D208 Exploratory Predictive Data Modeling Performance Task 2

Logistic Regression for Predictive Modeling

Sean Simmons

WDU Data Analytics

MSDA D208

January 2023

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.

2. Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

1. “What variables are most important in whether a customer will Churn?“ is my research question. This is relevant to the real-world organizational setting because churning is a key metric for this organization. Churn is a categorical variable with Yes or No values. Yes to Churn is when a customer has left within the last month. As a result, the organization would be highly interested in determining whether a customer will leave them or not. Logistic regression is used with these categorical values and can be used to determine the probability of a certain event occurring (in this case, Churn being Yes).
2. The objective of the data analysis is to determine the factors that are more important in determining what makes a customer Churn and use this information to predict what a customer that is about to Churn or will Churn looks like (in terms of the information in the data set). That is to say, we must use logistic regression to determine the factors that make the probability of a “Yes” event occurring for Churn. This objective is informed by the organization’s need to retain customers.

Part II: Method Justification

B. Describe logistic regression methods by doing the following:

1. Summarize the assumptions of a logistic regression model.

2. Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.

3. Explain why logistic regression is an appropriate technique to analyze the research question summarized in Part I.

1. The assumptions of a logistic regression model are as follows:
    a. The dependent response variable must be binary (Churn is Yes/No for example).
    b. All observations are independent from another.
    c. Multicollinearity does not exist among explanatory variables (there is no high correlation between them)
    d. There is a lack of errors and severe outliers in a sufficiently large data set

2. The benefits of using Python are shown in each part. Firstly, Jupyter notebook is an interface system for python that provides a user-friendly coding environment, visualization abilities, and code extraction. Secondly, Python has packages that make handling large data sets easier and more efficient. The ability to visualize all of our necessary plots and data for univariate, bivariate, logistic regression, and confusion matrix is also readily available in Python. These abilities and my experience with Python make it the tool of choice for the entirety of this project. More specifically, python can use packages to view and perform our univariate and bivariate analysis, as well as

perform a user-friendly logistic regression, confusion matrix, and data analysis for understanding our data set.

3. Logistic regression is an appropriate technique to analyze the research question because our dependent variable is a binary variable. After passing the first criteria, logistic regression is still our best tool because it is a predictive analysis that is used to describe the relationship between the dependent variable and several independent variables (IBM 2023). This function of logistic regression will aid us in answering our research question of “what are the variables that make a customer more likely to Churn?” and predict churning (Swaminathan 2019).

Part III: Data Preparation

C. Summarize the data preparation process for logistic regression by doing the following:

1. Describe your data preparation goals and the data manipulations that will be used to achieve the goals.

2. Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.

3. Explain the steps used to prepare the data for the analysis, including the annotated code.

4. Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.

5. Provide a copy of the prepared data set.

1. My data preparation goals are to clean the data set and perform univariate and bivariate analysis on the variables I intend to consider for the regression model. The data manipulations I will do are listed below:
    a. Import the dataset into the Python jupyter notebook environment
    b. Evaluate the structure of the data to gain a better understanding of the variables and data types using print commands
    c. Rename variables that need to be more descriptive (items 1 - 8)
    d. Check for missing values and mitigate them using central tendency
    e. Check for outliers and make a decision about their mitigation
    f. Create numeric columns for all of our categorical data
    g. Perform univariate and bivariate analysis with visualizations to search for problematic, dirty, or misleading data.

No missing values are present, as shown in the following output:

```
Out[42]: CaseOrder                0
         Customer_id              0
         Interaction              0
         UID                      0
         City                     0
         State                    0
         County                   0
         Zip                      0
         Lat                      0
         Lng                      0
         Population               0
         Area                     0
         TimeZone                 0
         Job                      0
         Children                 0
         Age                      0
         Income                   0
         Marital                  0
         Gender                   0
         Churn                    0
         Outage_sec_perweek       0
         Email                    0
         Contacts                 0
         Yearly_equip_failure     0
         Techie                   0
         Contract                 0
         Port_modem               0
         Tablet                   0
         InternetService          0
         Phone                    0
         Multiple                 0
         OnlineSecurity           0
         OnlineBackup             0
         DeviceProtection         0
         TechSupport              0
         StreamingTV              0
         StreamingMovies          0
         PaperlessBilling         0
         PaymentMethod            0
         Tenure                   0
         MonthlyCharge            0
         Bandwidth_GB_Year        0
         item1_responses          0
         item2_fixes              0
         item3_replacements       0
         item4_reliability        0
         item5_options            0
         item6_respectfulness     0
         item7_courteous          0
         item8_listening          0
         dtype: int64
```

2. The target variable is Churn, here represented as numeric data in "Churn_numeric." The predictor variables are in the list below. I will exclude Lat, Lng, Caseorder, Zip, Customer_id, Population, UID, City, State, County, Area, Timezone, Job, and Marital from my analysis, however, they are still a part of the dataset and will be cleaned along with the dependent and independent variables of question so are shown below as well:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Lat                 10000 non-null  float64
 1   Lng                 10000 non-null  float64
 2   Income              10000 non-null  float64
 3   Outage_sec_perweek  10000 non-null  float64
 4   Tenure              10000 non-null  float64
 5   MonthlyCharge       10000 non-null  float64
 6   Bandwidth_GB_Year   10000 non-null  float64
dtypes: float64(7)
memory usage: 547.0 KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CaseOrder             10000 non-null  int64
 1   Zip                   10000 non-null  int64
 2   Population            10000 non-null  int64
 3   Children              10000 non-null  int64
 4   Age                   10000 non-null  int64
 5   Email                 10000 non-null  int64
 6   Contacts              10000 non-null  int64
 7   Yearly_equip_failure  10000 non-null  int64
 8   Item1                 10000 non-null  int64
 9   Item2                 10000 non-null  int64
 10  Item3                 10000 non-null  int64
 11  Item4                 10000 non-null  int64
 12  Item5                 10000 non-null  int64
 13  Item6                 10000 non-null  int64
 14  Item7                 10000 non-null  int64
 15  Item8                 10000 non-null  int64
dtypes: int64(16)
memory usage: 1.2 MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 27 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Customer_id       10000 non-null  object
 1   Interaction       10000 non-null  object
 2   UID               10000 non-null  object
 3   City              10000 non-null  object
 4   State             10000 non-null  object
 5   County            10000 non-null  object
 6   Area              10000 non-null  object
 7   TimeZone          10000 non-null  object
 8   Job               10000 non-null  object
 9   Marital           10000 non-null  object
 10  Gender            10000 non-null  object
 11  Churn             10000 non-null  object
 12  Techie            10000 non-null  object
 13  Contract          10000 non-null  object
 14  Port_modem        10000 non-null  object
 15  Tablet            10000 non-null  object
```

```
 16  InternetService   10000 non-null  object
 17  Phone             10000 non-null  object
 18  Multiple          10000 non-null  object
 19  OnlineSecurity    10000 non-null  object
 20  OnlineBackup      10000 non-null  object
 21  DeviceProtection  10000 non-null  object
 22  TechSupport       10000 non-null  object
 23  StreamingTV       10000 non-null  object
 24  StreamingMovies   10000 non-null  object
 25  PaperlessBilling  10000 non-null  object
 26  PaymentMethod     10000 non-null  object
dtypes: object(27)
memory usage: 2.1+ MB
None
   CaseOrder Customer_id                           Interaction  \
0          1     K409198  aa90260b-4141-4a24-8e36-b04ce1f4f77b
1          2     S120509  fb76459f-c047-4a9d-8af9-e0f7d4ac2524

                                UID         City State                 County  \
0  e885b299883d4f9fb18e39c75155d990  Point Baker    AK  Prince of Wales-Hyder
1  f2de8bef964785f41a2959829830fb8a  West Branch    MI                 Ogemaw

     Zip       Lat        Lng  ...  MonthlyCharge Bandwidth_GB_Year Item1  \
0  99927  56.25100 -133.37571  ...     172.455519        904.536110     5
1  48661  44.32893  -84.24080  ...     242.632554        800.982766     3

  Item2  Item3  Item4  Item5 Item6 Item7 Item8
0     5      5      3      4     4     3     4
1     4      3      3      4     3     4     4

[2 rows x 50 columns]
```

There are no outliers that need removal, as all values should be kept to preserve the integrity of the data set. Finally, the measures of central tendency for the variables are as follows: Please note all of the categorical values have been changed to numeric and are represented by the same naming system of “original title_numeric”. Yes = 0, No = 1.

<u>Mean</u>

```
CaseOrder             5000.500000
Zip                  49153.319600
Lat                     38.757567
Lng                    -90.782536
Population            9756.562400
Children                 2.087700
Age                     53.078400
Income               39806.926771
Outage_sec_perweek      10.001848
```

```
Email                    12.016000
Contacts                  0.994200
Yearly_equip_failure      0.398000
Tenure                   34.526188
MonthlyCharge           172.624816
Bandwidth_GB_Year      3392.341550
item1_responses           3.490800
item2_fixes               3.505100
item3_replacements        3.487000
item4_reliability         3.497500
item5_options             3.492900
item6_respectfulness      3.497300
item7_courteous           3.509500
item8_listening           3.495600
Churn_numeric             0.735000
Area_numeric              1.000000
Marital_numeric           2.017500
Gender_numeric            0.571800
Contract_numeric          1.034000
PaymentMethod_numeric     1.700300
InternetService_numeric   0.772100
Techie_numeric            0.832100
Port_modem_numeric        0.516600
Tablet_numeric            0.700900
Phone_numeric             0.093300
Multiple_numeric          0.539200
OnlineSecurity_numeric    0.642400
OnlineBackup_numeric      0.549400
DeviceProtection_numeric  0.561400
TechSupport_numeric       0.625000
StreamingTV_numeric       0.507100
StreamingMovies_numeric   0.511000
PaperlessBilling_numeric  0.411800
dtype: float64
```

Median

```
CaseOrder                5000.500000
Zip                     48869.500000
Lat                        39.395800
Lng                       -87.918800
Population               2910.500000
Children                    1.000000
Age                        53.000000
Income                  33170.605000
Outage_sec_perweek         10.018560
Email                      12.000000
Contacts                    1.000000
Yearly_equip_failure        0.000000
Tenure                     35.430507
MonthlyCharge             167.484700
Bandwidth_GB_Year        3279.536903
```

```
item1_responses            3.000000
item2_fixes                4.000000
item3_replacements         3.000000
item4_reliability          3.000000
item5_options              3.000000
item6_respectfulness       3.000000
item7_courteous            4.000000
item8_listening            3.000000
Churn_numeric              1.000000
Area_numeric               1.000000
Marital_numeric            2.000000
Gender_numeric             1.000000
Contract_numeric           1.000000
PaymentMethod_numeric      2.000000
InternetService_numeric    1.000000
Techie_numeric             1.000000
Port_modem_numeric         1.000000
Tablet_numeric             1.000000
Phone_numeric              0.000000
Multiple_numeric           1.000000
OnlineSecurity_numeric     1.000000
OnlineBackup_numeric       1.000000
DeviceProtection_numeric   1.000000
TechSupport_numeric        1.000000
StreamingTV_numeric        1.000000
StreamingMovies_numeric    1.000000
PaperlessBilling_numeric   0.000000
dtype: float64
```

3. The steps to prepare the data for analysis are inside of the annotated code file below and summarized as follows:

    i. Import dataset to Python.

    ii. Change the name of the columns of the responses to the organization’s survey to easily recognizable descriptions (ex: "Item1" to "item1_responses").

    iii. Pull a description of the data set, structure (columns & rows) & data types with central tendencies.

    iv. View summary statistics of the data.

    v. Check for records with missing data & impute missing data with meaningful measures of central tendency (mean, median or mode) or simply remove outliers that are several standard deviations above the mean. This step might not be necessary if it is determined we will not be removing outliers.

vi. Create numeric variables in order to encode categorical, yes/no data points into 1/0 numerical values.

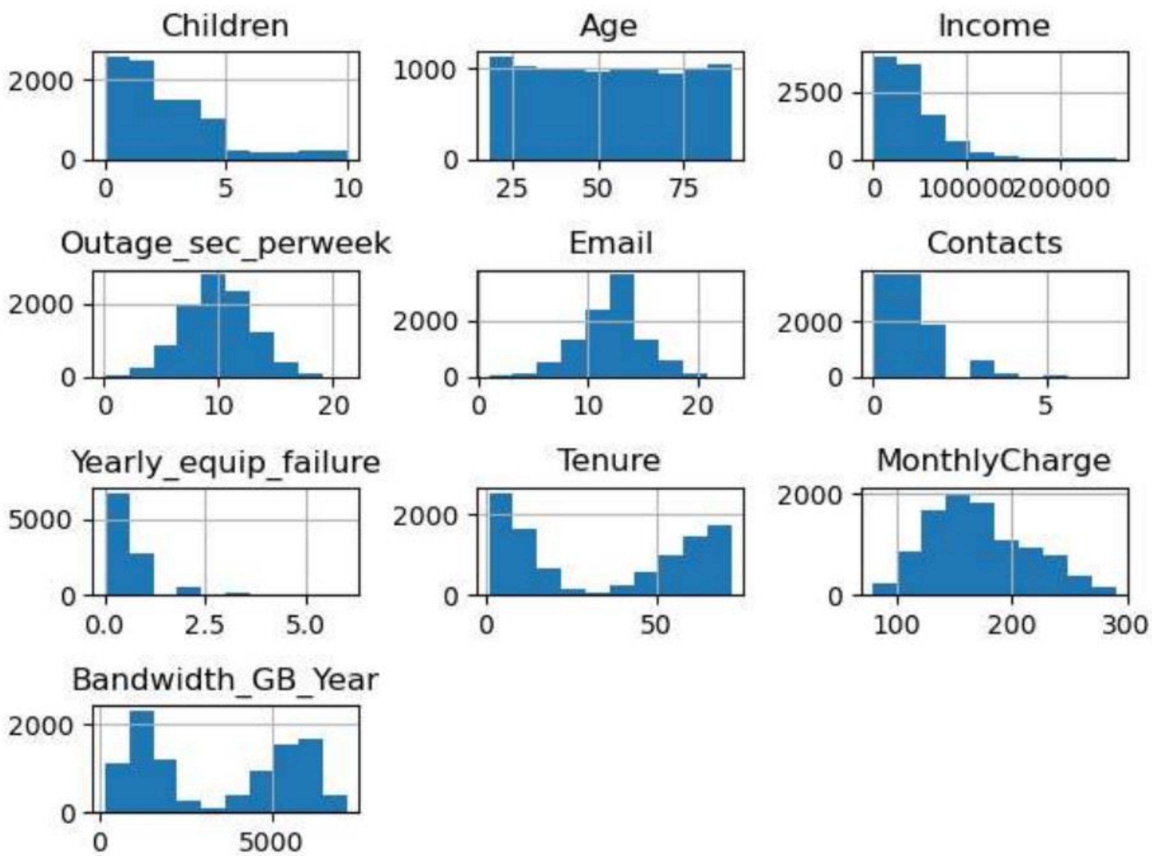
vii. View univariate & bivariate visualizations.

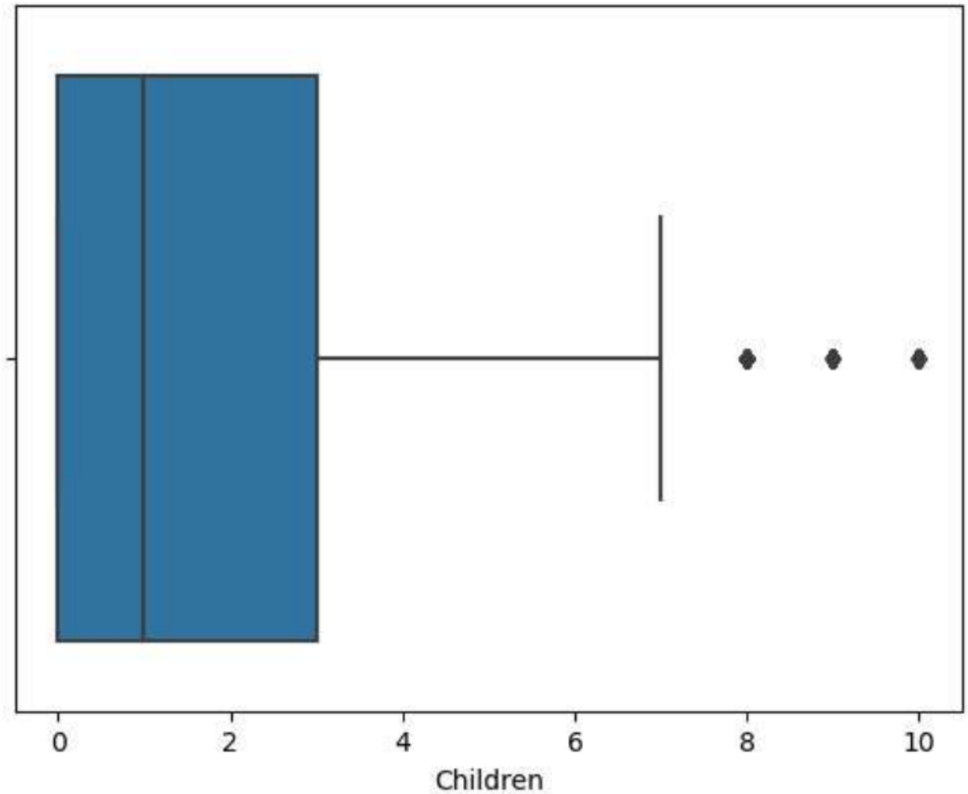
viii. Finally, the prepared dataset will be extracted & provided as "churn_Task2.csv"

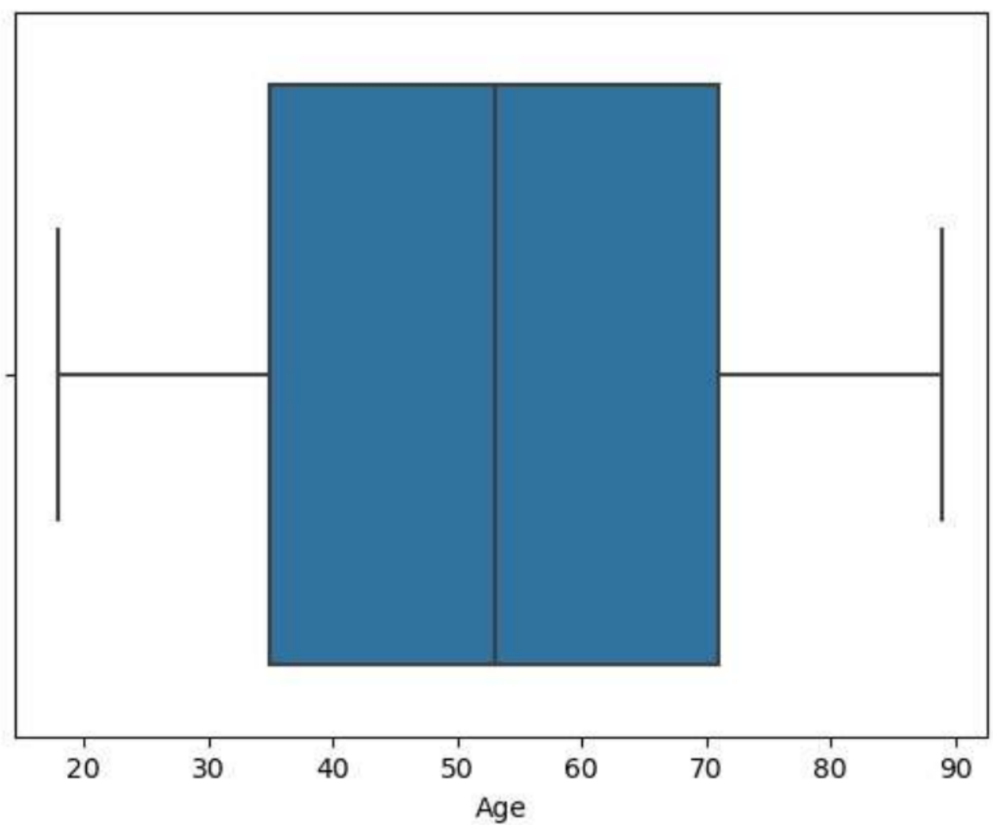
b. The annotated code can be found in “PA_D208_Code_Task2”

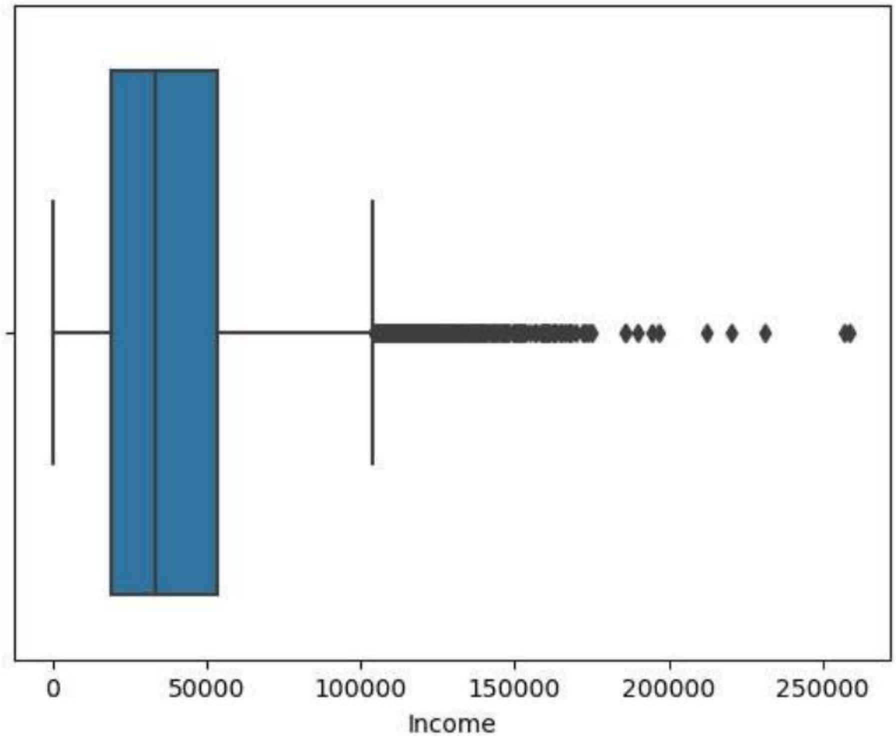
4. The visualizations are as follows:

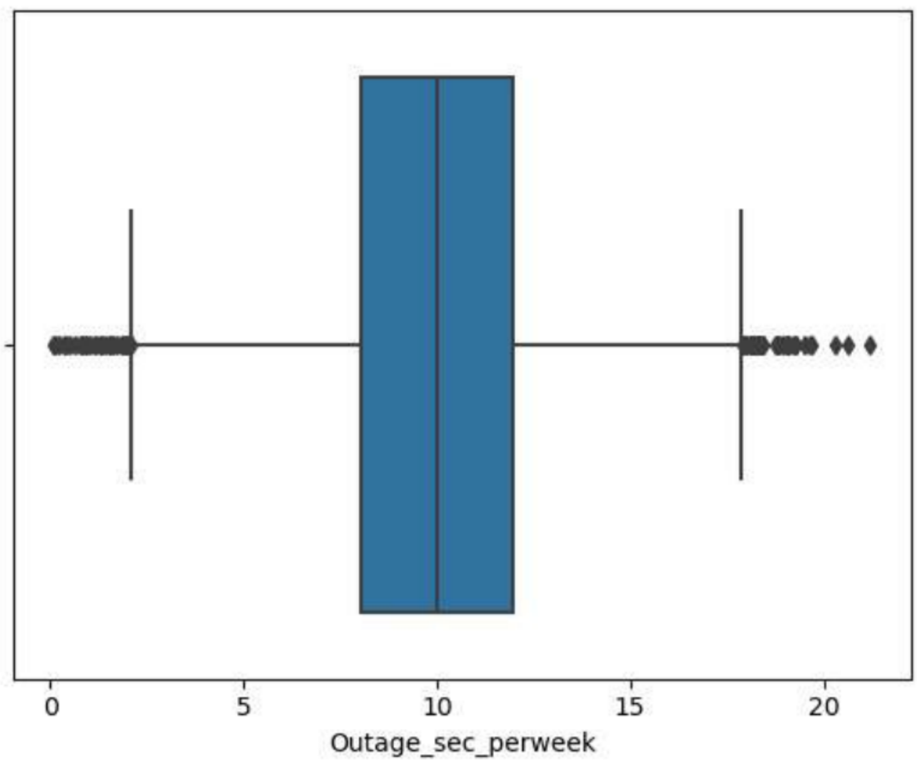
   a. Univariate

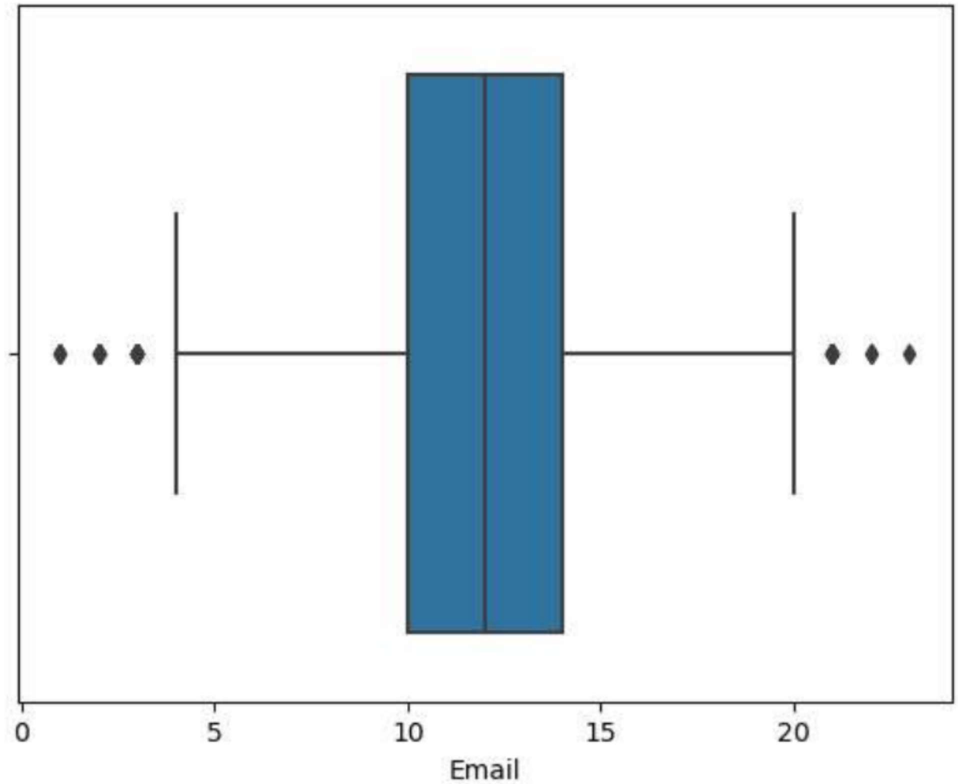

Children
Age
Income
Outage_sec_perweek
Email
Contacts
Yearly_equip_failure
Tenure
MonthlyCharge
Bandwidth_GB_Year

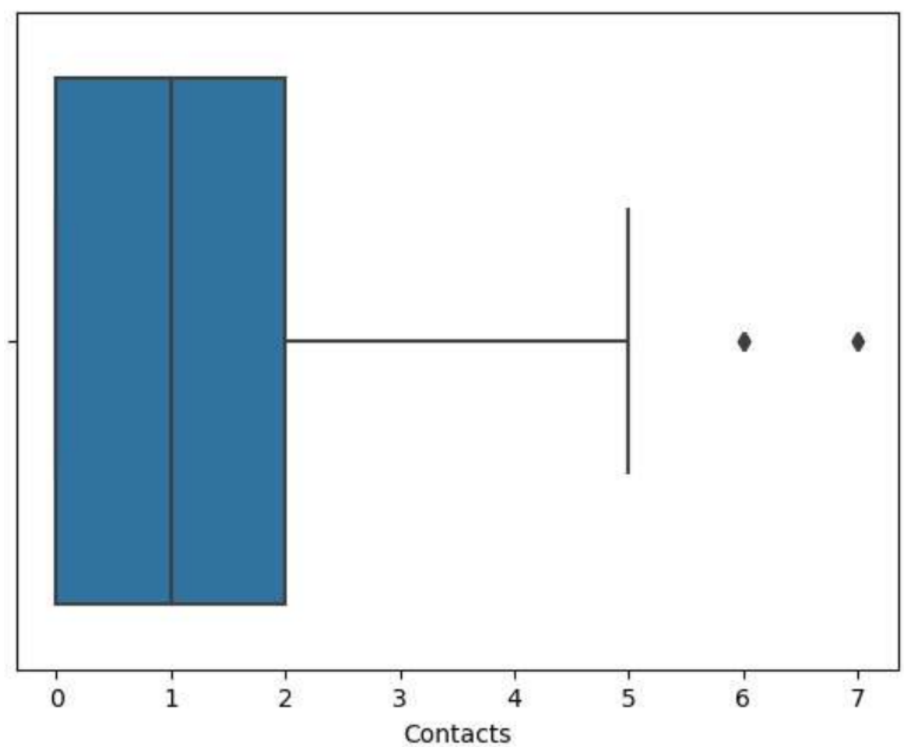

0
2
4
6
8
10
Children

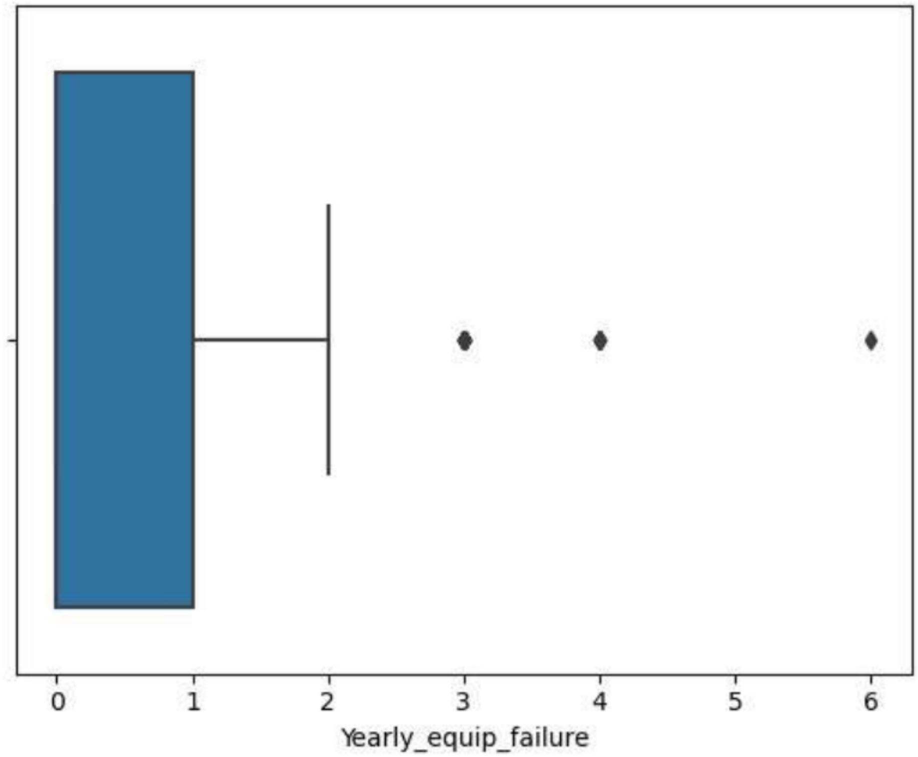

20
30
40
50
60
70
80
90
Age

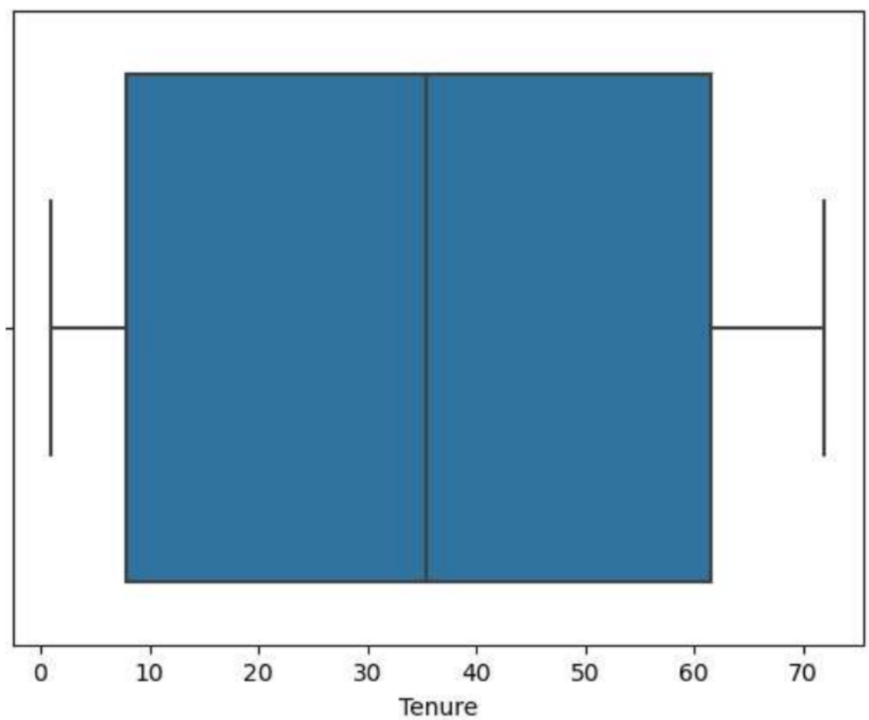

0
50000
100000
150000
200000
250000
Income

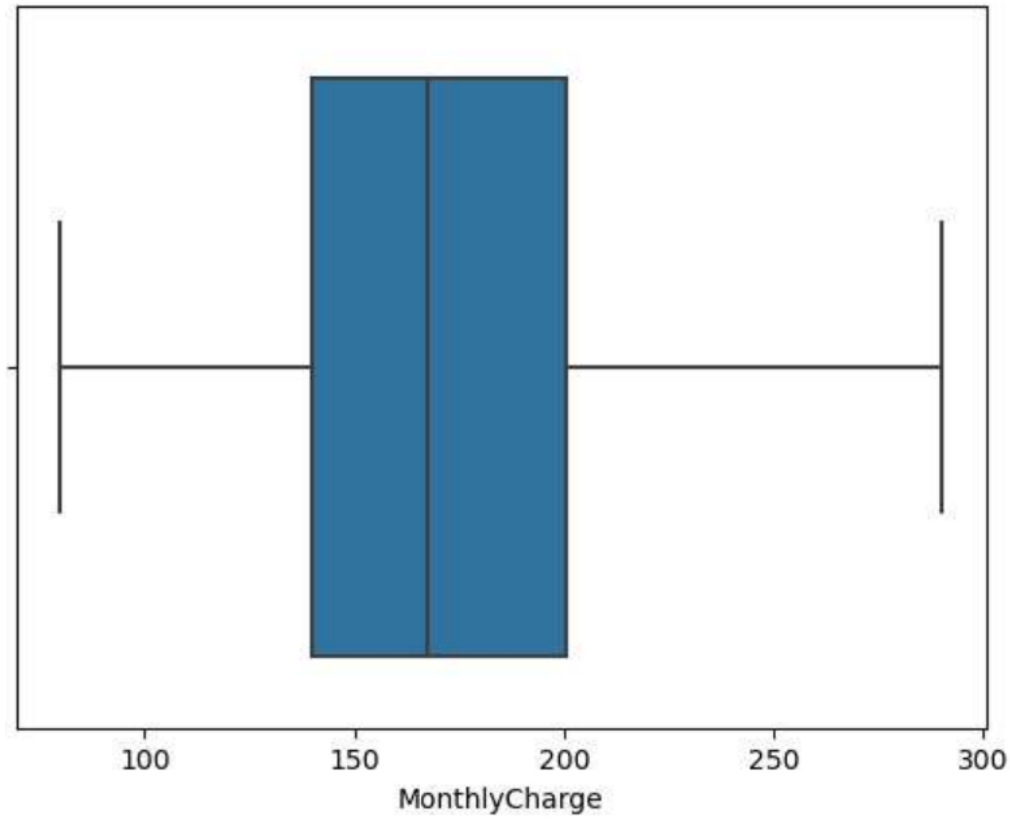

0
5
10
15
20
Outage_sec_perweek

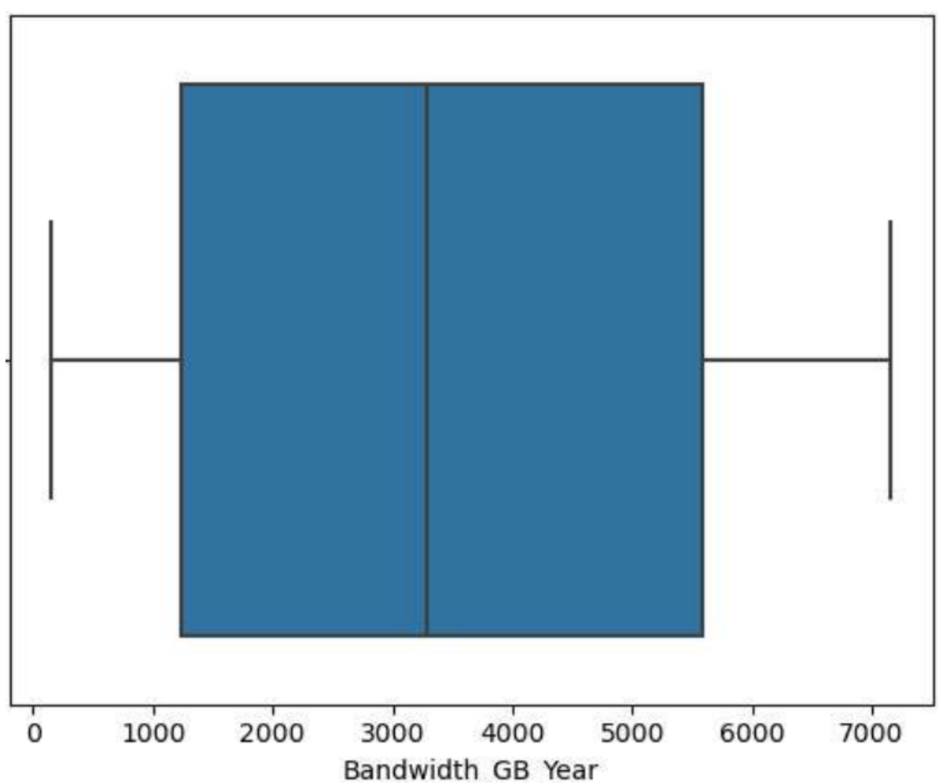

0
5
10
15
20
Email


0
1
2
3
4
5
6
7
Contacts


0
1
2
3
4
5
6
Yearly_equip_failure

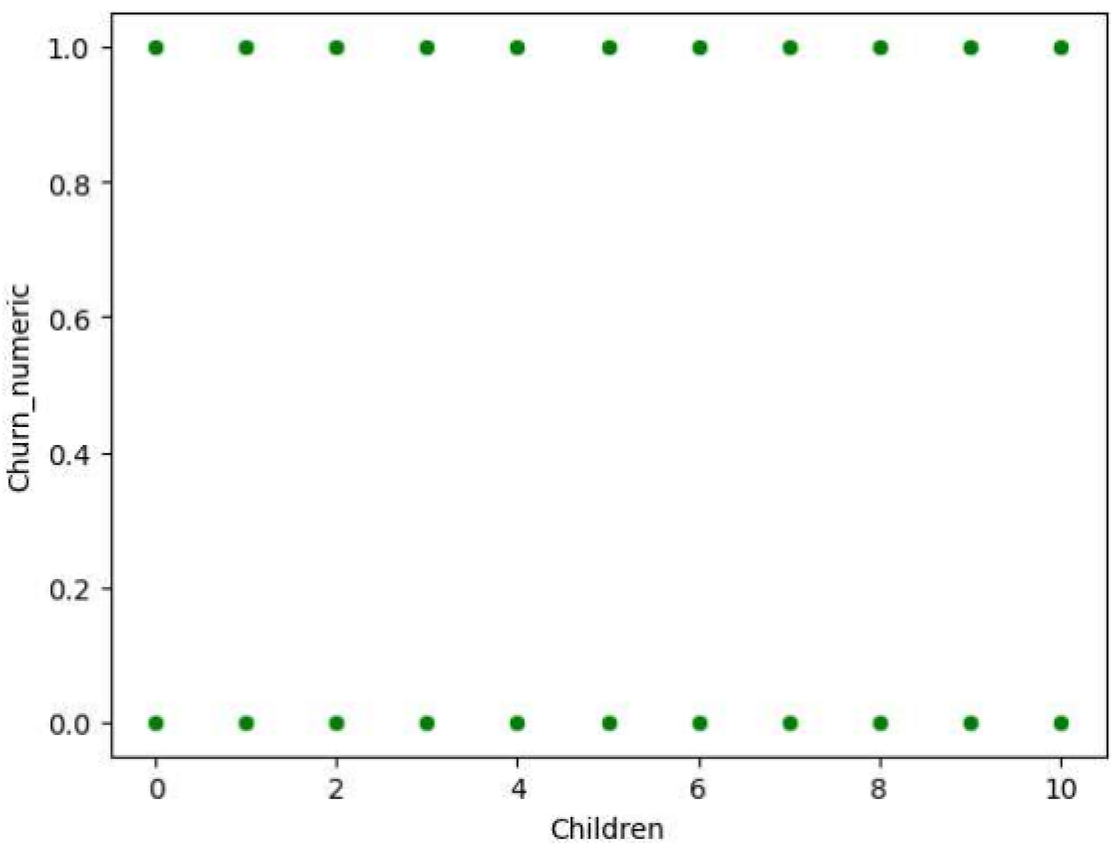

0
10
20
30
40
50
60
70
Tenure

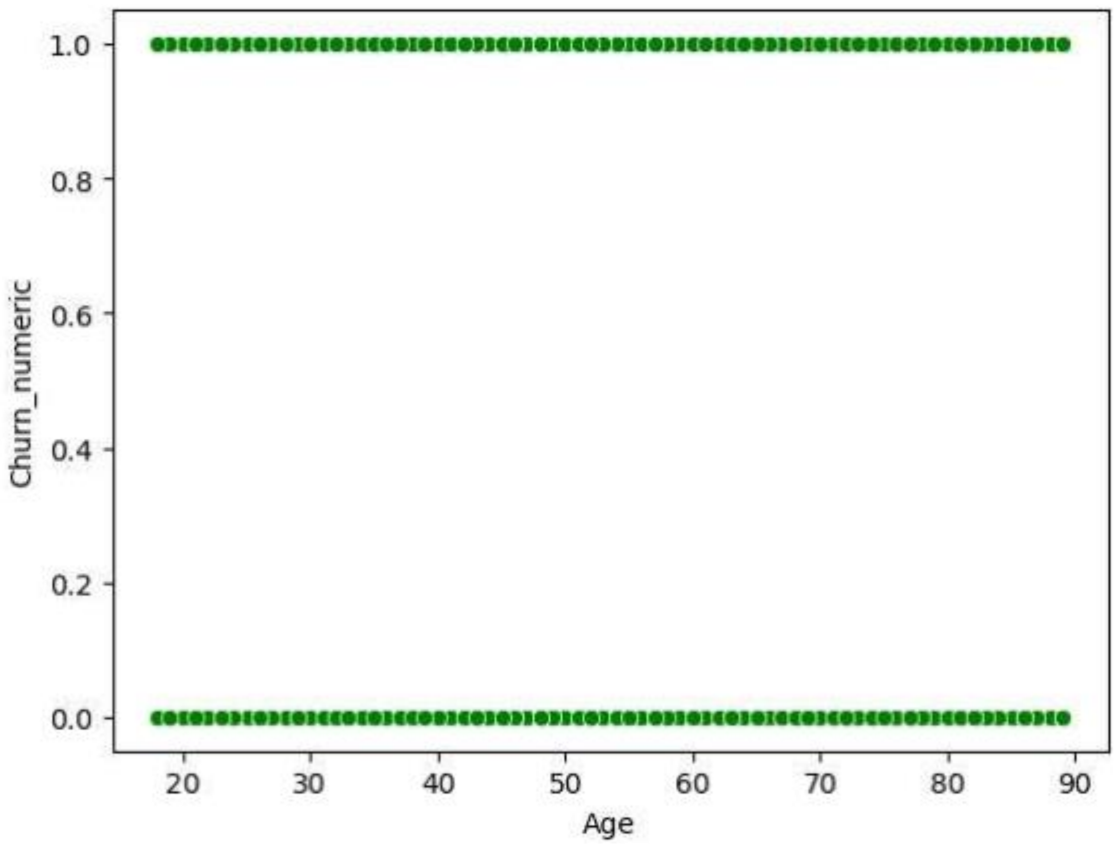

100
150
200
250
300
MonthlyCharge


0
1000
2000
3000
4000
5000
6000
7000
Bandwidth_GB_Year

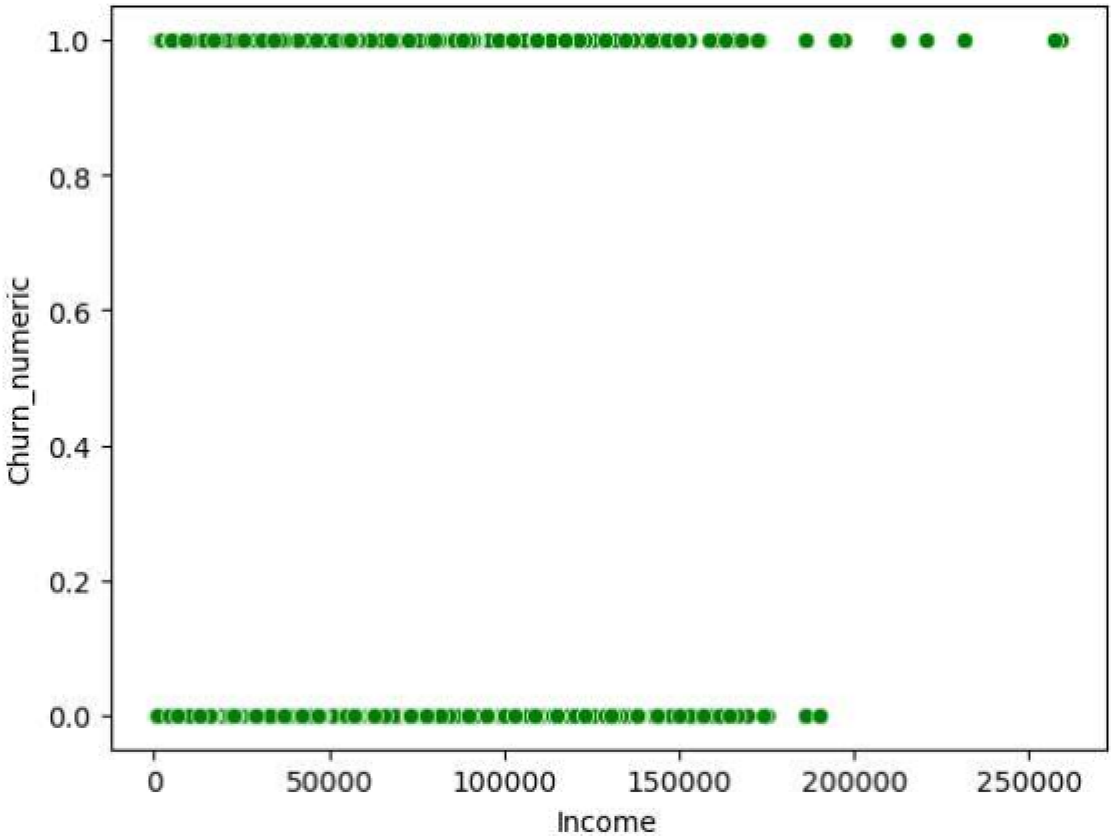
Although these boxplots show the existence of values that would be considered outliers, they will be kept in the dataset to preserve the integrity of the customer data and are assumed to not be human error.

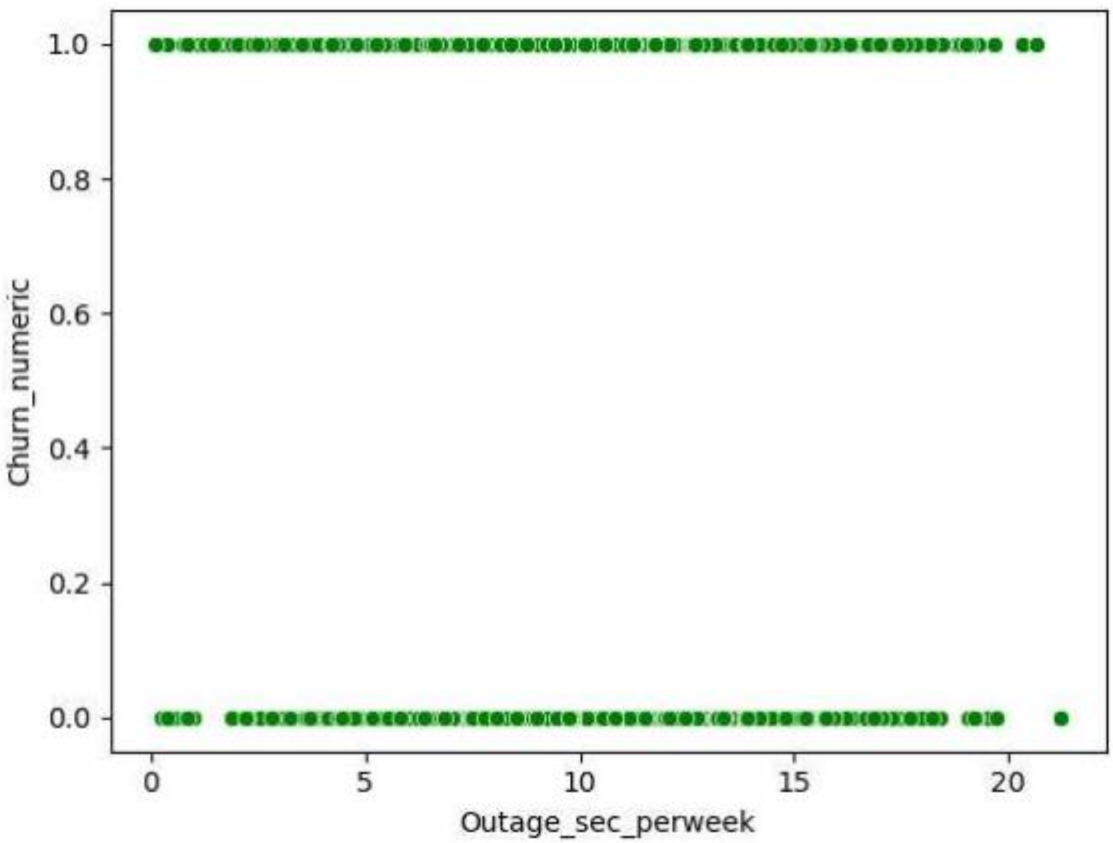
b. Bivariate

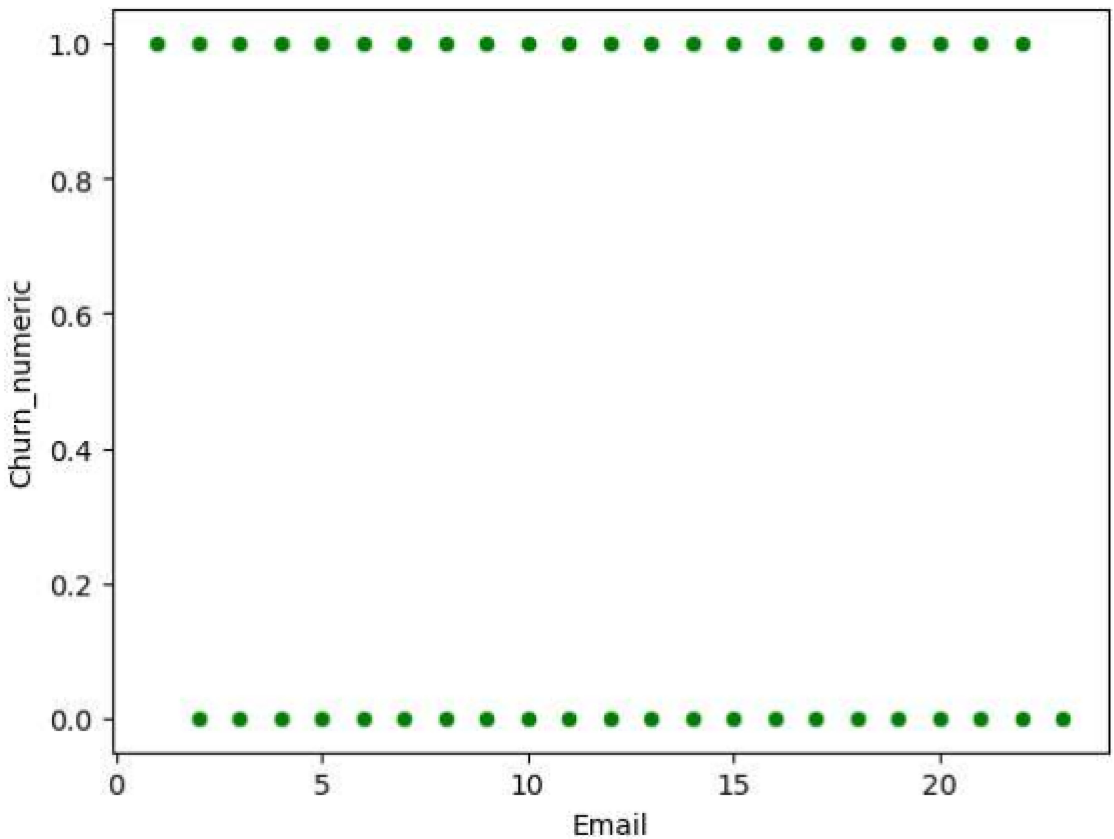


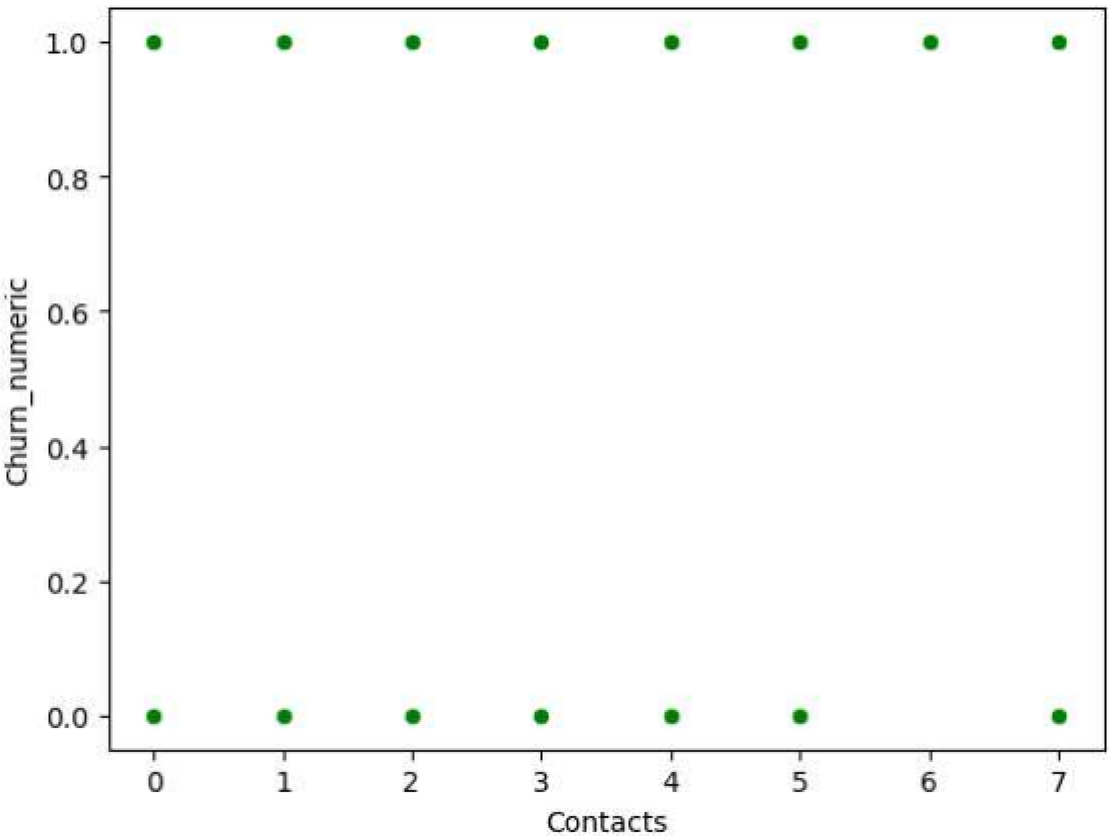


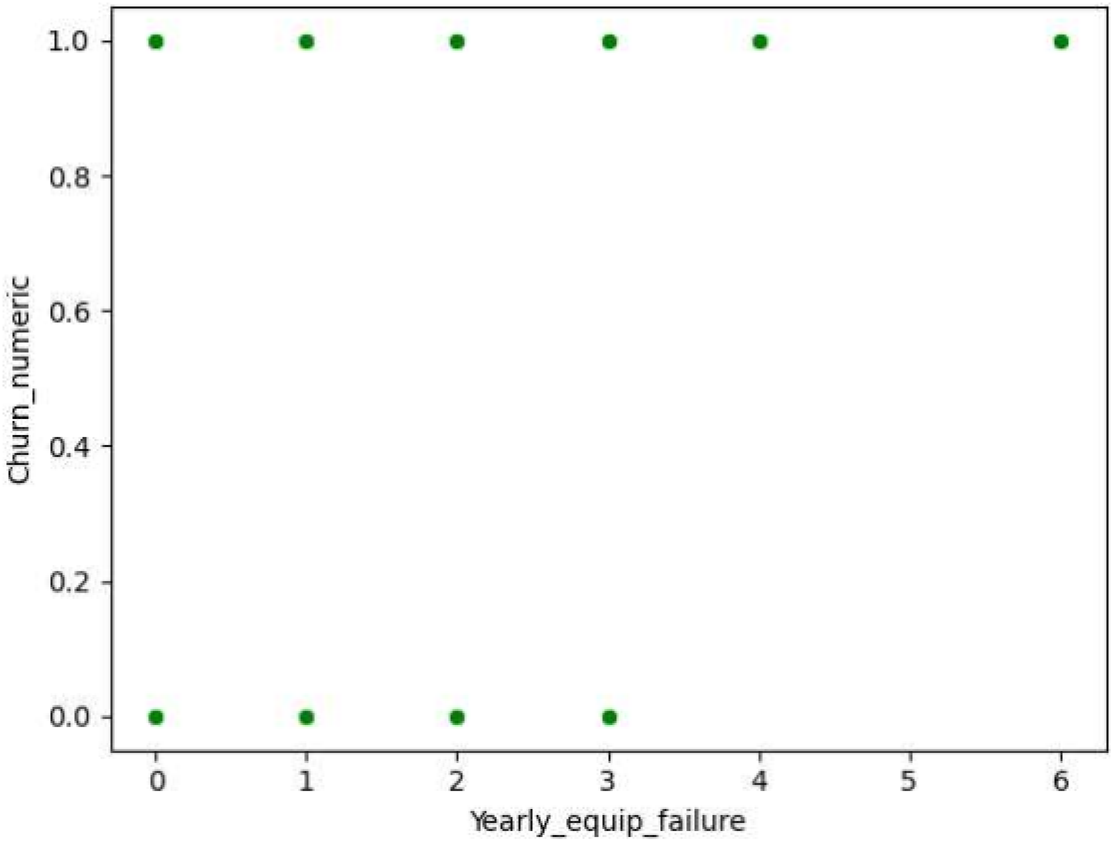
1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
20
30
40
50
60
70
80
90
Age

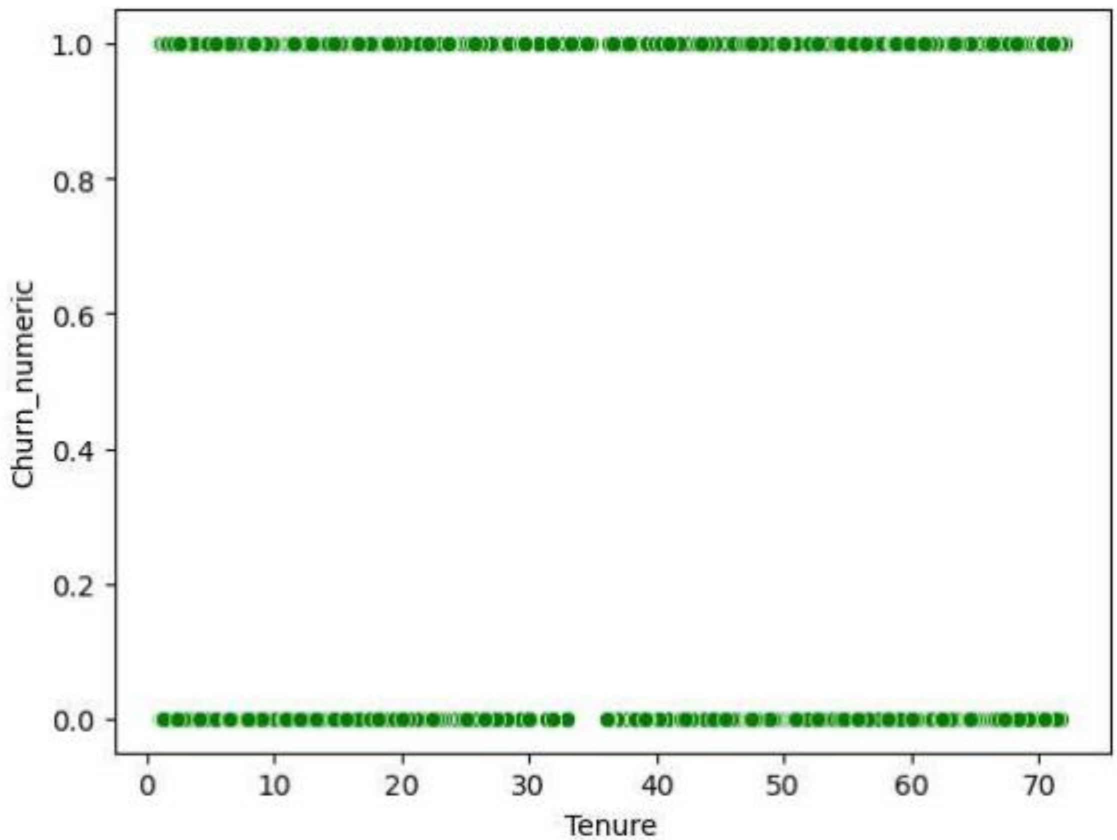

1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0
50000
100000
150000
200000
250000
Income

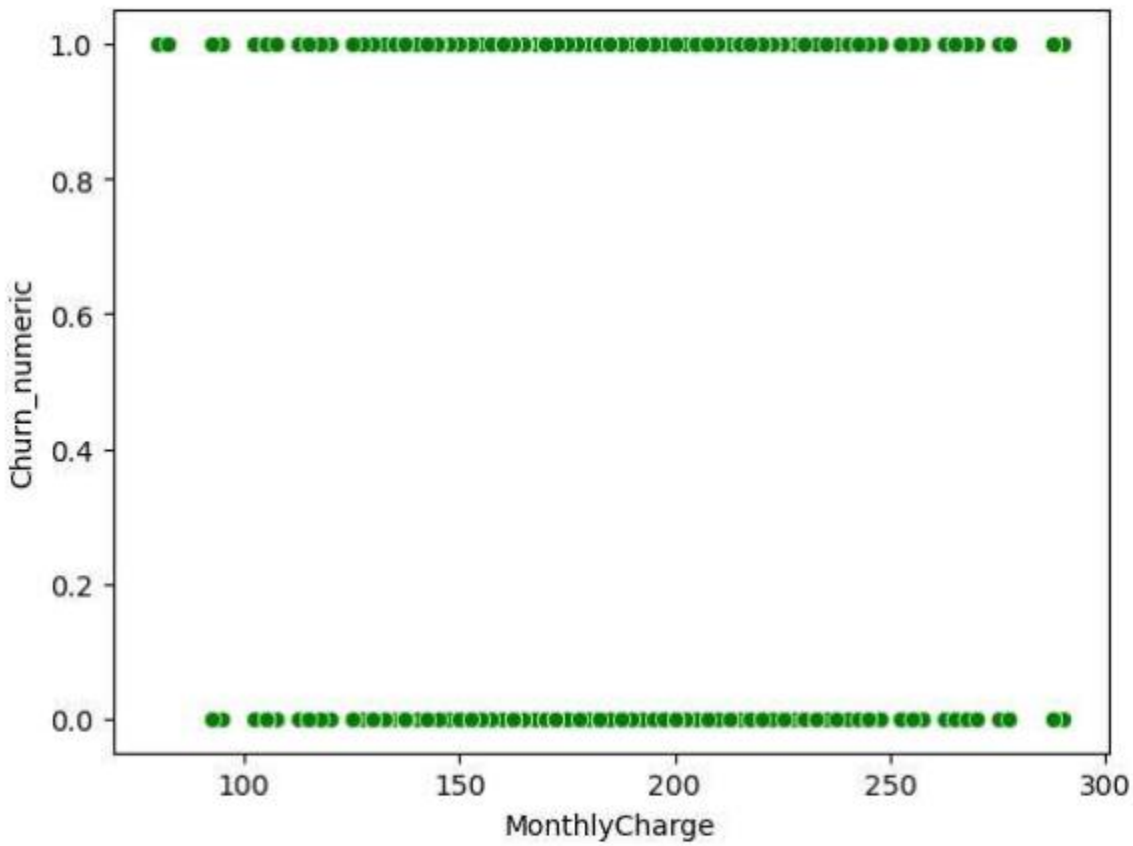

1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0
5
10
15
20
Outage_sec_perweek


1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0
5
10
15
20
Email

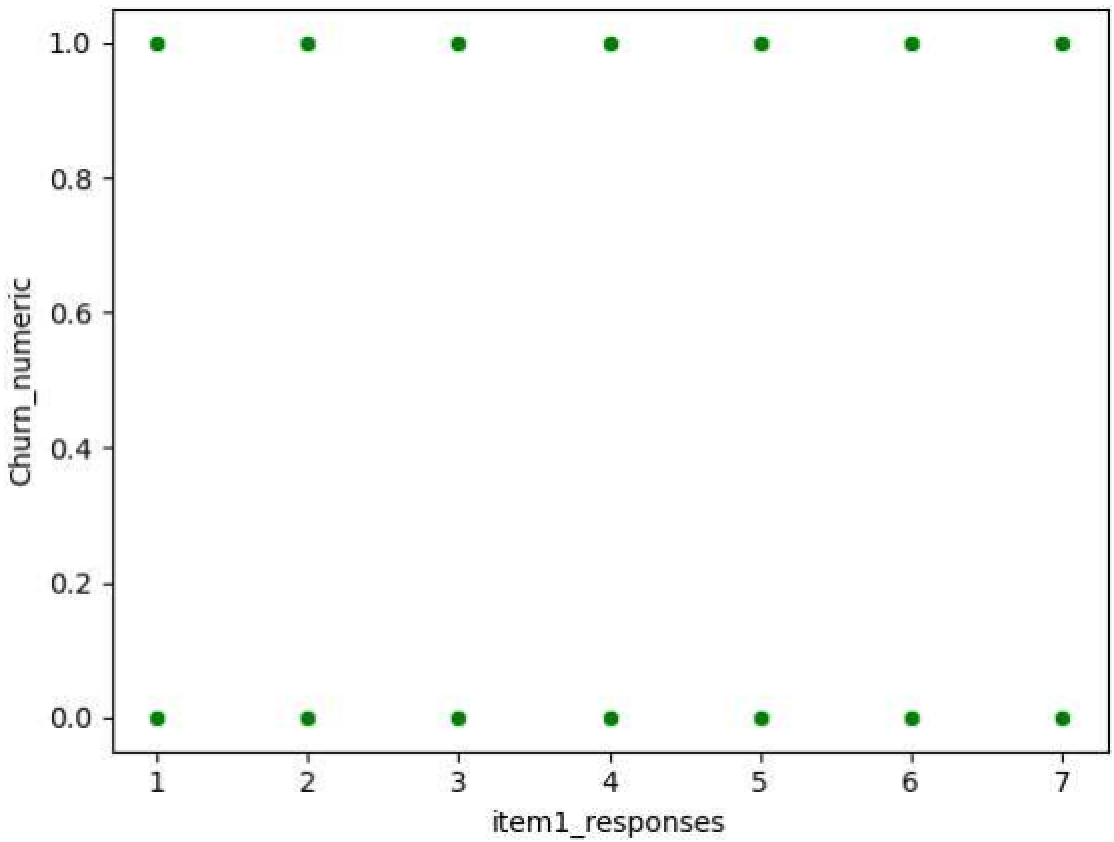

Churn_numeric
Contacts
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7

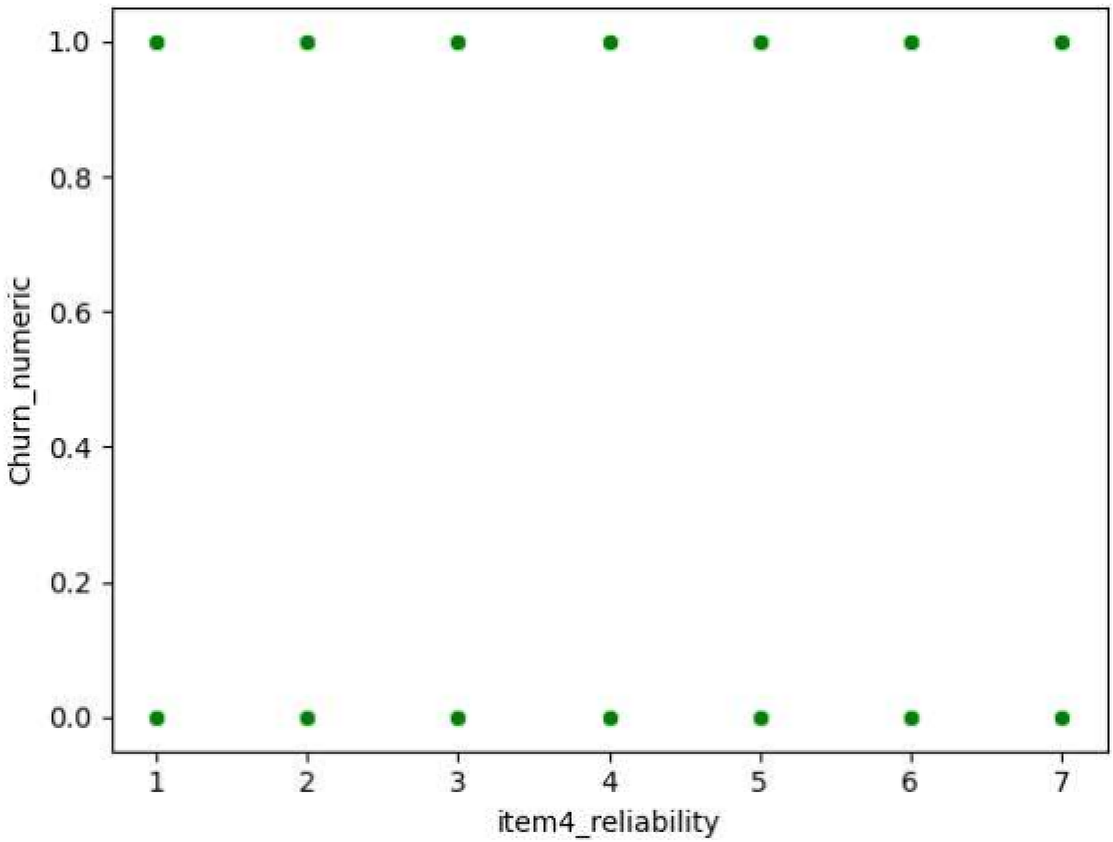

1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0
1
2
3
4
5
6
Yearly_equip_failure

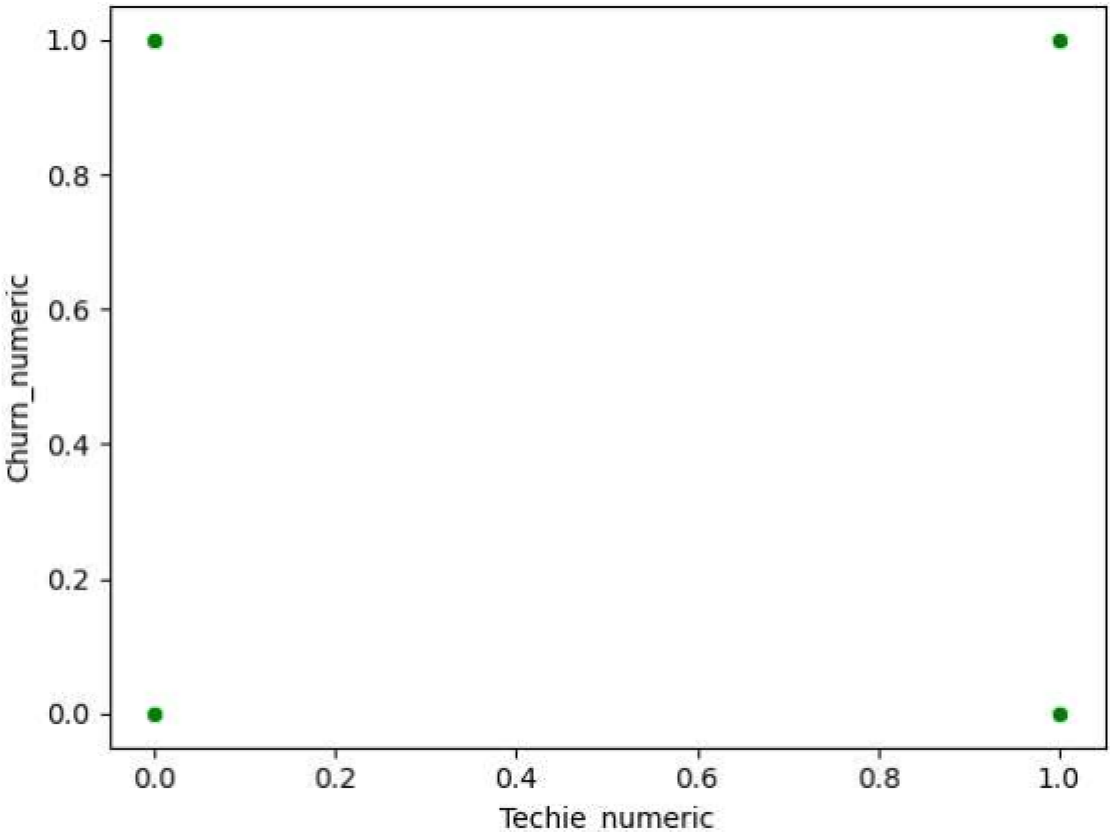

1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0
10
20
30
40
50
60
70
Tenure

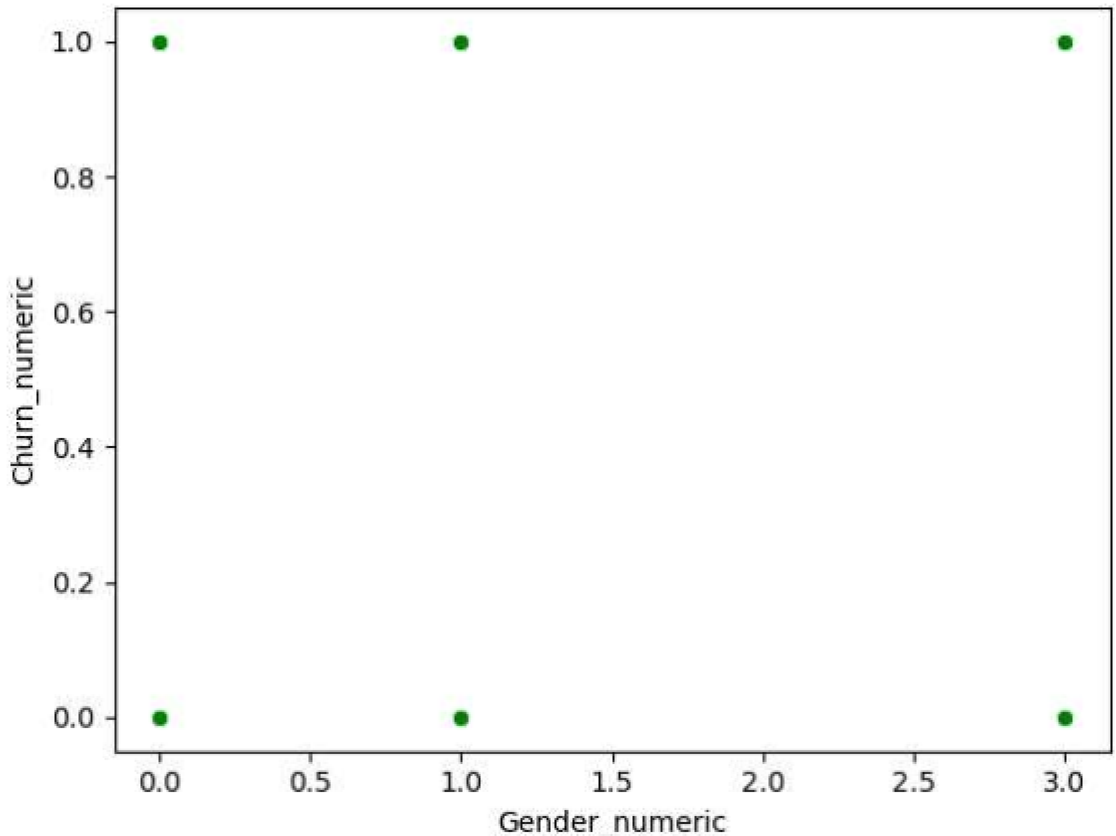


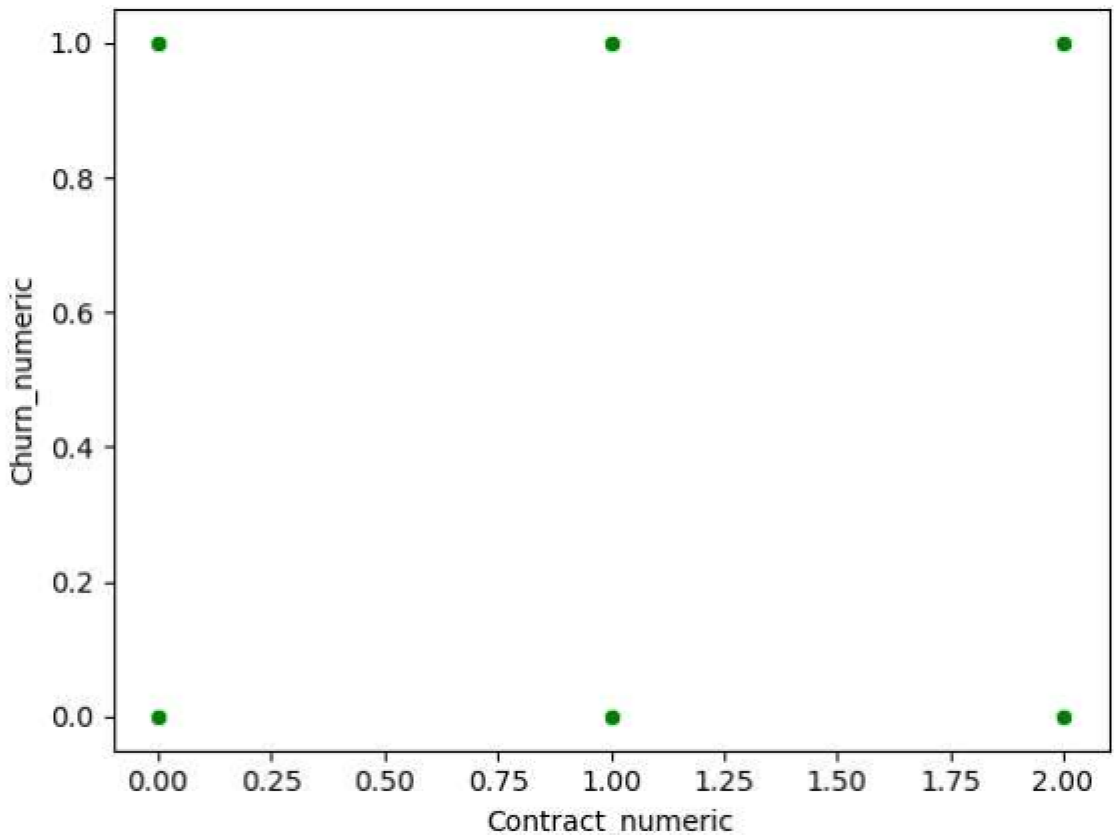
Churn_numeric
MonthlyCharge
1.0
0.8
0.6
0.4
0.2
0.0
100
150
200
250
300

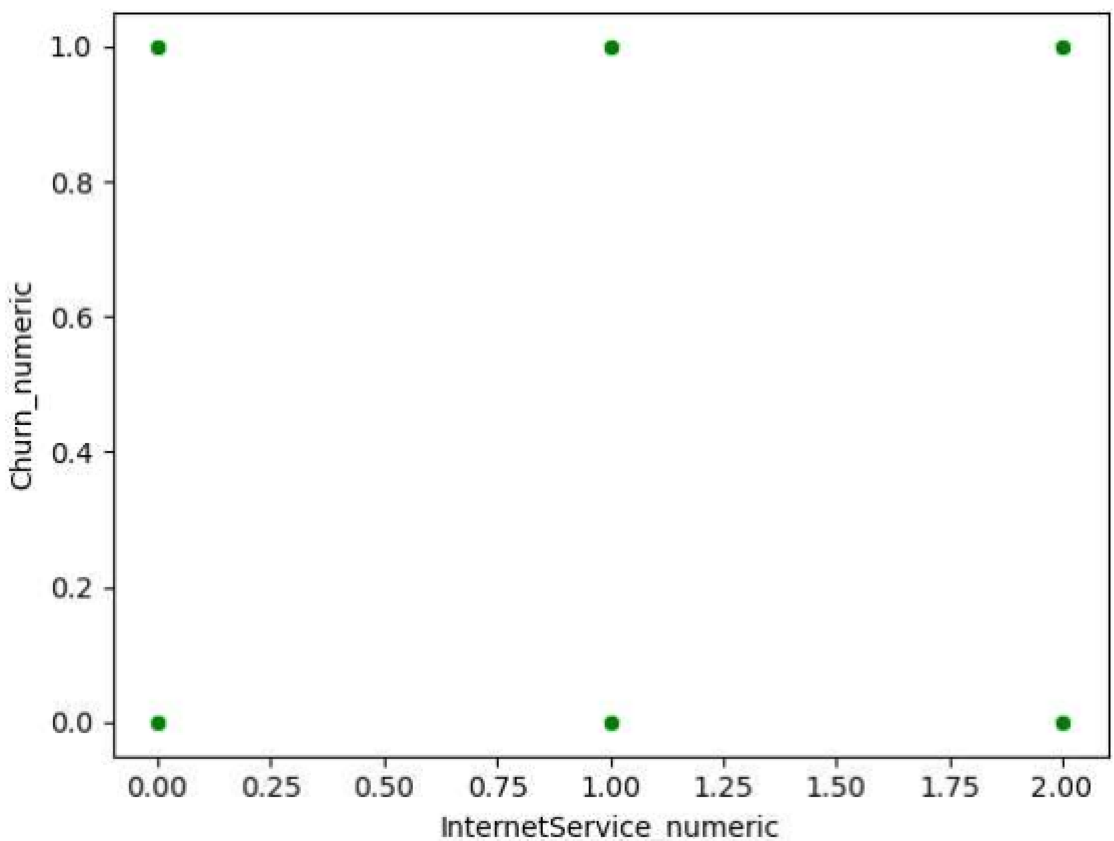

1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
1
2
3
4
5
6
7
item1_responses


1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
1
2
3
4
5
6
7
item4_reliability


1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0.0
0.2
0.4
0.6
0.8
1.0
Techie_numeric


Churn_numeric
Gender_numeric
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0


1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Contract_numeric


1.0
0.8
0.6
0.4
0.2
0.0
Churn_numeric
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
InternetService_numeric

5. The prepared data set is included as “D208_cleaned_task2”

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced logistic regression model by doing the following:

1. Construct an initial logistic regression model from *all* predictors that were identified in Part C2

2. Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.

3. Provide a reduced logistic regression model.

1. Initial logistic regression model from all predictors:
    a. Output of the model is as follows:

```
Optimization terminated successfully.
         Current function value: 0.303048
         Iterations 8
Intercept                    3.504301e+00
Children                     8.012789e-02
Age                         -1.042392e-02
Income                      -5.988466e-07
Outage_sec_perweek           2.370309e-03
Email                        7.550732e-04
Contacts                    -2.838988e-02
Yearly_equip_failure         3.481197e-02
Bandwidth_GB_Year           -2.411986e-03
MonthlyCharge               -2.182571e-02
item1_responses              9.495239e-03
item2_fixes                 -5.104751e-03
item3_replacements           1.171219e-02
item4_reliability            3.396220e-02
item5_options                3.862956e-02
item6_respectfulness         1.650727e-02
item7_courteous              1.573994e-02
item8_listening              2.267211e-03
Tenure                       2.823891e-01
Techie_numeric               7.302896e-01
Port_modem_numeric           1.200425e-01
Tablet_numeric              -5.718566e-02
Phone_numeric               -2.719007e-01
Multiple_numeric             2.670015e-01
OnlineSecurity_numeric      -3.882211e-01
OnlineBackup_numeric        -1.361517e-01
DeviceProtection_numeric    -2.024367e-01
TechSupport_numeric         -6.572530e-02
StreamingTV_numeric          5.774312e-01
StreamingMovies_numeric      8.474911e-01
PaperlessBilling_numeric     8.509312e-02
InternetService_numeric     -3.452051e-01
Contract_numeric             1.076693e-01
Gender_numeric               2.506459e-02
dtype: float64
                           Logit Regression Results
==============================================================================
Dep. Variable:          Churn_numeric   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9966
Method:                           MLE   Df Model:                           33
Date:                Sun, 29 Jan 2023   Pseudo R-squ.:                  0.4759
Time:                        18:21:34   Log-Likelihood:                -3030.5
converged:                       True   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================
```

```
                          coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-------
Intercept               3.5043      1.182      2.964      0.003       1.187
5.822
Children                0.0801      0.016      4.971      0.000       0.049
0.112
Age                    -0.0104      0.002     -6.204      0.000      -0.014
-0.007
Income              -5.988e-07   1.16e-06     -0.518      0.604   -2.86e-06
```

```
Income                -5.988e-07    1.16e-06     -0.518      0.604    -2.86e-06
```



```
1.67e-06
Outage_sec_perweek         0.0024      0.011      0.218      0.827      -0.019
0.024
Email                      0.0008      0.011      0.070      0.944      -0.020
0.022
Contacts                  -0.0284      0.033     -0.863      0.388      -0.093
0.036
Yearly_equip_failure       0.0348      0.051      0.678      0.498      -0.066
0.135
Bandwidth_GB_Year         -0.0024      0.000    -14.382      0.000      -0.003
-0.002
MonthlyCharge             -0.0218      0.004     -5.526      0.000      -0.030
-0.014
item1_responses            0.0095      0.046      0.205      0.838      -0.081
0.100
item2_fixes               -0.0051      0.044     -0.117      0.907      -0.091
0.080
item3_replacements         0.0117      0.040      0.295      0.768      -0.066
0.090
item4_reliability          0.0340      0.035      0.963      0.336      -0.035
0.103
item5_options              0.0386      0.037      1.045      0.296      -0.034
0.111
item6_respectfulness       0.0165      0.038      0.437      0.662      -0.058
0.091
item7_courteous            0.0157      0.036      0.437      0.662      -0.055
0.086
item8_listening            0.0023      0.034      0.066      0.947      -0.065
0.069
Tenure                     0.2824      0.014     19.609      0.000       0.254
0.311
Techie_numeric             0.7303      0.084      8.704      0.000       0.566
0.895
Port_modem_numeric         0.1200      0.065      1.853      0.064      -0.007
0.247
Tablet_numeric            -0.0572      0.070     -0.812      0.417      -0.195
0.081
Phone_numeric             -0.2719      0.109     -2.488      0.013      -0.486
-0.058
Multiple_numeric           0.2670      0.145      1.837      0.066      -0.018
0.552
```

OnlineSecurity_numeric	-0.3882	0.070	-5.532	0.000	-0.526	-0.251
OnlineBackup_numeric	-0.1362	0.111	-1.222	0.222	-0.354	0.082
DeviceProtection_numeric	-0.2024	0.083	-2.429	0.015	-0.366	-0.039
TechSupport_numeric	-0.0657	0.084	-0.787	0.431	-0.229	0.098
StreamingTV_numeric	0.5774	0.191	3.023	0.003	0.203	0.952
StreamingMovies_numeric	0.8475	0.225	3.764	0.000	0.406	1.289
PaperlessBilling_numeric	0.0851	0.066	1.292	0.196	-0.044	0.214
InternetService_numeric	-0.3452	0.079	-4.352	0.000	-0.501	-0.190
Contract_numeric	0.1077	0.049	2.216	0.027	0.012	0.203
Gender_numeric	0.0251	0.052	0.478	0.633	-0.078	

y = 3.504301e+00 + (Children * 8.012789e-02) + (Age * -1.042392e-02) + (Income * -5.988466e-07) + (Outage_sec_perweek * 2.370309e-03) + (Email * 7.550732e-04) + (Contacts * -2.838988e-02) + (Yearly_equip_failure * 3.481197e-02) + (Bandwidth_GB_Year * -2.411986e-03) + (MonthlyCharge * -2.182571e-02) + (item1_responses * 9.495239e-03) + (item2_fixes * -5.104751e-03) + (item3_replacements * 1.171219e-02) + (item4_reliability * 3.396220e-02) + (item5_options * 3.862956e-02) + (item6_respectfulness * 1.650727e-02) + (item7_curteous * 1.573994e-02) + (item8_listening * 2.267211e-03) + (Tenure * 2.823891e-01) + (Techie_numeric * 7.302896e-01) + (Port_modem_numeric * 1.200425e-01) + (Tablet_numeric * -5.718566e-02) + (Phone_numeric * -2.719007e-01) + (Multiple_numeric * 2.670015e-01) + (OnlineSecurity_numeric * -3.882211e-01) + (OnlineBackup_numeric * -1.361517e-01) + (DeviceProtection_numeric * -2.024367e-01) + (TechSupport_numeric * -6.572530e-02) + (StreamingTV_numeric * 5.774312e-01) + (StreamingMovies_numeric * 8.474911e-01+ (PaperlessBilling_numeric * 8.509312e-02) + (InternetService_numeric * -3.452051e-01) + (Contract_numeric * 1.076693e-01) + (Gender_numeric * 2.506459e-02)

The Pseudo-R squared is 0.4759, which indicates a 47.6% strength of prediction, which is not a very strong predictor and indicates this model can be reduced and refined to create a better model. In order to reduce, we will use stepwise reduction and remove all variables whose P values were larger than .05 from the initial model. We will then perform a Variance inflation factor (VIF) analysis and remove all variables who have a greater than 3 value for VIF.

1. The statistically based variable selection procedure is backwards stepwise reduction removing all variables whose P value is larger than .05 in the initial model. I will also use Variance Inflation Factor (VIF) to aid in this. When we look at the initial regression model, the variables whose P value are greater than .05 and therefore according the idea of backwards stepwise reduction, were removed from the dataset are:
    - Income
    - Outage_sec_perweek
    - Email
    - Contacts
    - Yearly_equip_failure
    - item1_responses
    - item2_fixes
    - item3_replacements
    - item4_reliability
    - item5_options
    - item6_respectfulness
    - item7_courteous
    - item8_listening
    - Port_modem_numeric
    - Tablet_numeric
    - Multiple_numeric
    - OnlineBackup_numeric
    - TechSupport_numeric

- PaperlessBilling_numeric
- Gender_numeric

The VIF Model is:

	VIF	variable
0	146.037288	Intercept
1	1.107329	Children
2	1.118770	Age
3	118.564331	Bandwidth_GB_Year
4	3.885663	MonthlyCharge
5	117.843491	Tenure
6	1.001073	Techie_numeric
7	1.001409	Phone_numeric
8	1.035909	OnlineSecurity_numeric
9	1.102843	DeviceProtection_numeric
10	1.999480	StreamingTV_numeric
11	2.423660	StreamingMovies_numeric
12	1.472667	InternetService_numeric
13	1.001065	Contract_numeric

We will remove Bandwidth_GB_Year, MonthlyCharge, and Tenure due to their large VIF values.

The reduction process leaves us with Children, Age, Techie_numeric, Phone_numeric, OnlineSecurity_numeric, DeviceProtection_numeric, StreamingTV_numeric, StreamingMovies_numeric, InternetService_numeric, and Contract_numeric as our predictors for our reduced model.

2. The reduced logistic regression model that includes both categorical and continuous variables is found on the next page.
    a. Output of the model:

```
Optimization terminated successfully.
         Current function value: 0.500103
         Iterations 6
Intercept                  -0.635203
Children                    0.004219
Age                        -0.000223
Techie_numeric              0.439192
Phone_numeric              -0.179682
OnlineSecurity_numeric     -0.093453
DeviceProtection_numeric    0.259570
StreamingTV_numeric         1.212397
StreamingMovies_numeric     1.505179
InternetService_numeric    -0.061330
Contract_numeric            0.095838
dtype: float64
                           Logit Regression Results
==============================================================================
Dep. Variable:          Churn_numeric   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9989
Method:                           MLE   Df Model:                           10
Date:                Sun, 29 Jan 2023   Pseudo R-squ.:                  0.1351
Time:                        20:27:48   Log-Likelihood:                -5001.0
converged:                       True   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
=====================================================================================
=======
                               coef    std err          z      P>|z|      [0.025
0.975]
-------------------------------------------------------------------------------------
-------
Intercept                   -0.6352      0.113     -5.637      0.000      -0.856
-0.414
Children                     0.0042      0.012      0.365      0.715      -0.018
0.027
Age                         -0.0002      0.001     -0.188      0.851      -0.003
0.002
Techie_numeric               0.4392      0.063      6.936      0.000       0.315
0.563
Phone_numeric               -0.1797      0.082     -2.179      0.029      -0.341
-0.018
OnlineSecurity_numeric      -0.0935      0.052     -1.809      0.071      -0.195
0.008
DeviceProtection_numeric     0.2596      0.049      5.249      0.000       0.163
0.356
StreamingTV_numeric          1.2124      0.051     23.739      0.000       1.112
1.312
StreamingMovies_numeric      1.5052      0.052     28.865      0.000       1.403
1.607
InternetService_numeric     -0.0613      0.032     -1.937      0.053      -0.123
0.001
Contract_numeric             0.0958      0.037      2.618      0.009       0.024
0.168
=====================================================================================
```

E. Analyze the data set using your reduced logistic regression model by doing the following:

1. Explain your data analysis process by comparing the initial and reduced logistic regression models, including the following elements:

- **the logic of the variable selection technique**
- **the model evaluation metric**

2. Provide the output and *any* calculations of the analysis you performed, including a confusion matrix.

Note: The output should include the predictions from the refined model you used to perform the analysis.

3. Provide the code used to support the implementation of the logistic regression models.

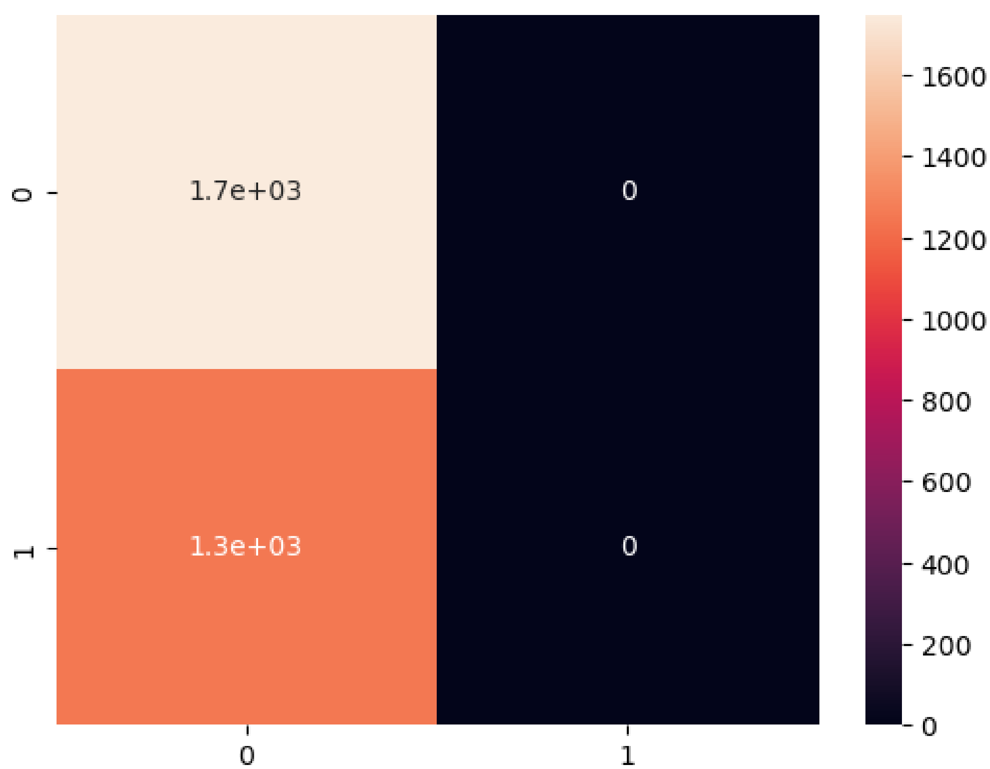
1. The data analysis summary
    a. Logic of the variable selection technique: Our target variable is a binary categorical column, Churn, and in our initial model we included all of the possible variables in the data set, excluding demographic information such as area, zip code, etc. as stated in the variable list. We used our statistical reduction metrics to choose the predictor variables for the reduced model. We used the pseudo-R value to compare the models and then a confusion matrix to evaluate the reduced model.
    b. Model evaluation metric: initial backwards stepwise reduction with VIF and classification report using the confusion matrix training for our reduced model.

2. The output of any and all calculations of the analysis are found throughout, specifically the confusion matrix and prediction is as follows:

```
Confusion Matrix
[[1749    0]
 [1251    0]]
Accuracy: 59.04 %
Standard Deviation: 0.07 %
Clasification Report
              precision    recall  f1-score   support

           0       0.58      1.00      0.74      1749
           1       0.00      0.00      0.00      1251

    accuracy                           0.58      3000
   macro avg       0.29      0.50      0.37      3000
weighted avg       0.34      0.58      0.43      3000
```



3. The code used to support implementation of the logistic regression models is found in "PA_D208_Code_Task2" and annotated to the respective part of this report.

Part V: Data Summary and Implications

F. Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following elements:

- **a regression equation for the reduced model**
- **an interpretation of coefficients of the statistically significant variables of the model**
- **the statistical and practical significance of the model**
- **the limitations of the data analysis**

2. Recommend a course of action based on your results.

1. Summary of results:

   a. Regression equation for the reduced model: y = -0.635203 + (0.004219 * Children) + (-0.000223 * Age) + (0.439192 * Techie_numeric) + (-0.179682 * Phone_numeric) + (-0.093453 * OnlineSecurity_numeric) + (0.259570 * DeviceProtection_numeric) + (1.212397 * StreamingTV_numeric) + (1.505179 * StreamingMovies_numeric) + (-0.061330 + InternetService_numeric) + (0.095838 + Contract_numeric)

b. Interpretation of coefficients of the statistically significant variables of the previous model: The coefficients suggest that for every unit of the following variables, churn will increase/decrease as described. For every one unit of…

Intercept, Churn_numeric will decrease by 0.635203

Children, Churn_numeric will increase by 0.004219

Age, Churn_numeric will decrease by 0.000223

Techie_numeric, Churn_numeric will increase by 0.439192

Phone_numeric, Churn_numeric will decrease by 0.179682

OnlineSecurity_numeric, Churn_numeric will decrease by 0.093453

DeviceProtection_numeric, Churn_numeric will increase by 0.259570

StreamingTV_numeric, Churn_numeric will increase by 1.212397

StreamingMovies_numeric, Churn_numeric will increase by 1.505179

InternetService_numeric, Churn_numeric will decrease by 0.061330

Contract_numeric, Churn_numeric will increase by 0.095838

c. The statistical significance of the model is in its ability to predict an outcome of a categorical binary target variable with several predictor variables. In this case, we have found predictor variables that can be used to predict Churn, which directly answers our research question. The pseudo-R-squared value for this model is 0.1351, which means it has a 13.51% strength of prediction. This is very low and lower than our initial model which included all of the variables in the dataset. The reduction in pseudo-R-square is highly irregular and calls for more data analysis and a review of the data acquisition and cleaning phase of this data set.

d. Limitations of data analysis: We need a larger data set to create more accurate models. This model also specifically shows prediction through correlation, not causation, so we can inform the organization of this but it is not concrete evidence to begin making organizational changes. More analysis is required and taking a look at the data acquisition phase of this project's life cycle may help us produce cleaner results. We reduced our predictor variables to statistically significant ones and used VIF to remove variables with multicollinearity, and our pseudo-R value greatly decreased. As a result of this, I would recommend the organization to increase the resources for data analysis and start the data analysis cycle over again with a larger data set.

2. The course of action I recommend is for the organization to put more resources into the variables that had a high coefficient with Churn, StreamingTV and StreamingMovies, to determine why exactly those variables have a strong relationship with Churn and what can be done to influence them. I would also recommend the organization to collect more data for more analysis on this topic and restart the data analytics cycle with more data acquisition. Additionally, the company should investigate the other statistically significant variables found in our reduced logistic regression model to evaluate the strength of prediction and if they are able to put resources into investigating the relationships found within the model. Finally, based on the reduction in pseudo-R square and low accuracy related to the confusion matrix, I would recommend the organization

put more resources into restarting the data analytic life cycle for this project and focus on gathering more data. The accuracy related to the confusion matrix is also low, indicating a problematic model or dataset.

Part VI: Demonstration

G. Provide a Panopto video recording that includes *all* of the following elements:

- **a demonstration of the functionality of the code used for the analysis**
- **an identification of the version of the programming environment**
- **a comparison of the two logistic regression models you used in your analysis**
- **an interpretation of the coefficients**

Link to the panopto presentation:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=fc34328b-45ae-4df1-be3b-af9b01629289

H. List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

References

Li, S. (2019, February 27). *Building a logistic regression in Python, step by step*. Medium. Retrieved January 27, 2023, from https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

References

What is logistic regression? IBM. (n.d.). Retrieved January 27, 2023, from https://www.ibm.com/topics/logistic-regression

Swaminathan, S. (2019, January 18). *Logistic regression - detailed overview*. Medium. Retrieved January 27, 2023, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

J. Demonstrate professional communication in the content and presentation of your submission.

This aspect cannot be summarized; however, I hope it has shown through in all aspects of this report.