QBUS2820 Predictive Analytics
Semester 2, 2018

Assignment 1

Key information

Required submissions: Written report (word or pdf format, through Turnitin submission), predictions for the test data (through Kaggle), and Jupyter Notebook (through Ed). Group leader needs to submit the Written report and Jupyter Notebook.

Due date: Monday 24th September 2018, 2pm (report and Jupyter notebook submission. Kaggle competition closure). The late penalty for the assignment is 10% of the assigned mark per day, starting after 2pm on the due date. The closing date **Monday 1st October 2018, 2pm** is the last date on which an assessment will be accepted for marking.

Note the due date is extended compared to the one in the UoS outline, to release your pressure and avoid you have mid-exam and assignment 1 due date on the same day. The mid-exam date is centralized by the Uni which is not controlled by the unit coordinator. If you are planning to take a holiday during the Uni common vacation week, please start your assignment work earlier and make sure you meet the required timeline.

Weight: 20 out of 100 marks in your final grade.

Groups: You can complete the assignment in groups of up to **three students**. There are no exceptions: if there are more than three you need to split the group.

Length: The main text of your report (including Task 1 and Task 2) should have a maximum of 15 pages. Especially for Task 2, 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.

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 has to count. Extraneous and/or wrong material will reduce your mark no matter the quality of the assignment.

Anonymous marking: As the anonymous marking policy of the University, please only include your student ID and group ID in the submitted report, and do **NOT** include your name. The file name of your report should follow the following format. Replace "123" with your group SID. Example: Group123Qbus2820Assignment1S22018.

Presentation of the assignment is part of the assignment. Markers might assign up to **10%** of the mark for clarity of writing and presentation. Numbers with decimals should be reported to the **third decimal point.**

Key rules:

- Carefully read the requirements for each part of the assignment.
- Please follow any further instructions announced on Canvas, particularly for submissions.
- You must use Python for the assignment.
- Reproducibility is fundamental in data analysis, so that you will be required to submit a Jupyter Notebook that generates your results. Unfortunately, Turnitin does not accept multiple files, so that you will do this through Ed instead. 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 1 (20 Marks)

Instructions

You will work on the **Boston housing dataset**.

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.

Suppose you are interested in using the house AGE (proportion of owner-occupied units built prior to 1940) as the first feature $\boldsymbol{x_1}$ and the full-value property-tax rate TAX as the second feature $\boldsymbol{x_2}$**,** to predict the median value of owner-occupied homes in $1000's as the target $\boldsymbol{y}$. Write code to extract these two features and the target from the dataset.

(a) (5 marks) Before doing the regression, you would like to visualize the loss function to get a rough idea of the potential optimal values of the parameters. Use two chosen features and the target as the new dataset to plot the loss function (**MSE**):

$$L(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(x_i)\right)^2 \qquad \text{with} \quad \hat{f}(x_i) = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

That is, we are using a linear regression model without the intercept term β_0. Hint: This is a 3D plot and you will need to iterate over a range of β_1 and β_2 values.

(b) (10 marks) Based on slides from 27 to 32 of lecture 5, write your own Gradient Ascend algorithm (you may build your solution based on the template of tutorial 6 task), to estimate the parameters of the given regression problem without the intercept.

- Find the approximate range of acceptable learning rates and explain why some learning rates are defective.

- Find the optimal learning rate in this range and explain why this is the optimal value.

Use $\boldsymbol{\beta} = [0,0]^T$ as your initialization point.

(c) (5 marks) For this task, you can use the linear regression model *LinearRegression* in the *scikit-learn* package.

Use "*train_test_split*" function to split 80% of the data as your training data, and the remaining 20% as your testing data. Construct the **centred** training dataset by conducting the following steps in Python based on the training data:

(i) Calculate the mean of all the target values, then deduct this mean from each of target values. Take the resulting target values as the new target values $\mathbf{y}_{new}$;

(ii) Calculate the mean of all the first feature values, then deduct this mean from each of first feature values. Take the result as the new first feature $\boldsymbol{x}_{1new}$;

(iii) Do the same for the second feature. The result is $\boldsymbol{x}_{2new}$;

Now build linear regression, with and without the intercept respectively, to fit to the new data $\boldsymbol{x}_{1new}$, $\boldsymbol{x}_{2new}$ and $\mathbf{y}_{new}$. Report and compare the coefficients and the intercept.

Compare the predictive performance (using **MSE**) of two models over the testing data. Note that, when you take your testing data into the model to calculate predictive performance scores, you should transform the testing features and targets.

Task 2 (40 Marks)

Moneyball

You will work on the **NBA salary dataset**.

Note: This task does not require prior knowledge of basketball. You should not add any personal subjective assumptions about the data based on your existing knowledge, e.g. deciding which variables are important for the salary prediction. This can lead to inaccurate results. You should use the techniques that we learnt and you discovered to get good models and complete the prediction task.

NBA glossary link below and the glossary Table at the end of the file can help you understand the meaning of the variables better:
https://stats.nba.com/help/glossary

1. Problem description

Find the most appropriate predictive models of your choice to predict NBA player salary from performance statistics.

As a consultant working for a sports data analytics company, the NBA league approached you to develop predictive models to predict NBA salaries based on state-of-art techniques from predictive analytics. To enable this task, you were provided with a dataset containing highly detailed performance of the NBA players. The response is the **SALARY($Millions**) column in the dataset.

As part of the contract, 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. Under Kaggle "Team" page, you can form teams on Kaggle, and the team must have **the same group members** as in Canvas. The group leader should create a team on Kaggle which must use the Canvas **group ID as the team name**, e.g. Group 16. The team leader can then invite your team members. Then all group members should be able to submit prediction results.

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 will otherwise default to the 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/d28207ceaf8b4ebbac38c613fee4afae
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 target values. Kaggle randomly splits the observations in validation-test data into validation (approximately 30% of the test data) and test cases (approximately 70% 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 groups. At the end of the competition, Kaggle will rank the groups based on the **test data 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. You can assume that the client has training in business analytics. The client's time is important. The client has a limited attention span. The client is not interested in minor details. 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 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 additional methods, only brief explanations of the models are needed.

Suggested outline:

1. 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. 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. Feature engineering.
4. Methodology (present your final 2 best models, your rationale, how you fit them, some interpretation, etc). Note: you may try models that are not covered in the lecture, while **at least one of the presented models** must be the models that we have covered. If your final 2 best models are both not covered in the unit, then your present your 3rd best or 4th best, etc.
5. Validation set Kaggle results and comparison with other approaches that you have tried.
6. 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 groups. 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:

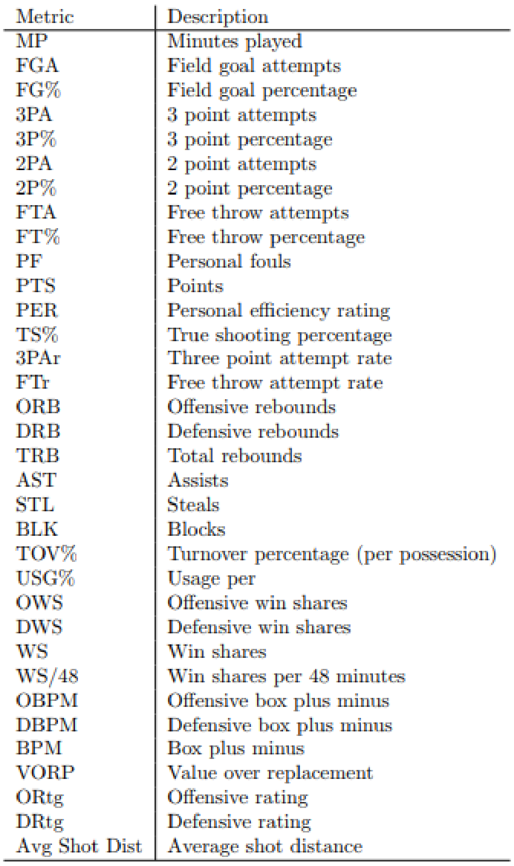
The ability to perform in a Kaggle competition is highly valued by employers. Some employers go as far as to set up a **Kaggle competition** just for recruitment.

Bonus marks:

The three teams with most accurate predictions for the **test data** will receive bonus marks. The first place will get 5 bonus marks for the unit (out of 100 final marks), the second place 3 marks, and third place 1 mark.

Later, dependent on how the competition and team participation go, we might have a benchmark method. Groups which 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.



Metric	Description
MP	Minutes played
FGA	Field goal attempts
FG%	Field goal percentage
3PA	3 point attempts
3P%	3 point percentage
2PA	2 point attempts
2P%	2 point percentage
FTA	Free throw attempts
FT%	Free throw percentage
PF	Personal fouls
PTS	Points
PER	Personal efficiency rating
TS%	True shooting percentage
3PAr	Three point attempt rate
FTr	Free throw attempt rate
ORB	Offensive rebounds
DRB	Defensive rebounds
TRB	Total rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV%	Turnover percentage (per possession)
USG%	Usage per
OWS	Offensive win shares
DWS	Defensive win shares
WS	Win shares
WS/48	Win shares per 48 minutes
OBPM	Offensive box plus minus
DBPM	Defensive box plus minus
BPM	Box plus minus
VORP	Value over replacement
ORtg	Offensive rating
DRtg	Defensive rating
Avg Shot Dist	Average shot distance

Sports Reference LLC, 2016a.