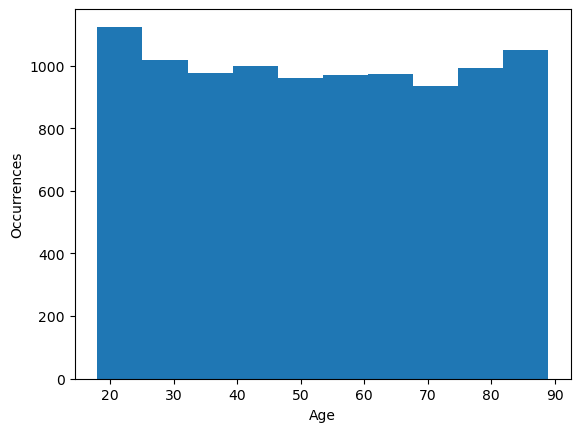
##### Item6



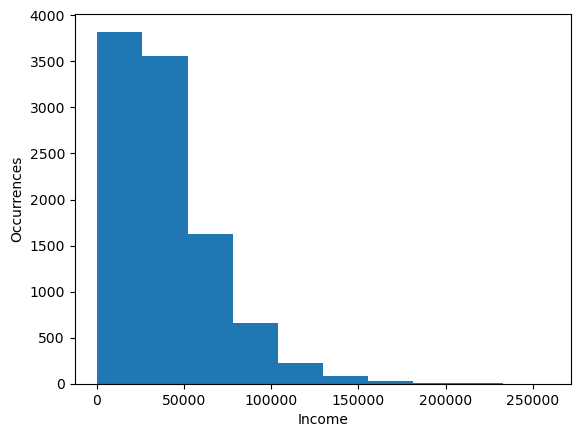
##### Gender



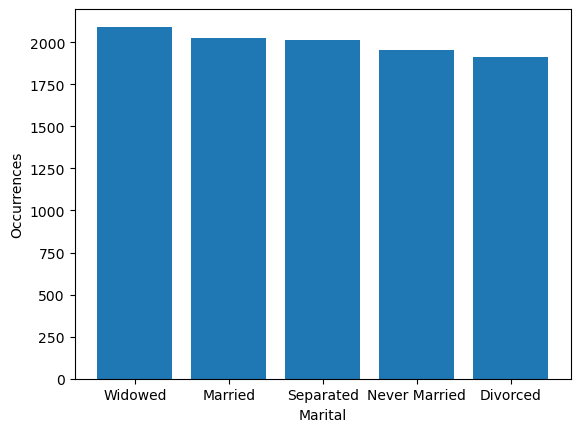
##### Age



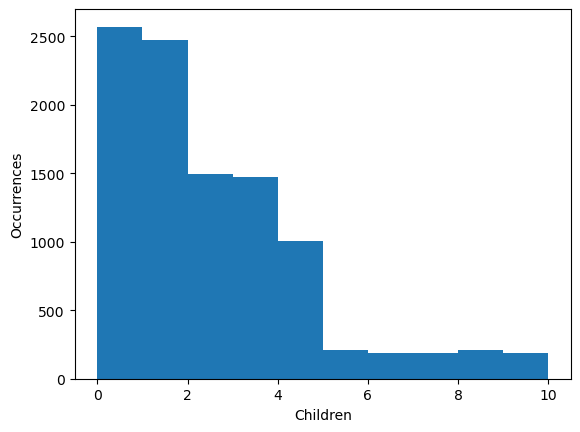
##### Income



##### Marital

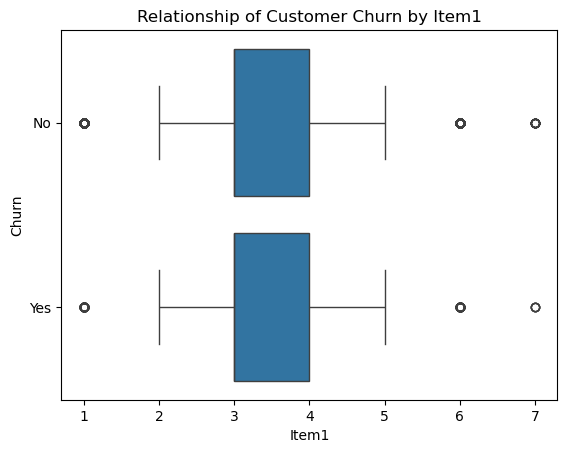


##### Children

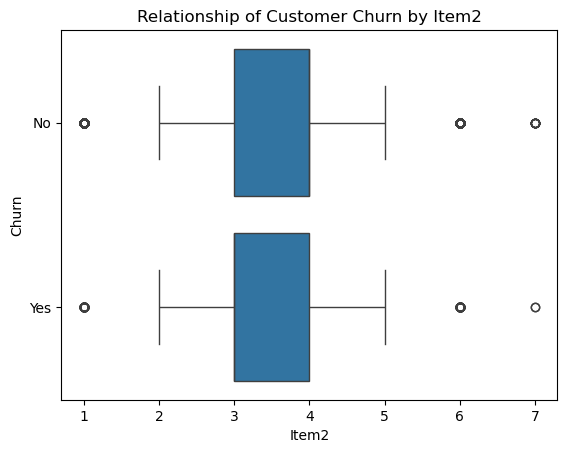


### Bivariate Analysis

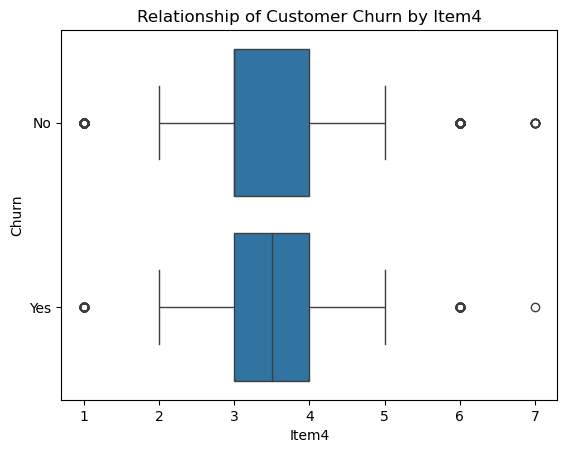
##### Churn vs Item1



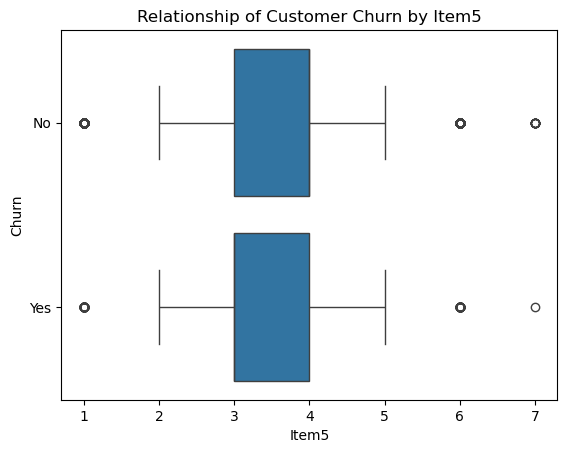
##### Churn vs Item2



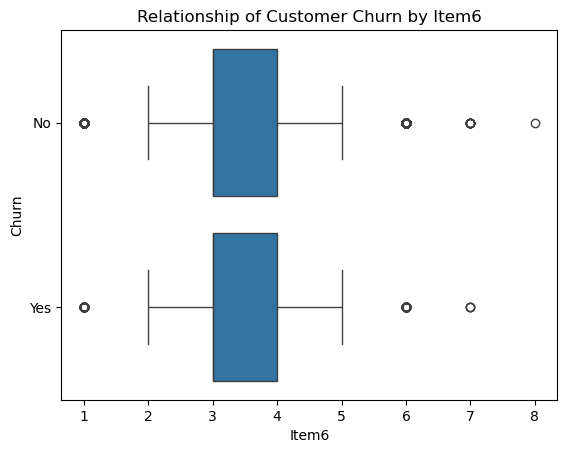
##### Churn vs Item4



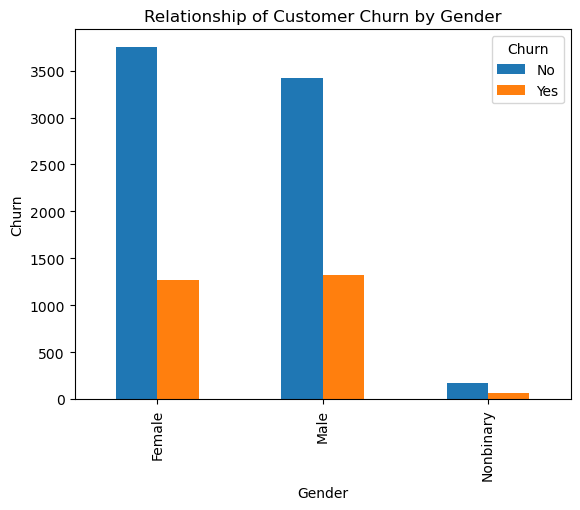
##### Churn vs Item5



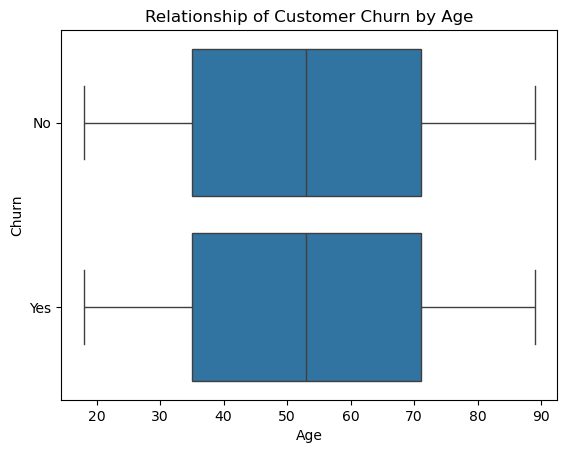
##### Churn vs Item6



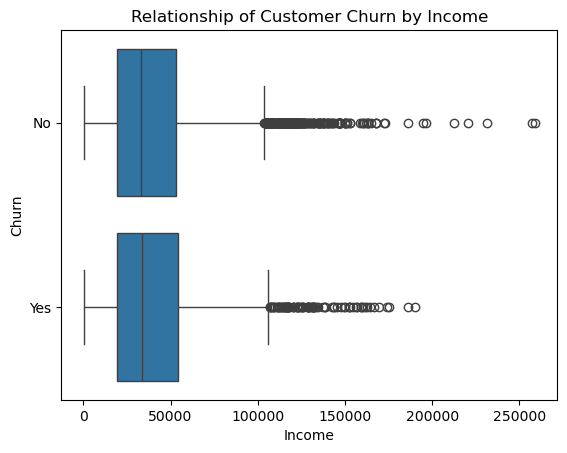
##### Churn vs Gender



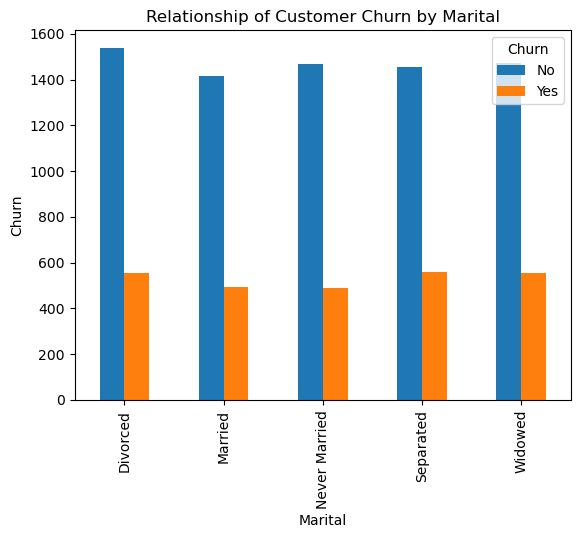
##### Churn vs Age



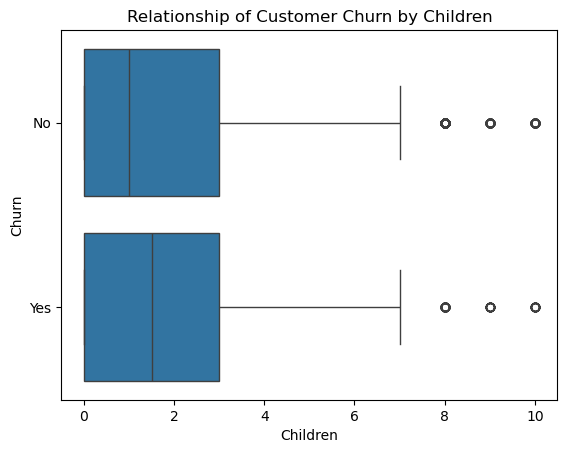
##### Churn vs Income



##### Churn vs Marital



##### Churn vs Children



## C4. Data Transformation

The selected variables featured several categorical variables with non-numeric values. The treatment for these needed to be considered on a case-by-case basis. Cursory analysis using:

categorical\_variables = ['Churn', 'Item1', 'Item2', 'Item4', 'Item5', 'Item6', 'Gender', 'Marital']

#Separate reports

print('##### Before Encoding ####')

#Examine values present in categorical\_variables to determine treatment

for variable in categorical\_variables:

print(f'{variable} : {df[variable].unique()}')

Yielded the following report:

##### Before Encoding ####

Churn : ['No' 'Yes']

Item1 : [5 3 4 6 2 1 7]

Item2 : [5 4 3 2 6 1 7]

Item4 : [3 4 2 5 7 6 1]

Item5 : [4 5 1 2 3 7 6]

Item6 : [4 3 5 2 6 7 1 8]

Gender : ['Male' 'Female' 'Nonbinary']

Marital : ['Widowed' 'Married' 'Separated' 'Never Married' 'Divorced']

Ordinal columns Item1, Item2, Item4, Item5, and Item6 were all numerically coded and required no treatment.

The depended variable, Churn, was dichotomous, so I elected for a mapping of 0 for ‘No’ and 1 for ‘Yes’ using the following code:

#Churn will be treated as dichotomous

df['Churn'] = df['Churn'].map({'No' : 0, 'Yes' : 1})

print(f'Churn : {df['Churn'].unique()}')

With the resulting outcome:

##### After Encoding ####

Churn : [0 1]

Gender and Married were both nominal variables, so ordinal encoding would not be appropriate for processing them. I decided to use one-hot encoding utilizing the OneHotEncoder function for Sci-Kit Learn. I created dummy variables, dropped on for each variable following the k-1 rule, concatenated the results to the data frame, and dropped the original variables.

I used the following code to do this:

#Perform One-Hot Encoding using Scikit-learn's OneHotEncoder

#https://www.geeksforgeeks.org/ml-one-hot-encoding/

#Initialize new encoder

encoder = OneHotEncoder(sparse\_output=False, drop='first')

#Encode values

one\_hot\_encoded = encoder.fit\_transform(df[one\_hot\_columns])

#Create a new data frame from encoded values

one\_hot\_df = pd.DataFrame(one\_hot\_encoded, columns=encoder.get\_feature\_names\_out(one\_hot\_columns))

#Verify new data frame

print('#### One-Hot Encoded Data Frame ####')

print(one\_hot\_df.head())

#Concatenate the original data frame with the one-hot encoded data frame

df = pd.concat([df, one\_hot\_df], axis=1)

#Drop the original columns

df = df.drop(one\_hot\_columns, axis=1)

df = df.drop('Unnamed: 0', axis=1)

Below is a printed report of the head() of the data frame created by one hot encoding:

#### One-Hot Encoded Data Frame ####

Gender\_Male Gender\_Nonbinary Marital\_Married Marital\_Never Married \

0 1.0 0.0 0.0 0.0

1 0.0 0.0 1.0 0.0

2 0.0 0.0 0.0 0.0

3 1.0 0.0 1.0 0.0

4 1.0 0.0 0.0 0.0

Marital\_Separated Marital\_Widowed

0 0.0 1.0

1 0.0 0.0

2 0.0 1.0

3 0.0 0.0

4 1.0 0.0

## C5. Prepared Data

*The prepared data is included in the file churn\_clean\_final.csv*

# Part IV: Model Comparison and Analysis

## D1. Initial Model

I used the following code to generate my initial model with *all* relevant variables outlined:

#Establish initial feature set

independent\_variables = ['Item1', 'Item2', 'Item4', 'Item5', 'Item6', 'Gender\_Male', 'Gender\_Nonbinary', 'Age', 'Income', 'Marital\_Married',

'Marital\_Never Married', 'Marital\_Separated', 'Marital\_Widowed', 'Children']

#Define dependent and independent variables

y = df['Churn']

x = df[independent\_variables]

#Perform initial regression and examine results

regression\_model = sm.Logit(y, x)

results = regression\_model.fit()

print(results.summary())

The result was:

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9986

Method: MLE Df Model: 13

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.0001305

Time: 09:04:21 Log-Likelihood: -5781.5

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 1.000

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item1 -0.0077 0.030 -0.257 0.797 -0.066 0.051

Item2 -0.0490 0.029 -1.666 0.096 -0.107 0.009

Item4 -0.0840 0.020 -4.156 0.000 -0.124 -0.044

Item5 -0.1106 0.019 -5.877 0.000 -0.147 -0.074

Item6 -0.0245 0.025 -0.988 0.323 -0.073 0.024

Gender\_Male 0.0962 0.045 2.121 0.034 0.007 0.185

Gender\_Nonbinary -0.0121 0.155 -0.078 0.938 -0.315 0.291

Age -0.0007 0.001 -0.684 0.494 -0.003 0.001

Income -7.146e-09 7.92e-07 -0.009 0.993 -1.56e-06 1.54e-06

Marital\_Married -0.0804 0.071 -1.136 0.256 -0.219 0.058

Marital\_Never Married -0.1267 0.071 -1.791 0.073 -0.265 0.012

Marital\_Separated 0.0176 0.069 0.256 0.798 -0.117 0.152

Marital\_Widowed 0.0006 0.069 0.009 0.993 -0.134 0.136

Children -0.0100 0.011 -0.948 0.343 -0.031 0.011

=========================================================================================

D2. Feature Selection The first step I chose to perform in feature selection was to remove variables with high probability of multicollinearity. I utilized VIF analysis with a threshold of 10 to perform this. I wrote a loop to examine the VIF off all features and remove each, one at a time, that had a score of 10 or higher. This was the code I used to achieve this:

#First step of feature elimination will be checking for multicolinearity utilizing VIF with a threshold of 10

vif\_threshold = 10;

#The process has been automated with a while loop providing feedback at each iteration

while True:

#Create a new data frame for VIF values

vif\_data = pd.DataFrame()

vif\_data['feature'] = df[independent\_variables].columns

#Calculate VIF for each feature

vif\_data['VIF'] = [variance\_inflation\_factor(df[independent\_variables].values, i) for i in range(len(df[independent\_variables].columns))]

#Print the VIF data frame

print(vif\_data)

#Identify the feature with the highest VIF

max\_vif = vif\_data['VIF'].max()

max\_vif\_feature = vif\_data.loc[vif\_data['VIF'].idxmax(), 'feature']

#Check if the highest VIF exceeds the threshold

if max\_vif >= vif\_threshold:

#Remove the feature with the highest VIF

independent\_variables.remove(max\_vif\_feature)

else:

#Exit the loop if all VIF values are below the threshold

break

The results were as follows:

feature VIF

0 Item1 23.184080

1 Item2 22.483626

2 Item4 10.554881

3 Item5 9.116717

4 Item6 15.936199

5 Gender\_Male 1.906583

6 Gender\_Nonbinary 1.045531

7 Age 6.951469

8 Income 2.914018

9 Marital\_Married 1.859731

10 Marital\_Never Married 1.878932

11 Marital\_Separated 1.898736

12 Marital\_Widowed 1.904732

13 Children 1.919185

feature VIF

0 Item2 13.908942

1 Item4 10.549356

2 Item5 8.851247

3 Item6 14.581393

4 Gender\_Male 1.906581

5 Gender\_Nonbinary 1.045479

6 Age 6.949856

7 Income 2.913996

8 Marital\_Married 1.858771

9 Marital\_Never Married 1.877783

10 Marital\_Separated 1.897462

11 Marital\_Widowed 1.904357

12 Children 1.918618

feature VIF

0 Item2 10.440646

1 Item4 8.699545

2 Item5 8.841584

3 Gender\_Male 1.900049

4 Gender\_Nonbinary 1.044660

5 Age 6.863119

6 Income 2.901755

7 Marital\_Married 1.852679

8 Marital\_Never Married 1.867043

9 Marital\_Separated 1.888118

10 Marital\_Widowed 1.896782

11 Children 1.915601

feature VIF

0 Item4 7.358326

1 Item5 7.512378

2 Gender\_Male 1.894856

3 Gender\_Nonbinary 1.044375

4 Age 6.704084

5 Income 2.879173

6 Marital\_Married 1.839653

7 Marital\_Never Married 1.853867

8 Marital\_Separated 1.872202

9 Marital\_Widowed 1.882094

10 Children 1.905491

I first removed Item1 as it had a VIF of 23.184080. After a second analysis, Item6 was shown to have a VIF of 14.581393, so I elected to remove it. Subsequent analysis resulted in the removal of Item2 with a VIF of 10.440646. After the removal of these variables the highest VIF was Item5 with a VIF of 7.512378, which is high, but without acceptable standards outlined in the process.

The primary process for feature selection I elected to perform was backward stepwise elimination. I wrote a loop to fit the model, examine the highest p-value above an alpha of 0.5, and remove it. When this was done, I printed a report of the final model. This is the code I wrote to perform the feature selection:

#Backwards stepwise elimination

#Establish significance level

significance\_level = 0.05

#Loop until model reaches exepected significance for all values

while True:

#Get current/updated list of independent variables

x = df[independent\_variables].assign(const = 1)

#Perform initial regression and examine results

regression\_model = sm.Logit(y, x)

results = regression\_model.fit()

print(results.summary())

#Get the p-values

p\_values = results.pvalues.drop('const')

#Find the highest p-value

max\_p\_value = p\_values.max()

#Get the variable with the highest p-value

highest\_variable = p\_values.idxmax()

#If the variable with the highest p-value is above significance, remove it, otherwise end elimination

if max\_p\_value > significance\_level:

independent\_variables.remove(highest\_variable)

print(f"I have decided to remove '{highest\_variable}' as it has a p-value {max\_p\_value}")

print(results.summary())

else:

#Reduced model

print(results.summary())

break

This is the result of the analysis:

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9988

Method: MLE Df Model: 11

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001484

Time: 09:04:21 Log-Likelihood: -5773.6

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.1033

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0226 0.025 -0.916 0.360 -0.071 0.026

Item5 -0.0407 0.025 -1.654 0.098 -0.089 0.008

Gender\_Male 0.1260 0.046 2.743 0.006 0.036 0.216

Gender\_Nonbinary 0.0117 0.154 0.076 0.939 -0.291 0.314

Age 0.0006 0.001 0.553 0.580 -0.002 0.003

Income 5.491e-07 8.01e-07 0.686 0.493 -1.02e-06 2.12e-06

Marital\_Married -0.0298 0.072 -0.413 0.680 -0.171 0.112

Marital\_Never Married -0.0749 0.072 -1.039 0.299 -0.216 0.066

Marital\_Separated 0.0712 0.070 1.014 0.311 -0.067 0.209

Marital\_Widowed 0.0547 0.070 0.777 0.437 -0.083 0.193

Children -0.0048 0.011 -0.449 0.654 -0.026 0.016

const -0.9103 0.173 -5.275 0.000 -1.248 -0.572

=========================================================================================

I have decided to remove 'Gender\_Nonbinary' as it has a p-value 0.9394441416673343

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9988

Method: MLE Df Model: 11

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001484

Time: 09:04:21 Log-Likelihood: -5773.6

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.1033

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0226 0.025 -0.916 0.360 -0.071 0.026

Item5 -0.0407 0.025 -1.654 0.098 -0.089 0.008

Gender\_Male 0.1260 0.046 2.743 0.006 0.036 0.216

Gender\_Nonbinary 0.0117 0.154 0.076 0.939 -0.291 0.314

Age 0.0006 0.001 0.553 0.580 -0.002 0.003

Income 5.491e-07 8.01e-07 0.686 0.493 -1.02e-06 2.12e-06

Marital\_Married -0.0298 0.072 -0.413 0.680 -0.171 0.112

Marital\_Never Married -0.0749 0.072 -1.039 0.299 -0.216 0.066

Marital\_Separated 0.0712 0.070 1.014 0.311 -0.067 0.209

Marital\_Widowed 0.0547 0.070 0.777 0.437 -0.083 0.193

Children -0.0048 0.011 -0.449 0.654 -0.026 0.016

const -0.9103 0.173 -5.275 0.000 -1.248 -0.572

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577364

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9989

Method: MLE Df Model: 10

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001483

Time: 09:04:21 Log-Likelihood: -5773.6

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.07106

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0226 0.025 -0.916 0.360 -0.071 0.026

Item5 -0.0407 0.025 -1.654 0.098 -0.089 0.008

Gender\_Male 0.1255 0.045 2.762 0.006 0.036 0.214

Age 0.0006 0.001 0.552 0.581 -0.002 0.003

Income 5.494e-07 8.01e-07 0.686 0.493 -1.02e-06 2.12e-06

Marital\_Married -0.0298 0.072 -0.413 0.680 -0.171 0.112

Marital\_Never Married -0.0749 0.072 -1.039 0.299 -0.216 0.066

Marital\_Separated 0.0712 0.070 1.014 0.311 -0.067 0.209

Marital\_Widowed 0.0547 0.070 0.778 0.437 -0.083 0.193

Children -0.0048 0.011 -0.448 0.654 -0.026 0.016

const -0.9097 0.172 -5.277 0.000 -1.248 -0.572

=========================================================================================

I have decided to remove 'Marital\_Married' as it has a p-value 0.6796977393738919

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9989

Method: MLE Df Model: 10

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001483

Time: 09:04:21 Log-Likelihood: -5773.6

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.07106

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0226 0.025 -0.916 0.360 -0.071 0.026

Item5 -0.0407 0.025 -1.654 0.098 -0.089 0.008

Gender\_Male 0.1255 0.045 2.762 0.006 0.036 0.214

Age 0.0006 0.001 0.552 0.581 -0.002 0.003

Income 5.494e-07 8.01e-07 0.686 0.493 -1.02e-06 2.12e-06

Marital\_Married -0.0298 0.072 -0.413 0.680 -0.171 0.112

Marital\_Never Married -0.0749 0.072 -1.039 0.299 -0.216 0.066

Marital\_Separated 0.0712 0.070 1.014 0.311 -0.067 0.209

Marital\_Widowed 0.0547 0.070 0.778 0.437 -0.083 0.193

Children -0.0048 0.011 -0.448 0.654 -0.026 0.016

const -0.9097 0.172 -5.277 0.000 -1.248 -0.572

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577373

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9990

Method: MLE Df Model: 9

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001468

Time: 09:04:21 Log-Likelihood: -5773.7

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.04900

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0225 0.025 -0.913 0.361 -0.071 0.026

Item5 -0.0409 0.025 -1.659 0.097 -0.089 0.007

Gender\_Male 0.1256 0.045 2.765 0.006 0.037 0.215

Age 0.0006 0.001 0.556 0.578 -0.002 0.003

Income 5.476e-07 8.01e-07 0.684 0.494 -1.02e-06 2.12e-06

Marital\_Never Married -0.0607 0.063 -0.957 0.339 -0.185 0.064

Marital\_Separated 0.0854 0.061 1.390 0.165 -0.035 0.206

Marital\_Widowed 0.0689 0.061 1.120 0.263 -0.052 0.189

Children -0.0048 0.011 -0.450 0.652 -0.026 0.016

const -0.9238 0.169 -5.467 0.000 -1.255 -0.593

=========================================================================================

I have decided to remove 'Children' as it has a p-value 0.6523847567206849

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9990

Method: MLE Df Model: 9

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001468

Time: 09:04:21 Log-Likelihood: -5773.7

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.04900

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0225 0.025 -0.913 0.361 -0.071 0.026

Item5 -0.0409 0.025 -1.659 0.097 -0.089 0.007

Gender\_Male 0.1256 0.045 2.765 0.006 0.037 0.215

Age 0.0006 0.001 0.556 0.578 -0.002 0.003

Income 5.476e-07 8.01e-07 0.684 0.494 -1.02e-06 2.12e-06

Marital\_Never Married -0.0607 0.063 -0.957 0.339 -0.185 0.064

Marital\_Separated 0.0854 0.061 1.390 0.165 -0.035 0.206

Marital\_Widowed 0.0689 0.061 1.120 0.263 -0.052 0.189

Children -0.0048 0.011 -0.450 0.652 -0.026 0.016

const -0.9238 0.169 -5.467 0.000 -1.255 -0.593

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577383

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9991

Method: MLE Df Model: 8

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001451

Time: 09:04:21 Log-Likelihood: -5773.8

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.03250

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0223 0.025 -0.905 0.365 -0.071 0.026

Item5 -0.0408 0.025 -1.655 0.098 -0.089 0.008

Gender\_Male 0.1255 0.045 2.763 0.006 0.036 0.215

Age 0.0006 0.001 0.570 0.569 -0.002 0.003

Income 5.444e-07 8.01e-07 0.680 0.497 -1.03e-06 2.11e-06

Marital\_Never Married -0.0607 0.063 -0.956 0.339 -0.185 0.064

Marital\_Separated 0.0854 0.061 1.390 0.164 -0.035 0.206

Marital\_Widowed 0.0689 0.061 1.121 0.262 -0.052 0.189

const -0.9354 0.167 -5.601 0.000 -1.263 -0.608

=========================================================================================

I have decided to remove 'Age' as it has a p-value 0.5688865468764898

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9991

Method: MLE Df Model: 8

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001451

Time: 09:04:21 Log-Likelihood: -5773.8

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.03250

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0223 0.025 -0.905 0.365 -0.071 0.026

Item5 -0.0408 0.025 -1.655 0.098 -0.089 0.008

Gender\_Male 0.1255 0.045 2.763 0.006 0.036 0.215

Age 0.0006 0.001 0.570 0.569 -0.002 0.003

Income 5.444e-07 8.01e-07 0.680 0.497 -1.03e-06 2.11e-06

Marital\_Never Married -0.0607 0.063 -0.956 0.339 -0.185 0.064

Marital\_Separated 0.0854 0.061 1.390 0.164 -0.035 0.206

Marital\_Widowed 0.0689 0.061 1.121 0.262 -0.052 0.189

const -0.9354 0.167 -5.601 0.000 -1.263 -0.608

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577399

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9992

Method: MLE Df Model: 7

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001423

Time: 09:04:21 Log-Likelihood: -5774.0

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.02128

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0222 0.025 -0.900 0.368 -0.070 0.026

Item5 -0.0408 0.025 -1.658 0.097 -0.089 0.007

Gender\_Male 0.1256 0.045 2.767 0.006 0.037 0.215

Income 5.424e-07 8.01e-07 0.677 0.498 -1.03e-06 2.11e-06

Marital\_Never Married -0.0606 0.063 -0.955 0.340 -0.185 0.064

Marital\_Separated 0.0852 0.061 1.386 0.166 -0.035 0.206

Marital\_Widowed 0.0687 0.061 1.118 0.264 -0.052 0.189

const -0.9024 0.157 -5.762 0.000 -1.209 -0.595

=========================================================================================

I have decided to remove 'Income' as it has a p-value 0.4981167825573355

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9992

Method: MLE Df Model: 7

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001423

Time: 09:04:21 Log-Likelihood: -5774.0

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.02128

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0222 0.025 -0.900 0.368 -0.070 0.026

Item5 -0.0408 0.025 -1.658 0.097 -0.089 0.007

Gender\_Male 0.1256 0.045 2.767 0.006 0.037 0.215

Income 5.424e-07 8.01e-07 0.677 0.498 -1.03e-06 2.11e-06

Marital\_Never Married -0.0606 0.063 -0.955 0.340 -0.185 0.064

Marital\_Separated 0.0852 0.061 1.386 0.166 -0.035 0.206

Marital\_Widowed 0.0687 0.061 1.118 0.264 -0.052 0.189

const -0.9024 0.157 -5.762 0.000 -1.209 -0.595

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577422

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9993

Method: MLE Df Model: 6

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001383

Time: 09:04:21 Log-Likelihood: -5774.2

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.01377

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0225 0.025 -0.914 0.360 -0.071 0.026

Item5 -0.0408 0.025 -1.657 0.097 -0.089 0.007

Gender\_Male 0.1249 0.045 2.751 0.006 0.036 0.214

Marital\_Never Married -0.0606 0.063 -0.955 0.340 -0.185 0.064

Marital\_Separated 0.0847 0.061 1.379 0.168 -0.036 0.205

Marital\_Widowed 0.0686 0.061 1.115 0.265 -0.052 0.189

const -0.8791 0.153 -5.754 0.000 -1.178 -0.580

=========================================================================================

I have decided to remove 'Item4' as it has a p-value 0.3604685309837652

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9993

Method: MLE Df Model: 6

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001383

Time: 09:04:21 Log-Likelihood: -5774.2

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.01377

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item4 -0.0225 0.025 -0.914 0.360 -0.071 0.026

Item5 -0.0408 0.025 -1.657 0.097 -0.089 0.007

Gender\_Male 0.1249 0.045 2.751 0.006 0.036 0.214

Marital\_Never Married -0.0606 0.063 -0.955 0.340 -0.185 0.064

Marital\_Separated 0.0847 0.061 1.379 0.168 -0.036 0.205

Marital\_Widowed 0.0686 0.061 1.115 0.265 -0.052 0.189

const -0.8791 0.153 -5.754 0.000 -1.178 -0.580

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577464

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9994

Method: MLE Df Model: 5

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001311

Time: 09:04:21 Log-Likelihood: -5774.6

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.009699

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item5 -0.0310 0.022 -1.398 0.162 -0.074 0.012

Gender\_Male 0.1263 0.045 2.785 0.005 0.037 0.215

Marital\_Never Married -0.0607 0.063 -0.955 0.339 -0.185 0.064

Marital\_Separated 0.0853 0.061 1.390 0.165 -0.035 0.206

Marital\_Widowed 0.0699 0.061 1.138 0.255 -0.051 0.190

const -0.9932 0.088 -11.260 0.000 -1.166 -0.820

=========================================================================================

I have decided to remove 'Marital\_Never Married' as it has a p-value 0.33935214514036127

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9994

Method: MLE Df Model: 5

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001311

Time: 09:04:21 Log-Likelihood: -5774.6

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.009699

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

Item5 -0.0310 0.022 -1.398 0.162 -0.074 0.012

Gender\_Male 0.1263 0.045 2.785 0.005 0.037 0.215

Marital\_Never Married -0.0607 0.063 -0.955 0.339 -0.185 0.064

Marital\_Separated 0.0853 0.061 1.390 0.165 -0.035 0.206

Marital\_Widowed 0.0699 0.061 1.138 0.255 -0.051 0.190

const -0.9932 0.088 -11.260 0.000 -1.166 -0.820

=========================================================================================

Optimization terminated successfully.

Current function value: 0.577510

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9995

Method: MLE Df Model: 4

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001232

Time: 09:04:21 Log-Likelihood: -5775.1

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.006555

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

Item5 -0.0307 0.022 -1.386 0.166 -0.074 0.013

Gender\_Male 0.1270 0.045 2.799 0.005 0.038 0.216

Marital\_Separated 0.1050 0.058 1.814 0.070 -0.008 0.219

Marital\_Widowed 0.0896 0.058 1.546 0.122 -0.024 0.203

const -1.0142 0.085 -11.864 0.000 -1.182 -0.847

=====================================================================================

I have decided to remove 'Item5' as it has a p-value 0.16582330272769474

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9995

Method: MLE Df Model: 4

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001232

Time: 09:04:21 Log-Likelihood: -5775.1

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.006555

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

Item5 -0.0307 0.022 -1.386 0.166 -0.074 0.013

Gender\_Male 0.1270 0.045 2.799 0.005 0.038 0.216

Marital\_Separated 0.1050 0.058 1.814 0.070 -0.008 0.219

Marital\_Widowed 0.0896 0.058 1.546 0.122 -0.024 0.203

const -1.0142 0.085 -11.864 0.000 -1.182 -0.847

=====================================================================================

Optimization terminated successfully.

Current function value: 0.577606

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9996

Method: MLE Df Model: 3

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001066

Time: 09:04:21 Log-Likelihood: -5776.1

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.006355

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

Gender\_Male 0.1277 0.045 2.814 0.005 0.039 0.217

Marital\_Separated 0.1053 0.058 1.818 0.069 -0.008 0.219

Marital\_Widowed 0.0878 0.058 1.516 0.130 -0.026 0.201

const -1.1212 0.037 -30.222 0.000 -1.194 -1.048

=====================================================================================

I have decided to remove 'Marital\_Widowed' as it has a p-value 0.12964055631668056

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9996

Method: MLE Df Model: 3

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.001066

Time: 09:04:21 Log-Likelihood: -5776.1

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.006355

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

Gender\_Male 0.1277 0.045 2.814 0.005 0.039 0.217

Marital\_Separated 0.1053 0.058 1.818 0.069 -0.008 0.219

Marital\_Widowed 0.0878 0.058 1.516 0.130 -0.026 0.201

const -1.1212 0.037 -30.222 0.000 -1.194 -1.048

=====================================================================================

Optimization terminated successfully.

Current function value: 0.577720

Iterations 5

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9997

Method: MLE Df Model: 2

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.0008682

Time: 09:04:21 Log-Likelihood: -5777.2

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.006602

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

Gender\_Male 0.1271 0.045 2.802 0.005 0.038 0.216

Marital\_Separated 0.0827 0.056 1.479 0.139 -0.027 0.192

const -1.0983 0.034 -32.525 0.000 -1.164 -1.032

=====================================================================================

I have decided to remove 'Marital\_Separated' as it has a p-value 0.1390820291759585

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9997

Method: MLE Df Model: 2

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.0008682

Time: 09:04:21 Log-Likelihood: -5777.2

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.006602

=====================================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------

Gender\_Male 0.1271 0.045 2.802 0.005 0.038 0.216

Marital\_Separated 0.0827 0.056 1.479 0.139 -0.027 0.192

const -1.0983 0.034 -32.525 0.000 -1.164 -1.032

=====================================================================================

## D3. Reduced Model

Below is the final reduced model:

Logit Regression Results

==============================================================================

Dep. Variable: Churn No. Observations: 10000

Model: Logit Df Residuals: 9998

Method: MLE Df Model: 1

Date: Sat, 09 Nov 2024 Pseudo R-squ.: 0.0006804

Time: 09:04:21 Log-Likelihood: -5778.3

converged: True LL-Null: -5782.2

Covariance Type: nonrobust LLR p-value: 0.005031

===============================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------

Gender\_Male 0.1272 0.045 2.805 0.005 0.038 0.216

const -1.0814 0.032 -34.094 0.000 -1.144 -1.019

===============================================================================

## E1. Comparison of Models

The feature reduction performed dramatically reduced the size of the model, with the final feature remaining in the model being Gender\_Male. While the reduced model is much narrower in scope, examining the LLR p-values between the models shows a significant increase in statistical relevance. The initial model showed an LLR p-value of 1.0 and the final was a value of 0.005031, which is well within the established alpha of 0.05. The result is a much smaller, much more significant reduced model.

## E2. Analysis Process and Output

I used the following code to generate a confusion matrix and accuracy score:

#Confusion matrix and accuracy

#https://stackoverflow.com/questions/39770376/scikit-learn-get-accuracy-scores-for-each-class

y\_prob = results.predict(x)

y\_pred = (y\_prob).astype(int)

matrix = confusion\_matrix(y, y\_pred)

print(matrix)

print(f'Accuracy: {matrix.diagonal()/matrix.sum(axis=1)}')

Below is the resulting output:

[[7350 0]

[2650 0]]

Accuracy: [1. 0.]

## E3. Code

*Code included in D208\_Task\_2.ipynb*

# Part V: Data Summary and Implications

## F1. Results of the Analysis

***In = -1.0814 + 0.1272Gender\_Male***

Keeping all things constant, for one unit increase in Gender\_Male, the changes log odds of Churn by 0.1272.

The statistical significance is unassailable with an accuracy score of 100% and a p-value of well below the established alpha. However, the practical significance of the model is limited. There is very little direct action that can be performed based solely on the information provided by the model and the reduction process cut the feature set down to the point that predictions based on the model have very little room for recourse.

In terms of the limitation of the analysis, the primary limitations are that we have established correlation between the independent and dependent variables, but we have not established causation. Causation cannot be ascertained from a simple survey of data and requires a rigorous process of experimentation that isolates the variables within the model and observe outcomes in a controlled environment. Another limitation comes down to the scope. In an effort to generate a highly accurate model, I ended up having to make a very small model. The size of the model and number of included features create an issue in terms of recommending prescriptive action based on the analysis.

## F2. Recommended Action

We have determined that a customer being male can potentially predict customer churn outcomes. The primary problem here is that there is limited action that can be taken without getting deeper into the issue. Therefore I propose two actions.

Conduct a second study with a scope that includes factors broad enough to capture the potential issues found with customer satisfaction. Gender should be an included metric, but men should not be specifically targeted for the study as that could introduce bias to any analysis performed. Random sampling should be done across the user population.

Conduct a second analysis with the existing data that casts a wider net. Starting with a larger set of features could potentially result in a larger reduced model that is more useful in predicting churn outcomes.

# Part VI: Demonstration

## G. Panopto Video

*A Panopto video has been included with the submission*

## H. Code Sources

Gerald Burke – D208 Task 1

<https://www.geeksforgeeks.org/plotting-multiple-bar-charts-using-matplotlib-in-python/>

https://stackoverflow.com/questions/39770376/scikit-learn-get-accuracy-scores-for-each-class

## I. Sources

Gerald Burke – D208 Task 1

Dr. Middleton’s D208 Part 1 and 2