Raquel Ocasio

D209 – Data Mining I, Task 1

October 23, 2023

Western Governors University

Part A1

Which variables can help determine the “Churn” classification for a customer?

Part A2

The goal of the data analysis is to be able to predict the “Churn” label for a customer.

Part B1

The K-Nearest Neighbors (KNN) classification method analyzes the selected data by using the similarity between data points to make predictions or classifications. The expected outcomes are either Yes or No.

Part B2

One key assumption of KNN is that similar data points are located near each other in the feature space. This assumption is fundamental to KNN's ability to make accurate predictions and classifications based on the distance between the data points.

Part B3

Library/package name	*How it supports the analysis*
pandas	Provides access to the read_csv() function.
numpy	Ability to perform mathematical calculations.
scipy	Provides the stats package for calculating z-scores.
sklearn.preprocessing	Provides the OneHotEncoder method for one-hot encoding of categorical variables.
sklearn.model_selection	Provides the train_test_split method for splitting datasets into training and test sets.
sklearn.neighbors	Provides the KNeighborsClassifier method for implementing KNN analysis.
sklearn.preprocessing	Provices MinMaxScaler method for scaling values across variables.
Sklearn	Provides metric method that is used in calculating AUC.
sklearn.metrics	Provides roc_auc_score method that is used in calculating AUC.

Part C1

One of the data preprocessing goals is to convert categorical values to numeric values.

Part C2

Variable name	*Variable description*	*Variable type*
Churn	Indicates if the custom has discontinued service with the last month	Categorical
Outage_sec_perweek	Indicates how long a system outage lasted in the customer's neighborhood	Numerical
Contacts	Indicates how many times the customer contacted technical support	Numerical
Yearly_equip_failure	Indicates how many times the customer's equipment had to be reset/replaced due to failure	Numerical
Contract	Indicates the contract term type of the customer	Categorical
Tenure	Indicates how many months the customer has been with the provider	Numerical
MonthlyCharge	Indicates the monthly amount charged to the customer	Numerical

Part C3

The data was prepared for analysis by detecting and treating duplicate values, missing values, outliers, and re-expressing categorical variables.

Duplicate values were detected using the combined .duplicated().value_counts() functions. The code segment is on lines 14 through 17. No duplicate values were found.

Missing values were detected using the combined .isnull().sum() functions. The code segment is on lines 19 through 22.

The missing values were treated by univariate imputation. The code segment is on lines 24 through 29.

Outlier values were detected by calculating the z-scores, then counting z-scores that were greater than three or less than negative three. The outlier values were retained. The code segment is on lines 31 through 50.

Before re-expressing categorical variables, the unique() method was used in a function to detect the unique values of each categorical variable.

After detecting the unique values, two methods were used to re-express the categorical variables. For categorical variables with two unique values, label encoding was applied using the factorize() function. For categorical variables with more than two unique values, one-hot encoding was applied using the get_dummies() function. The values that were one-hot encoded were converted from Boolean to integer format using the astype(int) function. The code segment is on lines 68 through 94.

Part C4

See attached file.

Part D1

See attached files. (Larose & Larose, 2019, p. 5.2.1)

Part D2

The KNN technique was used to analyze the data. KNN predicts the label of a datapoint using unforeseen points based on the values of the closest existing points. A preset number of values (k) is used to determine how many of the existing points will be used in the algorithm.

I did not perform any intermediate calculations.

Part D3

See attached code file.

Part E1

The accuracy of the KNN prediction classifier on the training data is 90%. The accuracy of the KNN prediction classifier on the testing data is 85%. The accuracy levels should be considered as very good since they are so close to being 100% accurate.

The Area Under The Curve (AUC) is 0.8847529644268775. The value of AUC ranges from zero to one. The closer the AUC value is to one, the better the classifier is at predicting a label value. Since the value of this classier is very close to one, it is a good classifier for the dataset.

Part E2

Results: When using the Outage_sec_perwoeek, Contacts, Yearly_equip_failure, Contract, Tenure and MonthlyCharge variables, the classification model correctly predicted the correct Churn label 85% of the time when using the test data.

Implications: The classification model can possibly be improved further by using a higher odd value for k, and/or using more labeled data to train the model.

Part E3

Since the outliers were retained, one limitation of the data analysis is that KNN is sensitive to outliers since it chooses the neighbors based on distance. (Genesis, 2018)

Part E4

It is recommended that the company use the classification model when determining the Churn label for customers. It is also recommended that the company try variations of the model by adding or removing variables, and changing the value of k, to achieve a more accurate result.

Part F

The demonstration can be viewed at
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=cb8305ff-9573-40f8-bcfb-b0a5002da827

Part G

Genesis. (2018, September 25). Pros and Cons of K-Nearest Neighbors - From The GENESIS. From the GENESIS. https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/

Part H

Larose, C. D., & Larose, D. T. (2019). *Data science using Python and R*. https://doi.org/10.1002/9781119526865