CO(7)3093 - Big Data & Predictive Analytics
Coursework 1 - Model building & Evaluation

Department of Informatics
University of Leicester

Assessment Information

Assessment Number	1
Contribution to Overall Mark	40%
Submission deadline	24 February 2020

Assessed Learning Outcomes

This second assessment aims at testing your ability to

- cleanse and visualise a dataset
- build up a predictive model and evaluate its performance
- communicate your findings on the data

How to submit

For this assignment, you need to submit the followings:

1. A short report (up to six pages) on your model building process for the given dataset and objective.
2. The Python source code written in order to complete the tasks set in the paper. It is recommended to submit a single Python code file, say `my_solution.py` for all the questions you have answered.

Please put your source codes, report and signed coursework cover into a zip file `CW1_YouremailID.zip` (e.g., `CW1_emt12.zip`) and then submit your assessment through the module's Blackboard site by the deadline.

Dataset

Consider this Housing dataset that records sale prices of properties in Manhattan from August 2012 to August 2013. For your convenience, the dataset can be downloaded from Blackboard along with a glossary of terms used in the data. The data include sale prices, neighborhood, building type category, square footage, and other types of variables.

Objective: Using the given dataset, we would like to build up a model that can predict the sale prices of houses in terms of some relevant features in the dataset.

Exploratory Data Analysis

Your first major task is to prepare the data – load the dataset and carry out data munging or cleansing. Answer the following questions:

- Load the data and conduct a descriptive analysis of the data. In particular, show a summary statitics and a summary of possible missing values in the dataset.
- Inspect the data types in the dataset; ensure that values you think are numerical are being treated as such and that dates are formatted correctly.
- Visualise the house prices and try to look for meaningful patterns in the dataset. For example, you may want to make comparisons accross neighborhoods or accross time.
- Cleanse the data by treating missing values and possible outliers. You may need to again visualise features of interest and their impact to the house prices. For example, a scatter matrix plot may help identify how some features correlate with house prices.
- Check the minimum and maximum values of numerical data. If the variation is too high in any column, you may carry out data normalization or scaling.
- Highlight features that may be more relevant for the house prices than others.

Model Building & Evaluation

1. Construct a linear model of sale prices with any predictors you feel are relevant. Make sure to withhold a subset of the data for testing. Justify why your model is appropriate to use.
2. Write down the mathematical equation of your fitted model and evaluate your model. You should aim for a model with a higher accuracy. To measure the accuracy, you may use the mean squarred error or the residual standard error. Show the graph of the predicted prices versus the actual prices.

Report

Write a short report (up to 6 pages) on your analyses of the given dataset. Your report should include a descriptive analysis, data visualisation & cleansing, and model building & evaluation. Discuss any decision that could be made or action that could be taken from this analysis.

Mark Scheme

The following areas are assessed:

1. exploring and visualising the dataset [40 marks]
2. building up and evaluating the predictive model [25 marks]

3. coding style [15 marks]

4. writing a report interpreting the results [20 marks]

Indicative weights on the assessed learning outcomes are given above. Note that for marking purpose, after the submission each student will show their work during the following Lab session and answer questions about their work. The following is a guide for the marking:

- First* ($\geq$ 86 to 100 marks): A complete coverage of data munging techniques exploring the dataset, a predictive model with a very high accuracy, a well documented and robust code, and a well written and structured report on the results obtained from the dataset and any decisions that may be recommended.

- First ($\geq$ 70 to 85 marks): A complete coverage of data munging techniques exploring the dataset, a predictive model with a good accuracy, a robust code, and a well written and structured report on the results obtained from the dataset and any decisions that may be recommended.

- Second Upper ($\geq$ 60 to 69 marks): A good coverage of data munging techniques exploring the dataset, a predictive model with an appreciable accuracy, a working code, and a well structured report on the results obtained from the dataset.

- Second Lower ($\geq$ 50 to 59 marks): Some techniques used for data munging are overlooked, a predictive model partially justified with an appreciable accuracy, and a good report of the findings about the dataset with few deficiencies.

- Third ($\geq$ 40 to 49 marks): Essential munging data techniques are covered, a predictive model is given with some justification, and a written report describing some of the work done.

- Fail ($\leq$ 39 marks): Not satisfy the pass criteria and will still get some marks in most cases.

- None-submission: A mark of 0 will be awarded.