**Performance Assessment D209 Task 1**

**CLASSIFICATION ANALYSIS**

AnnMarie Stagnitta

Student ID: 012258082

MS, Data Analytics

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

A1. Purpose one question relevant to a real-world organizational situation that you will answer using the KNN method

My research question for my D209 performance assessment is: What are the major predictor variables when predicting customer churn? I will be utilizing the k-nearest neighbor method to answer my 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 identify whether using the k-nearest neighbor classification method can accurately determine which customers are more susceptible to churn. I will compare the variables of customers who have churned and those who have 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 the KNN classification method analyzes the selected dataset. Include expected outcomes.

K-NN is described as a voting system “where the majority class label determines the class label of a new data point among its nearest ‘k’ neighbors in the feature space” (Shafi, 2023). Data labeling helps us to identify the difference between two or more data points by labeling the key features of the data points and looking for similarities between them. The expected outcome is to classify test data according to its closest neighbor. I would expect the data to be classified based on the Churn status of Y or N.

B2. Summarize one assumption of the KNN classification method

As discussed in our course resources, one assumption of KNN is that if a data point is far away from another group, it is dissimilar to those data points. This also means that data points close to each other will be highly similar.

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.preprocessing import scale | To scale the features |
| Sklearn.model\_selection import train\_test\_split | For splitting data into train and test sets |
| Sklearn.pipeline import Pipeline | Pipeline is to assemble several steps that can be performed together while setting different parameters. |
| Statsmodels.stats.outliers\_influence import variance\_inflation\_factor | VIF to determine multicollinearity |
| Statsmodels.graphics.mosaicplot import mosaic | Mosaic graphs for bivariant visualization |
| Sklearn.neighbors import KNeighborsClassifier | KNN classification algorithm |
| 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 |
|  |  |

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 KNN classification because we are dealing with categorical variables. These variables are non-numerical and cannot be used directly in calculations for KNN classifications. 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 *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

Before completing the KNN, we had to complete the preprocessing steps. I have listed the screenshots below:

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A graph of different sizes and shapes

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 screenshot of a computer

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 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.

[file:///Users/annmarie/Library/CloudStorage/OneDrive-WesternGovernorsUniversity/D209/Task 1/Prepared\_Churn\_data](file:///Users/annmarie/Library/CloudStorage/OneDrive-WesternGovernorsUniversity/D209/Task%201/Prepared_Churn_data)

A close-up of a sign

AI-generated content may be incorrect.

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

I used the 0.25 split, meaning 75% of my data will be used to train the KNN algorithm, and the remaining 25% 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.

I will use StandardScaler to standardize our large numerical ranges to ensure that the data aligns with the rest of the data points. I will then use GridSearchCV to find an optimal K value. The K values range from 1 to 50. If there is a low K value, KNN will look only at the neighbors closest to that data point. If it is a high K value, the algorithm uses more data to make its decision, reducing performance and predictive power.

A screenshot of a computer

AI-generated content may be incorrect.

Our data returns a K value of 22, which means that GridSearchCV will perform 22 iterations of training and evaluation for each combination of hyperparameters we are testing. It also returned a mean value of 0.8208221626452189. The mean determines the average performance metric obtained during cross-validation for a specific set of hyperparameters.

To visualize the model’s accuracy at each level of K, I would want to plot the results for both the train and test sets.

A graph of a number of neighbors

AI-generated content may be incorrect.

I would now want to determine the area under the curve, or AUC. I will use the roc\_auc\_score function to find the AUC. The AUC measures the ability of a classifier to distinguish between classes. The higher the AUC, the better the performance of the model. We currently have an AUC score of 0.8885683750868909. This indicated that 88% falls within the excellent discrimination level, which is the model’s ability to distinguish between the positive and negative classes. We can confirm this with the graph plotted based on this data. A curve above the line reflects a reasonable prediction rate.

A close-up of a computer screen

AI-generated content may be incorrect.

A line graph with a blue line

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 graph

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

E1.  Explain the accuracy and the area under the curve (AUC) of your classification model.

The accuracy of the model is 0.8208221626452189, and the AUC is 0.8885683750868909. This concludes that the model can correctly predict the class of 82% of instances in the test data. If an AUC is closer to 0.5, the model is randomly guessing, and closer to 1.0 indicates a perfect classification. Our model has an AUC of 0.8885, which is closer to 1.0. The model can distinguish between positive and negative classes based on the AUC. The AUC and accuracy scores indicate the model could improve, but it is a good starting point and can help identify the relationship between variables and churn.

E2.  Discuss the results and implications of your classification analysis

Using the SelectKBest approach to find which variables were the most significant in the model, I was able to identify Tenure, MonthlyCharge, DummyTechie, DummyGender, and DummyTechSupport. We can use these variables to help determine the factors that influence churn. To confirm my findings, I utilized the classification report to see how accurate our model is. With a K value of 22, we have an overall accuracy of 82%. Our negative class of 0 has a precision of 0.84, which signifies the correctly predicted to the total predicted. It also has a recall of 0.93, which is the ratio of correctly predicted to all data in the actual class. We also have an F1 score of 0.88. Our positive class of 1 has a precision of 0.72, a recall of 0.53, and an F1 score of 0.61. Although this is lower than our negative class, this model is still closer to 1 and suitable for use.

The AUC score of 0.888 falls within the “excellent discrimination” level (Zach Bobbitt, 2021). Our model can correctly classify true positives and true negatives 88% of the time. With the added overall accuracy of 82%, the model performs well without overfitting.

E3.  Discuss one limitation of your data analysis

One limitation of our data analysis is that the KNN method does not typically work well with large datasets. The dataset we are utilizing for this analysis can be considered large. It can result in slow prediction speed and can be memory-intensive when storing the entire training dataset in memory for predictions.

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, it seems this model can be used to 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=a937152a-31ac-426e-a008-b30b0050f208>

References:

**G:**

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/

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:**

Shafi, A. (2023, February 20). *K-Nearest Neighbors (KNN) classification with scikit-learn*. DataCamp. https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

Zach BobbittHey there. My name is Zach Bobbitt. I have a Masters of Science degree in Applied Statistics and I’ve worked on machine learning algorithms for professional businesses in both healthcare and retail. I’m passionate about statistics. (2021, September 9). *What is considered a good AUC score?*. Statology. https://www.statology.org/what-is-a-good-auc-score/