D208 Predictive Modeling Performance Assessment

Author: Eric Colwell

Western Governors University

This report deals with the rubric for Western Governors University course D208 Predictive Modeling and answers all the items in rubric order.

# **Part I:**

**A1:**

The research question for this analysis is “What variables influence Tenure”?

**A2:**

The objective of this data analysis is to determine which variables or factors have the greatest influence on Tenure through Multiple Linear Regression (MLR). For this analysis I have chosen Tenure as the dependent or target variable. The independent or predictor variables are Population, Area, Children, Age, Income, Marital, Gender, Churn, Outage\_sec\_perweek, Email, Contacts, Yearly\_equip\_failure, Techie, Contract, Port\_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod, MonthlyCharge, Bandwidth\_GB\_Year, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8. Variables not included in the initial model are CaseOrder, Customer\_id, Interaction(UID), City, State, County, Zip, Lat, Lng, TimeZone, and Job. These variables have high cardinalities and would not add predictive power to the model.

# **Part II:**

**B1:**

A multiple regression model assumes that there is a linear relationship between the dependent (target) variable and the independent (predictor) variables and that they are not highly correlated with each other. The observations should be selected independently and randomly from the population. The residuals should be normally distributed with a mean of zero and explanatory power should increase with an increase in variables. (Sewell, 2022)

**B2:**

The benefits of using the Python language for this analysis are that it is a capable tool for data science, statistics, data exploration, data analysis, and predictive analytics. Also, I am more familiar with Python. For these reasons I chose this language for the Multiple Linear Regression Analysis. I imported pandas to import the dataset, data manipulation, and analysis. I imported numpy for numerical calculations. I imported matplotlib.pyplot for plotting data. I imported stats from scipy for statistical analysis and outlier detection. I imported plotnine for plotting data. I imported missingno to visually check for missing values. I imported seaborn for visualizations. I imported LinearRegression for model creation. I imported train\_test\_split for model testing. I imported metrics for additional statistical output. I imported linear\_model for model creation. I imported variance\_inflation\_factor from statsmodels to check VIF scores. I imported statsmodels.api for model creation.

**B3:**

Multiple linear regression (MLR) is commonly used for predictive analysis. MLR is a statistical approach to model a relationship between a dependent (target) variable and a set of independent (predictor) variables. MLR can be used to discover which factors have the highest impact on predicted outcomes and how variables relate to each other (YahRaj5, 2022). In this case, MLR is used to answer the research question posed in A1.

# **Part III:**

**C1:**

My data preparation goals follow the Data Analytics Life Cycle. The research question is the Business Understanding phase. Importing the .csv file into my Jupyter notebook is the Data Acquisition phase. For the Data Cleaning Phase I checked for duplicated rows or instances then checked for null values. Next, using z-scores, I checked for outliers and remedied them using proper techniques. In the Data Exploration phase, I identified categorical (nominal and ordinal) values, and encoded them properly. These steps are necessary for the Predictive Modeling phase to begin.

**C2:**

Below are the summary statistics for the target variable and all predictor variables including screenshots of output showing the count of instances (10,000), mean (average), standard deviation (amount of variation of a set of data), minimum (least amount), quartile ranges (measure of variability of the data), maximum (largest amount) for continuous variables. Unique values, and value counts for categorical variables.

**Tenure (Target Variable):** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the number of months the customer has stayed with the provider (range is roughly 1 to 72 months).

Table

Description automatically generated with medium confidence

**Population:** Datatype is int64, quantitative, continuous, and has no null values. The value reflects the population within a mile radius of the customer, based on census data (range from 0 to 111,850).

Table

Description automatically generated

**Area:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the area type (Rural, Urban, Suburban), based on census data.

Text, table

Description automatically generated

**Children:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the number of children in the customer’s household as reported in sign-up information (ranging from 0 to 10).

Table

Description automatically generated

**Age:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the age of the customer as reported in sign-up information (ranging from 18 to 89).

Table

Description automatically generated

**Income:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the annual income of the customer (or invoiced person) as reported at time of sign-up (ranging from roughly 348 to 258,900).

Text

Description automatically generated with low confidence

**Marital:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the marital status of the customer as reported in sign-up information (Divorced, Married, Never Married, Separated, Widowed).

Text

Description automatically generated

**Gender:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the customers self-identification as male, female, or nonbinary.

Text

Description automatically generated

**Churn:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer discontinued service within the last month (Yes, No).

Graphical user interface, text

Description automatically generated

**Outage\_sec\_perweek:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the average number of seconds per week of system outages in the customer’s neighborhood (range from roughly 0.1 to 21.2).

Text

Description automatically generated

**Email:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the number of emails sent to the customer in the last year (range from 1 to 23).

Text, table

Description automatically generated

**Contacts:** Datatype is float64, quantitative, discrete, and has no null values. The value reflects the number of times the customer contacted technical support (range from 0 to 7).

Table

Description automatically generated

**Yearly\_equip\_failure:** Datatype is float64, quantitative, discrete, and has no null values. The value reflects the number of times customers equipment failed and had to be reset/replaced in the past year (range from 0 to 6).

Table

Description automatically generated

**Techie:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer considers themselves technically inclined (Yes, No).

Text

Description automatically generated

**Contract:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the contract term of the customer (Month-to-month, One Year, Two Year).

Text

Description automatically generated with medium confidence

**Port\_modem:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has a portable modem (Yes, No).

**Text, table

Description automatically generated**

**Tablet:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer owns a tablet such as iPad, Surface, etc. (Yes, No).

**Text

Description automatically generated**

**InternetService:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the customers internet service provider (DSL, Fiber Optic, or None).

**Text

Description automatically generated**

**Phone:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has a phone service (Yes, No).

**Text

Description automatically generated**

**Multiple:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has multiple lines (Yes, No).

Text

Description automatically generated

**OnlineSecurity:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has an online security add-on (Yes, No).

**Text

Description automatically generated**

**DeviceProtection:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has device protection add-on (Yes, No).

**Graphical user interface, text

Description automatically generated**

**TechSupport:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has a technical support add-on (Yes, No)

**Text

Description automatically generated**

**StreamingTV:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has streaming TV (Yes, No).

**Text

Description automatically generated**

**StreamingMovies:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has streaming movies (Yes, No).

**Text

Description automatically generated**

**PaperlessBilling:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has paperless billing (Yes, No).

Text

Description automatically generated

**PaymentMethod:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the customer’s payment Method (Bank Transfer, Credit Card, Electronic Check, Mailed Check).

**Text

Description automatically generated**

**MonthlyCharge:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the amount charged to the customer monthly. This is an average for the customer. For brand new customers, this value is the average for other customers who fit the new customers profile (range from roughly 80 to 290).

Text

Description automatically generated

**Bandwidth\_GB\_Year:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the average amount of data used, in GB, in a year by the customer (range from roughly 155 to 7159).

Text

Description automatically generated

The following variables (Item1 - Item8) represent responses to an eight-question survey asking customers the importance of various factors on a scale of 1 to 8 where 1 is most important and 8 is least important. The datatypes are float64, quantitative, continuous, and have no null values.

**Item1:** Timely response

Table

Description automatically generated

**Item2:** Timely fixes

Table

Description automatically generated

**Item3:** Timely replacements

Text, table

Description automatically generated

**Item4:** Reliability

Table

Description automatically generated

**Item5:** Options

Table

Description automatically generated

**Item6:** Respectful response

Table

Description automatically generated

**Item7:** Courteous exchange

Table

Description automatically generated

**Item8:** Evidence of active listening

Table

Description automatically generated

**C3:**

I imported the churn.csv file using pandas. Next, I checked the dataset shape and information. There are zero null values, 10,000 rows, and 50 columns. I checked for duplicated rows and found there were none. I made a copy of the original to preserve it. I made a new dataset that included the variables or columns that I used for the analysis. I checked the uniqueness of each of those chosen variables for validity. I used z-scores to identify outliers in the continuous variables, set the outliers equal to nan, then used median or mean imputation for replacement. I made a new dataframe (df2) that had been cleaned and includes only the variables being utilized for the MLR. In the Data Exploration phase I created univariate visualizations, and bivariate visualizations that are shown in C4. For the data wrangling phase I used ordinal encoding for the variables with Yes/No values. Then I used pd.get\_dummies with k-1 to encode the nominal variables. The dataset is now prepared for MLR.

The Jupyter notebook file (.ipynb format) containing the complete annotated code for this project will be uploaded with the submission.

**C4:**

Below are the univariate and bivariate visualizations for the cleaned dataset.

**Tenure (Target variable):**

Chart, histogram

Description automatically generated

**Population (univariate):**

Chart

Description automatically generated

**Population/Tenure (bivariate):**

Chart, scatter chart

Description automatically generated

**Area (univariate):**

Chart, bar chart

Description automatically generated

**Area/Tenure (bivariate):**

Chart, histogram

Description automatically generated

**Children (univariate):**

**Chart, histogram

Description automatically generated**

**Children/Tenure (bivariate):**

**Table

Description automatically generated**

**Age (univariate):**

Chart, histogram

Description automatically generated

**Age/Tenure (bivariate):**

A screenshot of a map

Description automatically generated with medium confidence

**Income (univariate):**

Chart, histogram

Description automatically generated

**Income/Tenure (bivariate):**

Chart, scatter chart

Description automatically generated

**Marital (univariate):**

**Chart

Description automatically generated**

**Marital/Tenure (bivariate):**

**Chart

Description automatically generated**

**Gender (univariate):**

Chart, bar chart

Description automatically generated

**Gender/Tenure (bivariate):**

Chart, histogram

Description automatically generated

**Churn (univariate):**

**Chart, bar chart

Description automatically generated**

**Churn/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**Outage\_sec\_perweek (univariate):**

Chart, histogram

Description automatically generated

**Outage\_sec\_perweek/Tenure (bivariate):**

Chart, scatter chart

Description automatically generated

**Email (univariate):**

**Chart, histogram

Description automatically generated**

**Email/Tenure (bivariate):**

**A picture containing chart

Description automatically generated**

**Contacts (univariate):**

Chart

Description automatically generated

**Contacts/Tenure (bivariate):**

Chart

Description automatically generated

**Yearly\_equip\_failure (univariate):**

Chart, bar chart

Description automatically generated

**Yearly\_equip\_failure/Tenure (bivariate):**

Chart

Description automatically generated

**Techie (univariate):**

**Chart, bar chart

Description automatically generated**

**Techie/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**Contract (univariate):**

Chart, bar chart

Description automatically generated

**Contract/Tenure (bivariate):**

Chart, histogram

Description automatically generated

**Port\_modem (univariate):**

**Chart, bar chart

Description automatically generated**

**Port\_modem/Tenure (bivariate):**

**Chart

Description automatically generated**

**Tablet (univariate):**

**Chart, bar chart

Description automatically generated**

**Tablet/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**InternetService (univariate):**

**Chart, bar chart

Description automatically generated**

**InternetService/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**Phone (univariate):**

**Chart, bar chart

Description automatically generated**

**Phone/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**Multiple (univariate):**

Chart, bar chart

Description automatically generated

**Multiple/Tenure (bivariate):**

Chart, histogram

Description automatically generated

**OnlineSecurity (univariate):**

**Chart, bar chart

Description automatically generated**

**OnlineSecurity/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**OnlineBackup (univariate):**

**Chart, bar chart

Description automatically generated**

**OnlineBackup/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**DeviceProtection (univariate):**

**Chart, bar chart

Description automatically generated**

**DeviceProtection/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**TechSupport (univariate):**

**Chart, bar chart

Description automatically generated**

**TechSupport/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**StreamingTV (univariate):**

**Chart, bar chart

Description automatically generated**

**StreamingTV/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**StreamingMovies (univariate):**

**Chart, bar chart

Description automatically generated**

**StreamingMovies/Tenure (bivariate):**

**Chart, histogram

Description automatically generated**

**PaperlessBilling (univariate):**

Chart, bar chart

Description automatically generated

**PaperlessBilling/Tenure (bivariate):**

Chart, histogram

Description automatically generated

**MonthlyCharge (univariate):**

Chart, histogram

Description automatically generated

**MonthlyCharge/Tenure (bivariate):**

Chart, scatter chart

Description automatically generated

**Bandwidth\_GB\_Year (univariate):**

Chart, histogram

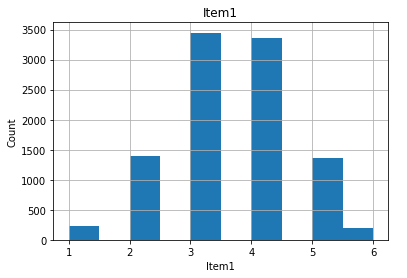
Description automatically generated

**Bandwidth\_GB\_Year/Tenure (bivariate):**

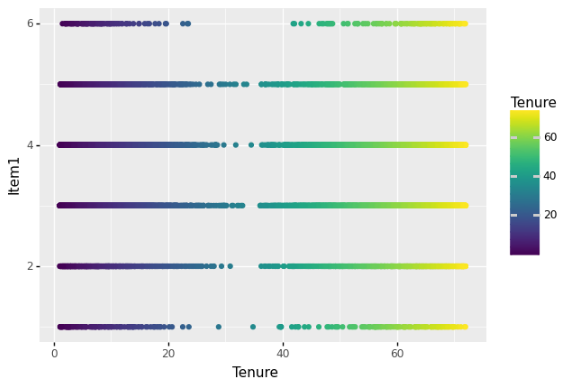
Chart, scatter chart

Description automatically generated

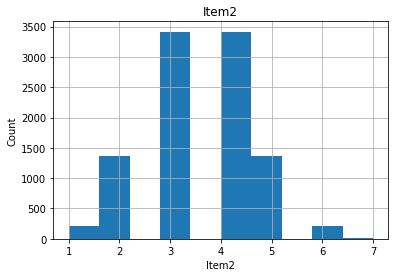
**Item1 (univariate):**

****

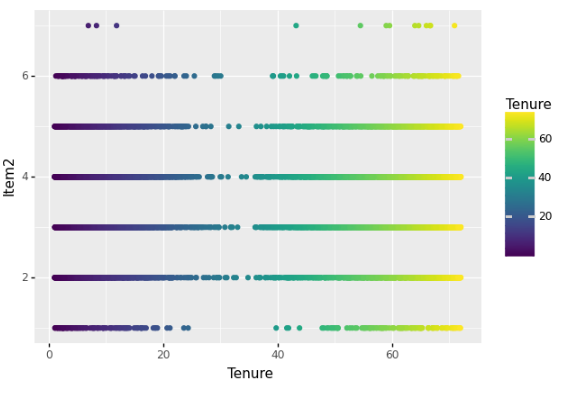
**Item1/Tenure (bivariate):**

****

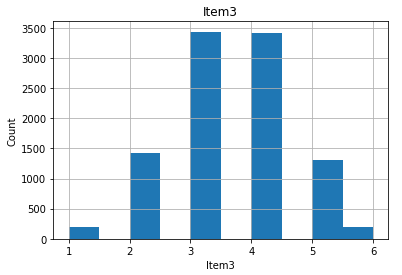
**Item2 (univariate):**

****

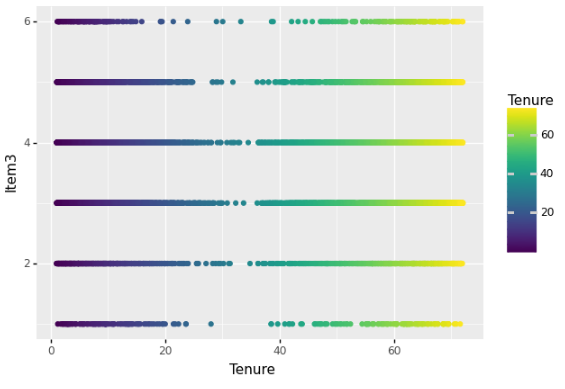
**Item2/Tenure (bivariate):**

****

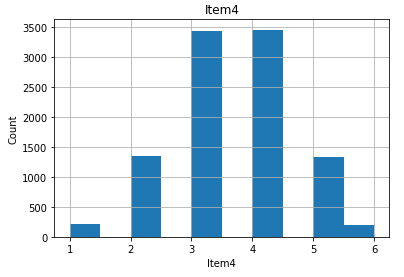
**Item3 (univariate):**

****

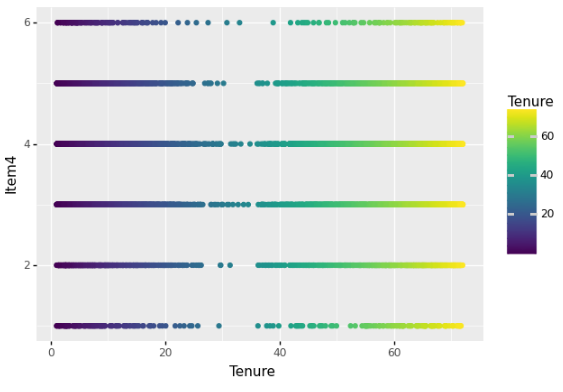
**Item3/Tenure (bivariate):**

****

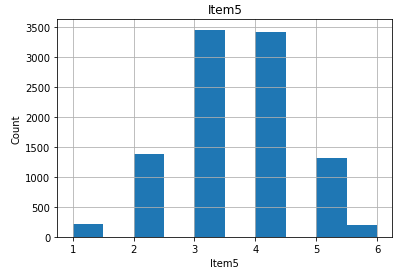
**Item4 (univariate):**

****

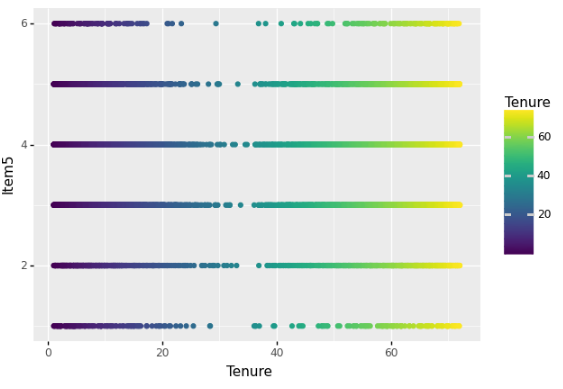
**Item4/Tenure (bivariate):**

****

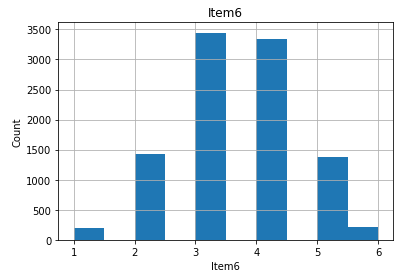
**Item5 (univariate):**

****

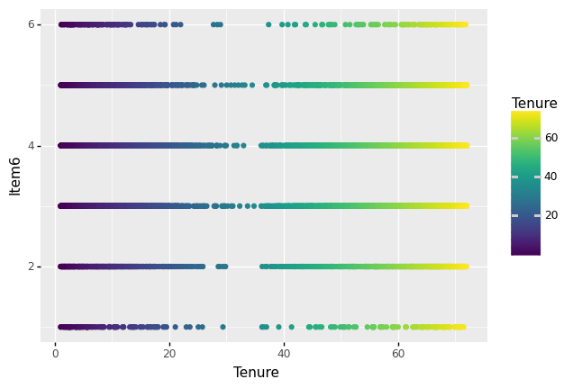
**Item5/Tenure (bivariate):**

****

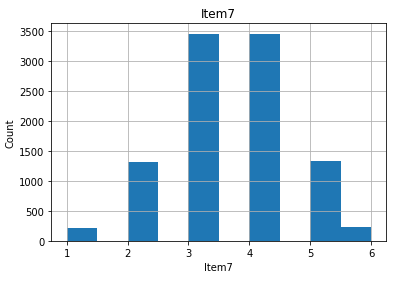
**Item6 (univariate):**

****

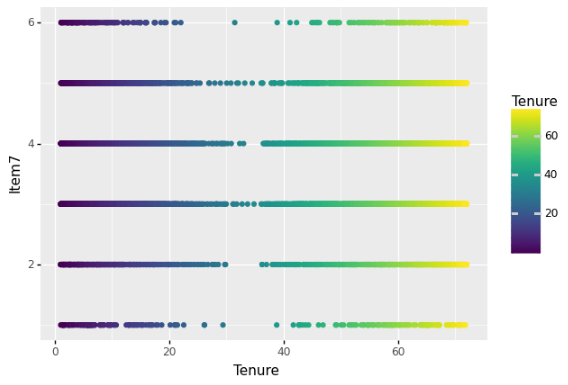
**Item6/Tenure (bivariate):**

****

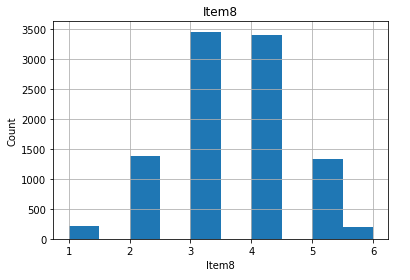
**Item7 (univariate):**

****

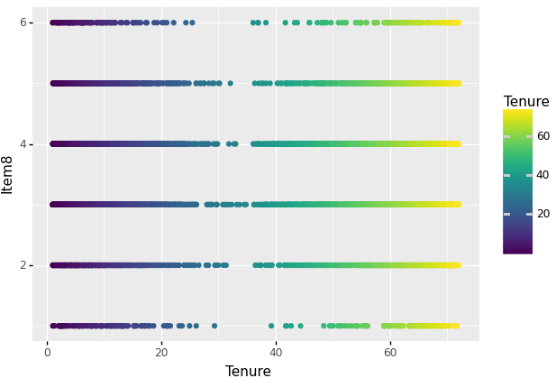
**Item7/Tenure (bivariate):**

****

**Item8 (univariate):**

****

**Item8/Tenure (bivariate):**

****

**C5:**

A copy of the cleaned and prepared dataset in .csv format will be uploaded with the submission.

# **Part IV:**

**D1:**

Initial Multiple Regression model with y-intercept below:

Text

Description automatically generated

Text

Description automatically generatedTable

Description automatically generatedA picture containing calendar

Description automatically generated

**D2:**

The method used for variable selection is the wrapper method backward stepwise elimination and the adjusted r-squared for the metric. First, I looked for high correlation between the independent variables and the target variable. Bandwidth\_GB\_Year had a very high correlation at 0.991495 (Max is 1.0), so I removed it. Before removal of that variable the r-squared value was 1.00. After removal, the r-squared value dropped to 0.315. Next, I checked VIF (Variance Inflation Factor) values and found that MonthlyCharge had a value well above 10 so I removed it. Then, using the backward stepwise elimination variable selection technique, I removed the least significant variable with the highest p-value, re-fit the model, and ran OLS (Ordinary Least Squares) again. I iterated through this process until there were no variables with a p-value greater than 0.05 and all VIF values were less than 10. Once I achieved these goals, the reduced model was complete.

Text, letter

Description automatically generated

**D3:**

Reduced model shown below:

**A picture containing graphical user interface

Description automatically generated**

**E1:**

The logic of the backward stepwise elimination variable selection technique is to remove the non-significant features iteratively according to the highest p-value. The p-values indicate whether there is a statistically significant relationship between the independent variable and the dependent variable. By removing the independent variables with p-value greater than 0.05 I created a model that is statistically significant.

I used R-squared for the model evaluation metric. This metric shows the percentage of the dependent variable variation the model explains. The R-squared value is between 0 and 100 percent. Zero percent means the model does not explain any of the variation in the dependent variable and 100 percent means the model explains all the variation in the dependent variable. For the reduced model, the R-squared value is 31 percent, meaning 31 percent of the variation in the dependent variable can be explained.

**Residual plots below:**

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

**E2:**

**Residual sum and mean:**

Text, letter

Description automatically generated

**Model coefficients:**

Text

Description automatically generated

**Model r-squared value:**

Graphical user interface, text, application

Description automatically generated

**Model residual value:**

Graphical user interface, text, application

Description automatically generated

**Model predictions:**

Chart, histogram

Description automatically generated

**E3:**

A Jupyter .ipynb file containing the full code for this project will be uploaded along with the Performance Assessment.

# **Part V: Data Summary and Implications**

**F1:**

The regression equation for the reduced model is:

y = 36.48 + (0.2 (Age) – 38.31 (Churn) + 2.32 (Techie) + 4.02 (Multiple) + 2.83 (OnlineBackup)

+ 8.19 (StreamingTV) + 9.62 (StreamingMovies) – 8.99 (Contract\_One year)

– 8.29 (Contract\_Two year) – 3.90 (InternetSerive\_Fiber Optic) – 4.26 (InternetServie\_None)

+ 1.50 (PaymentMethod\_Credit Card automatic) + 1.84 (PaymentMethod\_Electronic Check)

+ 1.33 (PaymentMethod\_Mailed Check)).

The coefficient of each independent variable gives the size of the effect that variable has on the dependent variable. For example, if there is one unit increase in Age there would be a 0.2 increase on average in y (Tenure). For the categorical variable coefficients, StreamingTV for instance, if the customer has the StreamingTV add-on there is an increase of 8.19 on average in y (Tenure). The sign (+ or -) gives you the direction of that effect. The coefficient for the constant (36.48) is the average expected value for the dependent variable when all the independent variables are equal to zero.

As explained above, there is statistical significance for this model. The model has 14 statistically significant variables due to p-values being less than 0.05 and an R-squared value of 31 percent. This means there can be conclusions drawn from the model that may have practical implications on the dependent variable. The practical significance of the model is that, given the coefficients of the independent variables, efforts could be made to retain customers by offering add-ons at a reduced price or free.

The limitations of the analysis are that overfitting can occur because of limited data or too many independent variables. When there are independent variables that have high correlation with the other independent variables multicollinearity can occur. This can cause p-values and coefficients to be unreliable (Sewell, 2022).

**F2:**

The analysis of the reduced multiple linear regression model shows insight to some independent variables having a significant impact on the outcome of Tenure. For example, StreamingMovies has a coefficient of +9.62. My recommendation is for further MLR analysis using different variable selection techniques and metrics to see if these same independent variables continue to show statistical significance. Using different techniques and returning similar results would support the significance of the independent variables. If these variables show significance in different models, then I would recommend a course of action where the company could potentially retain customers at a higher rate with bonus offers. One example would be to offer free Streaming Movie add-on for customers who sign up for a one-year contract.

**F:**

The Panopto video recording URL will be uploaded as part of the submission.

References

Sewell, W. (2022). D208 Predictive Modeling Webinar. Retrieved from <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b7ead95b-c392-4973-aa9c-ad1901031ab1>

YashRaj5 (2022, September 1). ML | Multiple linear regression using python. <https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/>