D208 Task 2

July 25, 2021

1 Part I: Research Question

1.1 A1: Question

What customer qualities or factors from the data can be used to predict the churn of customers?

1.2 A2: Objective and Goals

The objective of an analysis of the data is to determine what features, if any, can significantly predict whether or not the customer will continue services with the organization.

2 Part II: Method Justification

2.1 B1: Summary of Assumptions

2.1.1 Assumption #1: Independence of Observations

In logistic regression, the normal distribution of the observational errors is not assumed, however, the independence of the observations themselves are still assumed. In simple terms, the values of each variable must not have an effect on the other values within the same variable. If errors become correlated or there are duplicate observations in our dataset the standard error cannot be relied upon. The independence of the observation is a scientific method issue and not a statistical one, but it should be validated in the dataset.

2.1.2 Assumption #2: Linearity in the Logit for Continous Independent Variables

Unlike in multiple linear regression, logistic regression does not require the independent variable to be linearly related to the dependent variable. It does, however, assume linearity in the logit for any continuous independent variables. This means that there should be a linear relationship between each independent variable and the log odds of the dependent variable.

2.1.3 Assumption #3: Absence of Muliticollinearity Among Independent Variables

Multicollinearity in the model occurs when two or more independent variables share the same correlation with the target variable. This usually means that the independent variables are related and explain the same variance with the target variable. A logistic regression model with highly correlated independent variables will usually result in large standard errors for the estimated beta coefficients (or slopes) of these variables. (Stoltzfus, 2011)

2.1.4 Assumption #4: No Heavily Influential Outliers

Outliers are any abnormal or unusual data values when compared to other date values within an independent variable. Normally, an outlier can be confirmed if it is three or more standard deviations away from a statistic, but this can be subjective depending on the data. Outliers in the dataset can create inaccuracies within the model and should either be removed, changed, or left untouched with model notation, depending on desired results.

2.2 B2: Benefits of Using Python

By using Python, data can be easily cleaned, explored, and prepared for use in predictive model building. The models themselves can be created using Python. Plots, charts, and graphs can be created to visualize the data and better understand relationships within datasets. This creates opportunities to provide detailed visual information for presentations. Python contains many packages built by data scientists that help with the previously mentioned tasks. Some packages that will be used are Numpy, Pandas, Matplotlib, Seaborn, Statsmodels, and Sklearn.

2.3 B3: Why Logistic Regression?

Since our dependent variable "Churn" is a dichotomous variable, it is appropriate to use the logistic regression model to achieve answering the question. The dataset contains many other continuous and categorical variables that can be used to build the logistic regression model. The difference between the multiple linear regression model and the logistic regression model is that instead of trying to predict the variables' outcome, the linear regression model predicts the *probability* of the outcome of the variable.

3 Part III: Data Preparation

3.1 C1: Data Preparation Goals and Manipulation

The overall goal of data preparation is to ensure that the data that will be used for the logistic regression model is complete, accurate, and efficiently used. If the data used to create and input into the model are garbage, garbage will be returned from the model. Some data manipulation tasks that need to be completed for data preparation to conduct logistic regression are:

- Import the dataset
- Identify and handle missing data
- Identify and handle outliers or strange values
- Transform categorical variables in numerical values and drop a selection for each categorical variable
- Ensure the target variable is categorical

3.2 C2: Summary Statistics

The below statistical factors of the models will be needed to help answer the research question and were derived from the course textbook:

- Coefficients of all predictor variables - the coefficients of each independent variable will need to be determined to build the logistic regression model.

- Log-Likelihood - Effective for comparing models of the same data. A model with a log-likelihood closer to zero is a better candidate for fit.
- p-values - these indicate the probability of observing the test statistic assuming the null hypothesis that the population coefficient is zero. Normally, p-values less than 0.5 indicate that the null hypothesis can be rejected and the statistic can be used in the model.
- Pseudo R-squared statistic: The Pseudo R-squared statistic is an analogy to linear regressions R-squared but does not measure the proportion of variation in the dependent variable explained by the model. It is instead computed based on the ratio of the maximized log-likelihood function for the null model and the full model. Computed to be between 0 and 1, values closer to 1 indicate a better fitting model.
- Variance inflation factor(VIF) - provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

Independent Variables - Population (Continuous)

- Area (Categorical)
- Children (Continuous)
- Age (Continuous)
- Income (Continuous)
- Marital (Categorical)
- Gender (Categorical)
- State (Categorical)
- Outage_sec_perweek (Continuous)
- Email (Continuous)
- Contract (Categorical)
- Contacts (Continuous)
- Yearly_equip_failure (Continuous)
- Techie (Categorical)
- Tenure (Continuous)
- Port_modem (Categorical)
- Tablet (Categorical)
- InternetService (Categorical)
- Phone (Categorical)
- Multiple (Categorical)
- OnlineSecurity (Categorical)
- OnlineBackup (Categorical)

- DeviceProtection (Categorical)
- TechSupport (Categorical)
- StreamingTV (Categorical)
- StreamingMovies (Categorical)
- PaperlessBilling (Categorical)
- PaymentMethod (Categorical)
- Tenure (Categorical)
- MonthlyCharge (Continuous)
- Bandwidth_GB_Year (Continuous)
- Item1 (Categorical)
- Item2 (Categorical)
- Item3 (Categorical)
- Item4 (Categorical)
- Item5 (Categorical)
- Item6 (Categorical)
- Item7 (Categorical)
- Item8 (Categorical)

Target Variable

- Churn (Categorical)

3.3 C3: Steps to Prepare the Data for Analysis

[202]:
```
#Load packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import statsmodels.api as sm
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
```

```
%matplotlib inline
```

3.3.1 1) Import the Original Dataset

[203]:
```
#Load the dataset
churn_clean = pd.read_csv("C:/Users/holtb/Data/WGU Datasets/churn_clean.csv")
```

[204]:
```
len(churn_clean.columns)
```

[204]: 50

3.3.2 2) Drop Unused Variables and Convert Catergorical Variables

[205]:
```
#Drop unused variables
churn_model_data = churn_clean.
↪drop(['CaseOrder','Customer_id','Interaction','UID','City','County','Lat','Lng',
                   'TimeZone','Job','Zip'], axis=1)
```

[206]:
```
len(churn_model_data.columns)
```

[206]: 39

[207]:
```
#Converting Item# variables to categorical variables
churn_model_data[['Item1','Item2','Item3',
                  'Item4','Item5','Item6',
                  'Item7','Item8']] = churn_model_data[['Item1','Item2','Item3',
                                                        'Item4','Item5','Item6',
                                                        'Item7','Item8']].
↪astype('category')
```

3.3.3 3) Identify Missing Data

[208]:
```
display(churn_model_data.isnull().any())
```

```
State                   False
Population              False
Area                    False
Children                False
Age                     False
Income                  False
Marital                 False
Gender                  False
Churn                   False
Outage_sec_perweek      False
Email                   False
Contacts                False
Yearly_equip_failure    False
```

```
Techie                  False
Contract                False
Port_modem              False
Tablet                  False
InternetService         False
Phone                   False
Multiple                False
OnlineSecurity          False
OnlineBackup            False
DeviceProtection        False
TechSupport             False
StreamingTV             False
StreamingMovies         False
PaperlessBilling        False
PaymentMethod           False
Tenure                  False
MonthlyCharge           False
Bandwidth_GB_Year       False
Item1                   False
Item2                   False
Item3                   False
Item4                   False
Item5                   False
Item6                   False
Item7                   False
Item8                   False
dtype: bool
```

3.3.4 4) Identify and Handle Outliers

[209]:

```
#Dropping categorical variables
churn_continuous_data = churn_model_data.
 ↪drop(['Area','Marital','Gender','Churn', 'Techie','Contract', 'Port_modem',
                                            'Tablet', 'InternetService', 'Phone',␣
 ↪'Multiple','OnlineSecurity', 'OnlineBackup',
                                           ␣
 ↪'DeviceProtection','TechSupport','TechSupport', 'StreamingTV',␣
 ↪'StreamingMovies',
                                           ␣
 ↪'PaperlessBilling','PaymentMethod','Item1','Item2','Item3','Item4','Item5','Item6',
                                            'Item7','Item8'], axis=1)
```

[210]:

```
churn_continuous_data.describe()
```

[210]:

```
       Population  Children       Age    Income  Outage_sec_perweek    Email  \
count    10000.00  10000.00  10000.00  10000.00            10000.00 10000.00
mean      9756.56      2.09     53.08  39806.93               10.00    12.02
```

```
std     14432.70      2.15     20.70   28199.92                2.98      3.03
min         0.00      0.00     18.00     348.67                0.10      1.00
25%       738.00      0.00     35.00   19224.72                8.02     10.00
50%      2910.50      1.00     53.00   33170.60               10.02     12.00
75%     13168.00      3.00     71.00   53246.17               11.97     14.00
max    111850.00     10.00     89.00  258900.70               21.21     23.00

       Contacts  Yearly_equip_failure   Tenure  MonthlyCharge  \
count  10000.00              10000.00 10000.00       10000.00
mean       0.99                  0.40    34.53         172.62
std        0.99                  0.64    26.44          42.94
min        0.00                  0.00     1.00          79.98
25%        0.00                  0.00     7.92         139.98
50%        1.00                  0.00    35.43         167.48
75%        2.00                  1.00    61.48         200.73
max        7.00                  6.00    72.00         290.16

       Bandwidth_GB_Year
count           10000.00
mean             3392.34
std              2185.29
min               155.51
25%              1236.47
50%              3279.54
75%              5586.14
max              7158.98
```

[212]:
```python
len(churn_continuous_data.columns)
```

[212]:
```
12
```

3.3.5 5)Transform Categorical Data

[213]:
```python
churn_categorical_data = churn_model_data.
↪drop(['Area','Marital','Gender','Churn', 'Techie','Contract', 'Port_modem',
                                          'Tablet', 'InternetService', 'Phone',␣
↪'Multiple','OnlineSecurity', 'OnlineBackup',
                                         ␣
↪'DeviceProtection','TechSupport','TechSupport', 'StreamingTV',␣
↪'StreamingMovies',
                                         ␣
↪'PaperlessBilling','Item1','Item2','Item3','Item4','Item5','Item6','Item7',
                                          'Item8'], axis =1)

churn_categorical_data.head()
```

[213]:

```
  State  Population  Children  Age    Income  Outage_sec_perweek  Email  \
0    AK          38         0   68  28561.99                7.98     10
1    MI       10446         1   27  21704.77               11.70     12
2    OR        3735         4   50   9609.57               10.75      9
3    CA       13863         1   48  18925.23               14.91     15
4    TX       11352         0   83  40074.19                8.15     16

   Contacts  Yearly_equip_failure             PaymentMethod  Tenure  \
0         0                     1   Credit Card (automatic)    6.80
1         0                     1  Bank Transfer(automatic)    1.16
2         0                     1   Credit Card (automatic)   15.75
3         2                     0              Mailed Check   17.09
4         2                     1              Mailed Check    1.67

   MonthlyCharge  Bandwidth_GB_Year
0         172.46             904.54
1         242.63             800.98
2         159.95            2054.71
3         119.96            2164.58
4         149.95             271.49
```

[214]:

```
#Transform categorical variables to numeric using dummy variables
churn_model_transdata = pd.get_dummies(churn_model_data, columns =␣
 ↪['State','Area','Marital','Gender','Churn','PaymentMethod',
                                                                        ␣
 ↪'Techie','Contract','Port_modem', 'Tablet',
                                                                        ␣
 ↪'InternetService','Phone','Multiple','OnlineSecurity',
                                                                        ␣
 ↪'OnlineBackup','DeviceProtection','TechSupport',
                                                                        ␣
 ↪'StreamingTV', 'StreamingMovies',
                                                                        ␣
 ↪'PaperlessBilling','Item1','Item2','Item3','Item4',
                                                                        ␣
 ↪'Item5','Item6','Item7','Item8'])
```

[215]:

```
#Show current columns
for col in churn_model_transdata.columns:
    print(col)
```

```
Population
Children
Age
Income
Outage_sec_perweek
Email
```

```
Contacts
Yearly_equip_failure
Tenure
MonthlyCharge
Bandwidth_GB_Year
State_AK
State_AL
State_AR
State_AZ
State_CA
State_CO
State_CT
State_DC
State_DE
State_FL
State_GA
State_HI
State_IA
State_ID
State_IL
State_IN
State_KS
State_KY
State_LA
State_MA
State_MD
State_ME
State_MI
State_MN
State_MO
State_MS
State_MT
State_NC
State_ND
State_NE
State_NH
State_NJ
State_NM
State_NV
State_NY
State_OH
State_OK
State_OR
State_PA
State_PR
State_RI
State_SC
State_SD
```

```
State_TN
State_TX
State_UT
State_VA
State_VT
State_WA
State_WI
State_WV
State_WY
Area_Rural
Area_Suburban
Area_Urban
Marital_Divorced
Marital_Married
Marital_Never Married
Marital_Separated
Marital_Widowed
Gender_Female
Gender_Male
Gender_Nonbinary
Churn_No
Churn_Yes
PaymentMethod_Bank Transfer(automatic)
PaymentMethod_Credit Card (automatic)
PaymentMethod_Electronic Check
PaymentMethod_Mailed Check
Techie_No
Techie_Yes
Contract_Month-to-month
Contract_One year
Contract_Two Year
Port_modem_No
Port_modem_Yes
Tablet_No
Tablet_Yes
InternetService_DSL
InternetService_Fiber Optic
InternetService_None
Phone_No
Phone_Yes
Multiple_No
Multiple_Yes
OnlineSecurity_No
OnlineSecurity_Yes
OnlineBackup_No
OnlineBackup_Yes
DeviceProtection_No
DeviceProtection_Yes
```

```
TechSupport_No
TechSupport_Yes
StreamingTV_No
StreamingTV_Yes
StreamingMovies_No
StreamingMovies_Yes
PaperlessBilling_No
PaperlessBilling_Yes
Item1_1
Item1_2
Item1_3
Item1_4
Item1_5
Item1_6
Item1_7
Item2_1
Item2_2
Item2_3
Item2_4
Item2_5
Item2_6
Item2_7
Item3_1
Item3_2
Item3_3
Item3_4
Item3_5
Item3_6
Item3_7
Item3_8
Item4_1
Item4_2
Item4_3
Item4_4
Item4_5
Item4_6
Item4_7
Item5_1
Item5_2
Item5_3
Item5_4
Item5_5
Item5_6
Item5_7
Item6_1
Item6_2
Item6_3
Item6_4
```

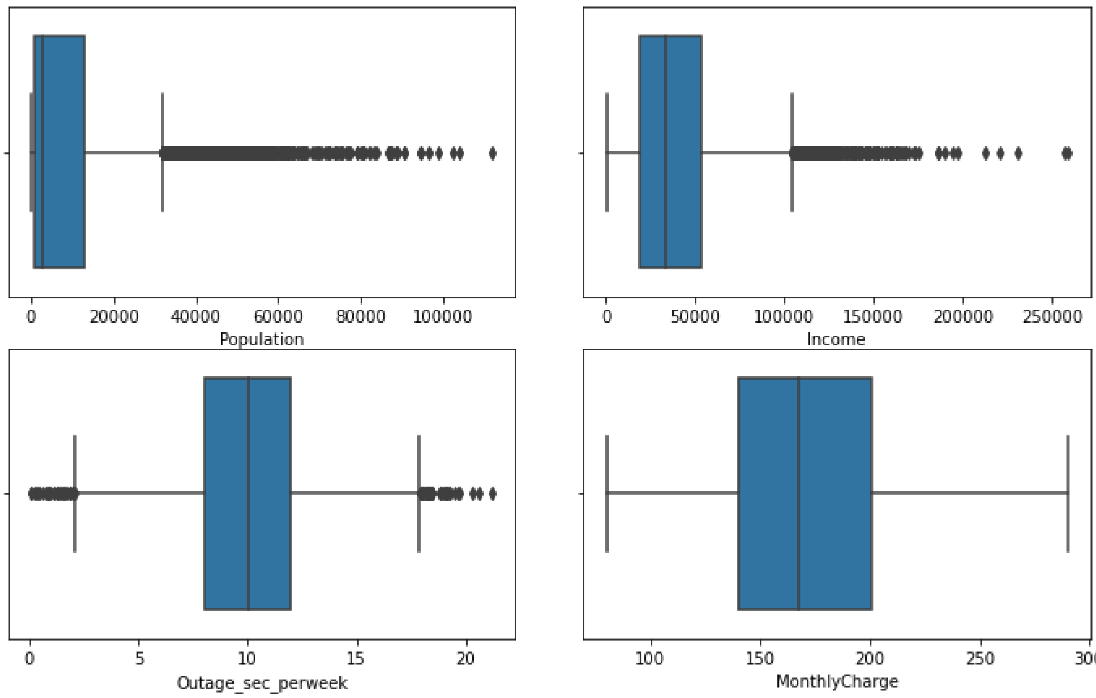
```
Item6_5
Item6_6
Item6_7
Item6_8
Item7_1
Item7_2
Item7_3
Item7_4
Item7_5
Item7_6
Item7_7
Item8_1
Item8_2
Item8_3
Item8_4
Item8_5
Item8_6
Item8_7
Item8_8
```

[216]:
```
#Dropping one column per catergorical variable to meet n-1 requirements
churn_LRM_data = churn_model_transdata.
 ↪drop(['State_AK','Area_Rural','Marital_Widowed','Gender_Nonbinary','Techie_No',
                                             'Contract_Two␣
 ↪Year','Port_modem_No','Tablet_No','InternetService_None',
                                            ␣
 ↪'Phone_No','Multiple_No','OnlineSecurity_No','DeviceProtection_No','TechSupport_No',
                                             'StreamingTV_No',␣
 ↪'StreamingMovies_No','PaperlessBilling_No','Item1_1','Item2_1',
                                            ␣
 ↪'Item3_1','Item4_1','Item5_1','Item6_1','Item7_1','Item8_1',
                                              'PaymentMethod_Bank␣
 ↪Transfer(automatic)','Churn_No','OnlineBackup_No'], axis = 1)
```

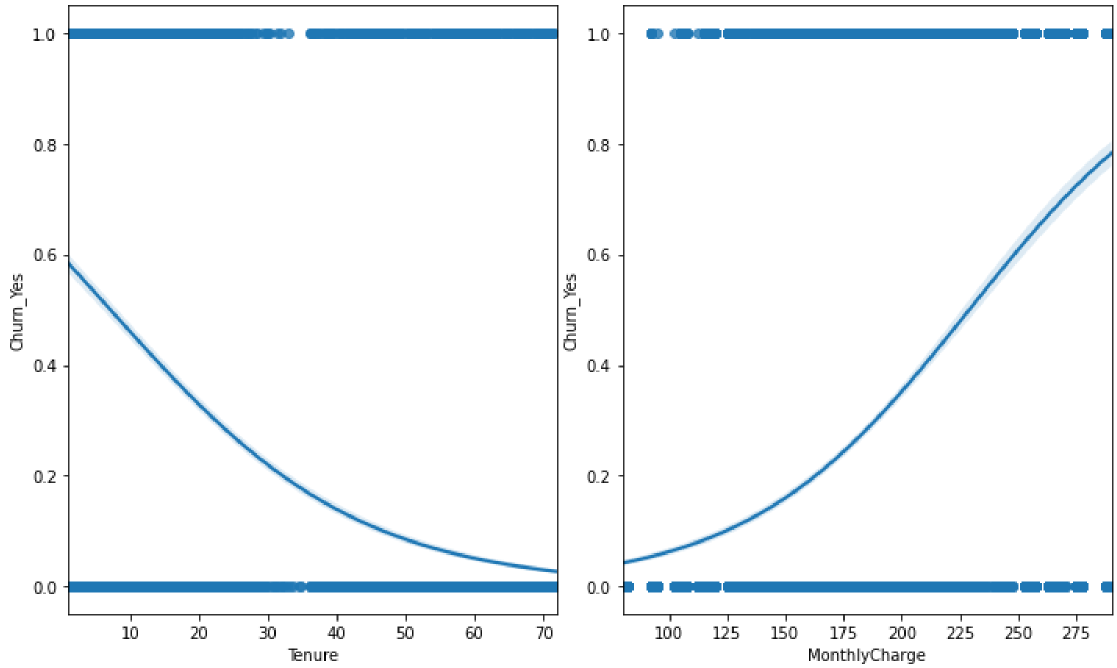
3.4 C4: Univariate and Bivariate Visualizations

[217]:
```
fig, axs = plt.subplots(2,2, figsize=(10,6))
plt.tight_layout()

sns.boxplot(x="Population",
            data = churn_continuous_data,
            ax = axs[0,0])
sns.boxplot(x="Income",
            data = churn_continuous_data,
            ax = axs[0,1])

sns.boxplot(x="Outage_sec_perweek",
```

```
            data = churn_continuous_data,
            ax = axs[1,0])

sns.boxplot(x="MonthlyCharge",
            data = churn_continuous_data,
            ax = axs[1,1]);
```



[218]:
```
fig, axs = plt.subplots(1,2, figsize=(10,6))
plt.tight_layout()

sns.regplot(x='Tenure',
            y='Churn_Yes',
            data=churn_LRM_data,
            logistic=True,
            ax = axs[0] )

sns.regplot(x='MonthlyCharge',
            y='Churn_Yes',
            data=churn_LRM_data,
            logistic=True,
            ax = axs[1]);
```



3.5 C5: Copy of Prepared Dataset

[219]:
```
churn_LRM_data.to_csv('C:/Users/holtb/Data/D208/Task 2/churn_LRM_data.csv')
```

4 Part IV: Model Comparison And Analysis

4.1 Initial Model:

[220]:
```
X = churn_LRM_data.drop(['Churn_Yes'], axis=1)
y = churn_LRM_data['Churn_Yes']
```

[221]:
```
#define the input
X2 = sm.add_constant(X)

#create an Logistic Regression Model
initial_model = sm.Logit(y.astype("float64"), X2.astype("float64"))

#fit the data
initial_est = initial_model.fit()

#Summarize the output
initial_est.summary()
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.212166
```

```
        Iterations: 35
C:\Users\holtb\anaconda3\lib\site-packages\statsmodels\base\model.py:566:
ConvergenceWarning: Maximum Likelihood optimization failed to converge. Check
mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
```

[221]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:              Churn_Yes   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9859
Method:                           MLE   Df Model:                          140
Date:                Sun, 25 Jul 2021   Pseudo R-squ.:                  0.6331
Time:                        15:46:34   Log-Likelihood:                -2121.7
converged:                      False   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
=========================
                                           coef    std err          z
P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
-------------------------
const                                   -9.9742      1.087     -9.174
0.000     -12.105      -7.843
Population                            1.177e-06   2.98e-06      0.395
0.693   -4.66e-06    7.01e-06
Children                                -0.0209      0.141     -0.149
0.882      -0.297       0.255
Age                                      0.0048      0.015      0.321
0.748      -0.025       0.034
Income                                3.907e-07    1.4e-06      0.279
0.780   -2.35e-06    3.13e-06
Outage_sec_perweek                      -0.0030      0.013     -0.223
0.824      -0.029       0.023
Email                                   -0.0129      0.013     -0.997
0.319      -0.038       0.013
Contacts                                 0.0605      0.040      1.514
0.130      -0.018       0.139
Yearly_equip_failure                    -0.0297      0.062     -0.476
0.634      -0.152       0.093
Tenure                                  -0.2208      0.372     -0.593
0.553      -0.950       0.509
MonthlyCharge                            0.0381      0.014      2.688
0.007       0.010       0.066
Bandwidth_GB_Year                        0.0012      0.005      0.271
0.786      -0.008       0.010
```

State_AL	-0.0831	0.501	-0.166	0.868	-1.065	0.899
State_AR	-0.4337	0.499	-0.870	0.384	-1.411	0.543
State_AZ	0.0178	0.574	0.031	0.975	-1.106	1.142
State_CA	-0.1793	0.436	-0.411	0.681	-1.034	0.676
State_CO	-0.0775	0.517	-0.150	0.881	-1.090	0.935
State_CT	-0.0802	0.597	-0.134	0.893	-1.250	1.090
State_DC	1.0902	0.951	1.146	0.252	-0.774	2.955
State_DE	-0.8449	0.897	-0.942	0.346	-2.603	0.913
State_FL	-0.4511	0.463	-0.973	0.330	-1.359	0.457
State_GA	0.1587	0.476	0.334	0.739	-0.774	1.091
State_HI	-0.1788	0.754	-0.237	0.813	-1.657	1.300
State_IA	-0.0230	0.460	-0.050	0.960	-0.924	0.878
State_ID	-0.1234	0.605	-0.204	0.838	-1.309	1.063
State_IL	-0.1329	0.441	-0.301	0.763	-0.998	0.732
State_IN	0.0528	0.472	0.112	0.911	-0.871	0.977
State_KS	0.1159	0.510	0.227	0.820	-0.883	1.115
State_KY	-0.0112	0.476	-0.024	0.981	-0.943	0.921
State_LA	-0.0989	0.503	-0.197	0.844	-1.084	0.886
State_MA	-0.3051	0.498	-0.612	0.540	-1.282	0.672
State_MD	0.1935	0.523	0.370	0.711	-0.831	1.218
State_ME	-0.0387	0.550	-0.070	0.944	-1.117	1.039
State_MI	0.0366	0.462	0.079	0.937	-0.868	0.942
State_MN	0.0351	0.467	0.075	0.940	-0.880	0.950
State_MO	0.1914	0.447	0.428			

				0.669	-0.685	1.068
State_MS	-0.0580	0.514	-0.113	0.910	-1.066	0.950
State_MT	0.5291	0.562	0.942	0.346	-0.572	1.630
State_NC	-0.1329	0.456	-0.292	0.770	-1.026	0.760
State_ND	0.0495	0.505	0.098	0.922	-0.940	1.039
State_NE	0.0365	0.503	0.073	0.942	-0.949	1.022
State_NH	-0.3308	0.610	-0.542	0.588	-1.527	0.865
State_NJ	-0.1267	0.492	-0.257	0.797	-1.091	0.838
State_NM	-0.6384	0.517	-1.235	0.217	-1.652	0.375
State_NV	-0.7622	0.801	-0.952	0.341	-2.332	0.807
State_NY	-0.3376	0.432	-0.782	0.434	-1.183	0.508
State_OH	-0.2495	0.447	-0.558	0.577	-1.126	0.627
State_OK	-0.2352	0.481	-0.489	0.625	-1.178	0.708
State_OR	0.4048	0.529	0.766	0.444	-0.631	1.441
State_PA	-0.2220	0.431	-0.515	0.607	-1.067	0.624
State_PR	-0.4585	0.779	-0.589	0.556	-1.984	1.068
State_RI	-4.5167	1.554	-2.906	0.004	-7.563	-1.470
State_SC	-0.0338	0.560	-0.060	0.952	-1.132	1.064
State_SD	-0.5758	0.576	-1.001	0.317	-1.704	0.552
State_TN	0.4684	0.503	0.931	0.352	-0.517	1.454
State_TX	0.0999	0.429	0.233	0.816	-0.741	0.940
State_UT	0.0832	0.660	0.126	0.900	-1.211	1.378
State_VA	0.0573	0.466	0.123	0.902	-0.855	0.970
State_VT	0.3642	0.599	0.608	0.543	-0.811	1.539

```
State_WA                                  0.2176      0.486       0.447
0.655      -0.736       1.171
State_WI                                  0.0289      0.469       0.062
0.951      -0.890       0.948
State_WV                                  0.6429      0.469       1.372
0.170      -0.276       1.562
State_WY                                 -0.2197      0.771      -0.285
0.776      -1.731       1.292
Area_Suburban                            -0.0406      0.098      -0.415
0.678      -0.232       0.151
Area_Urban                                0.0569      0.097       0.586
0.558      -0.133       0.247
Marital_Divorced                         -0.2908      0.123      -2.356
0.018      -0.533      -0.049
Marital_Married                          -0.1524      0.126      -1.207
0.227      -0.400       0.095
Marital_Never Married                    -0.2520      0.126      -1.998
0.046      -0.499      -0.005
Marital_Separated                        -0.1423      0.125      -1.141
0.254      -0.387       0.102
Gender_Female                             0.1332      0.292       0.456
0.649      -0.440       0.706
Gender_Male                               0.3036      0.478       0.635
0.525      -0.633       1.241
PaymentMethod_Credit Card (automatic)     0.2321      0.121       1.919
0.055      -0.005       0.469
PaymentMethod_Electronic Check            0.6479      0.109       5.964
0.000       0.435       0.861
PaymentMethod_Mailed Check                0.2634      0.119       2.210
0.027       0.030       0.497
Techie_Yes                                1.0994      0.105      10.437
0.000       0.893       1.306
Contract_Month-to-month                   3.6052      0.130      27.626
0.000       3.349       3.861
Contract_One year                         0.1173      0.141       0.832
0.406      -0.159       0.394
Port_modem_Yes                            0.1471      0.079       1.856
0.063      -0.008       0.302
Tablet_Yes                               -0.0682      0.087      -0.787
0.431      -0.238       0.102
InternetService_DSL                       0.5525      1.714       0.322
0.747      -2.807       3.912
InternetService_Fiber Optic              -1.1013      0.477      -2.311
0.021      -2.035      -0.167
Phone_Yes                                -0.3256      0.136      -2.390
0.017      -0.593      -0.059
Multiple_Yes                              0.4160      0.206       2.014
```

```
0.044       0.011        0.821
OnlineSecurity_Yes                        -0.3412      0.320     -1.067
0.286      -0.968        0.286
OnlineBackup_Yes                          -0.1255      0.185     -0.678
0.498      -0.488        0.237
DeviceProtection_Yes                      -0.1066      0.240     -0.445
0.656      -0.576        0.363
TechSupport_Yes                           -0.1956      0.178     -1.101
0.271      -0.544        0.153
StreamingTV_Yes                            1.0952      0.523      2.094
0.036       0.070        2.120
StreamingMovies_Yes                        1.3079      0.373      3.509
0.000       0.577        2.038
PaperlessBilling_Yes                       0.1490      0.080      1.852
0.064      -0.009        0.307
Item1_2                                   -0.0108      0.305     -0.035
0.972      -0.609        0.587
Item1_3                                   -0.2005      0.307     -0.652
0.514      -0.803        0.402
Item1_4                                   -0.1747      0.319     -0.547
0.584      -0.801        0.451
Item1_5                                   -0.1424      0.341     -0.418
0.676      -0.811        0.526
Item1_6                                   -0.0601      0.441     -0.136
0.891      -0.924        0.804
Item1_7                                    0.1279      1.117      0.114
0.909      -2.061        2.316
Item2_2                                    0.2879      0.315      0.913
0.361      -0.330        0.906
Item2_3                                    0.1875      0.316      0.593
0.553      -0.432        0.807
Item2_4                                    0.2701      0.324      0.832
0.405      -0.366        0.906
Item2_5                                    0.1517      0.344      0.441
0.659      -0.522        0.826
Item2_6                                    0.1974      0.436      0.452
0.651      -0.658        1.053
Item2_7                                    2.7383      1.560      1.755
0.079      -0.319        5.796
Item3_2                                   -0.1426      0.309     -0.462
0.644      -0.748        0.463
Item3_3                                   -0.0702      0.306     -0.229
0.819      -0.670        0.529
Item3_4                                   -0.1589      0.311     -0.511
0.609      -0.769        0.451
Item3_5                                   -0.0055      0.329     -0.017
0.987      -0.651        0.640
```

```
Item3_6                                          0.3683      0.424      0.869
0.385      -0.462        1.199
Item3_7                                         -1.3696      1.251     -1.094
0.274      -3.822        1.083
Item3_8                                        -21.3967    1.5e+04     -0.001
0.999   -2.95e+04    2.94e+04
Item4_2                                          0.0762      0.280      0.272
0.785      -0.473        0.625
Item4_3                                         -0.1900      0.271     -0.702
0.483      -0.721        0.341
Item4_4                                         -0.1034      0.274     -0.377
0.706      -0.640        0.433
Item4_5                                         -0.1464      0.291     -0.503
0.615      -0.717        0.424
Item4_6                                          0.1519      0.380      0.399
0.690      -0.594        0.898
Item4_7                                         -1.6345      1.209     -1.352
0.177      -4.005        0.736
Item5_2                                          0.0230      0.302      0.076
0.939      -0.569        0.615
Item5_3                                         -0.1331      0.296     -0.450
0.653      -0.713        0.447
Item5_4                                         -0.1490      0.300     -0.496
0.620      -0.738        0.440
Item5_5                                         -0.0754      0.319     -0.236
0.813      -0.701        0.550
Item5_6                                         -0.0108      0.409     -0.026
0.979      -0.813        0.791
Item5_7                                         -2.3750      1.192     -1.992
0.046      -4.712       -0.039
Item6_2                                          0.6039      0.327      1.849
0.064      -0.036        1.244
Item6_3                                          0.5050      0.322      1.566
0.117      -0.127        1.137
Item6_4                                          0.5662      0.329      1.723
0.085      -0.078        1.210
Item6_5                                          0.4069      0.344      1.183
0.237      -0.267        1.081
Item6_6                                          0.2536      0.423      0.599
0.549      -0.576        1.083
Item6_7                                          1.0309      1.407      0.733
0.464      -1.727        3.789
Item6_8                                        -18.7552    1.5e+04     -0.001
0.999   -2.94e+04    2.93e+04
Item7_2                                          0.2036      0.295      0.690
0.490      -0.374        0.782
Item7_3                                          0.2963      0.284      1.043
```

```
0.297      -0.260       0.853
Item7_4                                    0.0773      0.287      0.269
0.788      -0.486       0.640
Item7_5                                    0.3124      0.305      1.025
0.305      -0.285       0.910
Item7_6                                    0.4443      0.397      1.120
0.263      -0.333       1.222
Item7_7                                   -1.6012      1.373     -1.167
0.243      -4.291       1.089
Item8_2                                   -0.1517      0.278     -0.545
0.586      -0.697       0.394
Item8_3                                   -0.0777      0.266     -0.292
0.770      -0.600       0.444
Item8_4                                   -0.1910      0.269     -0.710
0.478      -0.718       0.336
Item8_5                                   -0.1390      0.284     -0.489
0.625      -0.696       0.418
Item8_6                                    0.0140      0.373      0.037
0.970      -0.717       0.745
Item8_7                                   -0.6697      0.953     -0.703
0.482      -2.537       1.198
Item8_8                                  -19.1840    1.5e+04     -0.001
0.999   -2.94e+04    2.93e+04
================================================================================
=========================

Possibly complete quasi-separation: A fraction 0.11 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
"""
```

4.1.1 Multicollinearity

The initial model had significant issues due to the multicollinearity between independent variables within the model. By calculating the variance inflation factors(VIF), variables with multicollinearity can be identified. Any variables calculated above 10 were removed one at a time and the VIFs were recalculated until all VIFs were below 10.

[222]:
```
# creating the data frames before and after removing variables that are␣
↪creating multicollinearity
churn_data_before = churn_LRM_data.drop(['Churn_Yes'], axis=1)

# adding a constant to the data frames as required for the VIF calculation
A1 = sm.tools.add_constant(churn_data_before)

# create the series
```

```
series_before = pd.Series([variance_inflation_factor(A1.values, i) for i in␣
↪range(A1.shape[1])], index=A1.columns)

# display the series
print('-'*100)
print('Data Before Removing Target and Redundant Variables')
print('-'*100)
pd.options.display.max_rows = 145
pd.set_option('display.float_format', '{:.2f}'.format)
display(series_before)
```

```
--------------------------------------------------------------------------------
--------------------
Data Before Removing Target and Redundant Variables
--------------------------------------------------------------------------------
--------------------

const                                     729.59
Population                                  1.16
Children                                   58.03
Age                                        61.10
Income                                      1.01
Outage_sec_perweek                          1.01
Email                                       1.01
Contacts                                    1.01
Yearly_equip_failure                        1.01
Tenure                                  61324.70
MonthlyCharge                             224.51
Bandwidth_GB_Year                       62386.64
State_AL                                    3.32
State_AR                                    3.25
State_AZ                                    2.45
State_CA                                    7.56
State_CO                                    2.99
State_CT                                    1.92
State_DC                                    1.19
State_DE                                    1.28
State_FL                                    5.14
State_GA                                    4.03
State_HI                                    1.46
State_IA                                    4.53
State_ID                                    2.05
State_IL                                    6.16
State_IN                                    4.06
State_KS                                    3.49
State_KY                                    4.02
State_LA                                    2.81
State_MA                                    3.21
```

State_MD	2.60
State_ME	2.44
State_MI	4.53
State_MN	4.35
State_MO	4.90
State_MS	2.63
State_MT	2.24
State_NC	4.55
State_ND	2.52
State_NE	3.32
State_NH	2.12
State_NJ	3.44
State_NM	2.47
State_NV	1.63
State_NY	7.85
State_OH	5.50
State_OK	3.59
State_OR	2.48
State_PA	7.75
State_PR	1.54
State_RI	1.25
State_SC	2.60
State_SD	2.30
State_TN	3.37
State_TX	8.39
State_UT	1.86
State_VA	4.60
State_VT	2.09
State_WA	3.25
State_WI	3.90
State_WV	4.13
State_WY	1.57
Area_Suburban	1.35
Area_Urban	1.35
Marital_Divorced	1.63
Marital_Married	1.59
Marital_Never Married	1.60
Marital_Separated	1.61
Gender_Female	12.96
Gender_Male	35.79
PaymentMethod_Credit Card (automatic)	1.55
PaymentMethod_Electronic Check	1.69
PaymentMethod_Mailed Check	1.59
Techie_Yes	1.01
Contract_Month-to-month	1.49
Contract_One year	1.49
Port_modem_Yes	1.01
Tablet_Yes	1.01

InternetService_DSL	422.44
InternetService_Fiber Optic	33.96
Phone_Yes	1.02
Multiple_Yes	6.05
OnlineSecurity_Yes	14.95
OnlineBackup_Yes	5.40
DeviceProtection_Yes	9.22
TechSupport_Yes	4.54
StreamingTV_Yes	43.81
StreamingMovies_Yes	21.44
PaperlessBilling_Yes	1.01
Item1_2	6.81
Item1_3	13.17
Item1_4	14.13
Item1_5	8.57
Item1_6	2.52
Item1_7	1.30
Item2_2	6.82
Item2_3	13.13
Item2_4	13.96
Item2_5	8.29
Item2_6	2.53
Item2_7	1.25
Item3_2	7.14
Item3_3	12.87
Item3_4	13.39
Item3_5	7.66
Item3_6	2.32
Item3_7	1.16
Item3_8	1.16
Item4_2	6.28
Item4_3	11.39
Item4_4	11.72
Item4_5	6.69
Item4_6	2.04
Item4_7	1.06
Item5_2	6.81
Item5_3	12.35
Item5_4	12.63
Item5_5	7.24
Item5_6	2.16
Item5_7	1.09
Item6_2	7.54
Item6_3	13.49
Item6_4	13.81
Item6_5	8.23
Item6_6	2.35
Item6_7	1.15

```
Item6_8                                        1.02
Item7_2                                        6.18
Item7_3                                       11.53
Item7_4                                       11.86
Item7_5                                        6.81
Item7_6                                        2.17
Item7_7                                        1.08
Item8_2                                        6.74
Item8_3                                       11.97
Item8_4                                       12.08
Item8_5                                        6.90
Item8_6                                        2.06
Item8_7                                        1.10
Item8_8                                        1.04
dtype: float64
```

[223]:
```
# creating the data frames before and after removing variables that are␣
 ↪creating multicollinearity
churn_data_after = churn_LRM_data.
 ↪drop(['Churn_Yes','Bandwidth_GB_Year','Gender_Female','Item1_3','Item2_4','Item5_4',
                                                    ␣
 ↪'Item8_4','Item6_3','Item3_3','Item4_3','Item7_3','MonthlyCharge'], axis=1)

# adding a constant to the data frames as required for the VIF calculation
A2 = sm.tools.add_constant(churn_data_after)

# create the series
series_after = pd.Series([variance_inflation_factor(A2.values, i) for i in␣
 ↪range(A2.shape[1])], index=A2.columns)

# display the series
print('-'*100)
print('Data After Removing Target and Redundant Variables')
print('-'*100)
pd.options.display.max_rows = 145
pd.set_option('display.float_format', '{:.2f}'.format)
display(series_after)
```

```
--------------------------------------------------------------------------------
--------------------
Data After Removing Target and Redundant Variables
--------------------------------------------------------------------------------
--------------------

const                                    220.79
Population                                 1.16
Children                                   1.01
Age                                        1.01
```

Income	1.01
Outage_sec_perweek	1.01
Email	1.01
Contacts	1.01
Yearly_equip_failure	1.01
Tenure	1.01
State_AL	3.32
State_AR	3.25
State_AZ	2.45
State_CA	7.56
State_CO	2.99
State_CT	1.92
State_DC	1.19
State_DE	1.28
State_FL	5.14
State_GA	4.03
State_HI	1.46
State_IA	4.53
State_ID	2.05
State_IL	6.15
State_IN	4.06
State_KS	3.49
State_KY	4.02
State_LA	2.81
State_MA	3.21
State_MD	2.59
State_ME	2.44
State_MI	4.53
State_MN	4.34
State_MO	4.90
State_MS	2.63
State_MT	2.24
State_NC	4.54
State_ND	2.52
State_NE	3.31
State_NH	2.12
State_NJ	3.44
State_NM	2.47
State_NV	1.63
State_NY	7.84
State_OH	5.50
State_OK	3.58
State_OR	2.48
State_PA	7.75
State_PR	1.54
State_RI	1.25
State_SC	2.60
State_SD	2.30

State_TN	3.37
State_TX	8.39
State_UT	1.86
State_VA	4.60
State_VT	2.09
State_WA	3.25
State_WI	3.90
State_WV	4.13
State_WY	1.57
Area_Suburban	1.35
Area_Urban	1.35
Marital_Divorced	1.63
Marital_Married	1.59
Marital_Never Married	1.60
Marital_Separated	1.61
Gender_Male	1.01
PaymentMethod_Credit Card (automatic)	1.55
PaymentMethod_Electronic Check	1.69
PaymentMethod_Mailed Check	1.58
Techie_Yes	1.01
Contract_Month-to-month	1.49
Contract_One year	1.49
Port_modem_Yes	1.01
Tablet_Yes	1.01
InternetService_DSL	1.74
InternetService_Fiber Optic	1.75
Phone_Yes	1.02
Multiple_Yes	1.01
OnlineSecurity_Yes	1.01
OnlineBackup_Yes	1.01
DeviceProtection_Yes	1.01
TechSupport_Yes	1.01
StreamingTV_Yes	1.01
StreamingMovies_Yes	1.01
PaperlessBilling_Yes	1.01
Item1_2	1.31
Item1_4	1.60
Item1_5	1.84
Item1_6	1.40
Item1_7	1.19
Item2_2	1.54
Item2_3	1.46
Item2_5	1.37
Item2_6	1.27
Item2_7	1.17
Item3_2	1.26
Item3_4	1.46
Item3_5	1.52

```
Item3_6                                    1.25
Item3_7                                    1.09
Item3_8                                    1.16
Item4_2                                    1.23
Item4_4                                    1.36
Item4_5                                    1.31
Item4_6                                    1.09
Item4_7                                    1.01
Item5_2                                    1.34
Item5_3                                    1.33
Item5_5                                    1.26
Item5_6                                    1.08
Item5_7                                    1.02
Item6_2                                    1.26
Item6_4                                    1.40
Item6_5                                    1.42
Item6_6                                    1.15
Item6_7                                    1.07
Item6_8                                    1.01
Item7_2                                    1.22
Item7_4                                    1.36
Item7_5                                    1.35
Item7_6                                    1.12
Item7_7                                    1.03
Item8_2                                    1.27
Item8_3                                    1.33
Item8_5                                    1.24
Item8_6                                    1.07
Item8_7                                    1.03
Item8_8                                    1.03
dtype: float64
```

[224]:

```
#Dropped target variable and indpendent variables causing multicollinarity␣
↪issues
X_next = churn_LRM_data.
↪drop(['Churn_Yes','Bandwidth_GB_Year','Gender_Female','Item1_3','Item2_4','Item5_4','Item8_
                        ␣
↪'Item6_3','Item3_3','Item4_3','Item7_3','MonthlyCharge'], axis=1)

y = churn_LRM_data['Churn_Yes']
```

[225]:

```
#define the input
X_next = sm.add_constant(X_next)

#create an Logistic Regression Model
next_model = sm.Logit(y.astype("float64"), X_next.astype("float64"))
```

```
#fit the data
next_est = next_model.fit()

#Summarize the output
next_est.summary()
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.216141
         Iterations: 35

C:\Users\holtb\anaconda3\lib\site-packages\statsmodels\base\model.py:566:
ConvergenceWarning: Maximum Likelihood optimization failed to converge. Check
mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
```

[225]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:              Churn_Yes   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9870
Method:                           MLE   Df Model:                          129
Date:                Sun, 25 Jul 2021   Pseudo R-squ.:                  0.6262
Time:                        15:48:09   Log-Likelihood:                -2161.4
converged:                      False   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
=========================
                                            coef    std err          z
P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
-------------------------
const                                    -6.7728      0.557    -12.169
0.000      -7.864      -5.682
Population                             1.572e-06   2.93e-06      0.536
0.592   -4.18e-06    7.32e-06
Children                                  0.0188      0.018      1.017
0.309      -0.017       0.055
Age                                       0.0013      0.002      0.686
0.493      -0.002       0.005
Income                                 5.912e-07   1.39e-06      0.425
0.671   -2.13e-06    3.32e-06
Outage_sec_perweek                       -0.0049      0.013     -0.370
0.711      -0.031       0.021
Email                                    -0.0120      0.013     -0.928
0.353      -0.037       0.013
Contacts                                  0.0591      0.040      1.494
0.135      -0.018       0.137
```

Yearly_equip_failure	-0.0432	0.062	-0.698	0.485	-0.164	0.078
Tenure	-0.1137	0.003	-39.631	0.000	-0.119	-0.108
State_AL	-0.2108	0.499	-0.423	0.673	-1.188	0.767
State_AR	-0.4225	0.495	-0.854	0.393	-1.392	0.547
State_AZ	-0.0173	0.567	-0.031	0.976	-1.128	1.094
State_CA	-0.1948	0.433	-0.450	0.653	-1.044	0.654
State_CO	-0.0680	0.513	-0.133	0.895	-1.073	0.937
State_CT	-0.0725	0.591	-0.123	0.902	-1.231	1.086
State_DC	0.9275	0.917	1.012	0.312	-0.869	2.724
State_DE	-0.9304	0.869	-1.070	0.285	-2.635	0.774
State_FL	-0.4432	0.459	-0.966	0.334	-1.342	0.456
State_GA	0.1220	0.472	0.258	0.796	-0.804	1.048
State_HI	-0.1164	0.747	-0.156	0.876	-1.580	1.347
State_IA	-0.0688	0.456	-0.151	0.880	-0.963	0.826
State_ID	-0.1106	0.594	-0.186	0.852	-1.276	1.054
State_IL	-0.1707	0.438	-0.389	0.697	-1.030	0.689
State_IN	0.0477	0.468	0.102	0.919	-0.869	0.965
State_KS	0.1143	0.505	0.226	0.821	-0.876	1.104
State_KY	-0.0817	0.473	-0.173	0.863	-1.008	0.845
State_LA	-0.1769	0.501	-0.353	0.724	-1.158	0.805
State_MA	-0.3176	0.496	-0.640	0.522	-1.290	0.655
State_MD	0.0900	0.516	0.174	0.862	-0.922	1.102
State_ME	-0.0381	0.543	-0.070	0.944	-1.103	1.027
State_MI	0.0344	0.459	0.075			

				0.940	-0.864	0.933
State_MN	0.0388	0.463	0.084	0.933	-0.868	0.946
State_MO	0.1471	0.443	0.332	0.740	-0.722	1.016
State_MS	-0.1339	0.510	-0.263	0.793	-1.133	0.866
State_MT	0.5013	0.561	0.893	0.372	-0.599	1.601
State_NC	-0.1538	0.452	-0.340	0.734	-1.040	0.733
State_ND	0.0070	0.504	0.014	0.989	-0.981	0.995
State_NE	0.0600	0.499	0.120	0.904	-0.919	1.039
State_NH	-0.3289	0.609	-0.540	0.589	-1.523	0.865
State_NJ	-0.1427	0.489	-0.292	0.770	-1.101	0.816
State_NM	-0.6876	0.510	-1.348	0.178	-1.688	0.312
State_NV	-0.5564	0.809	-0.687	0.492	-2.143	1.030
State_NY	-0.3996	0.428	-0.933	0.351	-1.239	0.439
State_OH	-0.2863	0.444	-0.645	0.519	-1.156	0.584
State_OK	-0.2107	0.478	-0.440	0.660	-1.148	0.727
State_OR	0.3056	0.524	0.583	0.560	-0.722	1.333
State_PA	-0.2038	0.428	-0.476	0.634	-1.043	0.636
State_PR	-0.4058	0.765	-0.531	0.596	-1.904	1.093
State_RI	-4.2002	1.488	-2.823	0.005	-7.116	-1.285
State_SC	-0.0280	0.554	-0.050	0.960	-1.113	1.057
State_SD	-0.5685	0.574	-0.990	0.322	-1.694	0.557
State_TN	0.4821	0.500	0.963	0.335	-0.499	1.463
State_TX	0.0654	0.426	0.154	0.878	-0.769	0.899
State_UT	-0.0219	0.652	-0.034	0.973	-1.300	1.256

```
State_VA                                      -0.0051      0.461     -0.011
0.991      -0.909       0.898
State_VT                                       0.3359      0.594      0.566
0.572      -0.828       1.499
State_WA                                       0.2517      0.482      0.522
0.601      -0.693       1.196
State_WI                                       0.0259      0.464      0.056
0.955      -0.883       0.935
State_WV                                       0.6860      0.464      1.479
0.139      -0.223       1.595
State_WY                                      -0.3832      0.756     -0.507
0.612      -1.865       1.099
Area_Suburban                                 -0.0424      0.097     -0.437
0.662      -0.232       0.148
Area_Urban                                     0.0573      0.096      0.595
0.552      -0.131       0.246
Marital_Divorced                              -0.2774      0.122     -2.268
0.023      -0.517      -0.038
Marital_Married                               -0.1323      0.125     -1.057
0.290      -0.378       0.113
Marital_Never Married                         -0.2172      0.125     -1.737
0.082      -0.462       0.028
Marital_Separated                             -0.1101      0.124     -0.890
0.374      -0.353       0.133
Gender_Male                                    0.2561      0.079      3.250
0.001       0.102       0.411
PaymentMethod_Credit Card (automatic)          0.2018      0.120      1.682
0.093      -0.033       0.437
PaymentMethod_Electronic Check                 0.6103      0.108      5.668
0.000       0.399       0.821
PaymentMethod_Mailed Check                     0.2439      0.119      2.058
0.040       0.012       0.476
Techie_Yes                                     1.1126      0.105     10.628
0.000       0.907       1.318
Contract_Month-to-month                        3.5286      0.126     27.961
0.000       3.281       3.776
Contract_One year                              0.1177      0.135      0.873
0.383      -0.147       0.382
Port_modem_Yes                                 0.1320      0.079      1.682
0.093      -0.022       0.286
Tablet_Yes                                    -0.0627      0.086     -0.728
0.467      -0.231       0.106
InternetService_DSL                            1.5433      0.114     13.571
0.000       1.320       1.766
InternetService_Fiber Optic                    0.1368      0.107      1.282
0.200      -0.072       0.346
Phone_Yes                                     -0.3436      0.134     -2.555
```

```
0.011      -0.607      -0.080
Multiple_Yes                              1.7309      0.086      20.079
0.000       1.562       1.900
OnlineSecurity_Yes                       -0.1492      0.082      -1.820
0.069      -0.310       0.012
OnlineBackup_Yes                          0.8334      0.081      10.313
0.000       0.675       0.992
DeviceProtection_Yes                      0.4704      0.080       5.910
0.000       0.314       0.626
TechSupport_Yes                           0.2938      0.081       3.630
0.000       0.135       0.452
StreamingTV_Yes                           3.0297      0.100      30.326
0.000       2.834       3.225
StreamingMovies_Yes                       3.5904      0.107      33.532
0.000       3.381       3.800
PaperlessBilling_Yes                      0.1460      0.080       1.831
0.067      -0.010       0.302
Item1_2                                   0.1665      0.130       1.281
0.200      -0.088       0.421
Item1_4                                   0.0185      0.104       0.177
0.859      -0.186       0.223
Item1_5                                   0.0400      0.152       0.263
0.793      -0.258       0.338
Item1_6                                   0.1790      0.314       0.570
0.569      -0.437       0.795
Item1_7                                   0.3143      1.036       0.303
0.762      -1.717       2.346
Item2_2                                  -0.0002      0.142      -0.002
0.999      -0.278       0.277
Item2_3                                  -0.1088      0.099      -1.096
0.273      -0.303       0.086
Item2_5                                  -0.1083      0.134      -0.805
0.421      -0.372       0.155
Item2_6                                  -0.0936      0.297      -0.315
0.753      -0.676       0.488
Item2_7                                   2.3875      1.477       1.617
0.106      -0.507       5.282
Item3_2                                  -0.0756      0.125      -0.606
0.544      -0.320       0.169
Item3_4                                  -0.0846      0.099      -0.851
0.395      -0.279       0.110
Item3_5                                   0.0662      0.142       0.467
0.641      -0.212       0.344
Item3_6                                   0.4552      0.298       1.530
0.126      -0.128       1.039
Item3_7                                  -0.9344      1.201      -0.778
0.437      -3.289       1.420
```

Item3_8	-21.0947	1.5e+04	-0.001	0.999	-2.95e+04	2.94e+04
Item4_2	0.2423	0.127	1.904	0.057	-0.007	0.492
Item4_4	0.0793	0.096	0.829	0.407	-0.108	0.267
Item4_5	0.0594	0.132	0.448	0.654	-0.200	0.319
Item4_6	0.3956	0.270	1.464	0.143	-0.134	0.925
Item4_7	-1.4749	1.179	-1.251	0.211	-3.786	0.836
Item5_2	0.1434	0.130	1.103	0.270	-0.111	0.398
Item5_3	-0.0074	0.094	-0.079	0.937	-0.193	0.178
Item5_5	0.0456	0.133	0.343	0.731	-0.215	0.306
Item5_6	0.1207	0.281	0.430	0.667	-0.430	0.671
Item5_7	-2.4299	1.156	-2.102	0.036	-4.696	-0.164
Item6_2	0.1478	0.124	1.191	0.234	-0.095	0.391
Item6_4	0.0842	0.098	0.857	0.392	-0.108	0.277
Item6_5	-0.0955	0.135	-0.707	0.479	-0.360	0.169
Item6_6	-0.2471	0.276	-0.894	0.371	-0.789	0.295
Item6_7	0.4625	1.324	0.349	0.727	-2.133	3.058
Item6_8	-19.4492	1.5e+04	-0.001	0.999	-2.94e+04	2.94e+04
Item7_2	-0.0964	0.129	-0.748	0.454	-0.349	0.156
Item7_4	-0.1945	0.095	-2.046	0.041	-0.381	-0.008
Item7_5	0.0384	0.133	0.288	0.773	-0.222	0.299
Item7_6	0.1699	0.284	0.599	0.549	-0.386	0.726
Item7_7	-1.7772	1.244	-1.428	0.153	-4.216	0.661
Item8_2	0.0324	0.130	0.250	0.803	-0.222	0.287
Item8_3	0.0917	0.095	0.964			

```
0.335      -0.095       0.278
Item8_5                                 0.0812      0.127      0.642
0.521      -0.167       0.329
Item8_6                                 0.2253      0.269      0.838
0.402      -0.302       0.752
Item8_7                                -0.4964      0.916     -0.542
0.588      -2.292       1.300
Item8_8                               -19.2036    1.5e+04     -0.001
0.999   -2.94e+04    2.93e+04
================================================================================
=========================
"""
```

4.1.2 P-Values > .05

Next, all variables with non-significant p-values were removed from the model. Starting with the highest p-value and using a p-value of .05 as the alpha, each variable was removed one at a time (backwards stepwise) and the model reran until all p-values were of significant value.

[226]:
```
X_next2 = churn_LRM_data.
↪drop(['Churn_Yes','Bandwidth_GB_Year','Gender_Female','Item1_3','Item2_4','Item5_4','Item8_
                                       ␣
↪'Item6_3','Item3_3','Item4_3','Item7_3','MonthlyCharge','Item8_8','State_VA',␣
↪'Item6_8',
                                       ␣
↪'Item3_8','Item7_5','Item8_2','Item2_2','State_AZ','State_ME','State_ND','State_UT','State_
                                       ␣
↪'State_HI','State_CT','Item5_3','Item1_4','Item1_5','Item1_7','Item7_6','Item4_5','Item2_6'
                                       ␣
↪'Item5_6','Item5_5','Population','Outage_sec_perweek','Income','State_ID','State_KY','State
                                       ␣
↪'State_LA','State_MD','State_PR','Item1_6','Item8_7','Item6_7','Item3_5','Item2_5','State_A
                                       ␣
↪'State_MI','State_MS','State_NJ','State_IA','State_MN','State_KS','State_NC','State_SC',
                                       ␣
↪'State_WI','State_NE','State_WY','State_GA','Item4_4','Item8_5','Item6_5','Item3_7','Item8_
                                       ␣
↪'Item8_3','Item8_3','Item2_7','Item6_6','Item7_2','Item3_2','Tablet_Yes','Contract_One␣
↪year',
                                       ␣
↪'State_CA','State_IL','State_NH','Area_Suburban','Yearly_equip_failure','Age','State_MA',
                                       ␣
↪'State_MO','State_OK','State_OH','State_TX','Marital_Separated','State_PA','State_NV',
                                       ␣
↪'Area_Urban','State_NV','State_DE','State_WA','State_PA','State_DC','Children','Email','Con
                                       ␣
↪'State_VT','State_AR','State_FL','State_MT','Marital_Married',
```

```
                              'PaymentMethod_Credit Card␣
↪(automatic)','InternetService_Fiber Optic','Item7_7','Item5_2',
                              ␣
↪'Item3_4','Item2_3','Item6_4','Item6_2','Item4_7','Item4_6','Item1_2','Item4_2',
                              ␣
↪'OnlineSecurity_Yes','Port_modem_Yes','PaymentMethod_Mailed␣
↪Check','Marital_Never Married',
                              ␣
↪'Marital_Divorced','State_OR','State_SD','State_NM','State_NY','PaperlessBilling_Yes',
                              ␣
↪'Item3_6','State_TN','State_RI','State_WV','Phone_Yes','Item5_7'], axis=1)

y = churn_LRM_data['Churn_Yes']

#define the input
X_next2 = sm.add_constant(X_next2)

# Split X and y into X_
X_train, X_test, Y_train, Y_test = train_test_split(X_next2, y, test_size=0.33,␣
↪random_state=1)
```

[227]:
```
#create an Logistic Regression Model
next2_model = sm.Logit(Y_train.astype("float64"), X_train.astype("float64"))

#fit the data
next2_est = next2_model.fit()

#Summarize the output
next2_est.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.219219
         Iterations 9
```

[227]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:              Churn_Yes   No. Observations:                 6700
Model:                          Logit   Df Residuals:                     6686
Method:                           MLE   Df Model:                           13
Date:                Sun, 25 Jul 2021   Pseudo R-squ.:                  0.6172
Time:                        15:48:37   Log-Likelihood:                -1468.8
converged:                       True   LL-Null:                       -3837.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
==================
```

```
                                      coef    std err          z      P>|z|
[0.025      0.975]
--------------------------------------------------------------------------------
------------------
const                              -6.7039      0.238    -28.132      0.000
-7.171      -6.237
Tenure                             -0.1118      0.003    -32.852      0.000
-0.118      -0.105
Gender_Male                         0.2758      0.094      2.920      0.003
0.091       0.461
PaymentMethod_Electronic Check      0.4425      0.099      4.472      0.000
0.249       0.636
Techie_Yes                          1.0323      0.126      8.211      0.000
0.786       1.279
Contract_Month-to-month             3.3991      0.129     26.394      0.000
3.147       3.652
InternetService_DSL                 1.3765      0.104     13.245      0.000
1.173       1.580
Multiple_Yes                        1.6914      0.102     16.530      0.000
1.491       1.892
OnlineBackup_Yes                    0.9048      0.097      9.311      0.000
0.714       1.095
DeviceProtection_Yes                0.4886      0.095      5.147      0.000
0.303       0.675
TechSupport_Yes                     0.1980      0.097      2.042      0.041
0.008       0.388
StreamingTV_Yes                     2.9585      0.119     24.886      0.000
2.725       3.191
StreamingMovies_Yes                 3.5135      0.128     27.523      0.000
3.263       3.764
Item7_4                            -0.1711      0.099     -1.733      0.083
-0.365       0.022
================================================================================
==================
"""
```

4.1.3 Reduced Model

With mulitcollinearity issues and insignificant variables removed the final model is created:

[228]:
```
X_final = X_next2
y = churn_LRM_data['Churn_Yes']

#define the input
X_final = sm.add_constant(X_next2)

# Split X and y into X_
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X_final, y, test_size=0.33,␣
↪random_state=1)
```

[229]:
```
#create an Logistic Regression Model
final_model = sm.Logit(Y_train.astype("float64"), X_train.astype("float64"))

#fit the data
final_est = final_model.fit()

#Summarize the output
final_est.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.219219
         Iterations 9
```

[229]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:              Churn_Yes   No. Observations:                 6700
Model:                          Logit   Df Residuals:                     6686
Method:                           MLE   Df Model:                           13
Date:                Sun, 25 Jul 2021   Pseudo R-squ.:                  0.6172
Time:                        15:48:53   Log-Likelihood:                -1468.8
converged:                       True   LL-Null:                       -3837.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
==================
                                     coef    std err          z      P>|z|
[0.025      0.975]
--------------------------------------------------------------------------------
------------------
const                             -6.7039      0.238    -28.132      0.000
-7.171      -6.237
Tenure                            -0.1118      0.003    -32.852      0.000
-0.118      -0.105
Gender_Male                        0.2758      0.094      2.920      0.003
0.091       0.461
PaymentMethod_Electronic Check     0.4425      0.099      4.472      0.000
0.249       0.636
Techie_Yes                         1.0323      0.126      8.211      0.000
0.786       1.279
Contract_Month-to-month            3.3991      0.129     26.394      0.000
3.147       3.652
InternetService_DSL                1.3765      0.104     13.245      0.000
1.173       1.580
Multiple_Yes                       1.6914      0.102     16.530      0.000
```

```
1.491       1.892
OnlineBackup_Yes                 0.9048      0.097      9.311      0.000
0.714       1.095
DeviceProtection_Yes             0.4886      0.095      5.147      0.000
0.303       0.675
TechSupport_Yes                  0.1980      0.097      2.042      0.041
0.008       0.388
StreamingTV_Yes                  2.9585      0.119     24.886      0.000
2.725       3.191
StreamingMovies_Yes              3.5135      0.128     27.523      0.000
3.263       3.764
Item7_4                         -0.1711      0.099     -1.733      0.083
-0.365       0.022
================================================================================
==================
"""
```

4.1.4 Part E: Analyze the Dataset Using the Reduced Logistic Regression Model

E1: The logic of the variable selection technique was explained in part IV as the variables were removed. The backwards stepwise method was used. Although not the best method to use according to Stoltzfus, it was chosen because of the requirement to have an intial model with all variables listed in part C2. In general, variables were removed to meet logistic model assumptions and ensure the variables were not overfitted by using only significant variables in the model.

E2: The following model evaluation metrics are used to compare the initial and reduced model:

- Converged: The initial model did not converge where as the reduced model did. This suggest that the reduced model is superior.
- Log-Likelihood: While still significantly well away from 0 in the reduced model, the log-likeihood is much closer to zero than in the initial model, indicating a better fit.
- Psuedo R-squared: The psuedo R-squared of the reduced model is slightly lower than the initial model, however, when combined with the other comparing factors above, the reducted value is insignificant.
- Model Accuracy Score: Both model's accuracy scores were similar, with a only slightly higher score of the initial model.
- Precision and Recall: Precision is the ability of a model not to label an instance positive that is actually negative, whereas Recall is the ability of a model to find all positive instances. Again, both models precision and recall scores were very similar.
- F1 Score: The F1 score is a weighted harmonic mean of precision and recall with 1.0 being the best and 0.0 being the worst. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. (Muthukrishnan, 2018) The inital model's F1 score is slightly high than the reduced model. However, other important previously mentioned factors need to be accounted for as well.
- Confusion Matrix: Both model's confusion matrixs were again relatively the same showing a high degree of prediction accuracy.

Initial Model Evaluation

[230]:
```
pipe = make_pipeline(StandardScaler(), LogisticRegression())
```

[231]:
```
pipe.fit(X2, y)
```

[231]:
```
Pipeline(steps=[('standardscaler', StandardScaler()),
                ('logisticregression', LogisticRegression())])
```

[232]:
```
initial_predictions = pipe.predict(X2)
```

[233]:
```
pipe.score(X2, y)
```

[233]:
```
0.9074
```

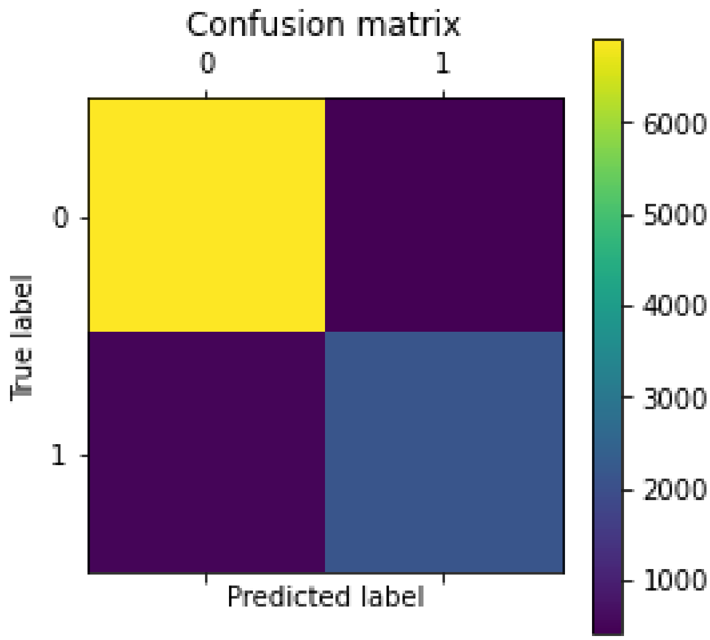
[234]:
```
print(classification_report(y, initial_predictions))
```

```
              precision    recall  f1-score   support

           0       0.93      0.94      0.94      7350
           1       0.84      0.81      0.82      2650

    accuracy                           0.91     10000
   macro avg       0.88      0.88      0.88     10000
weighted avg       0.91      0.91      0.91     10000
```

[235]:
```
print(confusion_matrix(y, initial_predictions))
```

```
[[6927  423]
 [ 503 2147]]
```

[236]:
```
plt.matshow(confusion_matrix(y, initial_predictions))
plt.title('Confusion matrix')
plt.colorbar()
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
```



Reduced Model Evaluation

[237]:
```
pipe.fit(X_train, Y_train)
pipe.score(X_train, Y_train)
```

[237]: 0.9046268656716417

[238]:
```
pipe.fit(X_test, Y_test)
pipe.score(X_test, Y_test)
```

[238]: 0.8966666666666666

[239]:
```
predictions = pipe.predict(X_test)
```

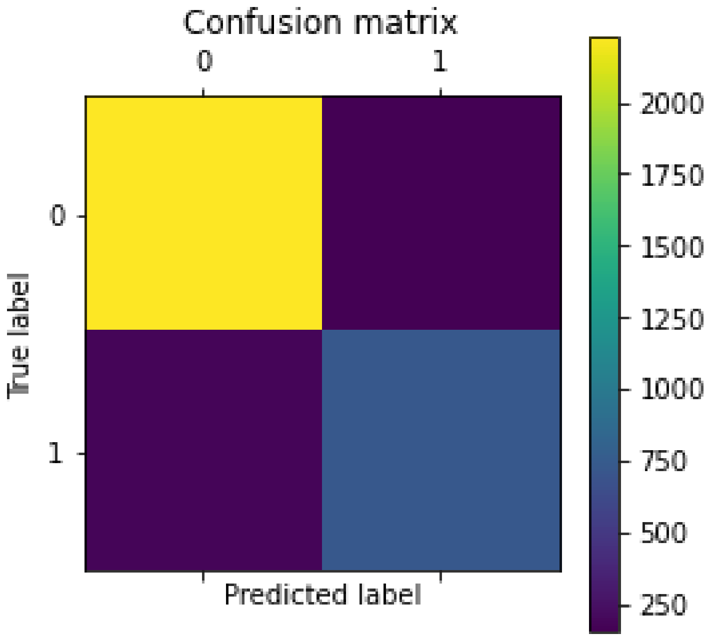
[240]:
```
print(classification_report(Y_test, predictions))
```

```
              precision    recall  f1-score   support

           0       0.92      0.93      0.93      2390
           1       0.82      0.80      0.81       910

    accuracy                           0.90      3300
   macro avg       0.87      0.87      0.87      3300
weighted avg       0.90      0.90      0.90      3300
```

[241]:
```
print(confusion_matrix(Y_test, predictions))
```

```
[[2232  158]
 [ 183  727]]
```

[242]:
```
plt.matshow(confusion_matrix(Y_test, predictions))
plt.title('Confusion matrix')
plt.colorbar()
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
```



5 Part V: Data Summary and Implications

5.0.1 F1: Results of the Data Analysis

Logistic Regression equation:

$$p = \frac{e^{-0.1118x_1+0.2758x_2+0.4425x_3+1.0323x_4+3.3991x_5+1.3765x_6+1.6914x_7+0.9048x_8+0.4886x_9+0.1980x_{10}+2.9585x_{11}+3.5135x_{12}-0.1711x_{13}-6.7039}}{1+e^{-0.1118x_1+0.2758x_2+0.4425x_3+1.0323x_4+3.3991x_5+1.3765x_6+1.6914x_7+0.9048x_8+0.4886x_9+0.1980x_{10}+2.9585x_{11}+3.5135x_{12}-0.1711x_{13}-6.7039}}$$

where

Interpretation of coefficients that were statistically significant:

x_1 = Tenure - As the only continuous variable in the model, this coefficient reduces the log(odds) that the customer will churn by -0.1118 for each month the customer remains a customer. For example, a customer with a tenure of 12 months will have a reduced log(odds) of churning -0.1118 * 12 = -1.3416 whereas a customer with 24 months tenure will have a reduced log(odds) of churning by -2.6832. This suggests that the longer a customer stays a customer, the probability of that customer churning decreases.

x_2 = Gender_Male - Male customers log(odds) of churning are increased by 0.2758.

x_3 = PaymentMethod_Electronic Check - Customers who use electronic check as a payment method log(odds) of churning are increased by 0.4425.

x_4 = Techie_Yes - Customer's who consider themselves "Techies" increase the log(odds) of churning by 1.0323.

x_5 = Contract_Month-to-month - Customers who are on month-to-month service contracts have an increased log(odds) of churning by 3.3991.

x_6 = InternetService_DSL - Customers with DSL service have an increased log(odds) of churning of 1.3765.

x_7 = Multiple_Yes - Customers with multiple phone lines have an increased log(odds) of churning of 1.6914.

x_8 = OnlineBackup_Yes - Customers with OnlineBackup service add-on have an increased log(odds) of churning of 0.9048.

x_9 = DeviceProtection_Yes - Customers with DeviceProtection service add-on have an increased log(odds) of churning of 0.4886

x_{10} = TechSupport_Yes - Customers with TechSupport service add-on have an increased log(odds) of churning of 0.1980

x_{11} = StreamingTV_Yes - Customers with StreamingTV service add-on have an increased log(odds) of churning of 2.9585.

x_{12} = StreamingMovies_Yes - Customers with StreamingMovies service add-on have an increased log(odds) of churning of 3.5135.

x_{13} = Item7_4 - Customers answering 4 on item 7 in the eight-question survey have a log(odds) increase of churning of .1711. (Not very useful)

Constant - The line of the log(odds) linear model crosses y at - 6.7039.

The statistical and practical significance of the model:

By reviewing the classification report it can be determined the model is statistically significant. The model is overall highly accurate with an approximate 90% accuracy rating and an F1 score of approximately 90%. As this is a model evaluating customers of a business, I believe an inaccuracy of 10% is acceptable.

This model is useful in a practical significance as well. The business can look at the variables that significantly increase or reduce the log(odds) of a customer churning. For example, the longer the tenure of a customer the lower the probability that the customer will churn. The organization can develop a model that predicts the customers' tenure to determine factors that increase or decrease the tenure of a customer. Another example is to look at the streaming services of the organization. The model suggests that customers with these services significantly increases the probability of churn. The business can look into the effectiveness of these services, attempt to determine why customers with these services have an increased likelihood of churn or evaluate whether or not the business should continue offering these services.

Limitations of the data analysis:

Due to the way the model was built (backward stepwise) the model may be missing crucial variables (such as MonthlyCharge). Because the model included all the variables at once, it is difficult to determine which variables were causing multicollinearity issues. Creating the model with one variable at first, then adding more as it was created, may be a better way of determining which variables cause multicollinearity issues. Building the model this way could also allow variables that were initially shown to have significant relationships to remain in the model.

5.0.2 F1: Recommended Course of Action

As stated above, the business can look at the variables that significantly increase or reduce the log(odds) of a customer churning. For example, the longer the tenure of a customer the lower the probability that the customer will churn. The organization can develop a model that predicts the customers' tenure to determine factors that increase or decrease the tenure of a customer. Another example is to look at the streaming services of the organization. The model suggests that customers with these services significantly increases the probability of churn. The business can look into the effectiveness of these services, attempt to determine why customers with these services have an increased likelihood of churn or evaluate whether or not the business should continue offering these services.

6 Part VI: Demonstration

6.1 G. Video

Link included as attachment to submission.

6.2 H. Code Sources

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2. (Publisher link).

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Python Software Foundation. Python Language Reference, version 3.7. Available at http://www.python.org

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.

W. McKinney, AQR Capital Management, pandas: a python data analysis library, http://pandas.sourceforge.net

6.3 I. References

Massaron, L., & Boschetti, A. (2016). Regression analysis with Python. Packt Publishing. ISBN: 9781785286315

Muthukrishnan. (2018, July 07). Understanding the Classification report through sklearn. Retrieved July 25, 2021, from https://muthu.co/understanding-the-classification-report-in-sklearn/

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. Academic Emergency Medicine, 18(10), 1099-1104. doi:10.1111/j.1553-2712.2011.01185.x