COMP 309/AIML421 — *Machine Learning Tools and Techniques*

Assignment 2: Data Exploration, Manipulation and Modelling

15% of Final Mark — Due: 11:59pm Wednesday 21st August 2024

1 Objectives

The goal of this assignment is to help you understand data manipulation and visualisation tools for machine learning. The purpose is to implement common data handling methods on real-world observations. To validate the effectiveness of the implemented methods, you are also required to perform data analysis tasks to draw useful conclusions. In particular, the following topics should be reviewed:

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- Exploratory Data Analysis (EDA)
- Data Preparation
- Feature Selection and Dimensionality Reduction

These topics are (to be) covered in weeks 4-6, but will also involve content from previous weeks. Research into online resources for AI and machine learning is encouraged. You are required to complete the following questions using data mining/machine learning tools introduced in the lectures - *Python* and/or *Orange*. For each part, make sure you finish reading all the questions before you start working on it, and your report for the whole assignment should *not exceed 8 pages* (note that this is a *maximum*, not a goal/target) with a font size no smaller than 10.

2 Question Description

In this assignment, we are working on building machine learning models to predict house prices. The information contained in the given dataset is what a typical home buyer would want to know before making a purchase. More specifically, with *79 explanatory variables/features* describing different aspects of residential homes such as location, number of room, neighborhood, etc, you are required to predict the final price of homes.

What makes the house price in a city high or low? In this assignment, let's try to answer this question by finding the most important features that affect a house price in the provided dataset. Your task in this assignment is to use data to discern the relationship between the house price ("SalePrice") with the given features. *The dataset and the description of the variables/features are available from the Assignments page of our course website.*

2.1 Part 1: Business and Data Understanding [40 marks]

The first part of this assignment is to explore the data. The task is to use CRISP-DM, EDA and data manipulation to define the machine learning tasks, understand, and prepare the data.

You should:

1. (20 marks) Perform an **initial EDA** on the given data to gain an understanding of the data. The analyses should explore the data from four different aspects including:

(a) Describe the *summary statistics* about the data including number of instances, number of features, how many categorical and numerical features, respectively.

(b) Find the *top 5 numerical features* highly correlated with the target variable ("SalePrice") according to the *pearson correlation*, report the correlation values.

(c) Plot the distributions of these 5 numerical features found in the previous question and the target variable using *histograms*, one for each feature/variable, describe how to determine *the number of bins*, the shape of their distributions with *skewness* and *kurtosis* (use Scipy for obtaining skewness and kurtosis values), and tell *two patterns* from the histograms accordingly.

(d) Check for missing values. Are there any missing values in the data? write a paragraph to briefly summarise the missing information regarding how many features contain missing values and at what percent.

Provide answers to these four questions. Show how you get the answers in your code (if you use Python)/workflow (if you use orange). Report your EDA methods and results in the report.

2. (5 marks) Investigate the business understanding questions based on your exploration of the data. Two key business understanding questions (or business objectives) are "what factors affect the house price?" and "how do these factors affect the house price?"/"in which way do the factors affect the house price?"

   (a) Translate the two business questions into two data mining goals;

   (b) Select two machine learning paradigms, e.g. classification, regression, dimensionality reduction and so forth, that can help you achieve these goals. Provide justifications of your decision.

3. (15 marks) EDA using clustering is very useful for understanding the important characteristics of the data. Provide a **further EDA** on the dataset using Hierarchical clustering on the 5 numerical features found in 1(b) to answer the question — "Does the house prices vary by neighbourhood?". Report the output dendrogram and any other plots and show how do they help you to answer the question.

2.2 Part 2: Data Preparation and Machine Learning [60 marks]

Address the business question of "what factors affect the house price and in which way?" using the provided dataset. Note that for supervised learning, it is important to partition the data before data preparation to avoid data leakage. Before answering the following questions, you need to split the data into a training set and a test set with a 70-30 splitting (use a random state=309).

You should:

1. (20 marks) *Determine and describe* the data preprocessing steps applied to the provided dataset, e.g. handle missing data, encoding categorical data, normalise the data if necessary, and/or remove any unnecessary instances, these could be redundant instances, outliers or non-effective instances and so forth. Show the process in your code/workflow. Submit *copies of the processed training and test data* (in CSV format).

2. (15 marks) Utilise two different dimensionality reduction techniques to identify which features are irrelevant and/or redundant to predicting the house price. Report the dimension reduction process and remove redundant/irrelevant data. Show the process in your code/workflow.

3. (25 marks) Now approach data mining goals on your preprocessed data using machine learning methods.

   (a) With the *two groups* of features selected in previous question, use the *ordinary linear regression* and *ridge regression* (with *alpha=0.5*) for predicting the house prices. Comparing their results regarding the *mean squared errors* on the *training* set and the *test* set. Present and analyse the *learnt regression models*, and highlight your observations. Submit your code or workflow.

(b) Using *Random Forest*, which is a more powerful ensemble regression method to predict the house price, compare with the results of linear regression and ridge regression, highlight your observations.

2.3 Part 3: Further Analyses (For AIML421 students only) [20 marks]

Use *two other regression techniques* to build prediction models for the house price with the same preprocessed data using in Part 2. Discuss the prediction results, compared with the regression techniques used in Part 2 and identify which technique is more suitable for this question and provide your justifications.

Relevant Data Files and Program Files

A soft copy of this assignment, the relevant data files are available from the course home page.

Assessment

Format: You can use any font to write the report, with a minimum of single spacing and 11 point size (hand writing is not permitted unless with approval from the lecturers). Reports are expected to be 2-8 pages for COMP309 students and 2-9 pages for AIML421 students. Reports exceeding the maximum page limit will be penalised.

Communication: A key skill required of a scientist is the ability to communicate effectively. No matter the scientific merit of a report, if it is illegible, grammatically incorrect, mispunctuated, ambiguous, or contains misspellings, it is less effective and marks will be deducted.

Late Penalties: You have three automatic extension days, which can be applied to any assignments throughout the course (but not the final project). Please note that these three days are for the whole course, not for each assignment. No formal application is required; instead, any remaining late hours will be automatically deducted when submitting assignments after the due date. You have the flexibility to use only a portion of your late day and retain the remainder for future use. The penalty for assignments that are handed in late without prior arrangement (or use of "late days") is one grade reduction per day. Assignments that are more than one week late will not be marked.

Plagiarism: Plagiarism in programming (copying someone else's code) is just as serious as written plagiarism, and is treated accordingly. Make sure you explicitly write down where you got code from (and how much of it) if you use any other resources asides from the course material. Using excessive amounts of others' code may result in the loss of marks, but plagiarism could result in zero marks!

Submission

You are required to submit a single *.pdf* report PLUS the python code file (.ipynb or .py) and/or the orange workflow file (.ows) through the web submission system from the COMP309/AIML421 course website *by the due time*. Provide a *README.txt* file if you use any non-standard python libraries.