HOW DO PREDICTIVE VARIABLES AFFECT MONTHLY CHARGES?

PREDICTIVE MODELING – D208

PERFORMANCE ASSESSMENT TASK 1

MULTIPLE REGRESSION FOR PREDICTIVE MODELING

STUDENT NAME: IBRAHIM SULEIMAN

DATA ANALYTICS

STUDENT NUMBER: 001429984

COLLEGE OF INFORMATION TECHNOLOGY, WESTERN GOVERNORS UNIVERSITY

NAME OF INSTRUCTOR: KEIONA MIDDLETON

September 11th, 2022

Multiple regression is a method used to generate regressions upon models alongside a sole subordinate variable and various self-sufficient variables. The distinction interpolated in the equation for linear regression alongside the equation for multiple regression is that the equation for multiple regression is pre-conditioned to have the capacity to handle numerous information rather than solely the single judgment of linear regression.

The equation for multiple regression entails the formation to report for this variation.

y = B\_1 \* x\_1 + B\_2 \* x\_2 + … + B\_n \* x\_n + A (Peter Grant, 2019).

**Part I: Research Question**

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

**1.  Summarize one research question relevant to a real-world organizational situation captured in the data set we have selected and that we will be using multiple regression.**

The summary includes one research question relevant to a realistic organizational situation and can be resolved with multiple regression and the chosen data set. Without a doubt, the study question pertains to the organizational environment: How do predictive variables monthly charge? Additionally, the chosen data set and multiple regression will be used to address this research question. The target variable used is the monthly charge variable and several explanatory variables.

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

***The submission defines the objectives or goals of the data analysis, and the objectives or goals are reasonable for the scope of the scenario and are represented in the available data.***

The purpose of this data analysis is to look at multiple regression models alongside the hypothesis over the standardized collection of the data set(cleaned\_churn) and the corresponding data variables to interpret organizational functions that could strengthen the organizational decision-making process when it comes to the variables that affect the monthly charge.

The objective is to apply the independent variables whose values are known to predict the value of the single dependent value.

Our goal is to assess the strength of the relations among our dependent variables (monthly charge) and predictive variables.

Our goal is to predict the value of the monthly charge based on the value of other predictive variables that affects the monthly charge.

**Part II: Method Justification**

**B.  Describe multiple regression methods by doing the following:**

**1.  Summarize the assumptions of a multiple regression model.**

***The submission accurately summarizes the assumptions of a multiple regression model****.*

A multiple regression model is a method of analysis we could apply to interpret the relation in the middle of various predictor variables alongside a response variable.

Principles underlying linear regression

- Linear relationship: dependent and independent variables are related linearly

- Multivariate normality: Has it that data should has a normal distribution.

- Multicollinearity: it assumes that there is correlation between one or two regressors in multiple regression.

- No autocorrelation: The residuals ought to be distinct from one another.

- Homoscedasticity: The regression line's residuals should be evenly spaced.

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

***The submission describes the benefits of using the tool(s) chosen in support of various phases of the multiple regression analysis, and the benefits logically align with the goal of the analysis.***

**Python**: is a programming language that is well-used to read, assemble and organize code capable of execution on several interfaces and platforms designated by workload management systems due to its powerful indentation.

**The Benefits of using Python for this data analysis**

· Eco-friendly environment – The user-friendly environment python entails provides a fast and easy way to adapt to the functions and perform code execution efficiently.

· Python is practical for implementing, reading code, detecting missing values and outliers in the data set, and creating visual representations to spot animalities and identify outliers during the data cleaning execution

· Python made it systematic for the 'Churn\_data.csv' file to be imported for the cleaning operations to begin.

In addition, Python allowed the importation of packages into the new environment, such as Pandas, NumPy, Matplotlib, Sklearn, and Seaborn. These packages offer a variety of features, such as creating visualizations of histograms, boxplots, and data tables. Without a doubt, these packages, alongside the programming languages, are user-friendly, ideal, and intuitive in providing data analysts with efficiency and error-free output in an innovative presentation opposing other tools (Michael Galarnyk,2018).

· **Pandas**: pandas is a software library drafted for Python operating systems for data handling alongside inquiry. Specifically, it presents data design alongside procedures for executing arithmetic groups.

**The Benefits of Pandas**

· It implements a quick and dynamic strategy to take care of data

· It is straightforward to treat data omitted values.

**NumPy**: NumPy is a Python library applicable for dealing with arrays. In addition, it contains operations for functioning in the domain of linear algebra alongside matrics.

**Benefits of Numpy**

· NumPy's arrays appear to be less in proportion compared to Python lists

· The quick execution is magnificent because it acts rapidly in computing than python lists

· **Matplotlib**: Matplotlib is a cross-structure, data representation, and graph plotting library for Python alongside its binary extension NumPy

**Benefit of Matplotlib**

· It presents the user with an interface to represent data by applying various sorts of plots to communicate the data effectively

· We can execute multiple sorts of plots (scatterplots, histograms, bar charts, error charts, boxplots, etc.) by executing a scanty line of code in Python

· **Sklearn**: Scikit-learn is a suitable and powerful library for machine learning in Python. It proffers a collection of powerful setups for machine learning and analytical modeling alongside distribution, regression, clustering, and dimensionality minimization over a consistent platform in Python.

**Benefits of Sklearn**

· Scikit-learn entails diversely supervised & unsupervised learning algorithms. Most significantly, its simplicities as well as the cleanest machine learning library

· It appears to be formative and unify distinctively along several Python libraries, including Matplotlib for charts, Numpy for arithmetic calculations, as well as Pandas for DataFrames

· **Seaborn**: Seaborn is a library regulated by Python that helps represent data and creates enormous and further analytical operations.

**Benefits of Seaborn**

· we could systematically represent our data on a plot

· This library is created to help us reflect on our data; without manipulating the inner technicalities.

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

***The submission accurately explains why multiple regression is an appropriate technique to analyze the research question from Part I.***

The multiple regression election procedure facilitates the analyst to gather a decreased group of variables from a bulkier group of predictors, therefore disposing of redundant predictors, clarifying data alongside intensifying predictive efficiency.

The logic why multiple regression is a functional approach is because it is the best strategy to use in figuring out the relationship between the outcome (Monthly Charge -dependent variable) and multiple predictive variables('Children', 'Age', 'Income', 'Marital', ‘'Gender', 'Churn', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Techie', 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', Tenure', 'Bandwidth\_GB\_Year', 'Timely response', 'Timely fixes', 'Timely replacements', 'Reliability', 'Options', 'Respectful response', 'Courteous exchange', 'Evidence of active listening Independent Variables).

The relevance of every predictive variable to the relations and the effects of several predictors would be statistically modified. Therefore, to predict the value of the monthly charge based on the value of other predictive variables that affect the monthly charge, the multiple regression method appears suitable for analyzing the research question from part 1 and the above variables can be used for multiple regression (Peter Grant, 2019).

**Part III: Data Preparation**

**C.  Summarize the data preparation process for multiple regression analysis by doing the following:**

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

***The submission describes the data preparation goals and the data manipulations that will be used to achieve the goals. The goals and manipulations align with each other and with multiple regression analysis and the research question.***

The goal of data preparation is to create a high quality data for analysis. Data preparation, including understanding the data accessible for analysis is the first step in data preparation. Data on customer churn comprises about 50 fields. A significant problem is deciding which fields can be used for regression analysis. Therefore we were able to narrow down to 33 fields after converting categorical variables numerical variable to enable us conduct a multiple regression.

Variables such as the  Case Order, Customer\_id, Internet Service, Contract, Interaction, UID, City, State, County, Payment Method, Zip, Lat, Lng, Population, Area, Time Zone, Job may not be necessary for analysis (related to customer service interactions).

We will also look for null values and outliers in the data; if any are detected, they must be adequately treated.

**Data Preparation Steps**

**Cleaning the data** - In the stage of data cleaning, we will perform multiple analytical techniques which go a long way in examining the variables, analytically detecting missing values and outliers alongside cleaning the data sets to be error-free using the below functions with the goal of correcting error in the data.

|  |
| --- |
| df.dropna()  df.fillna(df.mean(), inplace=True),  boxplot=sns.boxplot(x='Varchar' (input () ),data=df)  outlierFilter=df['Varchar' (input () )] (less than) int(input () )  df = df[outlierFilter]  boxplot=sns.boxplot(x='Varchar' (input () ),data=df)  df.duplicated() |

**Explore the data** - This step illustrates central tendencies, correlations, and variations to illuminate organizational resolution alongside conducting parametric hypothesis examination using the below functions with the goal of making easier understanding of the data.

|  |
| --- |
| df.head()  df.describe()  df.dtypes  df.info()  df.isna()  boxplot=sns.boxplot(x='Varchar' (input () )  data=df)  df.nunique()  df['Varchar' (input () )].hist()  %matplotlib inline  grouped%matplotlib inline  grouped'Varchar' (input () ).plot.bar()  sns.scatterplot(x='Varchar' (input () ), y='Varchar' (input () ), data=df)  plt.show()  print(list(df.columns))  df.head() |

**Wrangle the data**- The goal of this stage is to improve the way the data is used via preprocessing, cleaning, modifying, and renaming variables in a standardized manipulative approach. The process of multiple regression analysis would assist us in finding answers to the typical problems associated with data processing..

|  |
| --- |
| df.fillna(df.mean(), inplace=True),  grouped'Varchar' (input () ) = df.groupby(by='Varchar' (input () ).size()  grouped'Varchar' (input () )  %matplotlib inline  grouped%matplotlib inline  grouped'Varchar' (input () ).plot.bar()  grouped'Varchar' (input () ) = df.groupby(by='Varchar' (input () ).size()  grouped'Varchar' (input () )  %matplotlib inline  grouped'Varchar' (input () ).plot.bar() |

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

***The submission accurately discusses the summary statistics and discusses the target variable and all predictor variables that need to be gathered from the data set to answer the selected research question.***

Summary statistics must be identified in order for the multiple linear regression to respond to the research question. P-values for the independent variables must be determined. Coefficients for the independent variables must also be determined. This will indicate which independent variables will have an effect on the target variable. The target variable is the monthly charge . The predictor variables are 'Children', 'Age', 'Income', 'Marital', ‘'Gender', 'Churn', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Techie', 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', Tenure', 'Bandwidth\_GB\_Year', 'Timely response', 'Timely fixes', 'Timely replacements', 'Reliability', 'Options', 'Respectful response', 'Courteous exchange', 'Evidence of active listening- Independent Variables. The summary statistics for such numeric variables are shown in the charts below. The table depicts the standard deviations of each numerical variable as well as the dispersion in the interquartile ranges.

Table

Description automatically generated

Table

Description automatically generated

**Table

Description automatically generated**

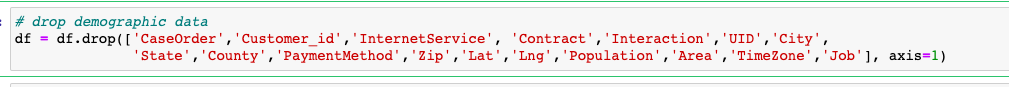
**Table

Description automatically generated**

To enable regression analysis and accurate data comparison, the categorical variables were also transformed into numerical variables.

**Justification for the Exclusion of Several Potential Predictor Variables**

Columns with more unique values will be ignored because they are dimensionality collectors. As a result, removing columns with different values frees up the extra storage space needed to store the data.



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

**The submission explains all the necessary steps to prepare the data for the analysis.**

***The steps include the annotated code and relate to preparing for multiple regression analysis.***

- **Define the Issue**

Any data analysis process's first stage is to define our objective. Data analytics jargon is sometimes called the 'problem statement. Defining our objective means developing a hypothesis and determining how to test it. What business issue are we seeking to resolve to begin with?

- **Collect the Data**

Once we have established our objective, we must create a strategy for collecting and aggregating the appropriate data. An essential part is determining which data we need to import using the python data frame. This could be qualitative (descriptive) data, like customer reviews, or quantitative (numeric) data, like sales figures.

- **Clean the data**

The next step after gathering our data is getting it ready for analysis. This procedure, known as "cleanup," ensures that our data has the finest quality. Essential data cleaning responsibilities include:

Remove null values, NAs, duplicates, and When integrating a variety of data sources using boxplots and df, outliers are a predictable concern. So the data cleaning is relevant (John Sullivan, 2018).

Means of identifying any missing values such as NAs and Null values

df.isna().any()

Means of dealing with missing values

df.dropna()

Means of identifying

# dictating outliers

boxplot=sns.boxplot(x='Income',data=df)

Means of dealing with outliers

# Dropping outliers systematically

outlierFilter=df['Income'] < 65000

df = df[outlierFilter]

# Dropping Rows containing outliers

df.drop\_duplicates(inplace = True)

- **Remove Irrelevant Observations** from our data that have no impact on the analysis we are conducting.

-**Structuring the data** - general fixing typos, will help us map and manipulate our data more efficiently.

- **Filling in gaps** - we might notice that essential data are missing. Once we have identified gaps, we can go about filling them.

- **Conducting an Exploratory Analysis**

Alongside purifying data, many analysts typically conduct exploratory data analysis employing scatter plots. This may enable us to refine our hypotheses when detecting preliminary trends and features. Similarly, let's use the hypothetical learning organization as an instance. through conducting an exploratory data analysis; maybe there is a pattern.

- **Analyzing the Information**

We have finally tidied up our data. The exciting part will now consist of evaluating it. Our goal does indeed have a significant impact on what type of data processing we execute.

- **Reporting Our Observations**

We have our perceptions. Communicating these findings with the clients constitutes the final stage in the data analytics procedure. This requires leveraging visualizations to visualize the results and convey data in a fashion that's digestible for various types of audiences, which can be more complex besides just sharing the story's actual data.

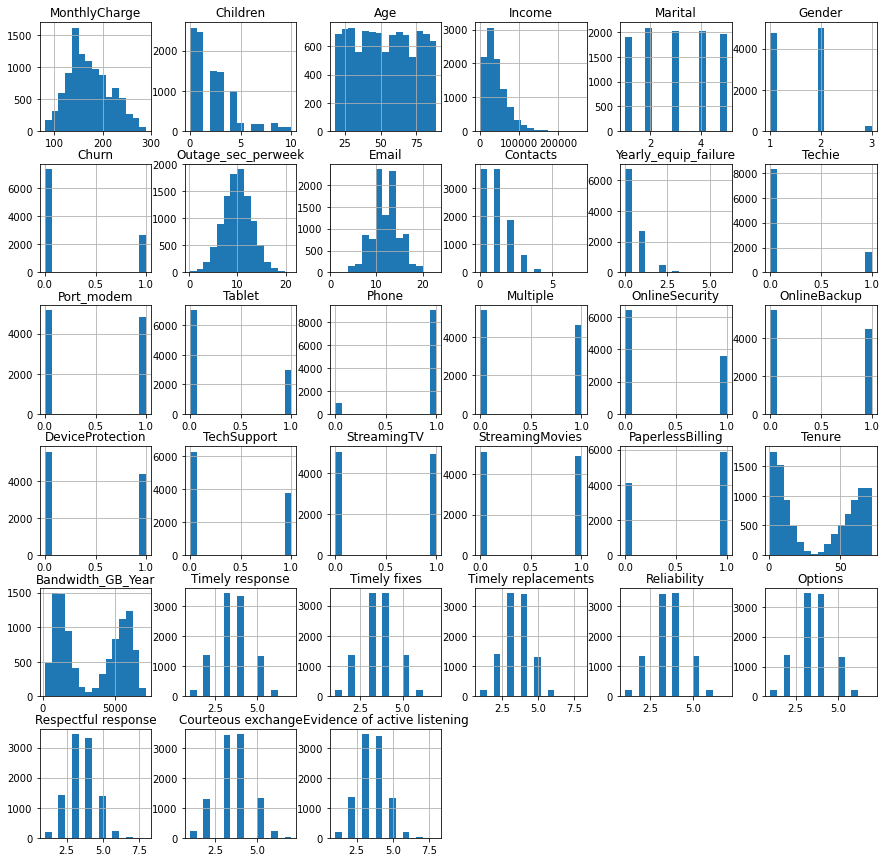
**Line of Codes**

|  |
| --- |
| **import numpy as np**  **import pandas as pd**  **from sklearn import linear\_model**  **import matplotlib.pyplot as plt**  **import seaborn as sns**  **%matplotlib inline**  **pd.set\_option('display.max\_columns', None)**  **import pylab**  **from pylab import rcParams**  **import statsmodels.api as sm**  **import statistics**  **from scipy import stats**  **import sklearn**  **from sklearn import preprocessing**  **from sklearn.linear\_model import LinearRegression**  **from sklearn.model\_selection import train\_test\_split**  **from sklearn import metrics**  **from sklearn.metrics import classification\_report**  **from scipy.stats import chisquare**  **from scipy.stats import chi2\_contingency**  **import os**  **os.getcwd()**  **df = pd.read\_csv("churn\_clean.csv")**  **df.dropna()**  **print(df.shape)**  **print(list(df.columns))**  **df.head()**  **df.rename(columns={'Item1':'Timely response','Item2':'Timely fixes','Item3':'Timely replacements','Item4':'Reliability','Item5':'Options','Item6':'Respectful response','Item7':'Courteous exchange','Item8':'Evidence of active listening'},inplace=True)**  **df.head()**  **df.info()**  **# summary statistics of character column**    **df.describe(include=['object'])**  **# summary statistics of character column**  **df.describe(include='all')**  **df.describe()**  **df.dtypes**  **df.shape**  **df.isna().any()**  **df.dropna()**  **df.fillna(df.mean(), inplace=True)**  **df.isna()**  **df.isna().sum()**  **#Search for columns containing duplicate data**  **df[df.duplicated()]**  **# Verify if there is any duplicate column - searching for False**  **df.columns.duplicated().any()**  **# Verify if there is any duplicate row - searching for False**  **df.duplicated().any()**  **# Some unwanted columns dropped**  **df = df.drop(['CaseOrder','Customer\_id','InternetService', 'Contract','Interaction','UID','City',**  **'State','County','PaymentMethod','Zip','Lat','Lng','Population','Area','TimeZone','Job'], axis=1)**  **# Confirm if columns were dropped**  **df.head()**  **#Transform yes/no to 1/0**  **df = df.replace(to\_replace = ['Yes','No'],value = [1,0])**  **df['Gender'] = df['Gender'].replace(['Male','Female','Nonbinary'],[1,2,3])**  **#Transform Marital to "Married/Not Married", then change to integer 1/0**  **df['Marital'] = df['Marital'].replace(['Married','Divorced','Widowed','Separated','Never Married','Not Married'],[1,2,3,4,5,6])**  **#convert Marital, OnlineSecurity, Contract, Phone, Gender and Churn to integers**  **df['Phone'] = df['Phone'].replace(['Yes','No'],[1,0])**  **df['Gender'] = df['Gender'].replace(['Male','Female'],[1,0])**  **df['Churn'] = df['Churn'].replace(['Yes','No'],[1,0])**  **df['OnlineSecurity'] = df['OnlineSecurity'].replace(['Yes','No'],[1,0])**  **df['Marital'] = df['Marital'].replace(['Married','Not Married'],[1,0])**  **df.info()**  **df.describe()**  **my\_list = df.columns.values.tolist()**  **print(my\_list)**  **#Transfer the target variable to the front of columns**  **df=df[['MonthlyCharge', 'Children', 'Age', 'Income', 'Marital', 'Gender', 'Churn', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Techie', 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', 'Tenure', 'Bandwidth\_GB\_Year', 'Timely response', 'Timely fixes', 'Timely replacements', 'Reliability', 'Options', 'Respectful response', 'Courteous exchange', 'Evidence of active listening']]**  **#Lets confirm if the target variable was moved**  **my\_list = df.columns.values.tolist()**  **print(my\_list)**  **#Extract the prepared dataset**  **df.to\_csv('churn\_prepared.csv', index = False)** |

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

***The submission accurately generates both univariate and bivariate visualizations of the distributions of variables in the cleaned data set. The bivariate visualizations include the target variable.***

Univariate statistics entails a single dependent variable and can include one or more independent variables. Therefore, histograms and box plots are some of the most commonly used univariate statistics methods to represent the data visually.

. 

|  |
| --- |
| boxplot=sns.boxplot(x=' **MonthlyCharge**',data=df)  Chart  Description automatically generated  boxplot=sns.boxplot(x='Income',data=df)    # Dropping outliers systematically  outlierFilter=df['Income'] < 65000  df = df[outlierFilter]  boxplot=sns.boxplot(x='Income',data=df)    boxplot=sns.boxplot(x='Age',data=df)  Chart  Description automatically generated |

|  |
| --- |
|  |

To ascertain a vivid understanding of the variables, we are systematically applying scatterplot and heatmap for bivariate analysis to unveil the relationship between two continuous variables(Monthly Charge and compared to the predictors), which would be measured on the ratio scales and intervals scales.

|  |
| --- |
|  |

**5.  Provide a copy of the prepared data set.**

***The submission provides a copy of the fully prepared data set.***

|  |
| --- |
| ***#exported prepared dataset has been uploaded as csv file.***  *df.to\_csv('churn\_prepared.csv', index = False)* |

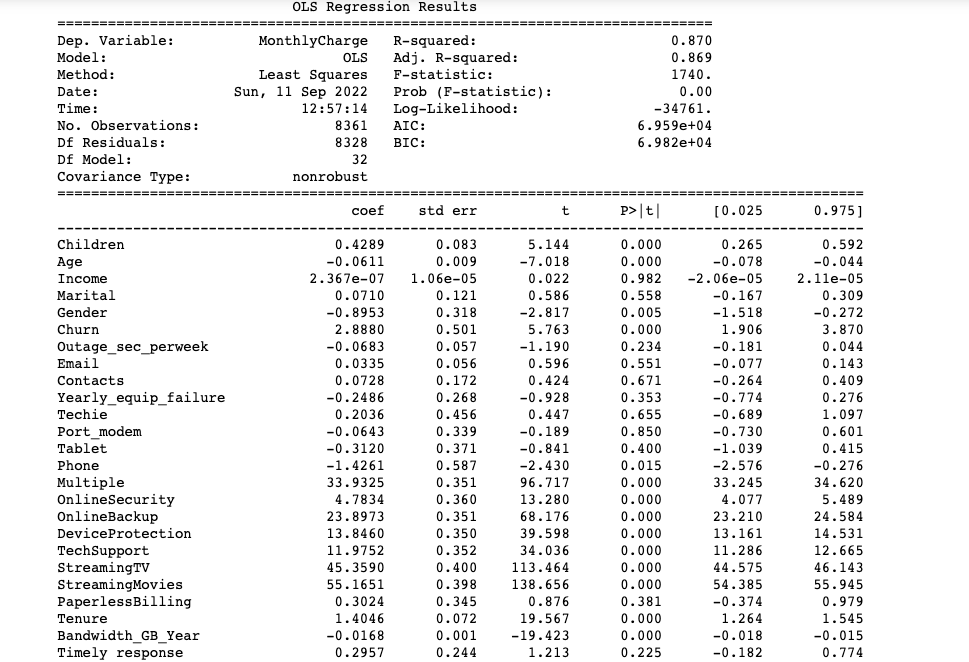
**Part IV: Model Comparison and Analysis**

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

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

***The submission provides an accurate initial multiple regression model from all predictors identified in Part C2.***

Potential predictor variables will be subjected to an initial regression. These are measured against the Initial days objective. The results of the OLS Regression are displayed below.



Table

Description automatically generated

We removed 7 variables in the first model().

Text

Description automatically generated

It appears that 87% of the variation could be represented by the first model, which has an R-squared value of (0.870). Since there are many conditions, there may not be a need for all of the variables because of severe multicollinearity. It is possible to identify areas of potential multicollinearity and begin to select the variables to include in the reduced model by using a heatmap and correlation matrix.

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

***The submission justifies a statistically based variable selection procedure and a model evaluation metric to reduce the initial model. The justification is in alignment with the research question.***

We selected the below variables that were used to conduct the heatmap for bivariate analysis.

|  |
| --- |
| Graphical user interface, table  Description automatically generated |

Graphical user interface, application, table, Excel

Description automatically generated

Table

Description automatically generated

**Table, Excel

Description automatically generated**

**Table

Description automatically generated**

**Table, Excel

Description automatically generated**

**A picture containing graphical user interface

Description automatically generated**

The monthly charge and the predictors are correlated using this heat map. This heat map shows us that a number of variables can be eliminated. The correlation matrix and heat map aid in identifying factors that may not make excellent predictors. We reduced the amount of variables as follows:

**Text

Description automatically generated**

**Graphical user interface, Teams

Description automatically generated**

It appears there is a weak linear relationship between monthly charge and evidence of active listening

**3.  Provide a reduced multiple regression model that includes both categorical and continuous variables.**

**Note: The output should include a screenshot of each model.**

***The submission provides a reduced multiple regression model that includes both categorical and continuous variables, and the reduced model is in alignment with the justification from part D2.***

Using the variables that were discovered earlier, a reduced multiple regression model can be used. According to the correlation matrix and heat map, the reduced OLS Regression Results are shown below:

**Table

Description automatically generated**

As you can see, the reduced model still accounts for 95.5% of the variance.

**E.  Analyze the data set using our reduced multiple regression model by doing the following:**

**1.  Explain our data analysis process by comparing the initial and reduced multiple regression models, including the following elements:**

***The submission accurately explains the data analysis process by comparing the initial and reduced multiple regression models, including all of the given elements.***

The correlation matrix's findings and a heatmap's mapping of the variables served as the basis for the variable selection technique. By doing so, the factors with the strongest correlations to the Initial days variable were found. The regression results with the model equation and analysis, including the R-squared values, are shown in the table above. This information is the model evaluation measure. Here is a picture of the model's residual plot.

Shape

Description automatically generated

**2.  Provide the output and any calculations of the analysis we performed, including the model’s residual error.**

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

***The submission provides the accurate output and calculations of the analysis performed, including the model’s residual error. The submissions includes all necessary output and calculations.***

The data and representations above depict the outcomes of the computations in addition to the residual error of the model.

Graphical user interface, table

Description automatically generated

In [34]:

Graphical user interface, table, Excel

Description automatically generated

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

***The submission provides the code used to support the implementation of the multiple regression models, and the code is complete and accurate****.*

**import numpy as np**

**import pandas as pd**

**from sklearn import linear\_model**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**%matplotlib inline**

**pd.set\_option('display.max\_columns', None)**

**import pylab**

**from pylab import rcParams**

**import statsmodels.api as sm**

**import statistics**

**from scipy import stats**

**import sklearn**

**from sklearn import preprocessing**

**from sklearn.linear\_model import LinearRegression**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn import metrics**

**from sklearn.metrics import classification\_report**

**from scipy.stats import chisquare**

**from scipy.stats import chi2\_contingency**

**import os**

**os.getcwd()**

**df = pd.read\_csv("churn\_clean.csv")**

**df.dropna()**

**print(df.shape)**

**print(list(df.columns))**

**df.head()**

**df.rename(columns={'Item1':'Timely response','Item2':'Timely fixes','Item3':'Timely replacements','Item4':'Reliability','Item5':'Options','Item6':'Respectful response','Item7':'Courteous exchange','Item8':'Evidence of active listening'},inplace=True)**

**df.head()**

**df.info()**

**# summary statistics of character column**

**df.describe(include=['object'])**

**# summary statistics of character column**

**df.describe(include='all')**

**df.describe()**

**df.dtypes**

**df.shape**

**df.isna().any()**

**df.dropna()**

**df.fillna(df.mean(), inplace=True)**

**df.isna()**

**df.isna().sum()**

**#Search for columns containing duplicate data**

**df[df.duplicated()]**

**# Verify if there is any duplicate column - searching for False**

**df.columns.duplicated().any()**

**# Verify if there is any duplicate row - searching for False**

**df.duplicated().any()**

**# Some unwanted columns dropped**

**df = df.drop(['CaseOrder','Customer\_id','InternetService', 'Contract','Interaction','UID','City',**

**'State','County','PaymentMethod','Zip','Lat','Lng','Population','Area','TimeZone','Job'], axis=1)**

**# Confirm if columns were dropped**

**df.head()**

**#Transform yes/no to 1/0**

**df = df.replace(to\_replace = ['Yes','No'],value = [1,0])**

**df['Gender'] = df['Gender'].replace(['Male','Female','Nonbinary'],[1,2,3])**

**#Transform Marital to "Married/Not Married", then change to integer 1/0**

**df['Marital'] = df['Marital'].replace(['Married','Divorced','Widowed','Separated','Never Married','Not Married'],[1,2,3,4,5,6])**

**#convert Marital, OnlineSecurity, Contract, Phone, Gender and Churn to integers**

**df['Phone'] = df['Phone'].replace(['Yes','No'],[1,0])**

**df['Gender'] = df['Gender'].replace(['Male','Female'],[1,0])**

**df['Churn'] = df['Churn'].replace(['Yes','No'],[1,0])**

**df['OnlineSecurity'] = df['OnlineSecurity'].replace(['Yes','No'],[1,0])**

**df['Marital'] = df['Marital'].replace(['Married','Not Married'],[1,0])**

**df.info()**

**df.describe()**

**my\_list = df.columns.values.tolist()**

**print(my\_list)**

**#Transfer the target variable to the front of columns**

**df=df[['MonthlyCharge', 'Children', 'Age', 'Income', 'Marital', 'Gender', 'Churn', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Techie', 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', 'Tenure', 'Bandwidth\_GB\_Year', 'Timely response', 'Timely fixes', 'Timely replacements', 'Reliability', 'Options', 'Respectful response', 'Courteous exchange', 'Evidence of active listening']]**

**#Lets confirm if the target variable was moved**

**my\_list = df.columns.values.tolist()**

**print(my\_list)**

**#Extract the prepared dataset**

**df.to\_csv('churn\_prepared.csv', index = False)**

|  |
| --- |
| **df['intercept'] = 1**  **lm\_MonthlyCharge = sm.OLS(df['MonthlyCharge'],df[['Children', 'Age', 'Income', 'Marital', 'Gender',**  **'Churn', 'Outage\_sec\_perweek', 'Email', 'Contacts',**  **'Yearly\_equip\_failure', 'Techie', 'Port\_modem', 'Tablet', 'Phone',**  **'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',**  **'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling',**  **'Tenure', 'Bandwidth\_GB\_Year', 'Timely response', 'Timely fixes',**  **'Timely replacements', 'Reliability', 'Options', 'Respectful response',**  **'Courteous exchange', 'Evidence of active listening','Intercept',]]).fit()**  **print(lm\_MonthlyCharge.summary())**  **#heatmap and correlatin matrix dataframe creation**  **Churn\_heatmap = df[['MonthlyCharge', 'Children', 'Age', 'Income', 'Marital', 'Gender', 'Churn', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Techie', 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', 'Tenure', 'Bandwidth\_GB\_Year', 'Timely response', 'Timely fixes', 'Timely replacements', 'Reliability', 'Options', 'Respectful response', 'Courteous exchange', 'Evidence of active listening']]**  **#Initial model heatmap**  **sns.heatmap(Churn\_heatmap.corr(), annot=False)**  **plt.show**  **Churn\_heatmap.corr()**  **#Narrowing the initial model, removing multiple variables**  **Churn\_heatmap = df[['MonthlyCharge','Timely response',**  **'Respectful response', 'Tablet','Yearly\_equip\_failure',**  **'Courteous exchange','Evidence of active listening']]**  **#Reduced Initial model heatmap**  **sns.heatmap(Churn\_heatmap.corr(), annot=True)**  **plt.show**  **#Reduced multiple regression model**  **df['intercept'] = 1**  **lm\_MonthlyCharge\_reduced = sm.OLS(df['MonthlyCharge'],df[['Timely response',**  **'Respectful response', 'Tablet','Yearly\_equip\_failure',**  **'Courteous exchange','Evidence of active listening']]).fit()**  **print(lm\_MonthlyCharge\_reduced.summary())**  **#load cleansed data for residual plot**  **Chur\_df = pd.read\_csv ('churn\_prepared.csv')**  **#Create residual plot**  **Chur\_df['intercept'] = 1**  **residuals = Chur\_df['MonthlyCharge']**  **lm\_MonthlyCharge\_reduced.predict(Chur\_df[['Timely response',**  **'Respectful response', 'Tablet','Yearly\_equip\_failure',**  **'Courteous exchange','Evidence of active listening']])**  **sns.scatterplot(x=Chur\_df['Evidence of active listening'],y=residuals,color='blue')**  **plt.show();**  **#import necessary libraries**  **import matplotlib.pyplot as plt**  **import statsmodels.api as sm**  **from statsmodels.formula.api import ols**  **#create residual vs. predictor plot for 'assists'**  **fig = plt.figure(figsize=(12,8))**  **fig = sm.graphics.plot\_regress\_exog(lm\_MonthlyCharge\_reduced, 'Evidence of active listening', fig=fig)**  **#create residual vs. predictor plot for 'Timely response'**  **fig = plt.figure(figsize=(12,8))**  **fig = sm.graphics.plot\_regress\_exog(lm\_MonthlyCharge\_reduced, 'Timely response', fig=fig)** |

**Part V: Data Summary and Implications**

**F.  Summarize our findings and assumptions by doing the following:**

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

***The submission accurately discusses the results of the data analysis, and the discussion addresses all of the given elements and is in alignment with the research question and the data analysis***

**A Regression Equation for the Reduced Model**

The concluded multiple linear regression equation for the reduced model is as follows:

Ŷ=96.3362+11.0046(Timely response)+ 9.8160( Respectful response)+ 8.7507(Tablet)+ 6.6654(Yearly\_equip\_failure)+11.4215(Courteous exchange)+13.7240 (Evidence of active listening)

• **Explanations of the Model's Statistically Meaningful Variables' Coefficients**

Following the coefficients of the statistical variables relevant in the modified model with correlation with the monthly charge, for every one unit increase in:

* Timely response – the monthly charge will be increased by 11.0046
* Respectful response – monthly charge increase by 9.816
* Tablet – monthly charge increase by 8.7507
* Yearly\_equip\_failure - monthly charge increase by 6.6654
* Courteous exchange - monthly charge increase by 11.4215
* Evidence of active listening - monthly charge increase by 13.7240

However, the p values of the above variables appears to be at 0.000 which is statistically significant.

• **The Statistical And Practical Significance of the Model**

The retention of a subscriber and indications of active listening are strongly correlated with the monthly fee. It is challenging to foresee whether a customer would leave. There is a good chance that the device usage and monthly fee will go up if a customer's year-end equipment fails.

• **The Limitations of the Data Analysis**

With a p-value of.000, or statistically significant, the six particular variables are the focus of this analysis. In this investigation, additional factors that might have influenced the findings were not taken into account. The findings may vary based on the method used because only one multiple regression model was performed.

**2.  Recommend a course of action based on our results.**

***The submission recommends an appropriate course of action based on the results as they relate to the research question.***

Conclusively, The possibility of a customer not receiving a prompt response or active listening, both of which are reliable indicators of the monthly fee, should be reduced by telecommunications firms. To ascertain whether other variables contribute to customer churn, a second multiple regression on those other variables should be run.

**Part VI: Demonstration**

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

•  a demonstration of the functionality of the code used for the analysis

•  an identification of the version of the programming environment

•  an interpretation of the coefficients.

***A Panopto video recording is provided that includes all of the given elements. For the duration of the presentation, the video captures both the presenter and the functioning code in a Panopto video recording.***

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0c433793-e36c-4b7d-bb7c-af0d01340eb3>

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

***The submission lists all web sources used to acquire data or segments of third-party code, and the web sources are reliable***

*Grant, P. (2019). Understanding Multiple Regression; The fundamental basis behind this commonly used algorithm.*

Medium***.***[***https://towardsdatascience.com/understanding-multiple-regression-249b16bde83e***](https://towardsdatascience.com/understanding-multiple-regression-249b16bde83e)

Deepanshu, B. (2020). *How to Import Data in Python.*

RSGB Business Consultant Pvt. Ltd. <https://www.listendata.com/2017/02/import-data-in-python.html#Import-CSV-files>

Pierre-Louis B. (2020). *Principle Components Analysis(PCA), Fundamentals, Benefits & Insights for Industry.*

Medium. <https://towardsdatascience.com/principal-components-analysis-pca-fundamentals-benefits-insights-for-industry-2f03ad18c4d7>

John S. (2018). *Data Cleaning with python and Pandas: Detecting Missing Values.*

Medium. <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

Angelica Lo D. (2021). *How to detect outliers with Python.*

Medium. <https://towardsdatascience.com/how-to-detect-outliers-with-python-pyod-aa7147359e4b>

Michael G. (2018). *Understanding Boxplots.*

Medium. <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

John S. (2018). *Data Cleaning with python and Pandas: Detecting Missing Values.*

Medium. <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

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

***The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available*.**

**Reference**

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. ISBN-13: 978-1-119-52684-1.