**Performance Assessment D209 Task 2**

**PREDICTIVE ANALYSIS**

AnnMarie Stagnitta

Student ID: 012258082

MS, Data Analytics

[AStagni@WGU.EDU](mailto:AStagni@WGU.EDU)

A1. Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following prediction methods

My research question for my D209 performance assessment is: What are the major predictor variables when predicting customer churn? I will be using a decision tree to answer the research question.

A2. Define one goal of the data analysis

My goal for this analysis is to identify the major predictors of churn. I will determine whether a decision tree can accurately classify which customers are more susceptible to churn by comparing the features or variables of customers who have churned with those who remained with the company. This information can help us identify patterns in customer behavior that can be used to increase customer retention, thus leading to decreased churn rates.

B1. Explain how a decision tree prediction method analyzes the selected dataset. Include expected outcomes.

A decision tree is a flowchart used for predictive modeling that can handle categorical and numerical data. It requires less computation and training time, is nonparametric, and is not sensitive to outliers. This method uses a tree-like structure to represent decisions and their potential outcomes by breaking the dataset into smaller subsets of instances with similar values. Each subset has a leaf node and a decision node. Where nodes represent features, an internal node represents an attribute, a branch is a decision, and each leaf represents a predicted outcome. This method predicts new data by traversing the tree from the root to a leaf node. The predicted result is then determined by the majority vote of the same in the corresponding leaf (Sehra, 2020). The expected outcome is a decision tree that we can use to predict churn based on the product categories.

B2. Summarize one assumption of the decision tree

As discussed in our course resources, one assumption of the decision tree algorithm is that it is non-parametric. This means the decision tree algorithm does not make distribution assumptions on the data.

B3. List the packages or libraries you have chosen for Python and justify how each item on the list supports the analysis.

|  |  |
| --- | --- |
| **Package/Libraries** | **Purpose** |
| Numpy | Provides array objects for calculations |
| Pandas | Importing data into Dataframe and data manipulations |
| Matplotlib | For visualizations |
| Seaborn | For visualizations |
| SciPy | Normalization and statistics |
| Sklearn.feature\_selection import SelectKBest,f\_classif | For feature selection. |
| Sklearn.model\_selection import train\_test\_split | For splitting data into train and test sets |
| Statsmodels.stats.outliers\_influence import variance\_inflation\_factor | VIF to determine multicollinearity |
| Statsmodels.graphics.mosaicplot import mosaic | Mosaic graphs for bivariant visualization |
| Sklearn.model\_selection import GridSearchCV | Grid search algorithm for hyperparameter tuning |
| Sklearn.metrics import roc\_curve | ROC curve for binary classification problem |
| Sklearn.metrics import classification\_report | Report that summarizes the performance of a classification task |
| Pandas.api.types import CategoricalDtype | Categorical data type with specified categories and ordering |
| Sklearn.tree import DecisionTreeClassifier, plot\_tree | Creates a decision tree to classify data. |
| Sklearn.metrics import  Mean\_squared\_error | For getting the mean squared error metric score for the model |
| Sklearn import preprocessing | Functions for preprocessing of the data |
| Sklearn.ensemble import AdaBoostClassifier | Ensemble method to improve classifier performance |
| Sklearn,metrics import roc\_auc\_score | Computes the ROC area under the curve |
| Sklearn.metrics import confusion matrix | Computes confusion matrix for a classification model |

C1. Describe one data preprocessing goal relevant to the classification method from part A1.

One data preprocessing goal relevant to the classification method from part A1 is one-hot encoding. One-hot encoding is appropriate in the preprocessing step for a decision tree because we are dealing with categorical variables. These variables are non-numerical and cannot be used directly in the model. One-hot encoding will create binary indicator variables for each category in a categorical variable.

C2. Identify the initial data set variables that you will use to perform the analysis for the classification question from part A1 and classify *each* variable as numeric or categorical.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Numeric or Categorical** | **Description** | **Example** |
| Churn | Categorical | Yes/No if customer canceled service | Yes |
| MonthlyCharge | Numeric | Amount in dollars the customer is charged per month | 172.455519 |
| Tenure | Numeric | Length of time in months the customer has maintained service | 6.795512947 |
| Population | Numeric | Population of customer residence | 38 |
| Children | Numeric | How many children live in the cusomters household | 2 |
| Age | Numeric | Customers age | 30 |
| Income | Numeric | Customers annual income | 28561.99 |
| Gender | Categorical | Self-identified gender of the customer | Female |
| Outage\_sec\_perweek | Numeric | Average number of seconds per week of system outages in customers neighborhood | 7.978322947 |
| Email | Numeric | Number of emails sent to customer over the past year | 10 |
| Contacts | Numeric | Number of times customer contacted technical support | 0 |
| Yearly\_equip\_failure | Numeric | Number of times customers equipment failed and replaced | 1 |
| Techie | Categorical | If a customer considers themselves technically inclined | Yes |
| TechSupport | Categorical | If customer has technical support add-on | No |

C3. Explain the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

The first step for this performance assessment is completing the preprocessing steps to ensure my data is ready for the decision tree. After my packages were loaded, I imported my primary dataset. To confirm the data loaded correctly, I used the df.info() function to display a summary of the dataframe. This function lists each column, the number of non-null values in that column, and the data type.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

My next step is to check for null values and duplicates. If any were found, we would clean the data and recheck to make sure the data is ready for our analyst.

A screenshot of a computer

AI-generated content may be incorrect.

I then went on to rename my columns to make them easily readable. I completed this step in every performance assessment and wanted to remain consistent throughout my time with WGU. Even though this step is unnecessary, it allows me to better understand what columns we are using. After renaming my columns with df.rename, I then confirmed my changes using the df.columns function. This displayed all the new columns that I updated in our last step.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

My next step was to create a histogram to show the distribution of our variables. Histograms provide a visual representation of the frequency distribution.

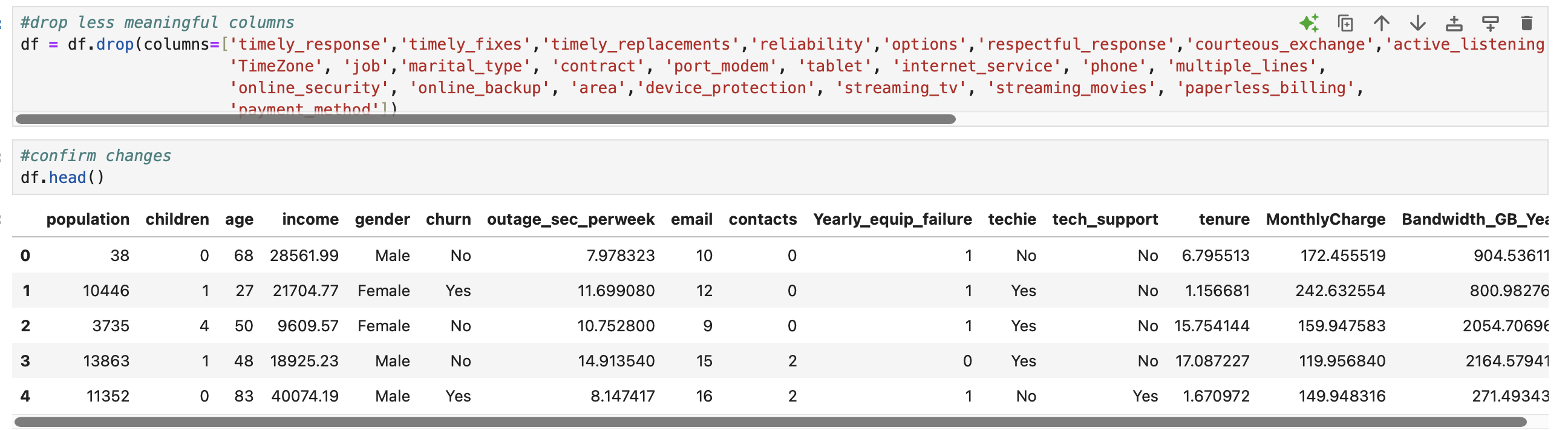
A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

The next step is to remove any less meaningful, unnecessary columns for our analysis. I used the df.drop function to perform this. I then confirmed my changes with the df.head function.



Now I want to review the statistical information for our remaining continuous variables. This helps us understand dataset patterns, trends, and relationships. It can be used to draw conclusions from numerical data.

A screenshot of a computer

AI-generated content may be incorrect.

Next, I identified the number of distinct values and filtered out duplicate or inconsistent entries. I used the df.nunique() function to find the number of unique values. This provides a measure of diversity, helps us gain insight into how the categories are distributed, and helps us maintain data quality and integrity within the data. I then used df.value\_counts() to display each unique value and its frequency to see how often each item appears.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Based on the number of unique values, I dropped the highly correlated variable, Bandwidth\_GB\_Year. Now that I have satisfied our unique values, I will use one-hot encoding to convert our categorical variables to numeric.

A screenshot of a computer

AI-generated content may be incorrect.

C4. Provide a copy of the cleaned data set.

I have attached a copy of the cleaned data set to this submission.

A white rectangular sign with red text

AI-generated content may be incorrect.[file:///Users/annmarie/Library/CloudStorage/OneDrive-WesternGovernorsUniversity/D209/Task 2/Prepared\_Churn\_data](file:///Users/annmarie/Library/CloudStorage/OneDrive-WesternGovernorsUniversity/D209/Task%202/Prepared_Churn_data)

D1. Split the data into training and test data sets and provide the files

To ensure the model has similar accuracy when predicting unseen data, I split the data into training and test datasets using an 80/20 split. 80% of the data will be used to train the algorithm, and the remaining 20% will be used to test the classification. I have attached the training and test data sets to this submission. I have also listed screenshots of my code below:

A screenshot of a computer

AI-generated content may be incorrect.

D2.  Describe the analysis technique you used to appropriately analyze the data.

To appropriately analyze the data, I used a variety of statistical measures. This includes accuracy, recall, precision, F1 Score, and Mean Squared Error. The first step I completed was defining the hyperparameters of the grid to search and create a decision tree classifier. I then calculated the accuracy and AUC of the model. Based on the output, we can see that the model will correctly predict the outcome 83.35% of the time. The AUC returned a value of 0.745357713002954, which is closer to 1, with 1 indicating a perfect classifier.

A close-up of a computer screen

AI-generated content may be incorrect.

I then used the parameters I found to fit the decision tree model and created a plot for visualization.

A diagram of a diagram

AI-generated content may be incorrect.

The next step was to use a classification report to identify the precision, recall, F1-score, and support of the model:

A number of numbers in a row

AI-generated content may be incorrect.

Class 0 reported that 85% of predictions were true positives. It also has a recall of 94%, indicating that the instances were correctly classified. The F1-score of 0.89 also demonstrates a good overall accuracy for this class. Class 1 indicates 77% of predications were true positives, and the recall indicated 55% were correctly classified. However, the F1-score of class 1 shows poor accuracy. The classification report shows an overall accuracy of 0.83; the model can correctly predict the class label for 83% of instances. An overall macro average F1-score of 0.77 indicates a reasonable overall performance with room for improvement.

I then generated a confusion matrix to identify the errors the model makes. The confusion matrix returned that we have 1367 instances where the model correctly classified as negative and 90 instances incorrectly classified as positive. We also see 243 instances where the model incorrectly classified an instance as negative, and 300 instances where the model correctly classified an instance as positive.

A screenshot of a computer code

AI-generated content may be incorrect.

A diagram of confusion matrix

AI-generated content may be incorrect.

Next, I calculated the accuracy of the training and test data sets. Both sets' accuracy was similar, indicating they are performing similarly. I also found the AUC-ROC score, which returned a value of 0.7453. An AUC-ROC score of 0.5 signifies random guessing; the score our model returned indicates the model is performing well.

A close-up of a sign

AI-generated content may be incorrect.



I then constructed an ROC curve, which is used for evaluating the performance of a binary classification model. An AUC-ROC score less than 0.6 is considered to be poor performing. If the curve is above the line, it reflects a reasonable prediction rate; if it is below the diagonal dotted line, it signifies a poor prediction rate. Our model returned a reasonable prediction rate.

A blue and orange line

AI-generated content may be incorrect.

My last step was calculating the Mean Squared and Root Mean Squared errors. I will review the results in E1 of this performance assessment.

A black text with numbers

AI-generated content may be incorrect.

D3.  Provide the code used to perform the classification analysis from part D2

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a diagram

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A graph of confusion matrix

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

A black text with numbers

AI-generated content may be incorrect.

E1.  Explain the accuracy and the MSE of your prediction model.

The model's accuracy is 0.8335, which means the model correctly predicted the class labels for 83.35% of the instances in the test set. Our model also returned an MSE of 0.1665. This suggests the models’ predictions were off by 0.1665 units from the actual value. This indicates better performance with the model being able to predict closer to the actual values. The RMSE returned a value of 0.4080, which suggests the model’s predictions were off by 0.4080 units from the actual values. This also indicates a better performance due to the low RMSE value.

E2.  Discuss the results and implications of your prediction analysis

I utilized the classification report to see how accurate our model is. An overall accuracy of 83% and an AUC score of 0.7453 helped us identify that our model is good at making predictions. Our negative class of 0 has a precision of 0.85, which signifies the correctly predicted to the total predicted. It also has a recall of 0.94, which is the ratio of correctly predicted to all data in the actual class. We also have an F1 score of 0.89. Our positive class of 1 has a precision of 0.77, a recall of 0.55, and an F1 score of 0.64. Although this is lower than our negative class, this model is still closer to 1 and suitable for use. The confusion matrix demonstrates that the model correctly predicted 1367 customers who did not churn and 243 who did.

The AUC score of 0.7453 falls within the “excellent discrimination” level (Zach Bobbitt, 2021). Our model can correctly classify true positives and true negatives 74.53% of the time. With the added overall accuracy of 83%, the model performs well without overfitting. The model correctly identifies a high percentage of customers who did not churn, but has room for improvement in identifying customers who did churn. The model accuracy can be improved by performing hyperparameter tuning on additional parameters to increase the model's accuracy. We see this because the hyperparameters were relatively simple and may be more susceptible to underfitting.

E3.  Discuss one limitation of your data analysis

One limitation of our data analysis is that decision trees can perform overfitting. This technique can tend to overfit the training data. If a tree becomes too complex, it may perform poorly on unseen data. To regularize the model, the growth must be restricted in some way, also known as pruning the tree.

E4. Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in E2.

Based on my analysis, this model can reliably predict which customers are at risk of churn. My recommended course of action would be to offer customers in danger of churn special promotional rates, bundles, or other incentives to entice them to remain with the company. This will help to retain customers and increase revenue.

F.  Panopto Recording

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c9195d4c-bf79-403a-b0a4-b33200ffff8f>

References:

**G:**

Attard, H. I. M. (2024, May 1). *8 key advantages and disadvantages of decision trees*. Inside Learning Machines. https://insidelearningmachines.com/advantages\_and\_disadvantages\_of\_decision\_trees/#8\_Disadvantages\_of\_Decision\_Trees

Bhandari, A. (2025, May 1). *Guide to AUC ROC curve in machine learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

GeeksforGeeks. (2024, July 19). *Performing feature selection with GRIDSEARCHCV in Sklearn*. https://www.geeksforgeeks.org/machine-learning/performing-feature-selection-with-gridsearchcv-in-sklearn/

GeeksforGeeks. (2025, January 9). *Limitations of Decision Tree*. https://www.geeksforgeeks.org/machine-learning/limitations-of-decision-tree/

Wijaya, C. Y. (2024, September 11). *Breaking down the classification report from Scikit-Learn - NBD Lite #6*. Breaking Down the Classification Report from Scikit-Learn - NBD Lite #6. https://www.nb-data.com/p/breaking-down-the-classification

**H:**

Sehra, C. (2020, November 30). *Decision trees explained easily*. Medium. https://chiragsehra42.medium.com/decision-tree-explained-easily-28f23241248

Zach Bobbitt (2021, September 9). *What is considered a good AUC score?*. Statology. https://www.statology.org/what-is-a-good-auc-score/