

|  |
| --- |
| PERFORMANCE ASSESSMENT OF  PREDICTIVE MODELING  TASK1  D208  BY KOFFI M. GANU |
| MSDA MAY 2022 AT WGU  INSTRUCTOR: DANIEL SMITH**h**  Phone  Email |

**Introduction**

People often say in business: “It takes a month to find a customer, but a second to lose one.” The primary goal of any company is to maintain its customers as long as possible. It is with this in mind that, especially with the advent of the use of data, these companies hire data specialists (Data analysts, Data scientists) for strategies and recommendations for the improvement of their company’s services. In the rest of our project, we will, on the one hand, do data analysis and try to highlight how different variables influence the churn of customers and, on the other hand, build a model to predict which customers will disconnect service.

**PART I/ RESEARCH QUESTION.**

A1-

The data set provided for our project has 50 variables. According to the data dictionary, tenure is the number of months the customer has stayed with the service provider, and it is a numerical variable. To predict the longevity of customers with their service provider, we will ask this question: our research question. Which variables most likely influence customer tenure?

A2-

In general, data analysis is the process of evaluating data using analytical or statistical tools to discover useful pieces of information. These information are used to make decisions or provide recommendations for business improvement. For our project especially, the main goal of our analysis is to find out which variables contribute to customer churn and how these variables impact the tenure of the customers. So, if we have the answer to these questions, we can eventually provide some recommendations to reduce customer churn.

**PART II/ METHOD JUSTIFICATION**

B1-

The mathematics behind regression makes certain assumptions, and these assumptions must be met satisfactorily before it is possible to draw any conclusions about the population based on the sample used for the regression. That is, the assumptions must be met in order to generate unbiased estimates of the coefficients such that on average, the coefficients derived from the sample will be the same as those that would be derived had the entire population been included in the analysis. Specifically, a multiple regression analysis should satisfy the following assumptions:

1. Continuous predictor variables.
2. Independence of residuals
3. Linearity
4. Homoscedasticity.
5. No correlation with variables not included in the regression model. There should be no other external variables that correlate highly with any of the predictors.
6. Normal distribution of errors.
7. No multicollinearity.
8. No significant outliers.

B2-

To analyze our data and provide a response to the asked research question, Python will be used. Python is an intuitive, simple-to-use programming language:  
• It offers versatile programming style and syntax  
• It provides mature packages for data science and machine learning.  
• Since Python is cross-platform, it will work well whether consumers of the analysis are using Windows PCs or a MacBook laptops.  
• It is fast when compared with other possible programming languages like R   
• There is strong support for Python as the most popular data science programming language in popular

literature and media.

**B3-**

Multiple linear regression estimates the relationship between two or more independent variables and one dependent variable. Therefore, multiple linear regression is appropriate to determine which variables contribute to customer churn. Indeed, one or many reasons may lead to customer churn.

**PART III/ DATA PREPARATION**

C1

Data preparation has several steps:  
step 1. Read the data set into Python using Pandas’ read\_csv command.  
step 2. Evaluate the data structure to better understand input data.  
step 3. Naming the dataset as the variable df.  
step 4. Examine potential misspellings, awkward variable naming & missing data.  
Step 5. Find outliers that may create or hide statistical significance using histograms.  
  
step 6. Imputing records of missing data with meaningful measures of central tendency (mean, median, or mode) or removing outliers that are several standard deviations above the mean.

C2

The data set provided for our project has 10000 records for 50 variables. For purposes of this analysis, some variables were removed from the data set such as CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital, and PaymentMethod. Also, binomial Yes/No or Male/Female, variables were encoded. This resulted in 34 remaining variables, including the target variable. The dataset appeared to be sufficiently cleaned, leaving no null, NAs, or missing data points.  
Measures of central tendency through histograms and boxplots revealed normal distributions for Monthly Charge, Outage\_sec\_perweek, and Email. The cleaned dataset no longer retained any outliers. Histograms for Bandwidth\_GB\_Year & Tenure displayed bimodal distributions, which demonstrated a direct linear relationship with each other in a scatterplot. The average customer was 53 years old with a standard deviation of 20 years. In addition, the average customer had two children with a standard deviation of 2 kids. The average income is 39,806, with a standard deviation of about 30,000. Furthermore, the average customer experienced ten outage seconds/week, contacted technical support less than once, and had less than one yearly equipment failure. Finally, the average customer has been with the company for 34.5 months, with a monthly charge of approximately 173 & uses 3,392 GBs/year.

C3

• Import the dataset to our directory.

  
• Rename columns of a survey to easily recognizable features, for example, Item1 to TimelyResponse).

Text

Description automatically generated

• Get information on the data set, structure (columns & rows) & data types.



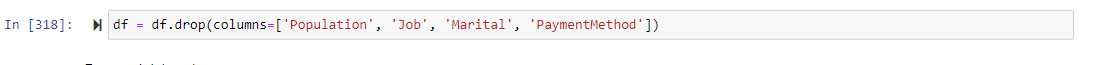
• View summary statistics.

Graphical user interface, text, application

Description automatically generated

• Drop some columns we have seen as not necessary for our project.





• Check for records with missing data & impute missing data with meaningful measures of central tendency (mean, median or mode)



Check and remove outliers that are several standard deviations above the mean.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated



C4

* Univariate visualization

Numerical variable

Diagram, schematic

Description automatically generated

Categorical variable

* Churn distribution

Chart, bar chart

Description automatically generated

* Techie distribution

Chart, bar chart

Description automatically generated

* Contract distribution

Chart, bar chart

Description automatically generated

* Marital distribution

Chart, bar chart

Description automatically generated

* Bivariate visualization

Income vs Tenure

Chart, scatter chart

Description automatically generated

Monthly Charge vs Tenure

Chart, scatter chart

Description automatically generated

Bandwidth\_GB\_Year vs Tenure

Chart, scatter chart

Description automatically generated

Tenure vs Marital

Chart, bar chart, box and whisker chart

Description automatically generated

Tenure vs Gender

Chart, box and whisker chart

Description automatically generated

Tenure vs Churn

Chart, box and whisker chart

Description automatically generated

Tenure vs Contract

Chart, box and whisker chart

Description automatically generated

Tenure vs Internet service

Chart, box and whisker chart

Description automatically generated

C5

A copy of the prepared data set

Table

Description automatically generated

Table

Description automatically generated

**PART IV/ MODEL COMPARISON AND ANALYSIS**

D1.

Initial Multiple Regression Model from all predictors that were identified in PART C2

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

D2

The R-squared and Adjusted R squared of our model are 1.00, indicating that it can explain 100% of the variance in the tenure. This is a clear indication that we have been able to create a very good model which is not underfitting the data. RMSE and MAE of train and test data are very close, which indicates that our model is not overfitting the train data. MAE indicates that our current model is able to predict Tenure within a mean error of 5.37 months on the test data set. The p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), any variable with a p-value less than 0.05 would be considered significant. Because the predictor variables which show a p-value greater than 0.05 do not significantly impact the target variable. Therefore, we will drop them. We will use this code to Select feature variables by dropping variables with a high p-value.

Graphical user interface, text

Description automatically generated

This is the output of this code.



The predictor variables dropped from 34 to 15. We will create the reduced logistic regression model with these variables selected.

D3

Reduced Multiple Regression Model from all selected predictors.

Text

Description automatically generated

Table

Description automatically generated

E1.

Our project objective is to build a logistic model to predict customer churn. Churn, according to the data dictionary, is defined as the percentage of customers who stopped using a provider’s product or service during a specific time frame. The data set provided has 10000 records and 50 columns or variables. After the preparation of the data set to resolve our problem, which is to build a multiple regression model, the prepared data set has only 34 variables. First, we will consider all 34 variables, and we will build the first model. After the result evaluation, we dropped the features that did not make an impact. The most important for this evaluation is the p-value noted as p > IzI. We will Focus on the p-values that are 0.05 and below. This value reduces the variable and trims the number of the variable from 34 to 15. Using these features reduces the complexity of the regression and provides better results. All parameters of the regression’s predicting capability were improved, including RMSE, MAE, and MAPE. The model evaluation is based on its predicting capability. These parameters, as noted above, are RMSE, MAE, and MAPE.

E2.

* Initial Multiple Regression Model training performance.

Text

Description automatically generated with low confidence

* Initial Multiple Regression Model testing performance

Text

Description automatically generated with low confidence

* Reduced Multiple Regression Model training performance.

A picture containing text

Description automatically generated

* Reduced Multiple Regression Model testing performance.

A picture containing table

Description automatically generated

Dropping the high p-value predictor variables has not adversely affected the model performance.

This shows that these variables do not significantly impact the target variables.

 We will test for linearity and independence by making a plot of fitted values vs residuals and checking for patterns. If there is no pattern, we say the model is linear, and residuals are independent. Otherwise, the model shows signs of non-linearity, and residuals are not independent.

* Actual, Fitted, and Residual values.

Table

Description automatically generated

* Fitted values vs residual.

Chart

Description automatically generated

We see two patterns in the plot above. So, the assumptions of linearity and independence are not satisfied.

* We will test for normality by checking the distribution of residuals, by checking the Q-Q plot of residuals, and by using the Shapiro-Wilk test. If the residuals follow a normal distribution, they will make a straight-line plot, otherwise not. If the p-value of the Shapiro-Wilk test is greater than 0.05, we can say the residuals are normally distributed.





Since p-value > 0.05, the residuals are homoscedastic.

E3

* Initial Multiple Regression Model

Graphical user interface, application

Description automatically generated with medium confidence

* Reduced Multiple Regression

Graphical user interface, application

Description automatically generated with medium confidence

**PART V/ DATA SUMMARY AND IMPLICATIONS**

F1-

The final model is expressed as follows:  
y = - 3.8023 + (-0.4011) \* Children + (0.0400) \* Age + (-0.0353) \* MonthlyCharge + ( 0.0122) \* Bandwidth\_GB\_Year + (-0.7941) \* Gender\_Male + (0.2726) \* Gender\_Nonbinary + ( 5.7504) \* InternetService\_FiberOptic +( 4.5975) \* InternetService\_None + (0.2712) \* Multiple\_Yes + (-0.8318) \* OnlineSecurity\_Yes + (- 0.3471) \* OnlineBackup\_Yes + (-0.5957) \* Device\_Protection\_Yes + (0.3843) \* TechSupport + (-1.2997) \* StreamingTV\_Yes + (-0.7147) \* StreamingMovies\_Yes

Interpretation of coefficients:

The coefficient of Children, MonthlyCharge, GenderMale, OnlineSecurity\_yes, OnlineBackup\_yes, StreamingTV\_Yes, StreamingMovies\_Yes are negative. So, we can say that these variables will not lead to tenure.

Coefficient of Age, Bandwidth\_GB\_Year, Gender\_Nonbinary, , InternetService\_Fiber Optic, InternetService\_None, TechSupport\_Yes, Multiple\_Yes, Device\_Protection\_Yes are positive. So, these variables will lead to tenure.

Limitation:

Our Multiple Regression Model is built on a sample of 10000 customer records. So, to confirm this result, we will need more data because the sample is not the entire population.

F2- Recommendations:

* After analysis of our multiple regression created, we have seen that some products line for the telecommunication companies, such as online security, online backup, streaming TV, and streaming movies, have a negative impact on tenure. On the other hand, some products line, such as internet services, multiple services, and device protection, have a positive impact on tenure. In addition, we have seen through our analysis that customers who get technical support don’t leave their telecommunication companies. Therefore, the model can be used for predictive purposes as it can predict tenure within 5%.

We recommend that telecommunication companies provide good technical support and device protection to keep their customers for as long as.

**PART VI/ DEMONSTRATION.**

G- PANOPTO VIDEO

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=96cb6e01-a776-41a8-9491-afaa011fdafe>

H-

Data camp course

D208 Predictive modeling WGU textbook

D208 Predictive modeling WGU course webinar

I-

No in-text references were used.

J-

Professional communication in the presentation of the submission was used.

.