CO3093 - Big Data and Predictive Analytics

CW Assignment

Assessment Information

Assessment Number	2
Contribution to overall mark	70%
Submission Deadline	28/03/2025 at 12:00

Assessed Learning Outcomes

This second assessment aims at testing your ability to

- carry out data cleansing and visualisation
- develop a predictive model and evaluate its performance
- perform appropriate and justified clustering of the data for local regressions
- communicate your findings on the data

How to submit

For this assignment, you need to submit the followings:

1. A short report (about 8 pages in pdf including all the graphs) on your findings in exploring the given dataset, a description of your model and its evaluation, a description of your clusters and its justification, local regressors based on your clusters as well as their evaluation.
2. The Python source code written in order to complete the tasks set in the paper. You should submit the Python code file with your username(s) in the name (e.g. if you are working individually submit a file called `ab123_solution.py` or `ab123_solution.ipynb`; if you are working in pairs only **one** of you should submit a file with both usernames: `ab123_cd456_solution.py` or `ab123_cd456_solution.ipynb`).
3. A signed coursework cover – this should include the names of all the students involved in the work submitted.

Please put your source code, report and signed coursework cover into a zip file `CW2_ab123.zip` (or `CW2_ab123_cd456.zip` if you are working in pairs) and then submit your assignment through the module's Blackboard site by the deadline. Note that to submit, you need to click on the Coursework link on Blackboard and then upload your zipped file. Remember it is **1 submission per group**!

Problem Statement

Consider the property dataset that records prices (per night) of AirBnb properties in London, last updated in June 2024. For your convenience, the dataset can be downloaded from Blackboard. The data include AirBnb prices, neighbourhood, room type category (e.g., private room, entire home,...), number of beds, review score rating, and other types of variables. ***The dataset `london_listings.csv` can be downloaded from Blackboard.***

Objective

Using the given dataset, we would like to build up a model that can predict the prices (per night) of AirBnb properties in terms of some relevant features in the dataset and use a K-Means approach to propose a nontrivial set of houses' clusters, which may improve the performance of the regression model by proposing clusters-based (or local) regression models.

Exploring the Data

Your first task is to prepare the data and carry out data munging or cleansing, bearing in mind the question you would like to answer. For example, what is the impact of neighbourhood, number of rooms, customer ratings, or other features on the AirBnb prices.

Part 1: Building up a basic predictive model

Load the dataset `london_listings.csv` into a pandas dataframe and carry out the following tasks. Organise your code bearing in mind robustness and maintainability.

Data cleaning and transformation

If you have a closer look at the dataset, you will see that there are missing values. We need to treat them and in this first model, we are going to follow a basic strategy, which you will improve for a better predictive model later on:

- Show the shape of the dataframe
- Create list of categorical variables and another for the numerical variables
- For price column, remove the `$`, or other non numerical characters, and then convert them to numeric.
- For each categorical variable, remove empty strings or lists and replace them with Nan
- Show a summary of all missing values as well as the summary statistics
- Drop unnecessary columns (e.g. `calendar_last_scraped`, `bathrooms_text`, `latitude`, `longitude`).
- Drop duplicates if any
- Drop rows with NaN values
- Identify and remove outliers if any
- Show the shape of the resulting dataframe
- Consider the log of the prices and normalise the data.

Data exploration

Consider the resulting dataframe. This first aggressive cleaning should give a smaller dataset, which you can start exploring by looking at relationships between the various features of the dataset.

- Visualise the prices accross neighbourhood
- Visualise the prices across number of possible tenants
- Does the average review rating affect the process?
- Show the scatter matrix plot and the correlation matrix
- Any further plots, which demonstrate your understanding of the data

Model building

Consider the resulting dataframe.

- Select the predictors that would have impact in predicting the AirBnb prices.
- Build up an initial linear model with appropriate predictors and evaluate it. Split the data into training and test sets; build up the model; and then show a histogram of the residuals.
- Evaluate your model by using a cross-validation procedure

Part 2: Improved model

This is an open-ended question and you are free to push your problem-solving skills in order to build up a useful model with higher performance.

1. Use the K-Means algorithm to cluster your cleansed dataset and compare the obtained clusters with the distribution found in the data. Justify your clustering and visualise your clusters as appropriate.
2. Build up local regressors based on your clustering and discuss how this clusters-based model regression compares to your regression model obtained in Part 2.1.
3. Consider the entire dataset given in this assignment. Develop an improved predictive model that predicts the prices of AirBnb properties. Make sure to validate your model. You should aim for a model with a higher performance while using a maximum of data points. This implies using more advanced models, different feature engineering, etc.

Marking Criteria

The following areas are assessed:

1. Cleansing, visualizing, and understanding the data **[30 marks]**
2. Building up and evaluating the predictive model **[20 marks]**
3. Improved model, clustering and evaluating/justifying cluster-based model **[20 marks]**
4. Coding style **[10 marks]**
5. Writing the report (about 8 pages) interpreting the results. **[20 marks]**

Indicative weights on the assessed learning outcomes are given above and can be found in the marking rubric on Blackboard. The following is a guide for the marking:

- **First++ (≥ 90 marks):** As in **First+** plus a classification model with excellent performance, excellent justification and visualisation of the clusters, great insights from the data, and a report of professional standards.
- **First+ (≥ 80 marks):** As in **First** plus a comprehensive coverage of data cleansing techniques leading demonstrating an excellent understanding of the data, a classification model of high performance and a well-structured, maintainable, and robust code usefully using functions.
- **First (≥ 70 marks):** As in **Second Upper** plus a well-justified predictive model by the data cleansing with sound evaluation techniques; well-justified clusters and a concise report containing any decisions that may be recommended.
- **Second Upper (60 to 69 marks):** A good coverage of data cleansing techniques exploring the dataset, a good visualisation of the clusters, a predictive model with an appreciable accuracy with a rationale behind it, a working code and a wellstructured report on the results obtained from the dataset.
- **Second Lower (50 to 59 marks):** Some techniques used for data cleansing are overlooked, a predictive model partially justified with an appreciable accuracy, a working clustering, a partially commented code with very few functions, and a narrative of the findings about the dataset with few deficiencies.
- **Third (40 to 49 marks):** Essential data cleansing techniques are covered, a predictive model is given with some justification, a working but basic block code with no clustering, and a written report describing some of the work done.
- **Fail (≤ 39 marks):** Doesn't satisfy the pass criteria and will still get some marks in most cases.
- **No submission:** A mark of 0 will be awarded.

Marking Group Work

If you work in pairs, both members will normally be given the same mark unless one member made little or no contributions. Any pair can be called for an interview during marking. Both students **must attend**, explain their contributions, and defend the work submitted.