D209 Data Mining I Performance Task 2

Sean Simmons

WDU Data Analytics

MSDA D209

February 2023

Part I: Research Question

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

1. Propose one question relevant to a real-world organizational situation that you will answer using one of the following prediction methods:

- **decision trees**
- **random forests**
- **advanced regression (i.e., lasso or ridge regression)**

2. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

1. Can we use the independent variables in the dataset to predict the Tenure length of a customer? Tenure is defined as the time, in months, the customer is with the organization (continuous values). I will use the decision tree regression method as the prediction method to answer this question. Decision tree regression is a supervised machine learning algorithm that can work with continuous variables to make predictions.
2. One goal of this analysis is to predict the tenure length of a customer using decision tree regression. This goal is reasonable within the organization because we can split our data into training and test sets and use the decision tree to predict the continuous value of Tenure with the other variables. The organization can benefit from this goal because it can inform their retention efforts and help inform decisions about how to keep customers with the organization. It can also inform the organization what characteristics or areas of the business they should put resources into to increase customer retention length.

Part II: Method Justification

B. Explain the reasons for your chosen prediction method from part A1 by doing the following:

1. Explain how the prediction method you chose analyzes the selected data set. Include expected outcomes.

2. Summarize one assumption of the chosen prediction method.

3. List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

1. A decision tree is a machine learning algorithm that uses past outcomes to predict a future outcome. It divides the dataset into smaller groups based on descriptive features. The data set is divided until it reaches a small enough sample to be described by a single label. Decision trees can use continuous variables as their target and can identify non-linear relationships through regression. For decision trees' feature variables, scaling is not necessary (standardization or normalization) and missing values do not present a large problem. As our target variable is continuous, we need to use decision tree regression. The root node is the beginning of the tree and a branch is the link between nodes.
2. One assumption of Decision Tree prediction is that the entire training dataset is considered as the root and all records are distributed recursively based on the attribute value (Vishalmendekarhere 2021).
3. The packages and their justification are as follows:
   a. Pandas - To load my dataset into my coding environment and perform basic manipulations.

b. Numpy - To perform mathematical operations with the dataset and scientific calculations.

c. Matplotlib, including subpackages - To perform variable analysis and provide visualizations for the models

d. Sklearn, including subpackages - To perform the decision tree regression functions and modeling, all of our data splitting, training, fitting, and modeling are done through sklearn.

e. Seaborn - To perform univariate and bivariate analysis - provides the visualization and calculations.

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:

1. Describe one data preprocessing goal relevant to the prediction method from part A1.

2. Identify the initial data set variables that you will use to perform the analysis for the prediction question from part A1, and group *each* variable as continuous or categorical.

3. Explain the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

4. Provide a copy of the cleaned data set.

1. One data preprocessing goal for our decision tree regression is to convert our categorical variables to numeric in order for them to be included in the analysis. The code for these can be seen in the file provided and changes all of the categorical values to original name_numeric with 0 and 1 for Yes/No and increasing counts for variables with more than 2 different values.

2. The initial data set variables we will use to perform the analysis of Tenure are as follows: Our target, y, value is Tenure. The other variables in the dataset are our feature variables, found below.

Variable Name	Data Type
Churn	Categorical
Outage_Sec_perweek	Continuous
Contract	Continuous
Tenure- Target	Continuous
MonthlyCharge	Continuous
Bandwidth_GB_year	Continuous
Email	Continuous
Yearly_equip_failure	Continuous
Contacts	Continuous
Children	Continuous
Age	Continuous
Income	Continuous
Gender	Categorical

DeviceProtection	Categorical
Phone	Categorical
Multiple	Categorical
OnlineSecurity	Categorical
OnlineBackup	Categorical
TechSupport	Categorical
StreamingTV	Categorical
StreamingMovies	Categorical
Techie	Categorical
Port_modem	Categorical
Tablet	Categorical
InternetService	Categorical
Paperless Billing	Categorical
Item 1	Categorical- Discrete Ordinal
Item 2	Categorical- Discrete Ordinal
Item 3	Categorical- Discrete Ordinal
Item 4	Categorical- Discrete Ordinal
Item 5	Categorical- Discrete Ordinal
Item 6	Categorical- Discrete Ordinal
Item 7	Categorical- Discrete Ordinal
Item 8	Categorical- Discrete Ordinal

```
Floats
<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
Integers
<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
```

```
Objects
<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
```

PA_D209_(

```
 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
```

```
Dataset Information
<bound method DataFrame.info of       CaseOrder Customer_id
Interaction  \
0             1     K409198  aa90260b-4141-4a24-8e36-b04ce1f4f77b
1             2     S120509  fb76459f-c047-4a9d-8af9-e0f7d4ac2524
2             3     K191035  344d114c-3736-4be5-98f7-c72c281e2d35
3             4      D90850  abfa2b40-2d43-4994-b15a-989b8c79e311
4             5     K662701  68a861fd-0d20-4e51-a587-8a90407ee574
...         ...         ...                                   ...
9995       9996     M324793  45deb5a2-ae04-4518-bf0b-c82db8dbe4a4
9996       9997     D861732  6e96b921-0c09-4993-bbda-a1ac6411061a
9997       9998     I243405  e8307ddf-9a01-4fff-bc59-4742e03fd24f
9998       9999     I641617  3775ccfc-0052-4107-81ae-9657f81ecdf3
9999      10000      T38070  9de5fb6e-bd33-4995-aec8-f01d0172a499

                                   UID         City State  \
0     e885b299883d4f9fb18e39c75155d990   Point Baker    AK
1     f2de8bef964785f41a2959829830fb8a  West Branch    MI
2     f1784cfa9f6d92ae816197eb175d3c71      Yamhill    OR
3     dc8a365077241bb5cd5ccd305136b05e      Del Mar    CA
4     aabb64a116e83fdc4befc1fbab1663f9   Needville    TX
...                                ...          ...   ...
9995  9499fb4de537af195d16d046b79fd20a  Mount Holly    VT
9996  c09a841117fa81b5c8e19afec2760104  Clarksville    TN
9997  9c41f212d1e04dca84445019bbc9b41c     Mobeetie    TX
9998  3e1f269b40c235a1038863ecf6b7a0df   Carrollton    GA
9999  0ea683a03a3cd544aefe8388aab16176 Clarkesville    GA

                    County    Zip       Lat        Lng  ...  MonthlyCharge  \
0     Prince of Wales-Hyder  99927  56.25100 -133.37571  ...     172.455519
1                   Ogemaw  48661  44.32893  -84.24080  ...     242.632554
2                  Yamhill  97148  45.35589 -123.24657  ...     159.947583
3                San Diego  92014  32.96687 -117.24798  ...     119.956840
4                Fort Bend  77461  29.38012  -95.80673  ...     149.948316
...                    ...    ...       ...        ...  ...            ...
9995               Rutland   5758  43.43391  -72.78734  ...     159.979400
9996            Montgomery  37042  36.56907  -87.41694  ...     207.481100
9997               Wheeler  79061  35.52039 -100.44180  ...     169.974100
9998               Carroll  30117  33.58016  -85.13241  ...     252.624000
9999             Habersham  30523  34.70783  -83.53648  ...     217.484000

     Bandwidth_GB_Year Item1 Item2  Item3  Item4  Item5 Item6 Item7 Item8
0           904.536110     5     5      5      3      4     4     3     4
```

1	800.982766	3	4	3	3	4	3	4	4
2	2054.706961	4	4	2	4	4	3	3	3
3	2164.579412	4	4	4	2	5	4	3	3
4	271.493436	4	4	4	3	4	4	4	5
...	...	...	...	...	...	...	...	...	...
9995	6511.252601	3	2	3	3	4	3	2	3
9996	5695.951810	4	5	5	4	4	5	2	5
9997	4159.305799	4	4	4	4	4	4	4	5
9998	6468.456752	4	4	6	4	3	3	5	4
9999	5857.586167	2	2	3	3	3	3	4	1

Missing Values:

```
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
```

Mean and then Median below:

```
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
```

```
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
```

```
                                   PA_[
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. Steps of the data preparation and code are written and annotated in "D209_Task2_Code.ipynb", I have added the steps here just for summary (letter a for example will be in the code as a comment "#a").
    a. Import the data into my coding environment
    b. View the data type and summary information to prepare for the modeling
    c. Rename nondescript columns of items
    d. Check for missing values
    e. Change categorical values into numeric counterparts (except for the discrete ordinal survey items columns)
    f. Check summary statistics, such as mean/median/mode
    g. Drop unnecessary columns that will not be included in our model, as described earlier in the report (mainly demographic and unique customer id information).
    h. Perform univariate and bivariate analysis of target/feature variables
    i. Extract data set to csv file. The data is now prepared.
4. The copy of the cleaned data set "D209_Task2_clean.csv"

Part IV: Analysis

D. Perform the data analysis and report on the results by doing the following:

1. Split the data into training and test data sets and provide the file(s).

2. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

3. Provide the code used to perform the prediction analysis from part D2.

1. The training set is found in “Task2_X_Train.csv”, “Task2_Y_Train”, “Task2_X_test.csv”, “Task2_Y_test.csv”and “Task2_Y_Predict.csv.” The data is split into 70% for training and 30% for testing.
2. The techniques to measure the accuracy of the model are included in the bulleted list below. The best parameters and score the model produced are also included in the output. The values (especially the best score) are peculiar and show more investigation is necessary for this dataset and target variable. Finally, I have included the graph of our predicted and test data overlap. While not wholly helpful, it does show peaks at either end of non-overlapped data with most of the model with the test and prediction data following the same pattern.
    - Mean Absolute Error (MAE)- This metric represents the difference between the original and predicted values. It works by extracting the averaged difference over the entire data set. The closer to 0 the better for this value and shows how accurate the model is by how far away the predicted values are from the original values. Our model’s MAE: 3.2214307902546984. While above 1, this number needs to be judged by the organization to see whether it is within acceptable bounds as the median for Tenure is 34, I would say this is reasonable for MAE.

- **Mean Square Error (MSE)-** This metric also represents the difference between the original and predicted values, but uses the squared average over the datasets, effectively showing us our level of error. MSE also shows how accurate the model is by how far away the predicted values are from the original values. The closer to 0 the more accurate the model. Our model's MSE: 17.49492067864168. 17.5 is a rather large number, but converting this to the Root mean square error can give us more insight. As it stands, 17 shows this model has accuracy issues.
- Root mean square error (RMSE)- This metric shows how far the data points are from the regression line. It is the square root of MSE. This model gives a better idea of the accuracy and ability to predict of our model by highlighting the error. Our model's RMSE: 4.18269299359177. This value should be as close to 0 as possible to be considered accurate. Usually values below 1 are considered accurate and anything above 1 is left up to the organization to decide if it is within acceptable variation.
- R squared- The R squared value is the coefficient of determination, it shows how well our values fit compared to the original values. This can be used for an accuracy measurement. The higher the value is the better. Our model's R squared: 0.9762896822926533. 97.6% is a great R squared and represents the model's ability to predict, however, this does not align with the relatively large RMSE value and indicates more investigation into this model is needed.

The outputs and screenshots are found below:

a. Output

#i: List features for analysis
Features **=** (list(churn_df.columns[:-1]))
print('Features for analysis include: \n', Features)
Features for analysis include:

['Children', 'Age', 'Income', 'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge', 'Bandwidth_GB_Year', 'item1_responses', 'item2_fixes', 'item3_replacements', 'item4_reliability', 'item5_options', 'item6_respectfulness', 'item7_courteous', 'item8_listening', 'Churn_numeric', 'Area_numeric', 'Marital_numeric', 'Gender_numeric', 'Contract_numeric', 'PaymentMethod_numeric', 'InternetService_numeric', 'Techie_numeric', 'Port_modem_numeric', 'Tablet_numeric', 'Phone_numeric', 'Multiple_numeric', 'OnlineSecurity_numeric', 'OnlineBackup_numeric', 'DeviceProtection_numeric', 'TechSupport_numeric', 'StreamingTV_numeric', 'StreamingMovies_numeric']

```
#:Fit dataframe to Decision Tree Regressor model
dt.fit(X_train, y_train)
```

```
DecisionTreeRegressor(max_depth=8, min_samples_leaf=0.1, random_state=1)
```

```
#: Calculate MSE for test set
mse_dt = MSE(y_test, y_pred)
print(mse_dt)
```

```
17.49492067864168
```

```
#: Now we can calculate RMSE
rmse_dt = mse_dt**(1/2)
print('RMSE calculation #1:', rmse_dt)
```

```
RMSE calculation #1: 4.18269299359177
```

```
#: Calculate & print the RMSE, a little redundant but I like to verify
RMSE = MSE(y_test, y_pred)**(1/2)
#n: Print the Root Mean Squared Error
print('Root Mean Squared Error calculation #2: {:.3f} '.format(RMSE))
```

```
Root Mean Squared Error calculation #2: 4.183
```

```
#n:Fit model to
dt_cv.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
GridSearchCV(cv=5, estimator=DecisionTreeRegressor(), n_jobs=-1,
             param_grid={'max_depth': [4, 6, 8],
                         'max_features': ['log2', 'sqrt'],
                         'min_samples_leaf': [0.1, 0.2]},
             scoring='neg_mean_squared_error', verbose=1)
```

```
#n: Print best parameters
print('Best parameters for this Decision Tree Regressor model: {}'.format(dt_cv.best_params_))
```

```
Best parameters for this Decision Tree Regressor model: {'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 0.1}
```

```
#n: Generate model best score
print('Best score for this Decision Tree Regressor model: {:.3f}'.format(dt_cv.best_score_))
```

```
Best score for this Decision Tree Regressor model: -227.066
```

```
#: Print R squared
score = dt.score(X_train, y_train)
print("R-squared:", score)
```

```
R-squared: 0.9762896822926533
```

```
#: Print MAE for another accuracy metric, MAE is the different between the original and predicted values.
mae = metrics.mean_absolute_error(y_test, y_pred)
print('MAE:',mae)
```

```
MAE: 3.2214307902546984
```

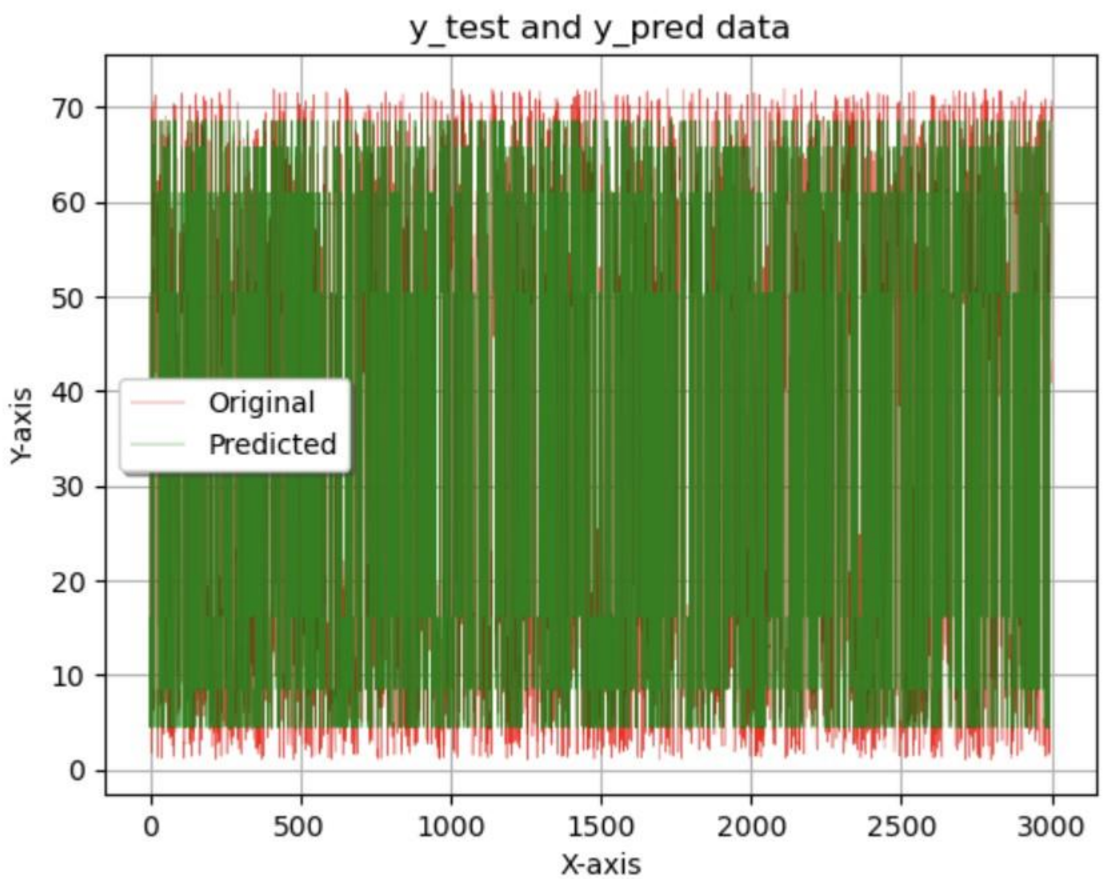
```
#: Calculate the coefficient of determination (R-squared)
scores = cross_val_score(dt, X, y, scoring='r2')
#: Print R-squared value
print('Cross validation R-squared values: ', scores)
```

```
Cross validation R-squared values:  [0.60953463 0.56320821 0.97535966 0.70298371 0.7110478 ]
```

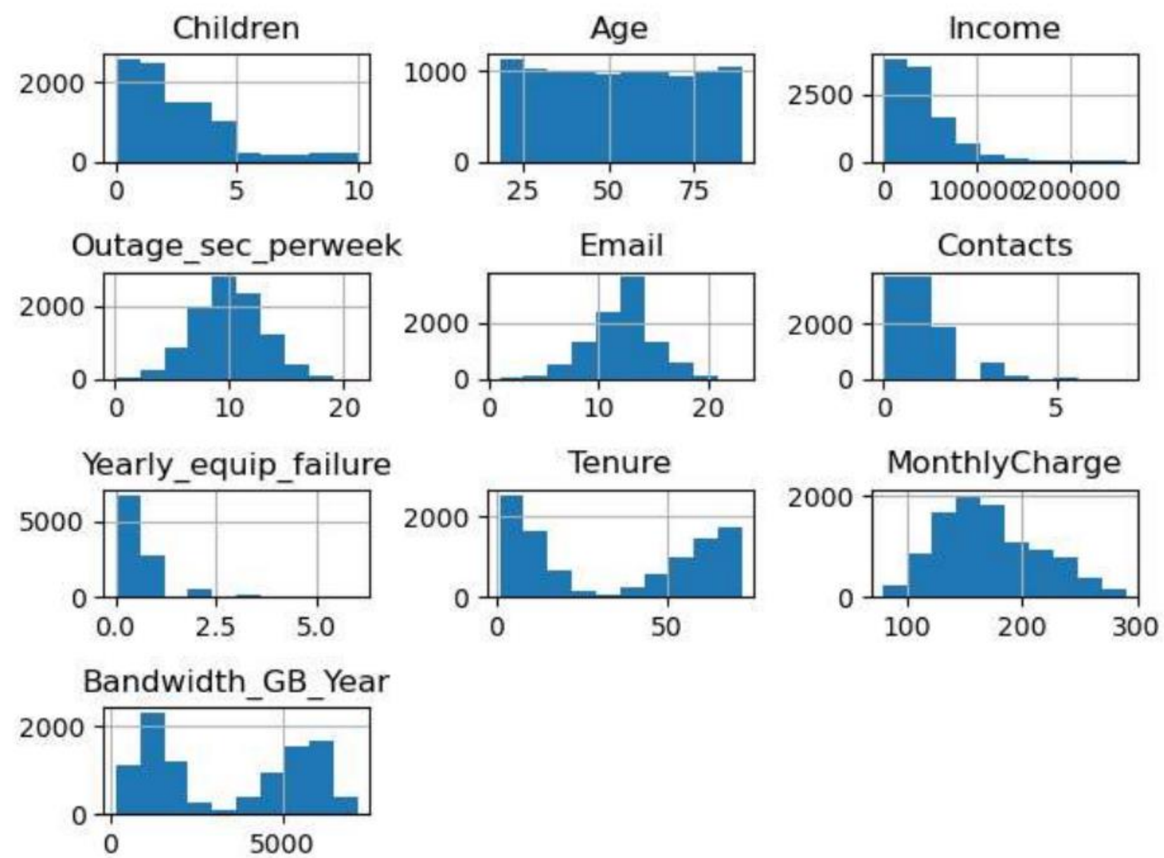
```
#: Get parameters of Decision Tree Regression model for cross validation
dt.get_params()
```

```
{'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': 8,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 0.1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'random_state': 1,
 'splitter': 'best'}
```

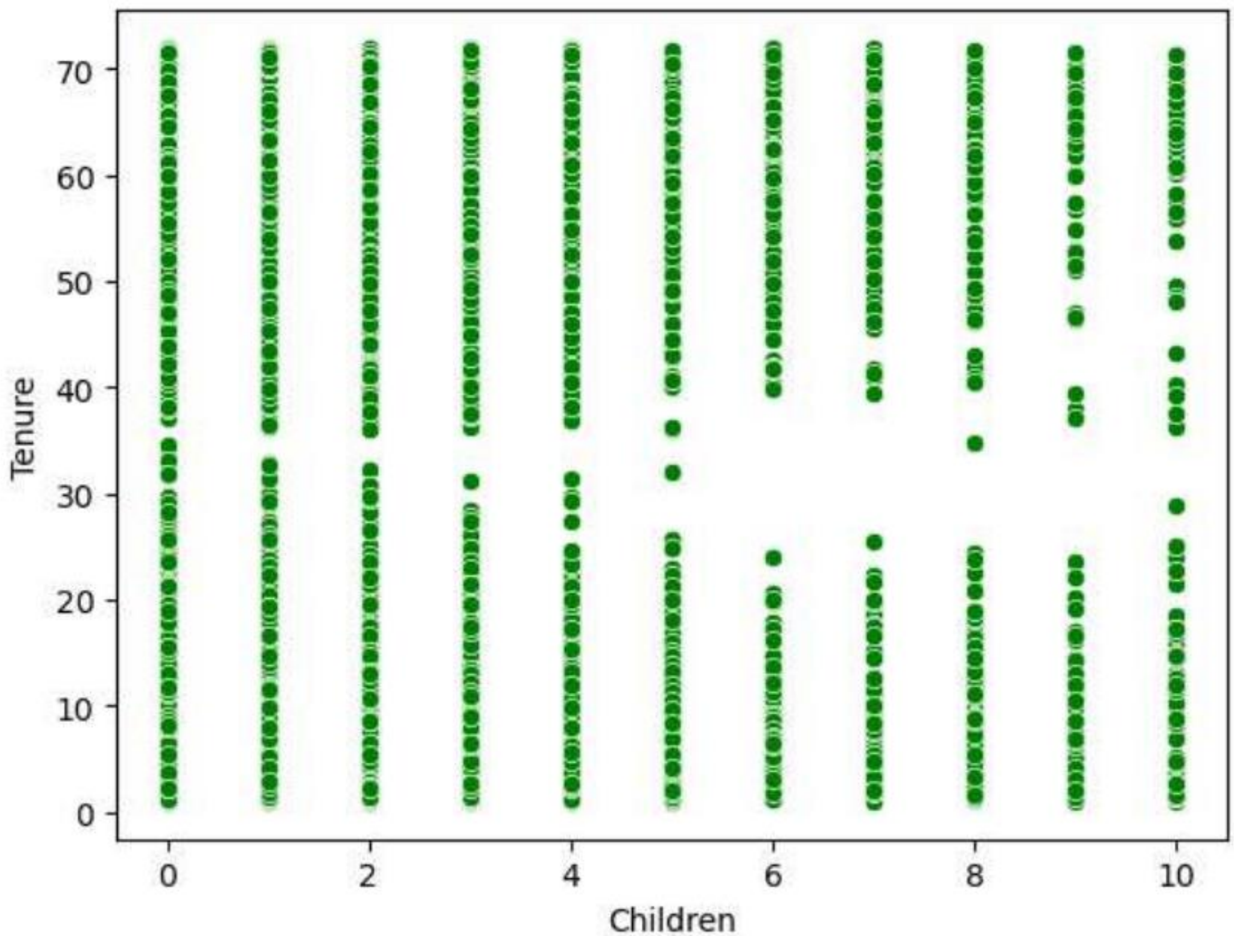
```
x = range(len(y_test))
mpl.plot(x, y_test, color='red', linewidth=.25, label="Original")
mpl.plot(x, y_pred, color='green', linewidth=.25, label="Predicted")
mpl.title("y_test and y_pred data")
mpl.xlabel('X-axis')
mpl.ylabel('Y-axis')
mpl.legend(loc='best',fancybox=True, shadow=True)
mpl.grid(True)
mpl.show()
```

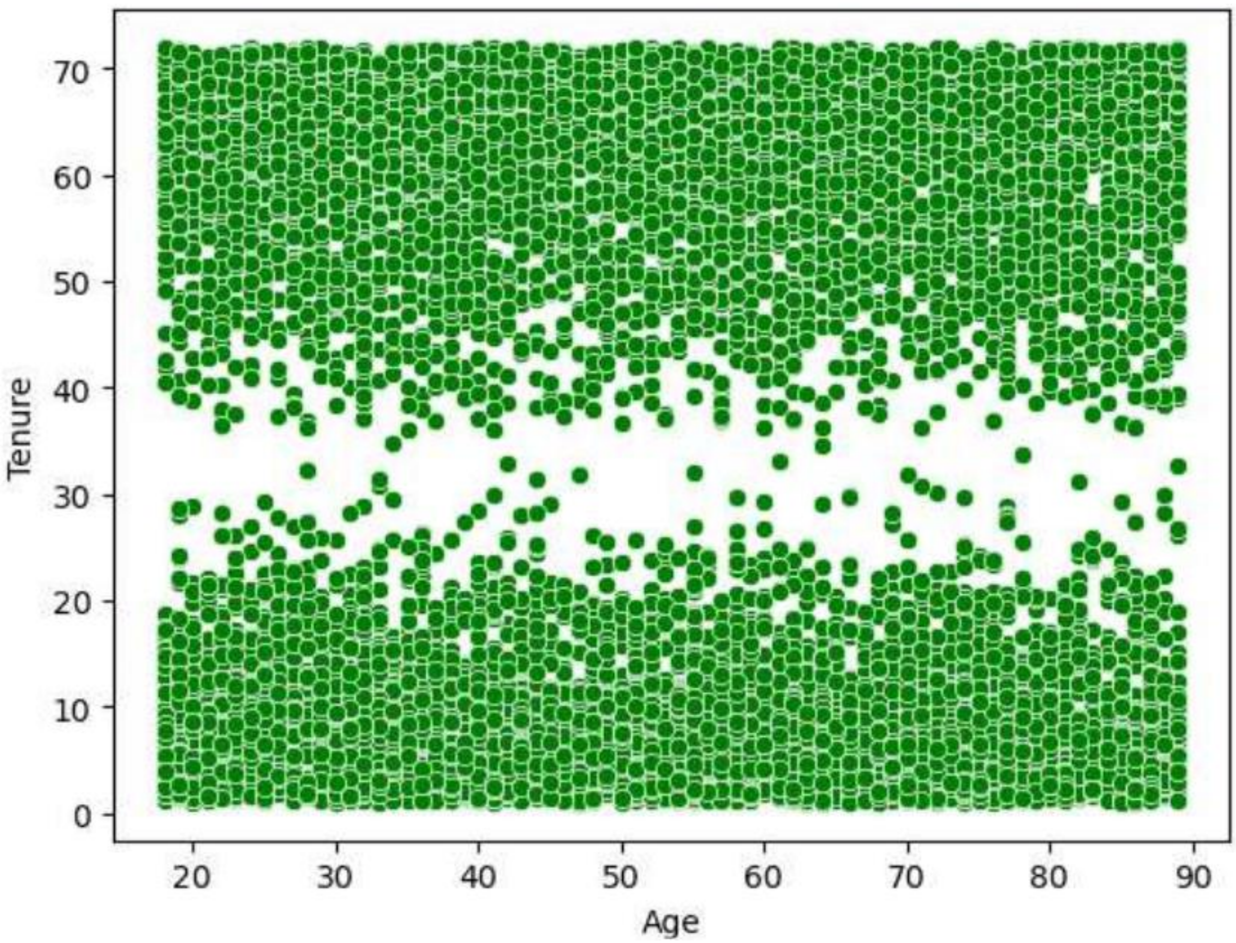


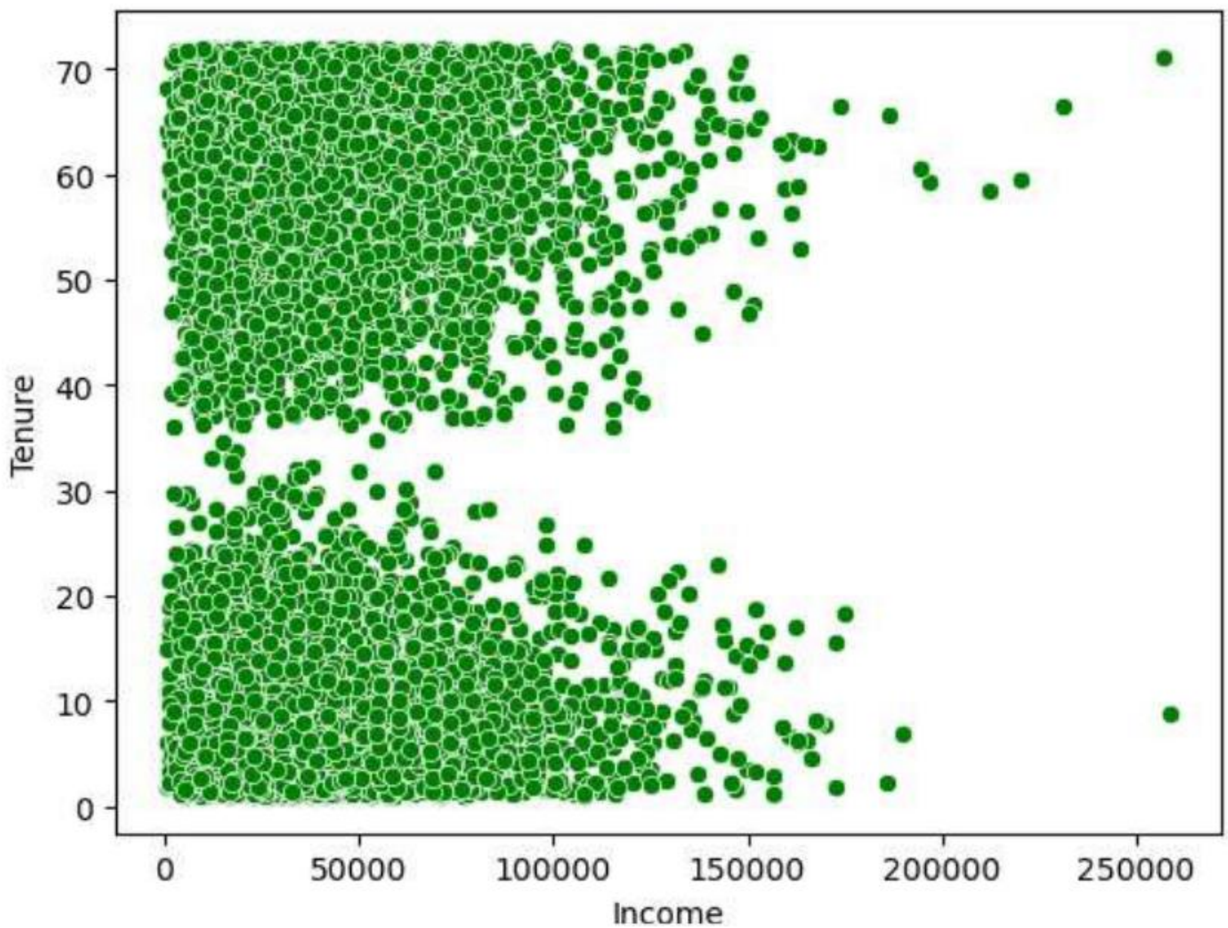
Univariate analysis

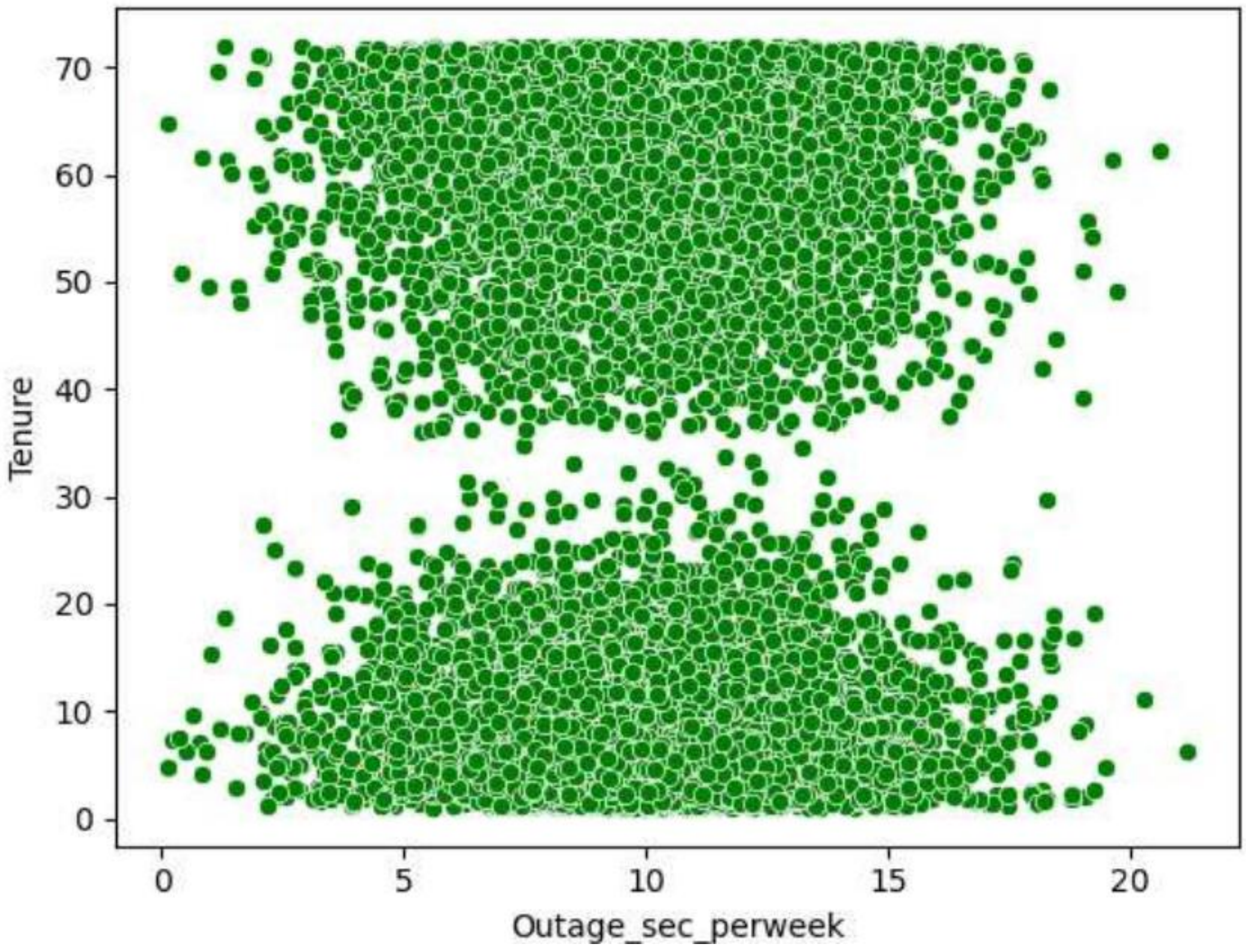


Bivariate Analysis:

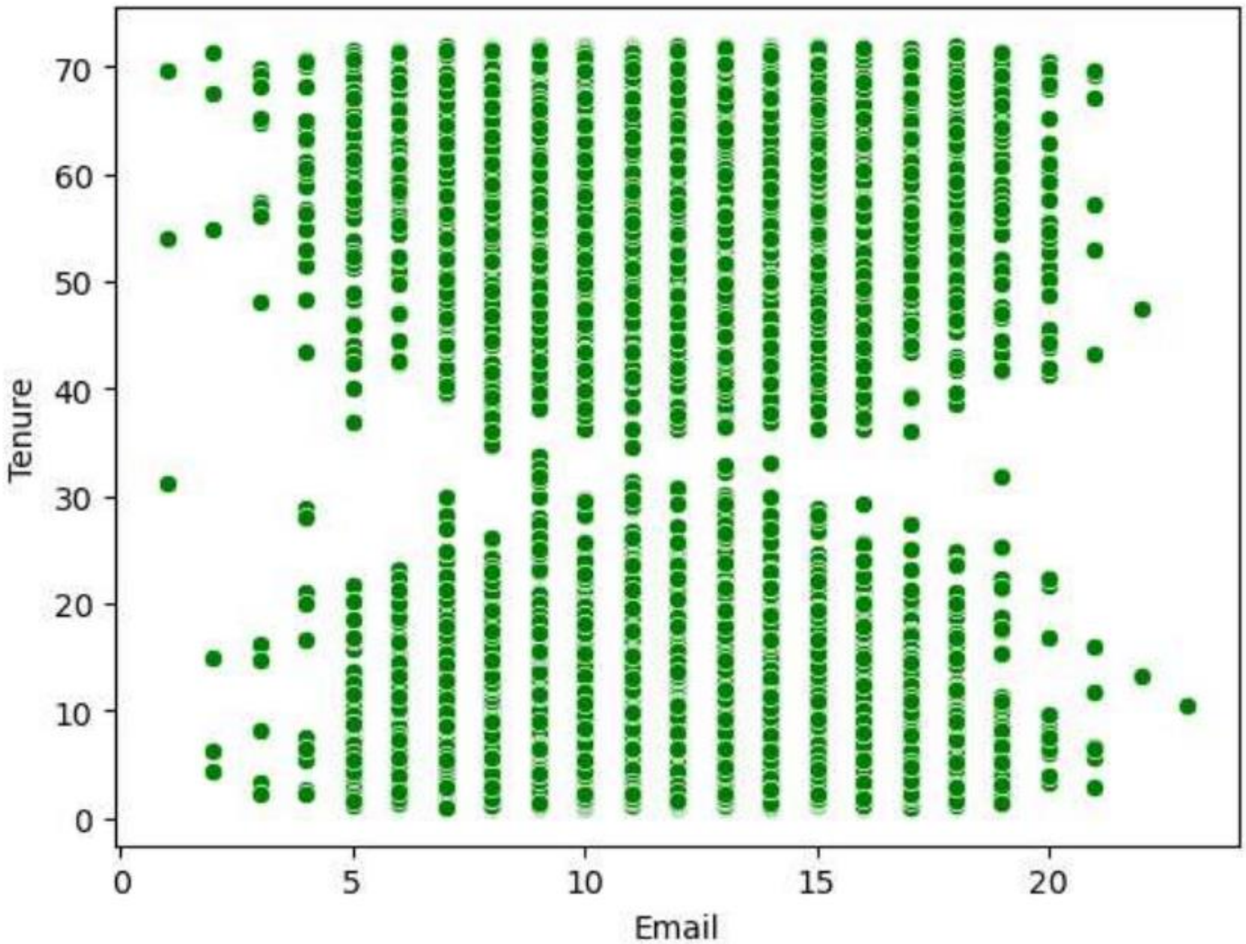


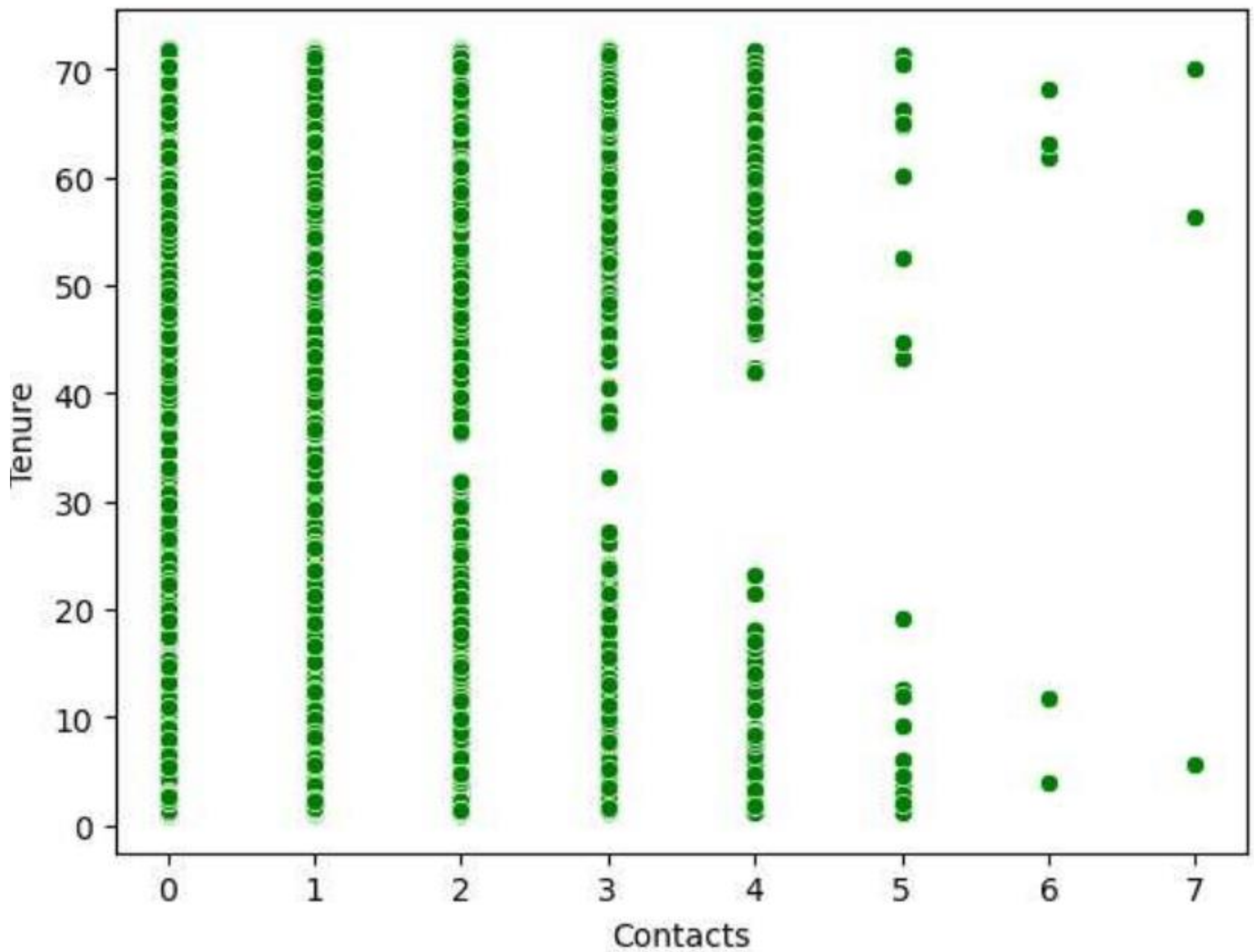

Tenure
Age
0
10
20
30
40
50
60
70
20
30
40
50
60
70
80
90

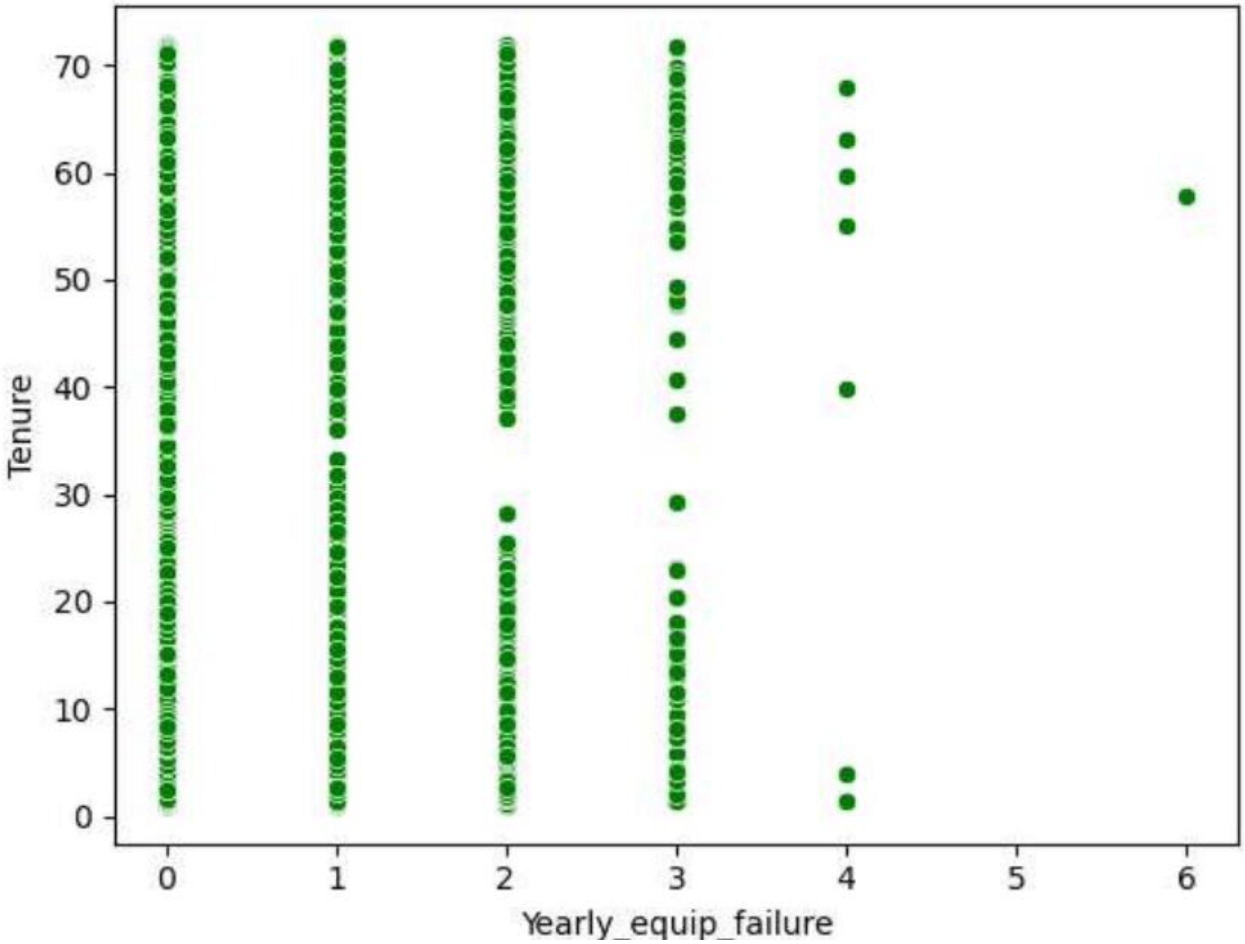

Tenure
Income
0
10
20
30
40
50
60
70
0
50000
100000
150000
200000
250000



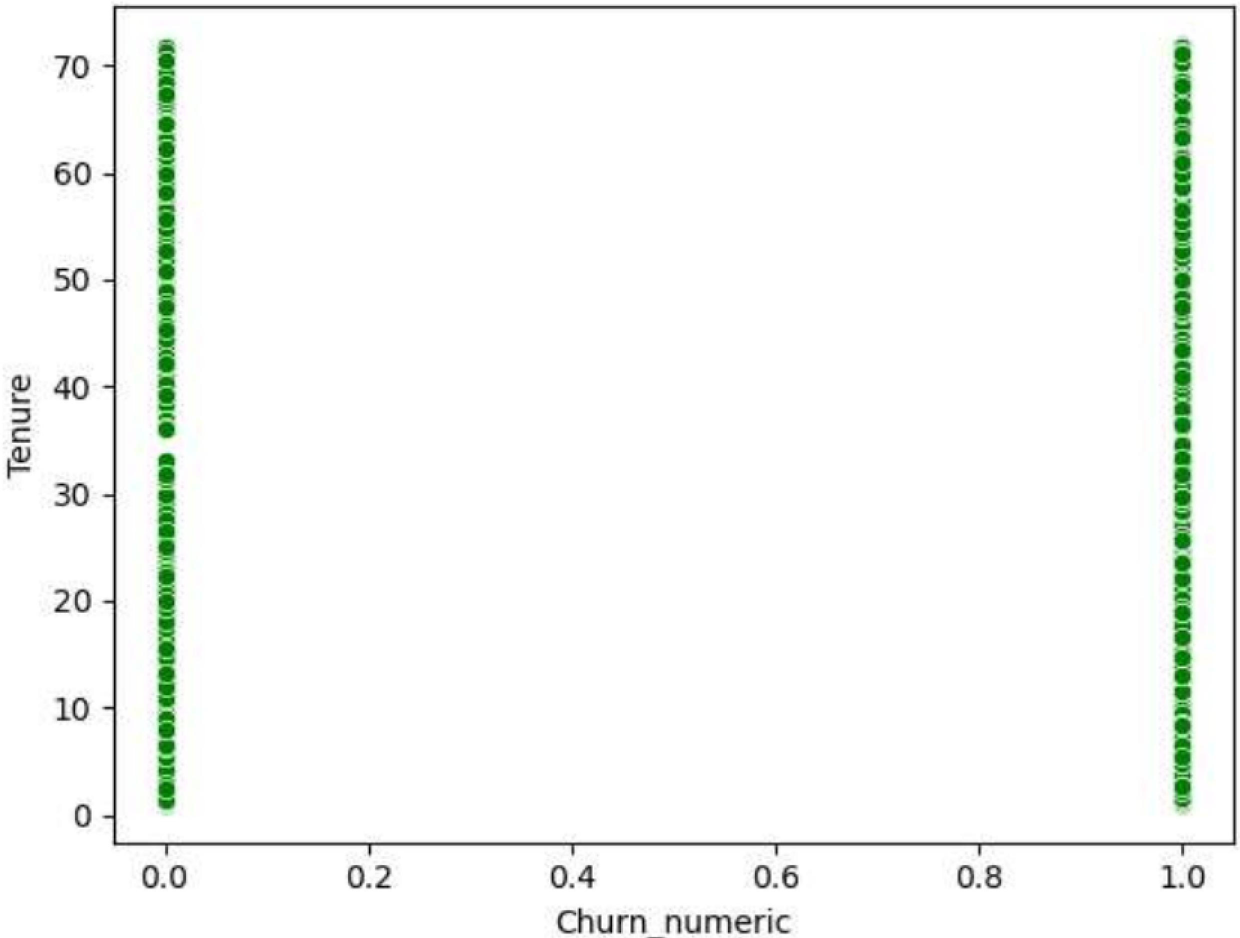
Tenure
Outage_sec_perweek
0
10
20
30
40
50
60
70
0
5
10
15
20

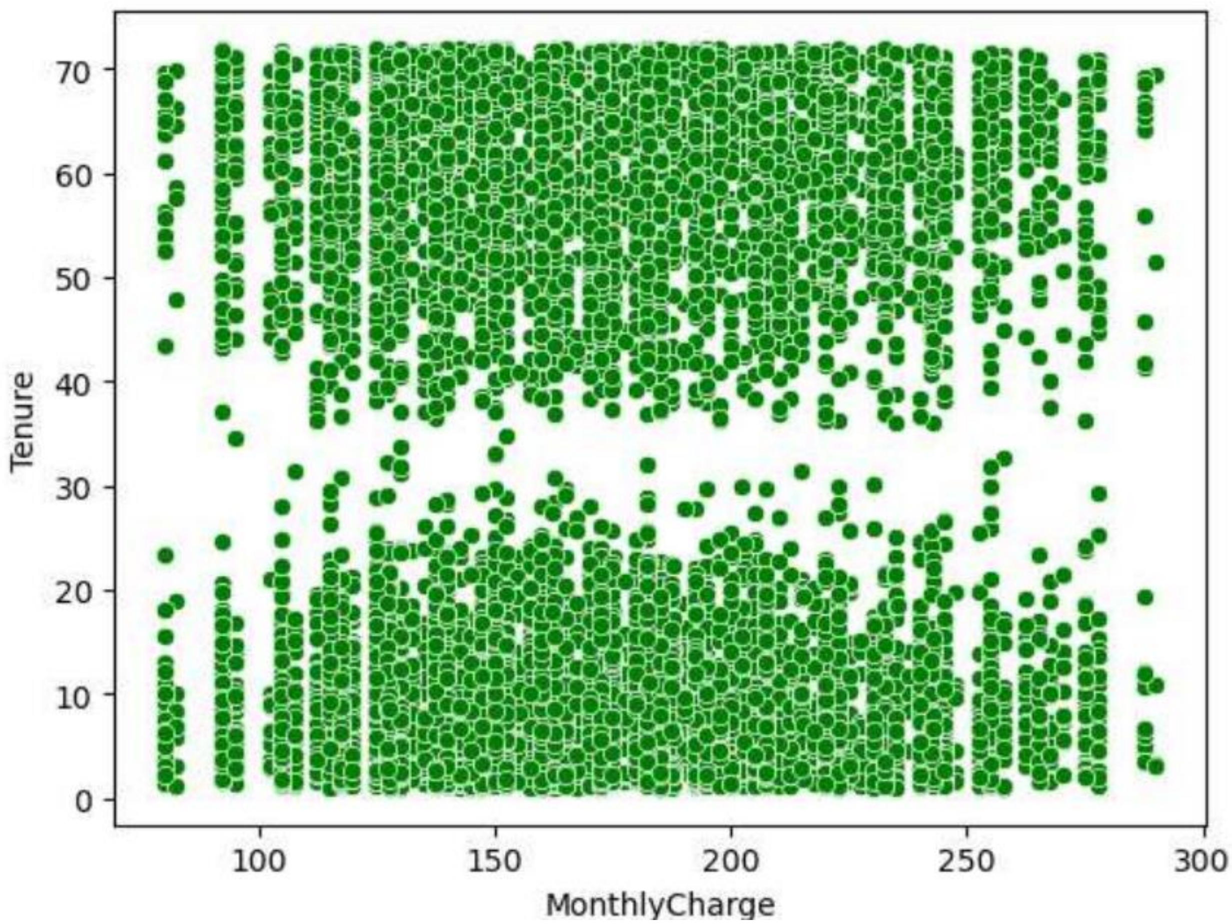

70
60
50
40
30
20
10
0
Tenure
0
5
10
15
20
Email

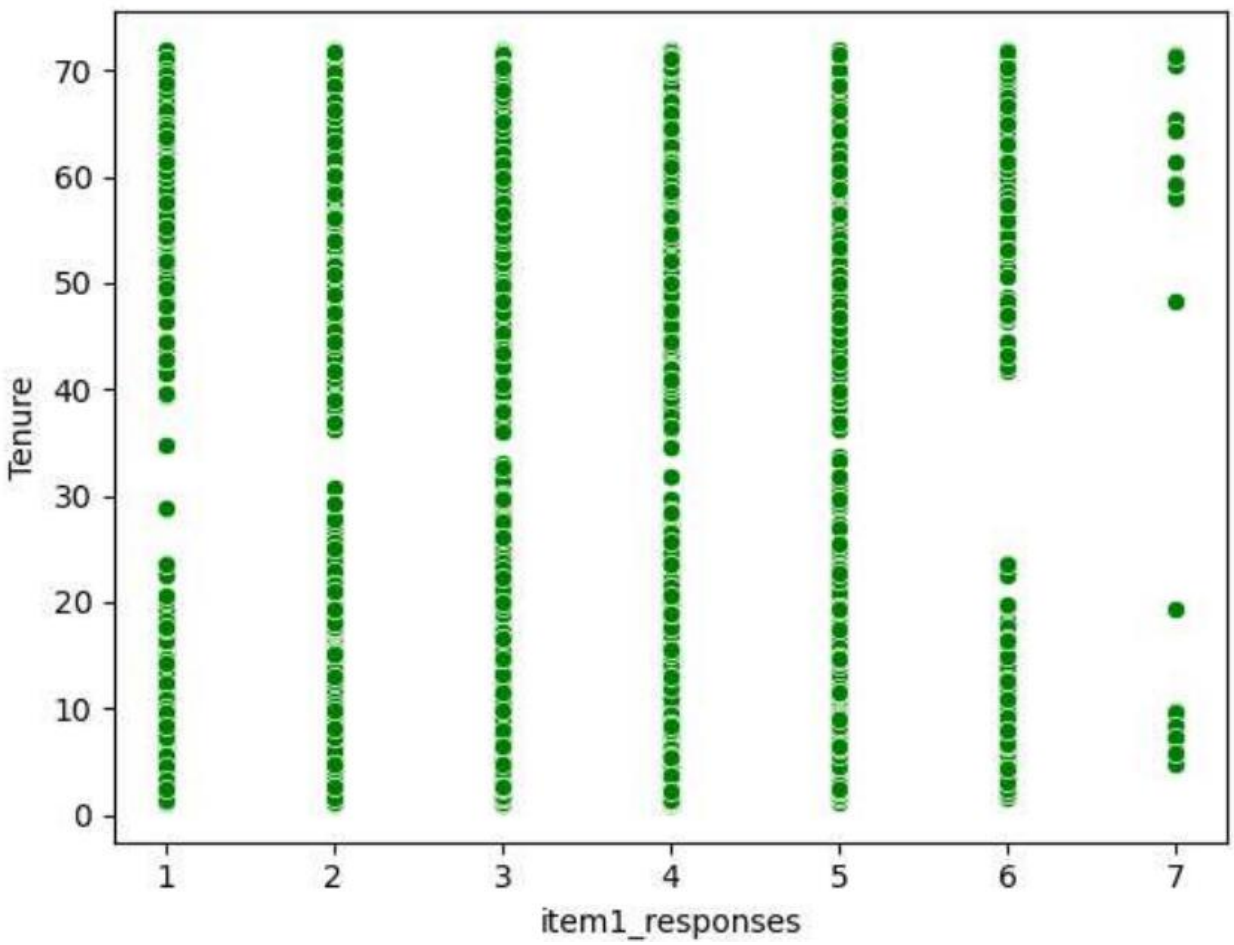

Tenure
Contacts
0
10
20
30
40
50
60
70
0
1
2
3
4
5
6
7

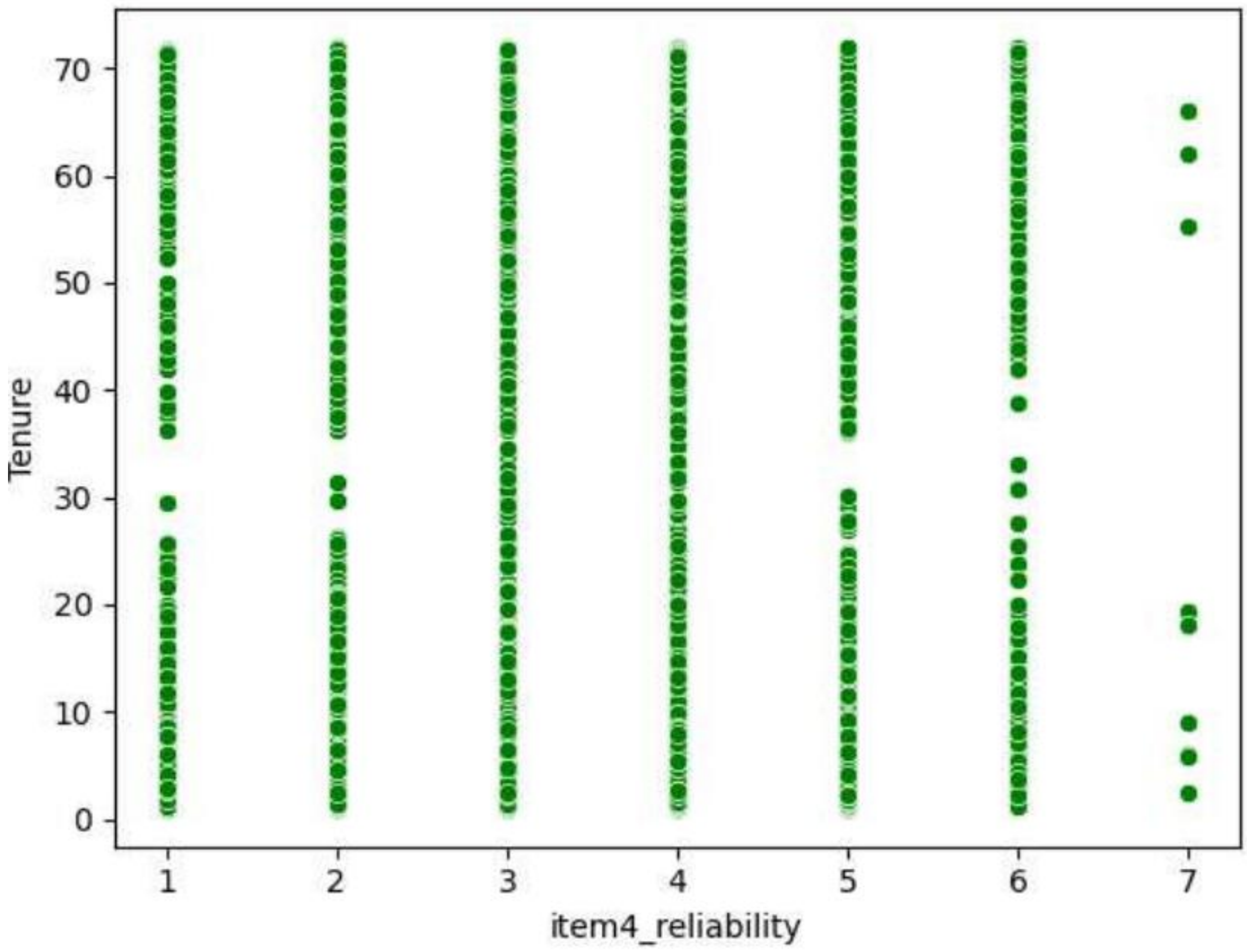


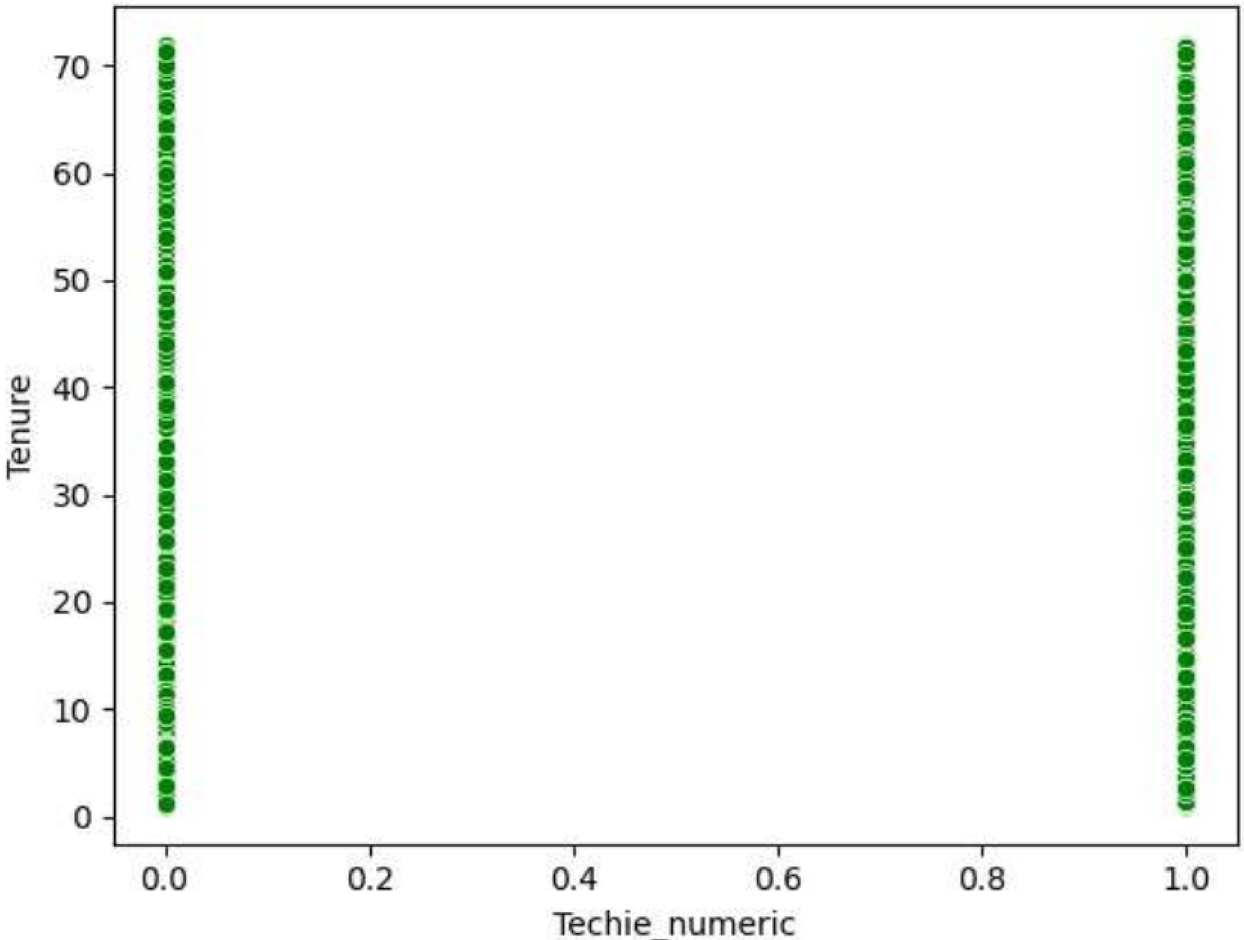
Tenure
Yearly_equip_failure
0
10
20
30
40
50
60
70
0
1
2
3
4
5
6

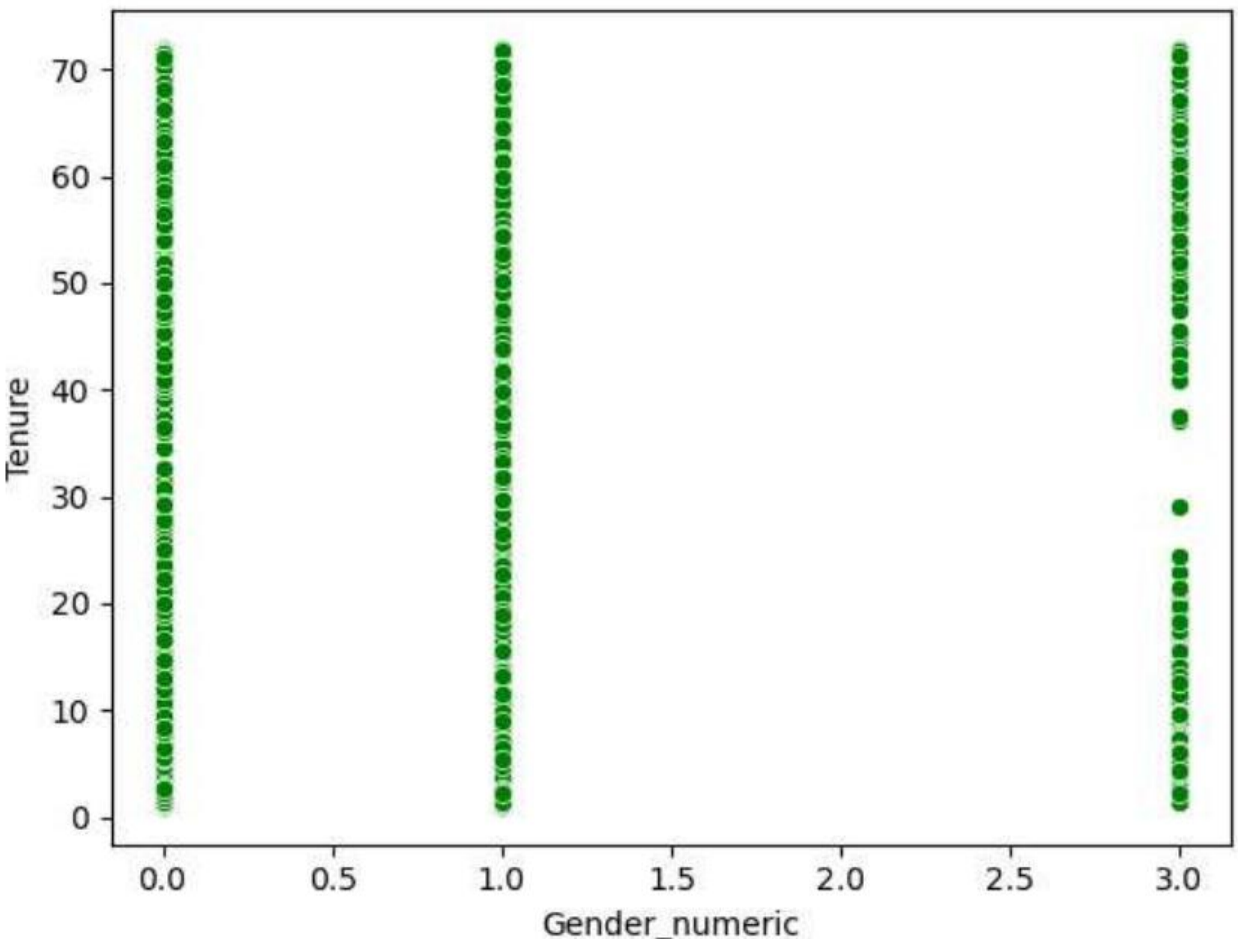

70
60
50
40
30
20
10
0
Tenure
0.0
0.2
0.4
0.6
0.8
1.0
Churn_numeric


70
60
50
40
30
20
10
0
Tenure
100
150
200
250
300
MonthlyCharge

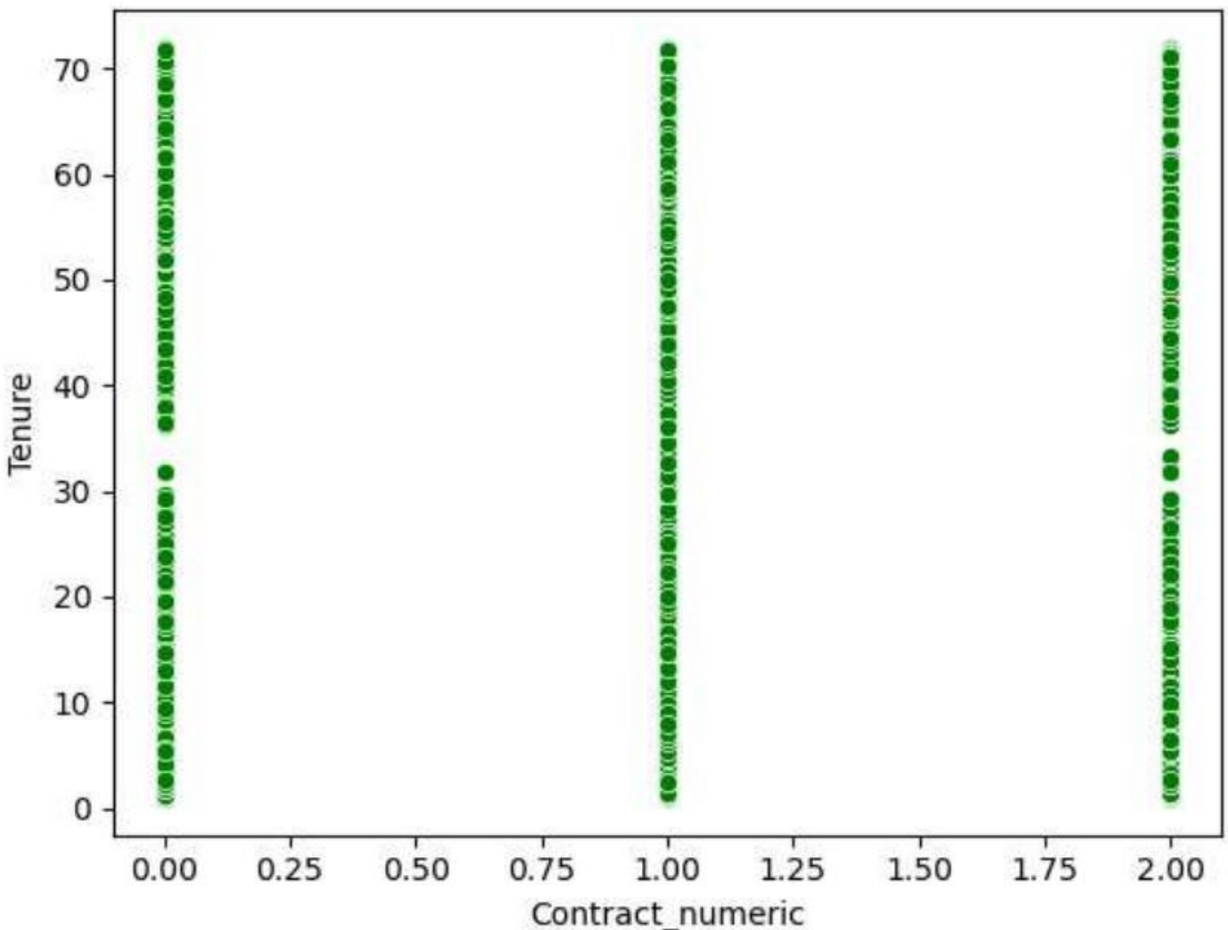

Tenure
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
item1_responses

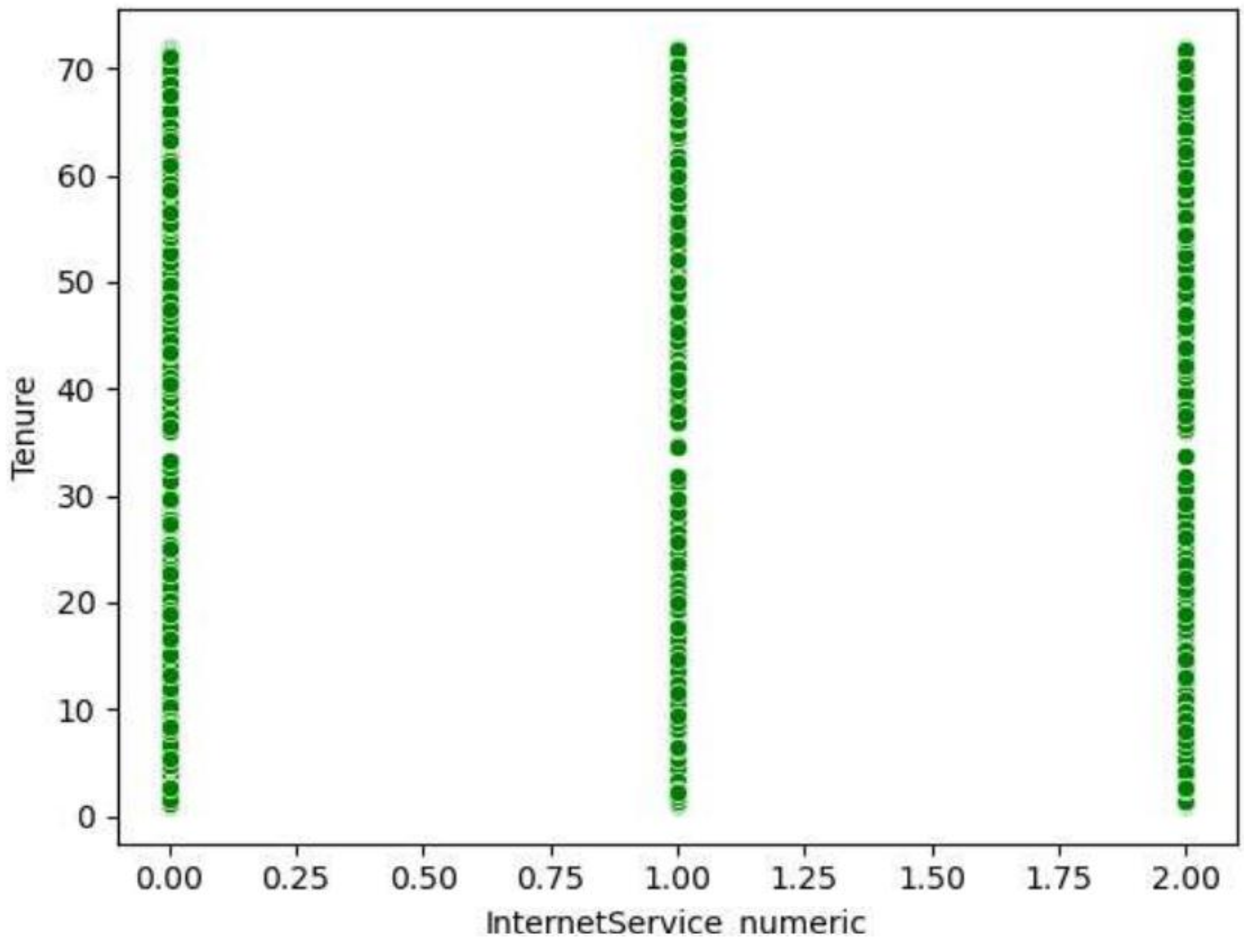

70
60
50
40
30
20
10
0
Tenure
1
2
3
4
5
6
7
item4_reliability


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



70
60
50
40
30
20
10
0
Tenure
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Gender_numeric


Tenure
70
60
50
40
30
20
10
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Contract_numeric



Tenure
0
10
20
30
40
50
60
70
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
InternetService_numeric

3. The code used to perform the classification analysis from part D2 is found in "D209_Task2_Code.ipynb"

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

1. Explain the accuracy and the mean squared error (MSE) of your prediction model.

2. Discuss the results and implications of your prediction analysis.

3. Discuss one limitation of your data analysis.

4. Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

1. The mean squared error (MSE) of the decision tree regression performed here is 17.5. The MSE is the test of accuracy for our model since it uses a continuous target by showing how far our predicted values are from the original ones. We also performed a test of the MAE, RMSE, and R squared value to show accuracy with 3.22, 4.183, and .97 or 97% strength of prediction accuracy that are explained in the previous section. However, to reiterate, these values for error are relatively large, as being closer to 0 is ideal, except for R squared. The R square value should be close to 1, 100%, and it is. The organization can decide whether the accuracy is good or bad and whether the MSE is

within an acceptable range, however; I believe these values indicate our model needs a larger data set and more investigation and resources to further explore Tenure as our target variable. Our MSE is very large and all of our metrics indicate this is not an incredibly accurate model.

2. The results of our analysis present a model that can predict the tenure length of a customer. Here, the large MSE value indicates the model has a large problem with error and simultaneously a high R squared value indicates it has a high strength of prediction. The implications of this are that it gives the organization a tool to do two things, predict Tenure length to inform their retention efforts, and gives them a model to further investigate to see what variables impact Tenure and what can be done to affect those variables and alter (increase) the tenure length of customers. The best parameters and score can be used to investigate these two aspects. The various metrics of accuracy and error show a conflicting story of the model and its reliability should be further tested and refined.
3. One limitation of my data analysis is that the branches are unrefined. In the model, there is no pruning or parameter testing method to remove branches with lower significance, so there may be bias in this data modeling project that may need to be dealt with by the organization at a later time. Specifically, the MSE value is large and would need to be reduced to produce a reliable and very accurate model.
4. A course of action I would recommend for this organization is to analyze the features that contribute to a shorter Tenure length and increase focus and efforts on the ones that lead to a longer tenure length. The model here can be used to help predict the tenure length of a customer and the organization should focus efforts on using it to identify the features of

long-tenured customers and try to replicate, support, or investigate those variables. I recommend to spend more time and effort into this project and continue to test different models to find the best one for predicting Tenure length and to work towards tuning the model to reduce the MSE.

Part VI: Demonstration

F. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

Panopto Video Link: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=22aecaf0-b7ae-4beb-940b-afa5016c44c3

G. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

Sources that informed the code used here:

Python | Decision Tree Regression using sklearn. (2018, October 4). GeeksforGeeks. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

scikit learn. (2009). *1.10. Decision Trees — scikit-learn 0.22 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/tree.html

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

References

Vishalmendekarhere. (2021, January 22). *It's All About Assumptions, Pros & Cons*. The Startup. https://medium.com/swlh/its-all-about-assumptions-pros-cons-497783cfed2d

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

This aspect of the rubric is evaluated through the entirety of this report and I hope professionalism has shown continuously.