Tonie Cooper

Master Data Analytics

[Tcoo158@wgu.edu](mailto:Tcoo158@wgu.edu)

D209 Task 1 Assessment

A1) What customers are at a high risk of churn using the K-Nearest Neighbors algorithm?

A2) Stakeholders will have confidence when making decisions on what customers they should focus their assets on to reduce the churn rate.

B1) The K-nearest neighbor algorithm is a supervised machine learning algorithm that is used for classification and regression analysis. It uses a new unlabeled point to make a prediction. This algorithm assumes that similar data points are close together and categorizes it based on that. It utilizes the idea of similarity combined with math to calculate the distance between the points on the graph. Then, to choose the correct K value for the data set this algorithm must be ran several times. We then choose the K value that reduces the number of errors while maintaining the precision needed for new data. We need to keep in my mind when making these reductions that as we decrease the value of K to 1 our prediction becomes less stable. Additionally, as we increase the K value, we achieve more accurate predictions. However, as we continue to increase K, we will see an increase in the number of errors. (Datagy 2022).

The expected outcome is the test data will be classified with the closest neighbors.

B2) An assumption of this method is that the data points that are nearby are like each other (Grant 2019).

B3) Packages for Python:

Pandas

Numpy

Scikit-learn

Matplotlib

Seaborn

Matplotlib, Pandas and Numpy are standard imports that provide statistical packages for reading, scoring and visualization. Seaborn contains descriptive graphs, matrix, and plots. Scikit-learn allows for splitting, fitting, predicting, and applying the metrics

C1) One data preprocessing goal is to change the qualitative binary data to a dummy variable 1/0

C2) Continuous predictor variables:

Children

Income

Outage\_sec\_perweek

Email

Contacts

Yearly\_equip\_failure

Tenure

MonthlyCharge

Bandwidth\_GB\_Year

Categorical predictor variables:

Techie

Contract

Tablet

InternetService

Phone

Multiple

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

C3) Steps to prepare the dataset:

Backup data to my computer – to prevent total loss of data

Read the csv file into Python – data needed to run analysis

Describe the data – find what data needs to be changed to a quantitative dummy variable to allow for calculations. This also allows for better understanding of the data available.

Name dataset as churn\_df and the dataframe as df

Search for misspellings, missing data and correct that data – ensure data are alike to prevent unknown errors and unintentional outliers. This will also allow us to utilize mean, median and mode to fill in missing data where applicable.

Use descriptive statistics to locate outliers by using histograms and boxplots – outliers can disrupt and cause a false calculation. I’ll remove them to have a more precise calculation on this analysis.

Remove less meaningful data – KNN will slow down when too much data is introduced to it so removing data that won’t play a role in this analysis is crucial. Removing less meaningful data also reduces the chances of errors accruing.

Extract cleaned dataset for use in the k-nearest neighbor model – extract dataset to allow for usage in the KNN algorithm.

Additionally, when cleaning the dataset dummy variables will be created to change qualitative data to quantitative data (1/0) to allow for calculations to take place and the data entered into the KNN analysis.

Lastly, discrete variables, shown as survey items, will have their names changed from “Items” to a name that represents what the Item is such as Response, Timely Fixes, Options, etc. This allows for better readability and analysis of the customers opinion on what they deem most important in a telecommunication service.

C4) See attachment

D1) See attachment

D2) For this classification I utilized the k-nearest neighbor method. With this method I was able to achieve an accuracy of 79%, up from 71%, at a precision of 84%. Looking at this model the KNN model allows us to check the sensitivity of the observations (Recall). A recall rating above 0.5 is what I’m aiming for, and I was able to achieve a result of 0.88 which is good.

Lastly, the model takes a weighted average of precision and recall producing the F1 score. This score considers the false positives and the false negatives and can be more useful than accuracy, especially when there’s an uneven distribution class. I observed and F1 score of 86%.

E1) The scaling improved model performance from an accuracy of 0.71 to 0.79 and precision of 0.78 to 0.84; The area under the curve is 0.7959.

E2) This algorithm resulted in an accuracy score of 0.79 with a precision of 0.84 and an F1 score of 0.86. This shows that the KNN model produced an accuracy of 79%. Meaning that the ratio of correctly predicted observations to the total observations is 79%. The precision ratio is the correctly predicted positive observations to the total predicted positive observations. The recall, 0.88, tells us if the model returned more relevant results; a recall of 88% is a great value for this data set.

The implications this classification method has is the large amount of memory this method requires. As this method may work for the data provided as we progress overtime and continue to add new data this method will become costly due to the large memory needed for storing the entire dataset. This makes this method less than ideal for very large datasets.

E3) A limitation of the k-nearest neighbors’ algorithm is its ability to handle large datasets. Since this algorithm is based off distance calculating it from a new point to existing points can drastically degrade the performance of the algorithm (Jain 2020).

E4) With a low score as stated above it’s recommended that the client understands the accuracy of this and should analyze the items that correlate with the customers leaving. They should attempt to reduce those grievances and suggest to more customers the available services that the company provides. It is in the best interest of client to improve grievances and up sale products available to each customer.

G) Code Citations

Yildirim, S. (2020, June 6). *Visualize missing values with Missingno*. Medium. Retrieved May 5, 2022, from https://towardsdatascience.com/visualize-missing-values-with-missingno-ad4d938b00a1

H) In-text Citations

Grant, P. (2019, July 21). *Introducing K-nearest neighbors*. Medium. Retrieved May 5, 2022, from https://towardsdatascience.com/introducing-k-nearest-neighbors-7bcd10f938c5

*K-Nearest Neighbor (KNN) algorithm in Python • datagy*. datagy. (2022, April 14). Retrieved May 5, 2022, from https://datagy.io/python-knn/