

|  |
| --- |
| PERFORMANCE ASSESSMENT OF  DATA MINING I  TASK2  D209  BY KOFFI M. GANU |
| MSDA MAY 2022 AT WGU  INSTRUCTOR: ERIC STRAW  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 using a decision tree algorithm to predict which customers will disconnect service.

**PART I/**

A1-

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 valuable pieces of information. This information is used to make decisions or provide recommendations for business improvement**.** For our project, the main goal of our data analysis is to develop a predictive model using the decision tree algorithm that accurately predicts the longevity of the customers with their telecommunication companies. This model can be used to identify variables that contribute to customer churn and develop strategies to reduce it and improve the business.

**PART II/**

B1

A decision tree is a supervised machine-learning algorithm that can be used for a regression or classification task. For our project, we will focus on regression tasks. Regarding how a decision tree analyzes the data set, we will say, first, the decision tree algorithm selects the variables that provide the most information gain and creates a decision node based on that variable. We mention that the information gain is a metric which measures the quality of the split. In addition, the algorithm creates a branch that leads to a child node for each possible value of that variable. Similarly, the same procedure repeats until a stopping criterion is met. We would like to mention that criterion is a measure used to determine the most effective split at each node of the tree. The outcome of our analysis is the model we have created to predict tenure for new instances based on the input features. We would like to mention that tenure is the number of months the customer has stayed with the provider. Furthermore, the decision tree algorithm also provides insights into the importance of different features in predicting tenure. The importance of each feature can be calculated based on how often it is used in the splits and how much it contributes to the reduction of variance in tenure.

B2-

The decision tree assumes that each node in the tree has only two branches called splits. This assumption is known as binary split because the split divides the data into two groups. In that assumption previously cited, we can add another one, which is recursive binary splitting which means that the decision tree recursively split the data into smaller groups until a stopping criterion is met.

B3-

For analyzing the data set provided using a decision tree algorithm in Python, we can use the following libraries:

* Pandas: Pandas provide a function to create a DataFrame by reading data from various files type. In addition, pandas help to manipulate and preprocess the data set provided by handling the missing value and converting categorical variables to numerical form.
* Numpy: This library can perform various mathematical operations.
* Matplotlib: This is a data visualization library that can be used to visualize the data and the results of the analysis.
* Scikit-learn: This is a tool for data analysis and modeling. It includes a DecisionTreeRegressor that can be used to build a decision tree model for the provided data set, train\_test\_split to split the data set into a train data set, and test data set, and mean\_squared\_error to get the model performance.

PART III/

C1

One data preprocessing goal relevant to the decision tree is encoding categorical variables and imputing missing values. Indeed, before putting our data through models, two steps that need to be performed on categorical data are encoding and dealing with missing nulls.

* Encoding is converting text or Boolean values to numerical values for processing. It is important to convert categorical data into numerical because the decision tree algorithm requires numerical data as input. We will use a one-hot encoding method to perform this task. As a result, all categorical variables will convert into a binary feature and take a value of 0 or 1.
* For the missing data, three ways were taught to handle null values in a data set. The first was to leave them in a case where the data was categorical and could be treated as a ‘missing’ or ‘NaN’ category. The second was to remove the data, either by row or column. Removing data is a slippery slope in which you do not want to remove too much data from your data set. If the feature with the missing values is irrelevant or correlates highly to another feature, then it would be acceptable to remove that column. Rows, on the other hand, are on a case-by-case basis. The third is to impute or replace with a placeholder value. As an example, we can mention mean imputation, which involves replacing the missing values with the mean of the variable. Our data set does not have any missing values. But, in general, it is important to treat the missing values because that helps to improve the accuracy and the reliability of our analysis through the model we have created.

C2

* The list of initial variables:

Text, letter

Description automatically generated

* List of the categorical variable:

Text

Description automatically generated

* List of the continuous variable:

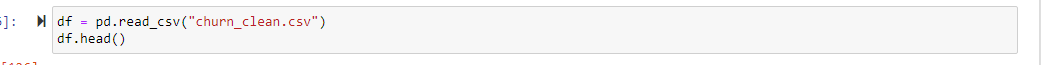
Text

Description automatically generated

C3

The following are the steps to prepare the data set provided for analysis.

* Import the necessary libraries: The necessary libraries for our analysis are Pandas, Numpy, Matplotlib, and Scikit-learn.
* Load the data set: We load the data set into our working directory using the read\_csv( ) fuction.



* Rename columns of a survey to easily recognizable features.

Text

Description automatically generated

* Get information on the data set, structure, and data types.



* Drop irrelevant columns: We can drop some columns we have seen as not necessary for our analysis.



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

Graphical user interface, text

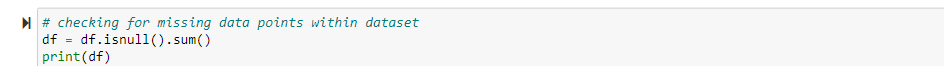
Description automatically generated

Graphical user interface, text, application, email

Description automatically generated



* Check for records with missing data and impute missing data with meaningful measures of central tendency.



C4

A copy of the prepared data set:

We will extract the prepared data set with this python code.



The data set will be attached and named “df\_clean\_209”.

D1

At this point in our project, we will split the data into training and testing sets using the train\_test\_split ( ) function from Scikit-learn. The Python code used for this purpose is:



X\_train and X\_test files will be attached and named respectively “X\_train\_209 csv” and “X\_test\_209 csv”.

D2

A decision tree is a supervised machine-learning algorithm that can be used for both classification and regression analysis. It has a hierarchical tree structure consisting of a root node, branches, internals nodes, and a terminal node called a leaf node. The decision tree starts with a root node which we can consider as the entire dataset, and at each step, it selects the variables that best split the data into the internal nodes. Based on the available variables, both nodes conduct evaluations to form a homogenous subset. The splitting process continues until a stopping criterion is met. Once the decision tree is built, it can be used to make a prediction for new input data. In our analysis using the decision tree algorithm method, we will split the data into training and test sets using train\_test\_split ( ) function from Scikit-learn. Then, we will use the Scikit-learn library to create a DecisionTreeRegressor object and fit it to the training data. Finally, we will use the trained regressor to make predictions on the test data and evaluate its performance using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (r2) score.

Here is the model performance:

Text

Description automatically generated

D3

* Split the data set into a training set and a testing set.



* Train a DecisionTreeRegressor.

Graphical user interface, text, application, chat or text message

Description automatically generated

* Make predictions on the train data.



* Validate predictions on the test data.



E1

Mean Squared Error (MSE) is a metric used to evaluate the performance of a regression model. For our model created by the decision tree algorithm, MSE measures the average squared difference between the predicted and actual target values.

E2

The MSE score for our model is 3.52. This score tells us that our model indicates better performance. However, MSE has a squared unit which can make it difficult to interpret. Therefore we will consider other metrics, such as Mean Absolute Error (MAE) or Root Mean Squared (RMSE).

E3

The decision tree is less effective in making predictions when the main goal is to predict the outcome of a continuous variable. Furthermore, the decision tree is unstable. Indeed, a small change in the data can result in a significant change in the structure of the decision tree.

E4

After analysis of our model created by the decision tree, it can be concluded that customers with a high risk of churning have short-term contracts and fair technical support. Therefore, we recommend that telecommunication companies sign at least a one-year contract with their new customers and provide good technical support to keep their customers for as long as.

**PART VI/ DEMONSTRATION.**

G- PANOPTO VIDEO

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=eb86343b-d95c-43b0-bb69-afd9011c8ec5

H-

Data camp course

D209 Predictive modeling WGU textbook

D209 Predictive modeling WGU course webinar

I-

No in-text references were used.

J-

We used professional communication in the presentation of the submission.