QBUS2820 Predictive Analytics
Semester 2, 2020

Individual Assignment 2

Key information

1. Required submissions:
    a. **ONE** written report (word or pdf format, through Canvas- Assignments- Report Submission (Individual Assignment 2)).
    b. **ONE** Jupyter Notebook .ipynb file (through Canvas- Assignments- Upload Your Program Code Files (Individual Assignment 2)).
    c. Kaggle competition- see Task A below for details.
2. Due date/time: **Friday 20-Nov-2020, 2pm** (report, Jupyter notebook submission and Kaggle competition **closure**). The Kaggle competition will be **closed** on this due date/time to make the competition fair to all students, while you can still submit your report and code file as late submission.
3. The late penalty for the assignment is 5% of the assigned mark per day, starting after 2pm on the due date. The closing date **Monday 30-November-2020, 2pm** is the last date on which an assessment will be accepted for marking.
4. Weight: **30%** of the final mark.
5. Length: The main text of your report (including Task A and Task B) should have a **maximum of 15 pages**. Especially for Task A, you should write a complete report including sections such as business context, problem formulation, data processing, EDA, and feature engineering, methodology, analysis, conclusions and limitations, etc, see "**Task_A_instructions**" file for more details.
6. If you wish to include additional material, you can do so by creating an appendix. There is no page limit for the appendix. Keep in mind that making good use of your audience's time is an essential business skill. Every sentence, table and figure have to count. Extraneous and/or wrong material will reduce your mark no matter the quality of the assignment.
7. **Anonymous marking**: As the anonymous marking policy of the University, please only include your student ID in the submitted report, and do **NOT** include your name. The file name of your report and code file should follow the following format. Replace "SID" with your Student ID. Example: **SIDQbus2820Assignment2S22020**.
8. Presentation of the assignment is part of the assignment. **Markers will assign 5 marks for clarity of writing and presentation**. Numbers with decimals should be reported to the **four-decimal point.**

Key rules:

- Carefully read the requirements for each part of the assignment.
- Please follow any further instructions announced on Canvas.
- You must use Python for the assignment. Use "*random_state= 1*" when needed, e.g. when using "*train_test_split*" function of Python. For all other parameters that are not specified in the questions, use the default values of corresponding Python functions.

- Reproducibility is fundamental in data analysis, so that you will be required to submit a Jupyter Notebook that generates your results. Not submitting your code will lead to **a loss of 50%** of the assignment marks.

- Failure to read information and follow instructions may lead to a loss of marks. Furthermore, note that it is your responsibility to be informed of the University of Sydney and Business School rules and guidelines, and follow them.

- Referencing: Harvard Referencing System. (You may find the details at: http://libguides.library.usyd.edu.au/c.php?g=508212&p=3476130)

Task A (35 Marks)

For this task, you can adapt and improve your **own** works from Assignment 1 Task A.

House Price Prediction

You will work on the **House Price Data.** The feature descriptions are presented in **"Kaggle_House_Data_Description".**

1. Problem description

Find the most appropriate predictive models of your choice to predict house prices from given data.

As a consultant working for a real estate analytics company, one client asked you to develop predictive models to predict house price based on machine learning techniques. To enable this task, you were provided with highly detailed housing data. The response variable is the "***SalePrice****"* column in the dataset.

As part of the work, you need to **write a report** according to the details below. The client will use a **test set** to evaluate your work.

2. Getting the data

Kaggle competition: you need to create a Kaggle account based on your **university e-mail address** in order to have access and make submissions. You name on the Kaggle competition must be your **SID**. **Students who do not following the naming rule on Kaggle will not be eligible for the bonus mark and may receive penalty for presentation.**

One student can only have one account, to make the number of submissions same & fair to all students.

The scoring metric is: **Root Mean Square Error (RMSE).**

The maximum submission per day is **10.**

You may select **2 final submissions** for final judging (the final private leader board rank). You can hand-select the eligible final submissions, or otherwise by default they will be your best public scoring submissions.

You can download the "*train.csv*" and "*test.csv*" data for the assignment and competition from Kaggle "Data" page, as well as read the further instructions, via the link below. "*sampleSubmission.csv*" is a sample submission file in the required format.

https://www.kaggle.com/t/f548e6df808c4920bdf26761c4f5ce36

Do **NOT** share this link outside of the unit.

3. Understanding the data

The information about the data is on the Kaggle page for the assignment. There are two data files, the training set and a validation-test set (simply called test on Kaggle). The latter omits the response "***SalePrice***" values. Kaggle randomly splits the observations in validation-test set into validation set (approximately **50%** of the test data) and test set (approximately **50%** of the test data), but you do not know which ones are in each set.

When you make a submission during the competition, you get a **public** score based on the **validation set**. The validation scores are visible to everyone and provide an ongoing ranking of submissions from students. At the end of the competition, Kaggle will rank the submissions based on the **test set only (private score)**. Be careful not to overfit the validation set in attempt to improve your ranking in the public leaderboard, as this may lead to a disappointing result for the test data (private leaderboard).

4. Written report

The purpose of the report is to describe, explain, and justify your solution to the client with polished presentation. Be concise and objective. Find ways to say more with less. When it doubts, put it in the appendix.

Requirement:

Your report must include the validation set (**public**) scores for **at least five different sets of predictions**, including your **final 2 best models**. You need to make a submission on Kaggle to get each validation score. You need to present your final 2 best models in details. For the other three methods, only brief explanations of the models are needed.

Suggested outline:

1. **(2 marks)** Introduction: write a few paragraphs stating the business problem and summarising your final solution and results. Use plain English and avoid technical language as much as possible in this section (it should be for a wide audience).
2. **(3 marks)** Data processing and exploratory data analysis: provide key information about the data, discuss potential issues, and highlight interesting facts that are useful for the rest of your analysis.
3. **(4 marks)** Feature engineering.

4. **(18 marks)** Methodology (present your final two best models, your rationale, how you fit them, model selection process, some interpretation, brief explanations of the other three models, etc). Note: you may try models that are not covered in the unit, while **at least one of the presented final two best models** must be the models that we have covered. If your Kaggle final two best models are both not covered in the unit, then your present your 3rd best or 4th best, etc.
5. **(5 marks)** Validation set Kaggle results.
6. **(3 marks)** Final analysis, conclusion, limitations and remarks (non-technical).

5. About Kaggle Competition

The purpose of the Kaggle competition is to incorporate feedback by allowing you to compare your performance with that of other students. Participation in the competition is part of the assessment. Your ranking in the competition will not affect your marks (apart from bonus marks and least performing deduction marks, as explained below), however we will assess your participation which represents the amount of genuine effort of producing good predictions and improving them.

Real world relevance:

The ability to perform in a Kaggle competition is highly valued by employers. Some employers go as far as to set up a **Kaggle competition** just for recruitment.

Bonus marks:

Students with most accurate predictions for the **test data (private leaderboard)** will receive bonus marks. The **top 5 ranked students** will get **bonus marks** for the assignment (in addition to the 65 marks of this assignment), with details as below.

Private leaderboard rank	**Bonus mark**
1st	15
2nd	13
3rd	11
4th	9
5th	7

Later, dependent on how the competition and team participation go, we might have a benchmark method. Students who have worse performance than the benchmark will potentially lose marks.

If multiple teams get same level of accuracy regarding the predictions, then the marker will decide the rank based on the metrics, including # of entries, methodologies employed, etc.

Task B (25 Marks)

In this task, you will use "*Visitors.csv*" data to **forecast 24 months of monthly number of visitors to a country following the last period** in the dataset.

Your objective in this assignment is to develop **univariate** forecasting models, i.e., only using the historical number of visitors, to address this problem.

You can download the dataset "*Visitors.csv*" from Canvas.

In this task, you need to:

- **(3 marks)** conduct exploratory data analysis
- **(20 marks)** select **two different forecasting models with justifications to complete the forecasting task**. At least one of the two forecasting models must be the models covered in the unit. For the presented two models, you need to present:
    - your rational,
    - methodology,
    - model diagnostics,
    - model validations,
    - forecasting results for 24 months of monthly number of visitors following the last period in the dataset.
- **(2 marks)** present conclusions, limitations and next step suggestions.