DA5020 – Practicum III

In this practicum you will use the k-nearest neighbor algorithm to predict a **continuous variable**. Each question in the practicum follows the CRISP-DM framework. The practicum was designed in this manner to help you to practice and conceptualize each phase, based on the requirements of an actual project.

This is a group practicum which means that you may choose to work in groups of up to three students. **You may fully collaborate with your team and submit the same work**. However, you must include all students' names on all submitted work. If a group member is not adequately contributing, the remaining team members may "vote to eject" the student from the team by emailing me the reason. In such an event, the team member who was "fired" must still complete the project individually by the due date.

If you are working in groups, please let us know who you are working with on the assignment!

Useful Concepts to Research:

- Normalizing Data with R
- kNN Algorithm using R
- KNN Algorithm: A Practical Implementation Of KNN Algorithm In R
- k-Nearest Neighbor: An Introductory Example
- Quick Guide to Creating Scatterplots in R with ggplot

Practicum Tasks

CRISP-DM: Business Understanding

The NYC Taxi and Limousine Commission (TLC) publishes a dataset on yellow and green taxi trip records which include: pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. For more information on the dataset, visit the following website and view the accompanying data dictionary for additional information.

Description of the Problem

You are hired as a Machine Learning Engineer, on the Data Insights and Analytics Team, for the NYC Taxi and Limousine Commission (TLC). Your first assignment is to analyze the trip data from the Green Taxis; more specifically, you need to evaluate where passengers use these cabs and how frequently. However, ***your main objective is to evaluate the factors that contribute toward cab drivers being incentivized*** (i.e. what determines whether or not they receive a tip). This will enable you to build a model that can be used to predict the tip amount for future trips.

In this use-case, you will conduct your analysis using the NYC Green Taxi Trip Records for **2018** and build a k-nn regression model to predict the tip amount. **You are free to use any libraries to support your analysis. You can find this data under assignment 3 in Canvas.**

Question 1 — (20 points) +10 optional points

CRISP-DM: Data Understanding

- **Load** the NYC Green Taxi Trip Records data into a data frame or tibble.
- **Data exploration:** explore the data to identify any patterns and analyze the relationships between the features and the target variable i.e. tip amount. At a minimum, you should analyze: 1) thedistribution, 2) the correlations 3) missing values and 4) outliers — provide supporting visualizations and explain all your steps.

Tip: remember that you have worked with this dataset in your previous assignments. You are free to reuse any code that support your analysis.

- **Feature selection**: identify the features/variables that are good indicators and should be used to predict the tip amount. Note: this step involves selecting a subset of the features that will be used to build the predictive model. If you decide to omit any features/variables ensure that you briefly state the reason.

- **Feature engineering: (+10 bonus points):** create a new feature and analyze its effect on the target variable (e.g. the tip amount). Ensure that you calculate the correlation coefficient and also use visualizations to support your analysis. Summarize your findings and determine if the new feature is a good indicator to predict the tip amount. If it is, ensure that you include it in your model. If it is not a good indicator, explain the reason.

  NOTE: If you attempt this bonus question, ensure that you create a meaningful feature (and nothing arbitrary). If you are unable to think about something meaningful, do not become fixated on this. There is another bonus question that you can attempt later in the practicum.

Question 2 — (20 points)

CRISP-DM: Data Preparation

- Prepare the data for the modeling phase and handle any issues that were identified during the exploratory data analysis. At a minimum, ensure that you:

- **Preprocess the data**: handle missing data and outliers, perform any suitable data transformation steps, etc. Also, ensure that you filter the data. The goal is to predict the tip amount, therefore you need to ensure that you extract the data that contains this information. Hint: read the data dictionary.

- **Normalize the data**: perform either max-min normalization or z-score standardization on the continuous variables/features.

- **Encode the data**: determine if there are any categorical variables that need to be encoded and perform the encoding.

- **Prepare the data for modeling**: shuffle the data and split it into training and test sets. The percent split between the training and test set is your decision. However, clearly indicate the reason.

Question 3 — (30 points)

CRISP-DM: Modeling

- In this step you will develop the k-nn regression model. Create a function with the following name and arguments: **knn.predict(data_train, data_test, k)**;

- **data_train** represents the observations in the training set,

- **data_test** represents the observations from the test set, and

- **k** is the selected value of k (i.e. the number of neighbors).

Perform the following logic inside the function:

- Implement the k-nn algorithm and use it to predict the tip amount for each observation in the test set i.e. data_test.
- **Note:** ***You are not required to implement the k-nn algorithm from scratch***. Therefore, this step may only involve providing the training set, the test set, and the value of k to your chosen k-nn library.
- Calculate the mean squared error (MSE) between the predictions from the k-nn model and the actual tip amount in the test set.
- The **knn-predict()** function should return the MSE.

Question 4 — (30 points)

CRISP-DM: Evaluation

- Determine the best value of k and visualize the MSE. This step requires selecting different values of k and evaluating which produced the lowest MSE. At a minimum, ensure that you perform the following:
- Provide at least 20 different values of k to the **knn.predict()** function (along with the training set and the test set).
  Tip: use a loop! Use a loop to call **knn.predict()** 20 times and in each iteration of the loop, provide a different value of k to **knn.predict()**. Ensure that you save the MSE that's returned.
- Create a line chart and plot each value of k on the x-axis and the corresponding MSE on the y-axis. Explain the chart and determine which value of k is more suitable and why.
- What are your thoughts on the model that you developed and the accuracy of its predictions? Would you advocate for its use to predict the tip amount of future trips? Explain your answer.

Question 5 — (10 optional/bonus points)

In this optional (bonus) question, you can: 1) use your intuition to create a compelling visualization that tells an informative story about one aspect of the dataset OR 2) optimize the k-nn model and evaluate the effect of the percentage split, between the training and test set, on the MSE. **Choose ONE of the following:**

- Create a compelling visualization that tells <u>an informative story</u> about how these cabs are used.

OR

- Evaluate the effect of the percentage split for the training and test sets and determine if a different split ratio improves your model's ability to make better predictions.

Ensure that you perform the steps of the bonus question in a new R chunk!

<u>Note: all charts that are displayed should have the following</u>:

- An informative title (and subtitle if applicable)

- Labels on the x-axis and y-axis that indicate the units of measurement.
- A caption that indicates the purpose of the chart.

Submission Details

- This practicum contains bonus points that can contribute to your practicum average.
- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.P3.FirstName.LastName.Rmd and your PDF/HTML DA5020.P3.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name. If you are submitting for a team, use the name of the person who is submitting.
- • The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.
- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the .Rmd file (or both) will result in a score of 0.