**CLASSIFICATION ANALYSIS (NVM3 TASK 2)– D209**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

***Part I: Research Question***

**A-1**

**“Can a decision tree model predict what customers will discontinue service withing the last month?”**

**A-2**

**The main goal of this decision tree model is to help identify customers that may discontinue service.**

***Part II: Method Justification***

**B-1**

**“Decision trees are preferred for many applications, mainly due to their high explainability, but also due to the fact that they are relatively simple to set up and train, and the short time it takes to perform a prediction with a decision tree (Almog, 2023).” The decision tree starts with the root at the top and makes comparisons of the features that form branches depending on the on the decision that is made at each feature. This continues until it reaches the prediction which is called a leaf node. The expected outcomes on the decision tree based on my A-1 question will be a prediction of yes, this customer is a churn customer or no, this customer is not a churn customer. The yes responses will show 1 and the no responses will show 0.**

**B-2**

**One of the few assumptions that is made with decision trees is that the root is the whole training data. This is the starting point of the decision tree and will be split based on features to create the tree.**

**B-3**

I will be using the Python langue for my analysis. Pandas is imported to allow for me to import the csv data that I will be using. From pandas I imported CategoricalDtype to help with memory efficiency with the categorical data I am using. Numpy was the next import. It was imported to have access to certain calculations. Matplotlib.pyplot and seaborn are imported to allow for visualizations that may be needed. From sklearn.model\_selection import train\_test\_split for splitting the dataset into training and testing sets. From sklearn.tree import DecisionTreeClassifier is for the decision tree algorithm. From sklearn.metrics import accuracy\_score, classification\_report allows to evaluate the model using accuracy score and detailed report including precision, recall, F1-score, and support for each class.

***Part III: Data Preparation***

**C-1**

**One data preprocessing goal that is needed to be done is encoding categorical data. This will require changing binary categories from “True” or “Yes” to 1 and “False” or “No” to 0. Categorical variables with more than two responses require similar changes to numerical values done with dummy variables for each of the possible responses.**

**C-2**

The following are the variables that will be used for my analysis:

* Children- Numeric
* InternetService- Categorical
* Contract- Categorical
* Gender- Categorical
* Techie- Categorical
* Multiple- Categorical
* OnlineBackup- Categorical
* DeviceProtection- Categorical
* TechSupport- Categorical
* StreamingTV- Categorical
* StreamingMovies- Categorical
* Bandwidth\_GB\_Year- Numeric
* Churn- Categorical

**C-3**

The following steps were taken to prepare the data for analysis:

* + Load the churn\_clean csv file. 
  + Check the columns for 10,000 non-null count and datatypes. 
  + Decision tree is not sensitive to outliers so they will not be checked for.
  + Create Boolean mapping and convert columns that need to be changed into Boolean columns. A screenshot of a computer program

    Description automatically generated
  + Create a new dataset with just the variables listed in C-2A screen shot of a computer

    Description automatically generated
  + Separate the explanatory variables (x) from the response variable (Y). A computer screen shot of a computer code

    Description automatically generated
  + Create dummy variables for my categorical variables. A screen shot of a computer code

    Description automatically generated
  + Insert dummy variables in with my explanatory variables and drop columns that dummy variables came from. A screen shot of a computer code

    Description automatically generated A screenshot of a computer program

    Description automatically generated
  + Decision tree is not sensitive to multicollinearity issues, so I will not be checking for them.
  + Decision trees evaluate feature importance during their construction, so I will not be using “SelectKBest” to select the important features.

**C-4**

Please see attached d209clean2.csv file for copy of clean data.

***Part IV: Analysis***

**D-1**

Please see attached csv files named Xtrain, Xtest, ytrain, and ytest for the split data.

**D-2**

Decision tree analysis is a non-parametric supervised learning method. The model makes a tree structure by breaking down a data set into smaller and smaller subsets through decision nodes. The data set starts at the top as the root node. The root node is also the first decision node. A decision node has two or more branches that lead to another decision node or a leaf node. This is done by a decision rule that is created at each node. A leaf node will give you the models final prediction. The prediction on my model will be either a “yes” or “no” due to my A-1 question. I did not preform any intermediate calculations as my decision tree model did not require any as I worked it.

**D-3**

Below I will provide screen shots of the code that I used starting after splitting and saving the data as that is when section D2 starts at. I am also uploading my full code on a Jupyter notebook if more is needed.

A screenshot of a computer program

Description automatically generated

A white background with black text

Description automatically generated

A close-up of a white background

Description automatically generated

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

**E-1**

**The decision tree model used for this assignment has an accuracy of 0.853. This is shown in the confusion matrix giving 1327 accurate predictions of not a churn customer and 379 accurate predictions of a churn customer. These equals to 1706 accurate predictions out of the total 2000 predictions which gives the 0.853. My model resulted in a mean squared error (MSE) of 0.147. This would suggest the model accuracy is good, but since my decision tree is for a binary classification, it is not a good metric to look at and is only included because the rubric specifically asked for it.**

**E-2**

The results from my decision tree model seems to show that it could be useful in predicting Churn. The accuracy score of 0.853 in the test set shows it predicts churn at a fairly high rate. This is higher than my rate from my KNN, even though I am not sure if this could be a direct comparison. Even though I am happy with the accuracy score I know it can be improved upon. The use of hyperparameter tuning was not required for this assignment but it could be used to optimize the model’s performance. I am not sure if including more labeled data will improve my model. I cannot see data points such as latitude and longitude of a customer being a useful indicator of Churn, but I have worked with data enough to know that sometimes unexpected things have a correlation.

**E-3**

The biggest limitation I believe is that this analysis is a snapshot in time for the customer’s information. I think having multiple snapshots of a customer’s activity would produce better results on a topic such as churn. I realize that this would possibly lead to different types of analysis that may be out of the scope of this class, but if this was work related and I was asked about limitations of my results this would be my first thing I would bring up.

**E-4**

Based the accuracy score I would recommend this k-NN model to predict customers that are at churn risk. I believe this could be useful for two different departments to have. The sales team could come up with ways such as promotions to keep the at-risk customers with the company. I also think informing the finance department would be useful as they could have an estimate of the money coming in from customers to set a better monthly budget for the company.

***Part VI: Demonstration***

**F**

Uploaded it to the Panopto drop box titled titled “Data Mining I - NVM2 | NVM3 | D209 (student creators) [assignments].” Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b9ce7825-0a69-48c6-85f6-b10900071622>

**G**

No web sources were used to acquire data or segments of third-party code.

**H**

Almog, U. (2023, May 8). Decision Trees Explained. Towards Data Science. <https://towardsdatascience.com/decision-trees-explained-d7678c43a59e>

**I**

The content in this Performance Assessment is set up and presented with the highest professional standards.