ENGG2112 Coding Assignment

Due on 23 April 2023, 11.59pm

10 APRIL 2023

Instructions

- This is an individual assignment and the submitted work must be your original work. You are allowed to discuss the method of solution with others, however the submitted code must be entirely written by you.
- Submit your work as a Python notebook in the template provided.
- Submissions must be made through Canvas only, and not by e-mail. The deadline will be strictly enforced: 11:59pm on 23 April 2023. (Students with disability adjustments will be contacted separately.)
- Please plan your time according to your own ability and schedule, seek help from the teaching team and peers in a timely fashion, and try not to ask for deadline extensions.
- Download the file House_Rent_Dataset.csv from the Canvas website. Use the notebook template ENGG 2112 coding assignment 2023 (1).ipynb to answer the following questions.

Description of the Dataset

In this Dataset, we have information on almost 4700+ Houses/Apartments/Flats Available for Rent with different parameters like BHK, Rent, Size, No. of Floors, Area Type, Area Locality, City, Furnishing Status, Type of Tenant Preferred, No. of Bathrooms, Point of Contact.

Dataset Glossary

BHK Number of Bedrooms, Hall, Kitchen

Rent Weekly Rent of the Property

Size Size of the Property in Square Feet

Floor Floor location of property and total number of floors in building (e.g. Ground out of 2, 3 out of 5, etc.)

Area Type Size of the property calculated on either Super Area, Carpet Area or Build Area.

Area Locality Locality of the Property

City City where the Property is located

Furnishing Status Furnished, semi-furnished or unfurnished

Tenant Preferred Type of tenant preferred by the owner or agent

Bathroom Number of bathrooms

Point of Contact Person to contact for more information

Problem 1

(This problem is worth 2 marks in each part, 6 marks in total.)

1. Find the minimum, maximum and average rent in the entire dataset. Assign these values to the variables `rent_min`, `rent_max` and `rent_avg` respectively.

2. Find the subset of data records that satisfies the following conditions:

   - The posted date is in June 2022 (i.e. 1st to 30th June 2022).
   - Information on the property should be obtained from the agent.
   - The size of the property is at least 1,000 square feet.

   Create the dataframe `df_q2` to hold the data, and determine the number of eligible records/samples. Put this value in the variable `num_rows`.

3. In the first cell, plot a histogram of property sizes. In the second cell, plot a scatter plot of property size versus date of posting. Ensure that the date is in ascending (i.e. chronological) order.

Problem 2

1. (3 marks) Use the columns “BHK", “Size", “Area Type" and “Bathroom" to build a linear regression model to predict the rent of a property. Convert all the categorical data into binary variables using one-hot encoding. Use 75% of the data for training, with the random state set to 2112. Find the coefficient of determination R^2 and the mean squared error, and store these values in the variables R2 and mse respectively.

2. (6 marks) Use the columns “BHK", “Size", “Floor", “Area Type" and “Bathroom" to build the following three classifiers to predict the furnished status of the property:

   a) Logistic regression with max_iter = 1000.

   b) Multi-layer perceptron with one hidden layer of 100 neurons, maximum number of iterations = 500, and random state = 2112.

   c) Gaussian Naïve Bayes

   Process the data as in Problem 2.1 above. In addition, for the “Floor" column, extract the information to two new columns “Floor_new" and “Total_floor”, containing the floor location of the property and the total number of floors in the building, respectively. Transform the “Floor” information as follows: Ground $\rightarrow 0$, Upper Basement $\rightarrow -1$, Lower Basement $\rightarrow -2$. Insert the two new columns into the dataframe and delete the original column “Floor".

   Compare the performance of the three classifiers on the test data. The evaluation metrics are f1 score and accuracy, both stored in the variable result.

Problem 3

(5 marks) Use the columns “BHK”, “Size” and “Bathroom” in a K-nearest neighbours (KNN) predictor of rent and furnished status. The model needs to be built from scratch, i.e. without using any pre-packaged function that implements KNN in existing libraries. The data should first be pre-processed using min-max normalization, i.e. replace each feature x_i, with minimum and maximum values across the dataset of $x_{i,\min}$ and $x_{i,\max}$, with the normalized feature

$$\tilde{x}_i = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}}.$$

Test your function knn using the two sample data records provided in the last two cells of the Python answer notebook.