Holt D208 Predictive Modeling Task 1

July 15, 2021

1 Part I: Research Question

1.1 A1: Question

What customer qualities or factors from the data set have a significant effect on a customer's length of tenure?

1.2 A2: Objective and Goals

The objective of an analysis of the data is to determine what features, if any, have a significant relationship with the length of tenure of a customer.

2 Part II: Method Justification

2.1 B1: Summary of Assumptions

2.1.1 1) Independent and Dependent Variables Have a Linear Relationship

Multiple linear regression analysis assumes that independent variables have a linear relationship with the target variable. If variables have a non-linear relationship with the target variable the regression results will under-estimate the true relationship. The preferable method to detect non-linearity is to examine residual plots (plots of the standardized residuals as a function of standardized predicted values). (Osborne, et al., 2002)

2.1.2 2) Assumption of Homoscedasticity

The multiple linear regression model assumes that the errors between the predicted value and the observed value are constant along the time series. Heteroscedasticity creates problems when using ordinary least-squares(OLS) regression because OLS attempts to minimize errors and gives equal weights across the errors. Observations with larger error variances would be given an unequal weight. Creating and chosing a model without checking for homoscedasticity can lead to negative consequences and could invalidate the model. Homoscedasticity can be checked by plotting the independent variable's residuals against the target variable.

2.1.3 3) No Multicollinearity

Multicollinearity in the model occurs when two or more independent variables share the same correlation with the target variable. In cases where a perfect correlation between two or more variables is present, multicollinearity can mean that no unique least-squares solution to a regression analysis can be computed. When less severe multicollinearity is present, as is more likely, the coefficient can be unreliable and the standard errors and confidence intervals for the coefficient estimates will be inflated leading to incorrect conclusions that a variable is statistically insignificant. (Williams, et al., 2013) Additionally, having two or more variables in the model with the same correlation factors can unnecessarily overcomplicate the model leading to inaccurate predictions. Using the variance inflation factor(VIF) results of each variable can help determine if multicollinearity exists within the model.

2.1.4 4) No Autocorrelation

Autocorrelation measures the relationship between a variable's current value and its past values. When variables have measurements that have a relationship or are dependent with previously measured values in the same variable it is deemed to be autocorrelated. This most often occurs in time series measured data where values that were measured closer together are more similar than values measured at a later time. Autocorrelation violates the linear model assumption that variable measurements are independent.

2.1.5 5) Normally distributed residuals (errors).

Finally, the multiple linear regression model assumes that independent variable residuals (errors) are normally distributed. With large sample sizes, normally distributed residuals are not required to estimate unbiased and efficient coefficients. However, with small samples sizes, it is the case that normally distributed residuals are required as significance tests and confidence intervals can be untrustworthy. (Williams, et al., 2013) A histogram of the variable's residuals or P-P plot can be created to observe the normality of the errors.

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 opportunites to provide detailed visual information for presentations. Python contains many packages built by data scientist that help with the previously mentioned tasks. Some packages that will be used are Numpy, Pandas, Matplotlib, Seaborn, Pyplot, Statsmodels, Sklearn and Yellowbrick.

2.3 B3: Why Multiple Linear Regression?

Using the question in part I, it was determined that multiple linear regression was an appropriate model to build. The predictor variable (Tenure) contains continuous values making it a prime target to use multiple linear regression. The dataset also contains many other continuous and categorical variables that can be used to build the model in relation to the predictor variable. While an initial model could be built using every variable in the dataset, by choosing specific variables using experience, sense, and previous familiarization with the dataset, unneeded variables can be weeded out to start with a simpler model. At this point in the workflow, however, we have yet to determine whether each assumption listed above has been determined to be true. There now is a clean dataset, a question to answer, and a decision of a model type to build. Now we move into the next portion of the analysis, data preparation.

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 we use for our linear 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 multiple linear regression are: - Identify and handle missing data - Identify and handle outliers or strange values - Ensure all variables are numerical and transform those that are not into numerical values - Ensure the target variable is numerical - Scale numeric variables if needed - Check if any linear model or multiple linear model assumptions are in violation

3.2 C2: Summary Statistics Needed

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

- Partial correlation correlations coefficients between all predictor variables and target variables: Along with visual scatterplots, the partial correlation matrix helps determine if there are colinear relationships between a predictor variable and the target variable and between each predictor variable.
- Coefficients of all predictor variables - the coefficients of each independent variable will need to be determined to build the linear regression model.
- R-squared statistic: The R-squared statistic is the proportion of variation in the dependent variable accounted for by the independent variable(s).
- Adjusted R-squared: The same statistic as the R-squared statistic but adjusted for the number of independent variables in the model. This statistic will always be lower than the R-squared.
- F-statistic - This statistic indicates whether a linear regression model provides a better fit to the data than a model that contains no independent variables.
- F-statistic probability - the probability that the F-statistic was obtained by chance. (p-value)
- 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

3.3 C3: Steps to Prepare the Data for Analysis & C4: Univariate and Bivariate Visualizations

```
1) Generate a list of potential variables; independent and dependent
2) Collect data on the variables (Import the Data), and transform as needed
3) Check the relationship between each independent variable and the dependent variable using sc
   and correlations
4) Check the relationships among the independent variables using scatterplots and correlations
5) Use the non-redundant independent variables in the analysis to find the best fitting model
```

The above list was generated from Brandon Foltz's website. (2014)

3.3.1 1) Generate a list of potential variables; independent and dependent

List of variables to be explored:

Independent Variables - Population (Continuous)

- Area (Categorical)
- Children (Continuous)
- Age (Continuous)
- Income (Continuous)
- Marital (Categorical)
- Gender (Categorical)
- Churn (Categorical)
- Outage_sec_perweek (Continuous)
- Email (Continuous)
- Contacts (Continuous)
- Yearly_equip_failure (Continuous)
- Techie (Categorical)
- Contract (Categorical)
- 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

- Tenure (Continuous)

3.3.2 2) Import the data and transform as required

[1]:
```
#Load packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
import pylab


import statsmodels.api as sm
from statsmodels.stats import diagnostic as diag
from statsmodels.stats.outliers_influence import variance_inflation_factor

from sklearn.linear_model import LinearRegression, Ridge, RidgeCV
from sklearn.model_selection import train_test_split, cross_val_predict
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error

from yellowbrick.regressor import AlphaSelection, PredictionError, ResidualsPlot

%matplotlib inline
```

[2]:
```
#Load the dataset
churn_clean = pd.read_csv("C:/Users/holtb/Data/WGU Datasets/churn_clean.csv")

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

```
#Verify there are no missing data and display data to verify correct variables␣
↪are loaded
display('-'*100)
display(churn_LRM_data.isnull().any())

display('-'*100)
churn_LRM_data.head(5)
```

```
'-----------------------------------------------------------------------------------------------
```

```
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

'-----------------------------------------------------------------------------------------------
```

[2]:

```
   Population       Area  Children  Age    Income    Marital  Gender Churn  \
0          38      Urban         0   68  28561.99    Widowed    Male    No
1       10446      Urban         1   27  21704.77    Married  Female   Yes
2        3735      Urban         4   50   9609.57    Widowed  Female    No
3       13863   Suburban         1   48  18925.23    Married    Male    No
4       11352   Suburban         0   83  40074.19  Separated    Male   Yes

   Outage_sec_perweek  Email  …  MonthlyCharge  Bandwidth_GB_Year Item1  \
0            7.978323     10  …     172.455519         904.536110     5
1           11.699080     12  …     242.632554         800.982766     3
2           10.752800      9  …     159.947583        2054.706961     4
3           14.913540     15  …     119.956840        2164.579412     4
4            8.147417     16  …     149.948316         271.493436     4

  Item2 Item3 Item4 Item5 Item6 Item7 Item8
0     5     5     3     4     4     3     4
1     4     3     3     4     3     4     4
2     4     2     4     4     3     3     3
3     4     4     2     5     4     3     3
4     4     4     3     4     4     4     5

[5 rows x 38 columns]
```

[3]:

```python
#Transform categorical variables to numeric using dummy variables and dropping␣
↪one column to meet n-1
churn_LRM_transdata = pd.get_dummies(churn_LRM_data, drop_first=True)

#Check that dummy variables properly transformed
churn_LRM_transdata.head(5)
```

[3]:

```
   Population  Children  Age    Income  Outage_sec_perweek  Email  Contacts  \
0          38         0   68  28561.99            7.978323     10         0
1       10446         1   27  21704.77           11.699080     12         0
2        3735         4   50   9609.57           10.752800      9         0
3       13863         1   48  18925.23           14.913540     15         2
4       11352         0   83  40074.19            8.147417     16         2

   Yearly_equip_failure     Tenure  MonthlyCharge  …  OnlineSecurity_Yes  \
0                     1   6.795513     172.455519  …                   1
1                     1   1.156681     242.632554  …                   1
2                     1  15.754144     159.947583  …                   0
```

```
3                      0  17.087227     119.956840  …                  1
4                      1   1.670972     149.948316  …                  0

   OnlineBackup_Yes  DeviceProtection_Yes  TechSupport_Yes  StreamingTV_Yes  \
0                 1                     0                0                0
1                 0                     0                0                1
2                 0                     0                0                0
3                 0                     0                0                1
4                 0                     0                1                1

   StreamingMovies_Yes  PaperlessBilling_Yes  \
0                    1                     1
1                    1                     1
2                    1                     1
3                    0                     1
4                    0                     0

   PaymentMethod_Credit Card (automatic)  PaymentMethod_Electronic Check  \
0                                      1                               0
1                                      0                               0
2                                      1                               0
3                                      0                               0
4                                      0                               0

   PaymentMethod_Mailed Check
0                           0
1                           0
2                           0
3                           1
4                           1

[5 rows x 47 columns]
```

Summary of data to obtain univariate statistics of each variable:

[4]:
```python
#Display univariate statistics of each numberical variable
churn_statistics = churn_LRM_transdata.describe()
churn_statistics.round(2)
```

[4]:
```
       Population  Children       Age     Income  Outage_sec_perweek  \
count    10000.00  10000.00  10000.00   10000.00            10000.00
mean      9756.56      2.09     53.08   39806.93               10.00
std      14432.70      2.15     20.70   28199.92                2.98
min          0.00      0.00     18.00     348.67                0.10
25%        738.00      0.00     35.00   19224.72                8.02
50%       2910.50      1.00     53.00   33170.60               10.02
75%      13168.00      3.00     71.00   53246.17               11.97
max     111850.00     10.00     89.00  258900.70               21.21
```

```
          Email  Contacts  Yearly_equip_failure    Tenure  MonthlyCharge  …  \
count  10000.00  10000.00              10000.00  10000.00       10000.00  …
mean      12.02      0.99                  0.40     34.53         172.62  …
std        3.03      0.99                  0.64     26.44          42.94  …
min        1.00      0.00                  0.00      1.00          79.98  …
25%       10.00      0.00                  0.00      7.92         139.98  …
50%       12.00      1.00                  0.00     35.43         167.48  …
75%       14.00      2.00                  1.00     61.48         200.73  …
max       23.00      7.00                  6.00     72.00         290.16  …

       OnlineSecurity_Yes  OnlineBackup_Yes  DeviceProtection_Yes  \
count            10000.00          10000.00              10000.00
mean                 0.36              0.45                  0.44
std                  0.48              0.50                  0.50
min                  0.00              0.00                  0.00
25%                  0.00              0.00                  0.00
50%                  0.00              0.00                  0.00
75%                  1.00              1.00                  1.00
max                  1.00              1.00                  1.00

       TechSupport_Yes  StreamingTV_Yes  StreamingMovies_Yes  \
count         10000.00         10000.00             10000.00
mean              0.38             0.49                 0.49
std               0.48             0.50                 0.50
min               0.00             0.00                 0.00
25%               0.00             0.00                 0.00
50%               0.00             0.00                 0.00
75%               1.00             1.00                 1.00
max               1.00             1.00                 1.00

       PaperlessBilling_Yes  PaymentMethod_Credit Card (automatic)  \
count              10000.00                               10000.00
mean                   0.59                                   0.21
std                    0.49                                   0.41
min                    0.00                                   0.00
25%                    0.00                                   0.00
50%                    1.00                                   0.00
75%                    1.00                                   0.00
max                    1.00                                   1.00

       PaymentMethod_Electronic Check  PaymentMethod_Mailed Check
count                        10000.00                    10000.00
mean                             0.34                        0.23
std                              0.47                        0.42
min                              0.00                        0.00
25%                              0.00                        0.00
```

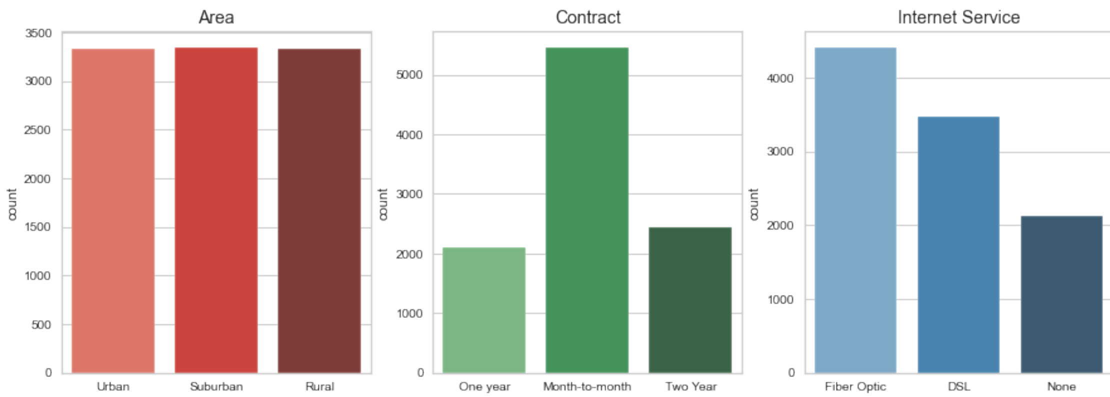
```
50%                                0.00                        0.00
75%                                1.00                        0.00
max                                1.00                        1.00

[8 rows x 47 columns]
```

[5]:
```
print("\n")
```

Bar charts of each categorical variable:

[6]:
```
fig, axs = plt.subplots(1,3, figsize=(15,5))
for i in range(3):
    ax = axs[i]
    if i == 0:
        title = ax.set_title("Area",
                             loc='center',
                             size = 14,
                             y=1.0),
    if i == 1:
        title = ax.set_title("Contract",
                             loc='center',
                             size = 14,
                             y=1.0)
    if i == 2:
        title = ax.set_title("Internet Service",
                             size = 14,
                             loc='center',
                             y=1.0)

sns.set_style('whitegrid')

#Create barcharts of each catergorical variable using the non-transformed␣
 ↪dataset
c1 = sns.countplot(x = 'Area',
                   data = churn_LRM_data,
                   palette="Reds_d",
                   ax = axs[0])
c2 = sns.countplot(x = 'Contract',
                   data = churn_LRM_data,
                   palette="Greens_d",
                   ax = axs[1])
c3 = sns.countplot(x = 'InternetService',
                   data = churn_LRM_data,
                   palette="Blues_d",
                   ax = axs[2])
```

```
#Remove xlabels
c1.set(xlabel=None);
c2.set(xlabel=None);
c3.set(xlabel=None);

print("\n")
```



3.3.3 3) Check relationships betweeen each independent variable and the dependent variable using scatterplots and correlations

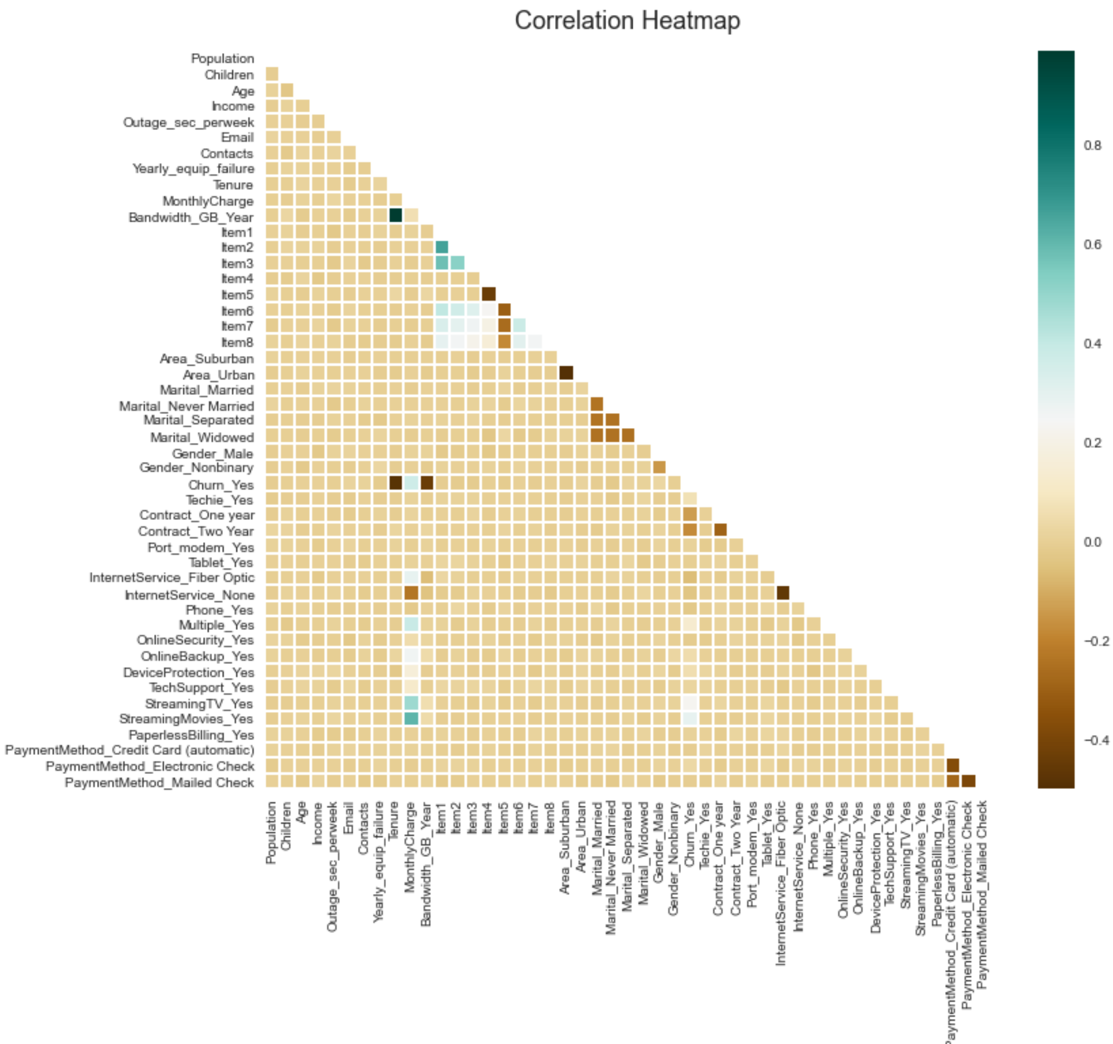
3.3.4 &

3.3.5 4) Check the relationships among the independent variables using scatterplots and correlations

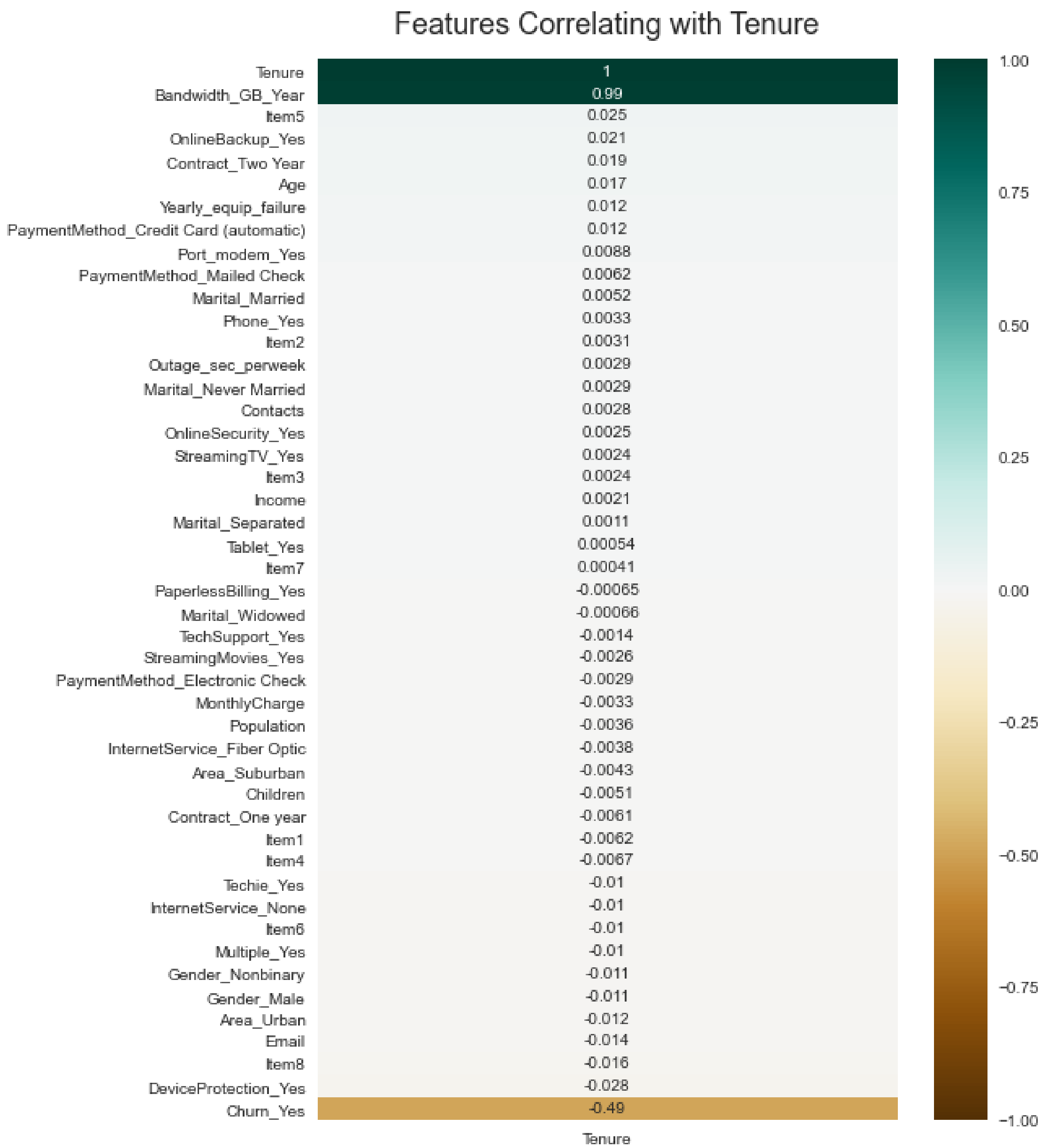
[27]:
```
#I marked out the correlation matrix as the matrix took up too much space in␣
↪the PDF as was not a requirement.

#corr = churn_LRM_transdata.corr()
#display(corr)
```

[8]:
```
mask = np.triu(np.ones_like(churn_LRM_transdata.corr()))
```

[9]:
```
plt.figure(figsize=(12, 10))
sns.heatmap(corr, xticklabels=corr.columns, yticklabels=corr.columns, mask =␣
↪mask, linewidth=.01,cmap='BrBG')
plt.title('Correlation Heatmap', fontdict={'fontsize':18}, pad=16);
```

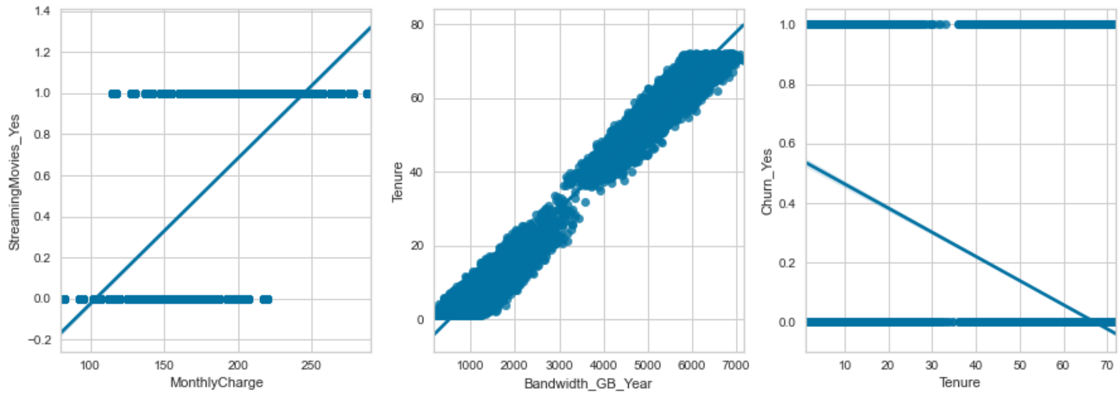


[10]:
```
plt.figure(figsize=(8, 12))
sns.heatmap(churn_LRM_transdata.corr()[['Tenure']].sort_values(by='Tenure',␣
→ascending=False), vmin=-1, vmax=1,
                     annot=True, cmap='BrBG')
plt.title('Features Correlating with Tenure', fontdict={'fontsize':18}, pad=16);
```



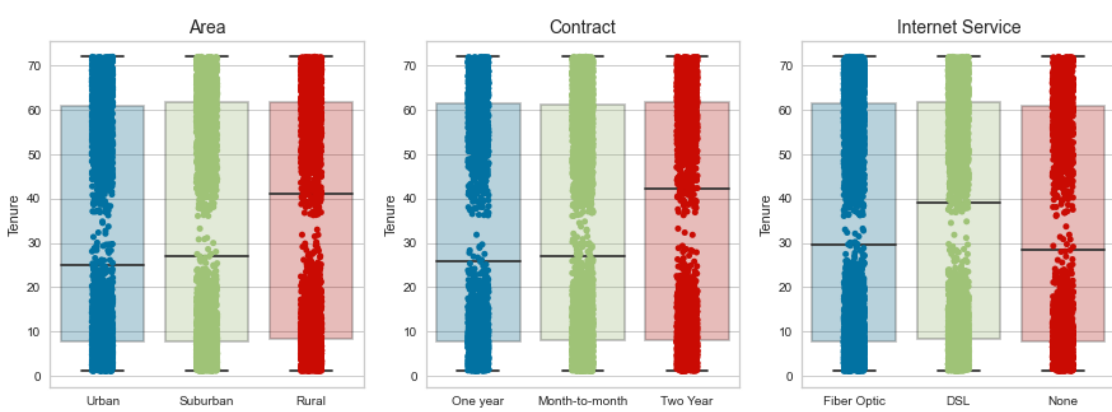
[11]:
```
fig, axs = plt.subplots(1,3,figsize=(15,5))

p1 = sns.regplot(x='MonthlyCharge',
                 y='StreamingMovies_Yes',
                 data=churn_LRM_transdata,
                 logistic = False,
                 ax = axs[0])

p2 = sns.regplot(x='Bandwidth_GB_Year',
                 y = 'Tenure',
```

```
                 data=churn_LRM_transdata,
                 ax = axs[1]);

p3 = sns.regplot(x='Tenure',
                 y='Churn_Yes',
                 data=churn_LRM_transdata,
                 logistic = False,
                 ax = axs[2])
```

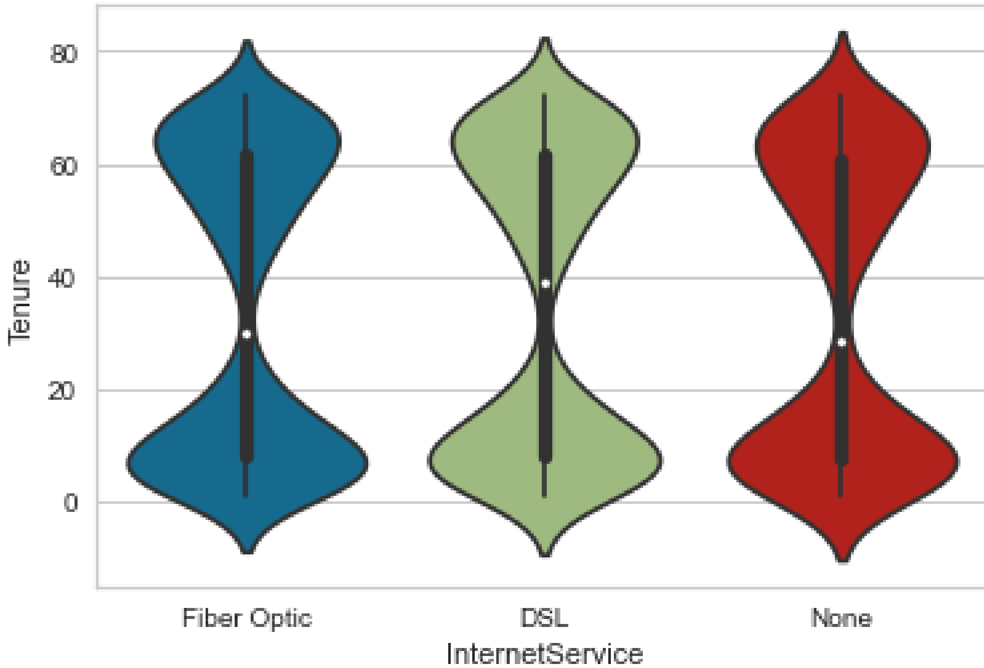


[12]:
```
fig, axs = plt.subplots(1,3, figsize=(15,5))
for i in range(3):
    ax = axs[i]
    if i == 0:
        title = ax.set_title("Area",
                             loc='center',
                             size = 14,
                             y=1.0),
    if i == 1:
        title = ax.set_title("Contract",
                             loc='center',
                             size = 14,
                             y=1.0)
    if i == 2:
        title = ax.set_title("Internet Service",
                             size = 14,
                             loc='center',
                             y=1.0)

#Create stripplots overlapped with boxplots of each catergorical variable
g1 = sns.boxplot(x='Area',
            y='Tenure',
            data = churn_LRM_data,
            boxprops=dict(alpha=.3),
```

```
              ax = axs[0])
sns.stripplot(x='Area',
              y='Tenure',
              data = churn_LRM_data,
              ax = axs[0])

g2 = sns.boxplot(x='Contract',
            y='Tenure',
            data = churn_LRM_data,
            boxprops=dict(alpha=.3),
            ax = axs[1])
sns.stripplot(x='Contract',
              y='Tenure',
              data = churn_LRM_data,
              ax = axs[1])

g3 = sns.boxplot(x='InternetService',
            y='Tenure',
            data = churn_LRM_data,
            boxprops=dict(alpha=.3),
            ax = axs[2])
sns.stripplot(x='InternetService',
              y='Tenure',
              data = churn_LRM_data,
              ax = axs[2])

#Remove xlabels
g1.set(xlabel=None);
g2.set(xlabel=None);
g3.set(xlabel=None);

print("\n")
```



[13]:
```
#Created a violin plot to better visualize the distibution of the␣
↪InternetService variable
sns.violinplot(x='InternetService',
               y='Tenure',
               data = churn_LRM_data);
```



3.3.6 5) Use the non-redudant independent variables in the analysis to find the best fitting model

[14]:
```
# creating the data frames before and after removing variables that are␣
↪creating multicollinearity
churn_data_before = churn_LRM_transdata.drop(['Tenure'], axis=1)
churn_data_after = churn_LRM_transdata.
↪drop(['Tenure','StreamingMovies_Yes','StreamingTV_Yes'], axis = 1)


# adding a constant to the data frames as required for the VIF calculation
X1 = sm.tools.add_constant(churn_data_before)
X2 = sm.tools.add_constant(churn_data_after)


# create the series for both
```

```python
series_before = pd.Series([variance_inflation_factor(X1.values, i) for i in␣
↪range(X1.shape[1])], index=X1.columns)
series_after = pd.Series([variance_inflation_factor(X2.values, i) for i in␣
↪range(X2.shape[1])], index=X2.columns)

# display the series
print('-'*100)
print('Data Before Removing Target and Redundant Variables')
print('-'*100)
display(series_before)

print('-'*100)
print('Data After Removing Target and Redundant Variables')
print('-'*100)
display(series_after)
```

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

const                                   238.984679
Population                                1.003163
Children                                  1.005223
Age                                       1.004917
Income                                    1.004313
Outage_sec_perweek                        1.005126
Email                                     1.004522
Contacts                                  1.004216
Yearly_equip_failure                      1.004513
MonthlyCharge                            24.383308
Bandwidth_GB_Year                         1.486010
Item1                                     2.218863
Item2                                     1.936285
Item3                                     1.608640
Item4                                     1.280036
Item5                                     1.376610
Item6                                     1.484337
Item7                                     1.315291
Item8                                     1.192026
Area_Suburban                             1.337302
Area_Urban                                1.339328
Marital_Married                           1.553599
Marital_Never Married                     1.562971
Marital_Separated                         1.572499
Marital_Widowed                           1.579742
Gender_Male                               1.029279
```

```
Gender_Nonbinary                        1.025488
Churn_Yes                               1.966203
Techie_Yes                              1.010727
Contract_One year                       1.191609
Contract_Two Year                       1.204735
Port_modem_Yes                          1.002303
Tablet_Yes                              1.006094
InternetService_Fiber Optic             2.680909
InternetService_None                    1.652055
Phone_Yes                               1.006150
Multiple_Yes                            4.435773
OnlineSecurity_Yes                      1.027237
OnlineBackup_Yes                        2.644752
DeviceProtection_Yes                    1.505744
TechSupport_Yes                         1.488305
StreamingTV_Yes                         6.779135
StreamingMovies_Yes                     9.881864
PaperlessBilling_Yes                    1.005103
PaymentMethod_Credit Card (automatic)   1.537269
PaymentMethod_Electronic Check          1.675824
PaymentMethod_Mailed Check              1.570644
dtype: float64

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

const                                 159.189361
Population                              1.003147
Children                                1.005126
Age                                     1.004854
Income                                  1.004180
Outage_sec_perweek                      1.005030
Email                                   1.004505
Contacts                                1.004130
Yearly_equip_failure                    1.004434
MonthlyCharge                           2.085597
Bandwidth_GB_Year                       1.485036
Item1                                   2.218140
Item2                                   1.935335
Item3                                   1.608234
Item4                                   1.279988
Item5                                   1.376347
Item6                                   1.484187
Item7                                   1.315162
Item8                                   1.191926
```

```
Area_Suburban                               1.337271
Area_Urban                                  1.339255
Marital_Married                             1.553586
Marital_Never Married                       1.562880
Marital_Separated                           1.572453
Marital_Widowed                             1.579559
Gender_Male                                 1.028943
Gender_Nonbinary                            1.025177
Churn_Yes                                   1.965478
Techie_Yes                                  1.010350
Contract_One year                           1.190871
Contract_Two Year                           1.204389
Port_modem_Yes                              1.002122
Tablet_Yes                                  1.005818
InternetService_Fiber Optic                 1.512357
InternetService_None                        1.313499
Phone_Yes                                   1.005838
Multiple_Yes                                1.239551
OnlineSecurity_Yes                          1.007713
OnlineBackup_Yes                            1.115532
DeviceProtection_Yes                        1.045398
TechSupport_Yes                             1.037459
PaperlessBilling_Yes                        1.004773
PaymentMethod_Credit Card (automatic)       1.536866
PaymentMethod_Electronic Check              1.675565
PaymentMethod_Mailed Check                  1.570557
dtype: float64
```

3.4 C5: Provide a copy of the data set:

[15]:
```
churn_LRM_transdata.to_csv('C:/Users/holtb/Data/D208/Task1/churn_LRM_transdata.
↪csv')
```

4 Part IV: Model Comparison Analysis

4.1 D1: Initial Linear Regression Model

The model is first built with all variables determined in part C3:

[16]:
```
# define our input variable (X) & output variable
churn_LRM_data_initial = churn_LRM_transdata
X = churn_LRM_data_initial.drop('Tenure', axis = 1)
Y = churn_LRM_data_initial[['Tenure']]

# Split X and y into X_
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20,␣
 ↪random_state=1)
```

[17]:
```
#define the input
X2 = sm.add_constant(X)

#create an OLS model
initial_model = sm.OLS(Y, X2)

#fit the data
initial_est = initial_model.fit()

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

[17]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.316e+07
Date:                Sun, 11 Jul 2021   Prob (F-statistic):               0.00
Time:                        14:33:57   Log-Likelihood:                 8140.4
No. Observations:               10000   AIC:                        -1.619e+04
Df Residuals:                    9953   BIC:                        -1.585e+04
Df Model:                          46
Covariance Type:            nonrobust
================================================================================
=========================
                                            coef    std err          t
P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
-------------------------
const                                    -3.8431      0.017   -231.335
0.000      -3.876      -3.811
Population                            -7.836e-08   7.46e-08     -1.051
0.293   -2.25e-07    6.78e-08
Children                                 -0.3755      0.001   -748.384
0.000      -0.377      -0.375
Age                                       0.0400    5.2e-05    768.285
0.000       0.040       0.040
Income                                 1.564e-08   3.82e-08      0.410
0.682   -5.92e-08    9.05e-08
Outage_sec_perweek                        0.0003      0.000      0.809
0.419      -0.000       0.001
Email                                 -6.71e-05      0.000     -0.189
```

```
0.850      -0.001       0.001
Contacts                               -0.0006      0.001     -0.545
0.586      -0.003       0.002
Yearly_equip_failure                    0.0002      0.002      0.113
0.910      -0.003       0.004
MonthlyCharge                          -0.0352      0.000   -284.931
0.000      -0.035      -0.035
Bandwidth_GB_Year                       0.0122   5.99e-07   2.04e+04
0.000       0.012       0.012
Item1                                   0.0023      0.002      1.503
0.133      -0.001       0.005
Item2                                  -0.0007      0.001     -0.484
0.628      -0.004       0.002
Item3                                   0.0010      0.001      0.758
0.448      -0.002       0.004
Item4                                   0.0004      0.001      0.377
0.706      -0.002       0.003
Item5                                  -0.0006      0.001     -0.483
0.629      -0.003       0.002
Item6                                  -0.0013      0.001     -1.007
0.314      -0.004       0.001
Item7                                  -0.0011      0.001     -0.898
0.369      -0.003       0.001
Item8                                7.959e-05      0.001      0.070
0.944      -0.002       0.002
Area_Suburban                          -0.0068      0.003     -2.593
0.010      -0.012      -0.002
Area_Urban                             -0.0033      0.003     -1.251
0.211      -0.008       0.002
Marital_Married                        -0.0006      0.003     -0.166
0.868      -0.007       0.006
Marital_Never Married                  -0.0011      0.003     -0.329
0.742      -0.008       0.006
Marital_Separated                       0.0026      0.003      0.785
0.432      -0.004       0.009
Marital_Widowed                       3.45e-05      0.003      0.010
0.992      -0.007       0.007
Gender_Male                            -0.7924      0.002   -362.915
0.000      -0.797      -0.788
Gender_Nonbinary                        0.2618      0.007     36.143
0.000       0.248       0.276
Churn_Yes                               0.0020      0.003      0.571
0.568      -0.005       0.009
Techie_Yes                          -1.104e-05      0.003     -0.004
0.997      -0.006       0.006
Contract_One year                       0.0011      0.003      0.389
0.697      -0.005       0.007
```

```
Contract_Two Year                           0.0023      0.003      0.855
0.393      -0.003       0.008
Port_modem_Yes                              0.0027      0.002      1.248
0.212      -0.002       0.007
Tablet_Yes                                  0.0007      0.002      0.277
0.781      -0.004       0.005
InternetService_Fiber Optic                 5.7542      0.004   1623.665
0.000       5.747       5.761
InternetService_None                        4.6006      0.003   1363.478
0.000       4.594       4.607
Phone_Yes                                   0.0017      0.004      0.457
0.648      -0.006       0.009
Multiple_Yes                                0.2684      0.005     59.120
0.000       0.260       0.277
OnlineSecurity_Yes                         -0.8313      0.002   -365.801
0.000      -0.836      -0.827
OnlineBackup_Yes                           -0.3552      0.004   -101.115
0.000      -0.362      -0.348
DeviceProtection_Yes                       -0.5971      0.003   -224.678
0.000      -0.602      -0.592
TechSupport_Yes                             0.3850      0.003    142.188
0.000       0.380       0.390
StreamingTV_Yes                            -1.2984      0.006   -232.008
0.000      -1.309      -1.287
StreamingMovies_Yes                        -0.7227      0.007   -106.938
0.000      -0.736      -0.709
PaperlessBilling_Yes                       -0.0037      0.002     -1.675
0.094      -0.008       0.001
PaymentMethod_Credit Card (automatic)       0.0021      0.003      0.628
0.530      -0.004       0.008
PaymentMethod_Electronic Check              0.0029      0.003      0.975
0.329      -0.003       0.009
PaymentMethod_Mailed Check                  0.0073      0.003      2.283
0.022       0.001       0.014
==============================================================================
Omnibus:                    34822.579   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1633.265
Skew:                          -0.034   Prob(JB):                         0.00
Kurtosis:                       1.021   Cond. No.                     8.31e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 8.31e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

4.2 D2: Linear Regression Model with Colinear Independent Variables Removed

As described in the notes section of the model, the condition number calculated indicates there are strong multicollinearity issues within the model. This can be confirmed by looking at the variance inflation factors (VIF) that were calculated in section C. VIF values above 5 to 10 are usually indicators of multicollinearity. Variables identified as having colinearity were dropped in each category until the VIF calculation returned to an acceptable value. In this case, the 'StreamingMovies_Yes' and 'StreamingTV_Yes' variables were dropped which reduced "MonthlyCharge" VIF to less than 2. These same variables will be dropped to create the next model.

[18]:
```
churn_LRM_nomulti = churn_LRM_data_initial.
 ↪drop(['StreamingMovies_Yes','StreamingTV_Yes'], axis=1)
next2_X = churn_LRM_nomulti.drop('Tenure', axis = 1)
next2_Y = churn_LRM_nomulti[['Tenure']]

# Split X and y into X_
next2_X_train, next2_X_test, next2_Y_train, next2_Y_test =␣
 ↪train_test_split(next2_X, next2_Y, test_size=0.20, random_state=1)
```

[19]:
```
#define the input
next2_X2 = sm.add_constant(next2_X)

#create an OLS model
next2_model = sm.OLS(next2_Y, next2_X2)

#fit the data
next2_est = next2_model.fit()

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

[19]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.320e+06
Date:                Sun, 11 Jul 2021   Prob (F-statistic):               0.00
Time:                        14:35:00   Log-Likelihood:                -3581.3
No. Observations:               10000   AIC:                             7253.
Df Residuals:                    9955   BIC:                             7577.
Df Model:                          44
Covariance Type:            nonrobust
================================================================================
=========================
                                            coef    std err          t
```

```
P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
-------------------------
const                                          -2.3651      0.044    -54.027
0.000      -2.451      -2.279
Population                                  -1.522e-08   2.41e-07     -0.063
0.950   -4.87e-07    4.57e-07
Children                                       -0.3741      0.002   -230.896
0.000      -0.377      -0.371
Age                                             0.0401      0.000    238.707
0.000       0.040       0.040
Income                                       9.543e-08   1.23e-07      0.774
0.439   -1.46e-07    3.37e-07
Outage_sec_perweek                              0.0002      0.001      0.134
0.894      -0.002       0.002
Email                                          -0.0005      0.001     -0.438
0.661      -0.003       0.002
Contacts                                        0.0002      0.004      0.065
0.948      -0.007       0.007
Yearly_equip_failure                           -0.0038      0.005     -0.696
0.486      -0.015       0.007
MonthlyCharge                                  -0.0542      0.000   -464.833
0.000      -0.054      -0.054
Bandwidth_GB_Year                               0.0122   1.93e-06   6305.369
0.000       0.012       0.012
Item1                                          -0.0062      0.005     -1.246
0.213      -0.016       0.004
Item2                                           0.0033      0.005      0.711
0.477      -0.006       0.012
Item3                                           0.0066      0.004      1.540
0.124      -0.002       0.015
Item4                                        4.772e-05      0.004      0.012
0.990      -0.007       0.008
Item5                                           0.0046      0.004      1.159
0.246      -0.003       0.012
Item6                                          -0.0011      0.004     -0.262
0.793      -0.009       0.007
Item7                                           0.0025      0.004      0.654
0.513      -0.005       0.010
Item8                                          -0.0031      0.004     -0.834
0.404      -0.010       0.004
Area_Suburban                                  -0.0031      0.009     -0.362
0.717      -0.020       0.014
Area_Urban                                      0.0021      0.009      0.247
0.805      -0.015       0.019
Marital_Married                                 0.0018      0.011      0.160
0.873      -0.020       0.023
```

```
Marital_Never Married                    -0.0089      0.011     -0.817
0.414      -0.030       0.013
Marital_Separated                        -0.0024      0.011     -0.223
0.823      -0.024       0.019
Marital_Widowed                          -0.0105      0.011     -0.964
0.335      -0.032       0.011
Gender_Male                              -0.8021      0.007   -113.809
0.000      -0.816      -0.788
Gender_Nonbinary                          0.2254      0.023      9.638
0.000       0.180       0.271
Churn_Yes                                -0.0181      0.011     -1.642
0.101      -0.040       0.004
Techie_Yes                                0.0055      0.009      0.593
0.553      -0.013       0.024
Contract_One year                        -0.0157      0.009     -1.688
0.091      -0.034       0.003
Contract_Two Year                        -0.0078      0.009     -0.875
0.382      -0.025       0.010
Port_modem_Yes                            0.0005      0.007      0.069
0.945      -0.013       0.014
Tablet_Yes                               -0.0063      0.008     -0.829
0.407      -0.021       0.009
InternetService_Fiber Optic               6.1306      0.009    713.355
0.000       6.114       6.147
InternetService_None                      4.3502      0.010    447.834
0.000       4.331       4.369
Phone_Yes                              7.915e-05      0.012      0.007
0.995      -0.023       0.024
Multiple_Yes                              0.8915      0.008    115.042
0.000       0.876       0.907
OnlineSecurity_Yes                       -0.7776      0.007   -107.001
0.000      -0.792      -0.763
OnlineBackup_Yes                          0.0799      0.007     10.852
0.000       0.065       0.094
DeviceProtection_Yes                     -0.3542      0.007    -49.549
0.000      -0.368      -0.340
TechSupport_Yes                           0.6221      0.007     85.224
0.000       0.608       0.636
PaperlessBilling_Yes                      0.0083      0.007      1.175
0.240      -0.006       0.022
PaymentMethod_Credit Card (automatic)     0.0051      0.011      0.483
0.629      -0.016       0.026
PaymentMethod_Electronic Check            0.0109      0.009      1.146
0.252      -0.008       0.029
PaymentMethod_Mailed Check                0.0131      0.010      1.263
0.207      -0.007       0.033
==============================================================================
```

```
Omnibus:                     1428.031   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1425.968
Skew:                          -0.859   Prob(JB):                    2.26e-310
Kurtosis:                       2.315   Cond. No.                     6.31e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.31e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

4.3 D2 Cont: Linear Regression Model Variables with P-Values greater than .05 Removed

A p-value determines the significance between a null hypothesis that a coefficient is zero and the alternate hypothesis that the coefficient is not equal to zero. In the model, the normal significance threshold of p <.05 was used. The variable with the largest p-value was removed one at a time and the model was reran using stepwise until all variable's p-values were less than .05.

[20]:
```
churn_LRM_rmvhighpvalues = churn_LRM_nomulti.
 ↪drop(['Population','Contacts','Item4','Phone_Yes',
                                                    'Port_modem_Yes',␣
 ↪'Marital_Separated','Outage_sec_perweek',
                                                    ␣
 ↪'Marital_Married','Item6','Area_Suburban','Yearly_equip_failure',
                                                    ␣
 ↪'Email','PaymentMethod_Credit Card (automatic)','Area_Urban',
                                                    ␣
 ↪'Income','Tablet_Yes','Techie_Yes','Item7','Item8','Item2',
                                                    'Contract_Two␣
 ↪Year','Marital_Never Married','Marital_Widowed',
                                                    'PaymentMethod_Electronic␣
 ↪Check','PaymentMethod_Mailed Check',
                                                    ␣
 ↪'PaperlessBilling_Yes','Item1','Churn_Yes','Contract_One year',
                                                    'Contract_One␣
 ↪year','Item3','Item5'], axis=1)
next3_X = churn_LRM_rmvhighpvalues.drop('Tenure', axis = 1)
next3_Y = churn_LRM_rmvhighpvalues[['Tenure']]

# Split X and y into X_
next3_X_train, next3_X_test, next3_Y_train, next3_Y_test =␣
 ↪train_test_split(next3_X, next3_Y, test_size=0.20, random_state=1)
```

[21]:
```
#define the input
next3_X2 = sm.add_constant(next3_X)

#create an OLS model
next3_model = sm.OLS(next3_Y, next3_X2)

#fit the data
next3_est = next3_model.fit()

#Summarize the output
next3_est.summary()
```

[21]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.472e+06
Date:                Sun, 11 Jul 2021   Prob (F-statistic):               0.00
Time:                        14:35:56   Log-Likelihood:                -3590.7
No. Observations:               10000   AIC:                             7209.
Df Residuals:                    9986   BIC:                             7310.
Df Model:                          13
Covariance Type:            nonrobust
================================================================================
===============
                                 coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
---------------
const                         -2.3378      0.019   -121.206      0.000
-2.376      -2.300
Children                      -0.3742      0.002   -231.428      0.000
-0.377      -0.371
Age                            0.0401      0.000    239.136      0.000
0.040       0.040
MonthlyCharge                 -0.0543      0.000   -540.082      0.000
-0.055      -0.054
Bandwidth_GB_Year              0.0122    1.6e-06   7616.847      0.000
0.012       0.012
Gender_Male                   -0.8025      0.007   -114.275      0.000
-0.816      -0.789
Gender_Nonbinary               0.2240      0.023      9.592      0.000
0.178       0.270
InternetService_Fiber Optic    6.1348      0.008    751.051      0.000
6.119       6.151
```

```
InternetService_None                4.3514      0.010    450.093      0.000
4.332       4.370
Multiple_Yes                        0.8924      0.008    115.759      0.000
0.877       0.908
OnlineSecurity_Yes                 -0.7775      0.007   -107.279      0.000
-0.792      -0.763
OnlineBackup_Yes                    0.0815      0.007     11.127      0.000
0.067       0.096
DeviceProtection_Yes               -0.3536      0.007    -49.622      0.000
-0.368      -0.340
TechSupport_Yes                     0.6231      0.007     85.781      0.000
0.609       0.637
==============================================================================
Omnibus:                     1442.743   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1436.185
Skew:                          -0.862   Prob(JB):                         0.00
Kurtosis:                       2.311   Cond. No.                     2.74e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.74e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

4.4 D2 Cont: Linear Regression Model Variables without Linear Relationships Removed

By plotting the scatterplots of the remaining variables against the target variable and by referencing the partial correlation matrix in section C, it can be determined that there are variables within the model that are violating the linear relationship assumption. Additionally, removing the variables has little affect on the R-squared value meaning that they can be removed without compromising the accuracy of the model. Some variables remained in the model to meet task requirement of having two continuous and two categorical variables in the model.

[23]:
```
churn_LRM_linearity = churn_LRM_rmvhighpvalues.
 ↪drop(['MonthlyCharge','Age','Multiple_Yes','Gender_Male',
                                                          ␣
 ↪'Gender_Nonbinary','InternetService_Fiber Optic','OnlineSecurity_Yes',
                                                          ␣
 ↪'InternetService_None','TechSupport_Yes'], axis=1)

final_X = churn_LRM_linearity.drop('Tenure', axis = 1)
final_Y = churn_LRM_linearity[['Tenure']]

# Split X and y into X_
```

```
final_X_train, final_X_test, final_Y_train, final_Y_test =␣
↪train_test_split(final_X, final_Y, test_size=0.20, random_state=1)
```

[24]:
```
#define the input
final_X2 = sm.add_constant(final_X)

#create an OLS model
final_model = sm.OLS(final_Y, final_X2)

#fit the data
final_est = final_model.fit()

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

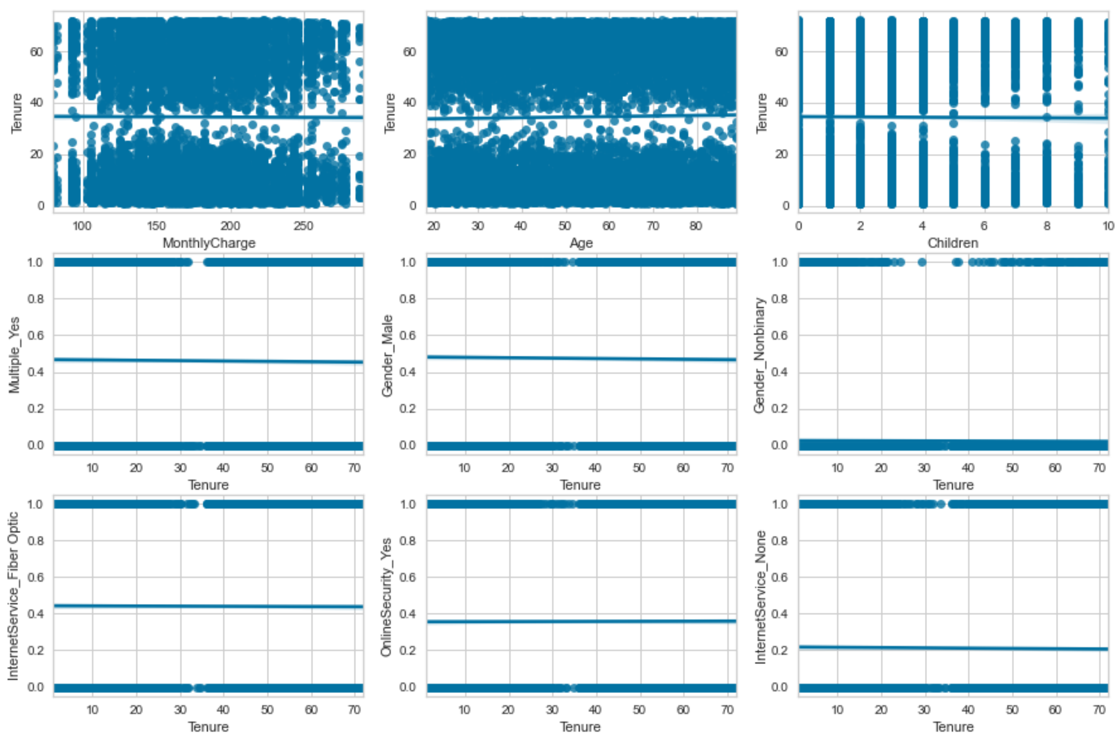
[24]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       0.985
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                 1.621e+05
Date:                Sun, 11 Jul 2021   Prob (F-statistic):               0.00
Time:                        14:37:50   Log-Likelihood:                -26002.
No. Observations:               10000   AIC:                         5.201e+04
Df Residuals:                    9995   BIC:                         5.205e+04
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
========
                           coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
--------
const                   -4.4968      0.078    -57.444      0.000      -4.650
-4.343
Children                -0.3743      0.015    -24.649      0.000      -0.404
-0.345
Bandwidth_GB_Year        0.0120   1.49e-05    804.600      0.000       0.012
0.012
OnlineBackup_Yes        -1.0870      0.066    -16.579      0.000      -1.216
-0.959
DeviceProtection_Yes    -1.0626      0.066    -16.177      0.000      -1.191
-0.934
==============================================================================
Omnibus:                      470.196   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              329.077
```

```
Skew:                          -0.337   Prob(JB):                     3.48e-72
Kurtosis:                       2.421   Cond. No.                     1.11e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.11e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[25]:

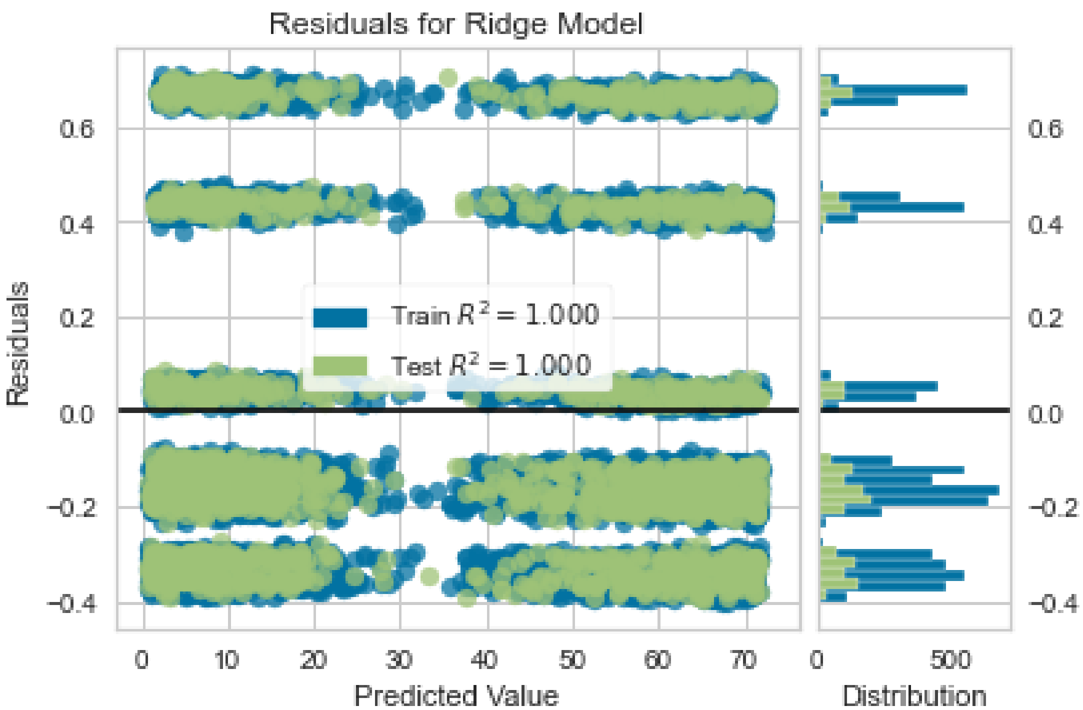
```
fig, axs = plt.subplots(3,3,figsize=(15,10))

pl1 = sns.regplot(x='MonthlyCharge',
                  y='Tenure',
                  data=churn_LRM_transdata,
                  ax = axs[0,0])

pl2 = sns.regplot(x='Age',
                  y = 'Tenure',
                  data=churn_LRM_transdata,
                  ax = axs[0,1]);

pl3 = sns.regplot(x='Children',
                  y='Tenure',
                  data=churn_LRM_transdata,
                  ax = axs[0,2])

pl4 = sns.regplot(x='Tenure',
                  y='Multiple_Yes',
                  data=churn_LRM_transdata,
                  logistic=True,
                  ax = axs[1,0])

pl5 = sns.regplot(x='Tenure',
                  y='Gender_Male',
                  logistic=True,
                  data=churn_LRM_transdata,
                  ax = axs[1,1])

pl6 = sns.regplot(x='Tenure',
                  y='Gender_Nonbinary',
                  logistic=True,
                  data=churn_LRM_transdata,
                  ax = axs[1,2])

pl7 = sns.regplot(x='Tenure',
```

```
                 y='InternetService_Fiber Optic',
                 logistic=True,
                 data=churn_LRM_transdata,
                 ax = axs[2,0])

pl8 = sns.regplot(x='Tenure',
                 y='OnlineSecurity_Yes',
                 logistic=True,
                 data=churn_LRM_transdata,
                 ax = axs[2,1])

pl9 = sns.regplot(x='Tenure',
                 y='InternetService_None',
                 logistic=True,
                 data=churn_LRM_transdata,
                 ax = axs[2,2])
```

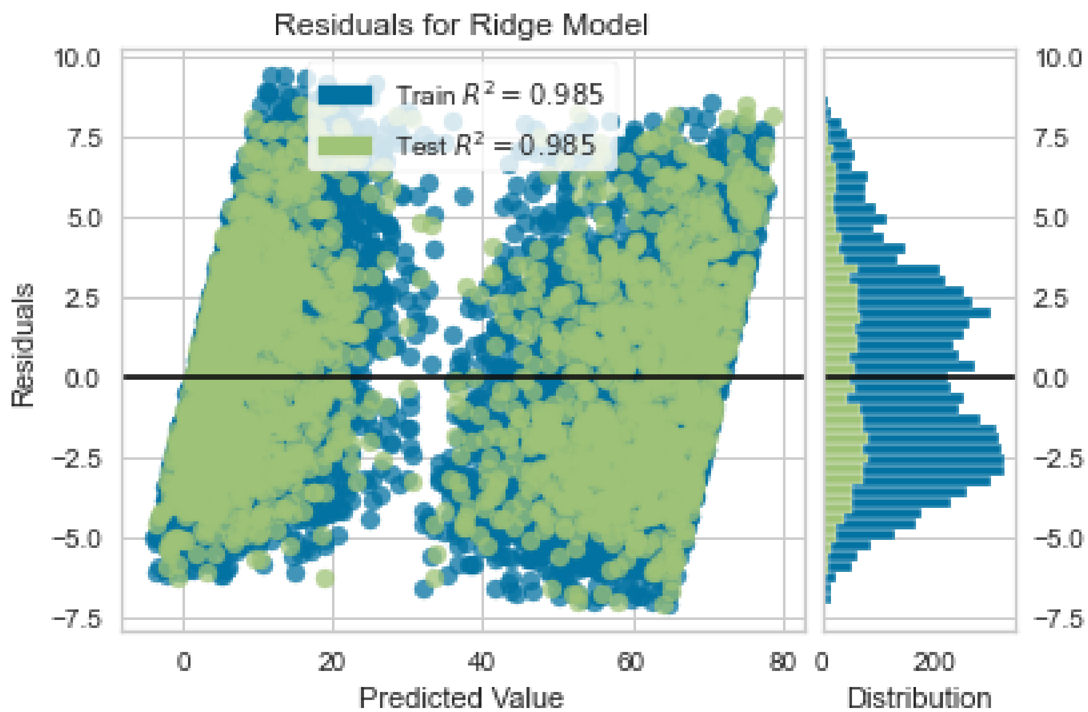


[22]:
```
#initial model residual plot
model = Ridge()
visualizer = ResidualsPlot(model)

visualizer.fit(next3_X_train, next3_Y_train)
visualizer.score(next3_X_test, next3_Y_test)
```

```
g = visualizer.poof()
```



[26]:
```
model = Ridge()
visualizer = ResidualsPlot(model)

visualizer.fit(final_X_train, final_Y_train)
visualizer.score(final_X_test, final_Y_test)
g = visualizer.poof()
```



4.5 Part E: Analyze the Dataset Using the Reduced Multiple Regression Model

The logic of the variable selection technique was to ensure each variable in the model met each of the multiple linear regression assumptions. If an assumption was violated it was removed. Each violation is described in section D2 as variables were removed.

With the exception of the amount of variables included, both the initial model and reduced multiple regression models are very similar. This is because the variable "Bandwidth_GB_Year", which is in both models accounts for the majority of the R-squared value. The independent variables that were removed had very little effect on the R-squared value but were removed to avoid assumption violations and legitimize the reduced model. Comparison of the initial and final models are made with the following metrics:

R-Squared/R-Squared adjusted: Both models have an extremely high R-Squared statistic that defines ~99% of the target variable's variance. The final model has a slightly lower R-squared value. These high R-squared indicate that both models are well fitted.

One concerning statistic in both models is the condition number. While significantly reduced in the final model these numbers are extremely large suggesting multicollinearity among the variables. However, the VIF numbers calculated early suggest that multicollinearity is insignificant. This large number may be caused by the near-perfect colinearity between the "Bandwidth_GB_Year" independent variable and the target variable.

As shown above using the Yellowbrick package, the inital models' errors were not normally distributed and did not meet the normally distributed errors assumption. After removing non colinear variables, the distribution normalized. However, the assumption of homoscedasticity among the residuals is not met. As visualized above between the predicted values of 20 and 40 the error distribution shrinks into a gap of errors between approximately 30 and 35. This violation of homoscedasticity suggests that the test results and confidence intervals are unreliable.

5 Part V: Data Summary and Implications

5.0.1 F1 Results of the Data Analysis

Regression equation:

$y = -0.3743x_1 + 0.120x_2 - 1.0870x_3 - 1.0626x_4 - 4.4968$

where

x_1 = Children

x_2 = Bandwidth_GB_Year

x_3 = OnlineBackup_Yes

x_4 = DeviceProtection_Yes

Coefficient interpretation

- Children: This coefficeient suggests that for each child the customer has reduce the estimated Tenure by .3743 months.
- Bandwidth_GB_Year: Even while being the smallest coefficient in the model, this variable has the most weight on the model. This is due to (1) the variable being the only continuous variable in the model and (2) the variable having the largest values between 155.51 and 7158.98 with a mean value of 3392.34. Ultimately, this independent variable is the best predictor variable for the target variable.
- OnlineBackup_Yes: As a categorical variable, this variable only has two potential values, 1 or 0. If the variable exists as a 1 in the observation it will decrease the target variable prediction (Tenure) by -1.0870 months.
- DeviceProtection_Yes: Again, as a categorical variable, this variable only has two potential values, 1 or 0. If the variable exists as a 1 in the observation it will decrease the target variable prediction (Tenure) by -1.0626 months.

Statistical and Practical Significance of the Model

Looking at the statistical significance of the model we determined that the R-squared value is .985. This means that the model can explain ~99% of the target variable's variance with a 95% confidence level. At a statistical level, the model can very accurately predict the tenure of a customer.

Practically, the model makes sense at first. Customers with a longer tenure would be much more likely to use a higher amount of data because they've been using the service longer. However, the Bandwidth_GB_Year variable isn't cumulative and continues the near-perfect positive colinear relationship with Tenure even past the 12-month mark. The model suggests there is a high correlation between the amount of data a customer uses and their tenure and further analysis suggests the causation is different than just the length of time the customer has used the service.

Limitations

A significant limitation of the model is the fact that the independent variable "Bandwidth_GB_Year" has such a large overshadowing effect over the other variables. The model is basically a simple linear model because of the strong correlation of this variable with the target variable. The other two remaining variables could be removed and there would be almost no change to the outcome of the model.

Another limitation of the model is, due to multicollinearity, some independent variables needed to be removed from the model because they were redundant. By returning to the initial model, redundant variables can be replaced to determine what other independent variables had similar effects on the target variables.

5.0.2 F2 Recommended Course of Action

The recommended course of action is to conduct further exploration and analysis on the correlation of the amount of data customers use over time and their length of tenure. Looking at the data, there is a cluster of customers who use a larger amount of data per year and stay with the customers longer. It would be beneficial to find out why these customers use more data to attempt to target potential customers with the same qualities.

6 Part VI: Demonstration

6.1 G. Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b066ba2f-6981-4845-8eab-ad6200cc72d3

6.2 H. Code Sources

Bengfort, B., & Bilbro, R. (2019). Yellowbrick: Visualizing the Scikit-Learn Model Selection Process. Journal is required!, 4(35).

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

Foltz, B. (2014, December 02). Statistics 101: Multiple Linear Regression [Video blog post]. Retrieved June 29, 2021, from https://www.bcfoltz.com/

Osborne, Jason W. and Waters, Elaine (2002) “Four assumptions of multiple regression that researchers should always test,” Practical Assessment, Research, and Evaluation: Vol. 8 , Article 2. DOI: https://doi.org/10.7275/r222-hv23

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

Williams, Matt N.; Grajales, Carlos Alberto Gomez; and Kurkiewicz, Dason (2013) “Assumptions of Multiple Regression: Correcting Two Misconceptions,” Practical Assessment, Research, and Evaluation: Vol. 18 , Article 11. DOI: https://doi.org/10.7275/55hn-wk47 Available at: https://scholarworks.umass.edu/pare/vol18/iss1/11