QBUS6850 Assignment 2

Due date/time: 2:00pm, Monday, 13-May-2019
Value: 10% of the final mark

Notes to Students

1. **Required submissions: ONE** written report (word or pdf format, through Canvas- Assignments- Report Submission (Assignment 2)) and **ONE** Jupyter Notebook .ipynb or .py code file (through Canvas- Assignments- Upload Your Program Code Files (Assignment 2)).
2. The assignment is due at **2:00pm on Monday, 13-May-2019**. The late penalty for the assignment is 5% of the assigned mark per day, starting after 2:00pm on the due date. The closing date **2:00pm on Monday, 20-May-2019** is the last date on which an assessment will be accepted for marking.
3. As anonymous marking policy, only include your Students ID in the report and do **NOT include your name**.
4. The name of the report and code file must follow: **123456_QBUS6850_Assignment2_ S12019. Replace "123456" with your actual student ID.**
5. Your answers shall be provided as a report giving full explanation and interpretation of any results you obtain. Output without explanation will receive **zero** marks.
6. Be warned that plagiarism between individuals is always obvious to the markers of the assignment and can be easily detected by Turnitin.
7. The data sets for this assignment can be downloaded from Canvas.
8. Presentation of the assignment is part of the assignment. **Markers will allocate up to 10% of the mark for writing in clarity and presentation**. You may insert small section of your code into the report for better interpretation when necessary. Think about the best and most structured way to present your work, summarise the procedures implemented, support your results/findings and prove the originality of your work.
9. Numbers with decimals should be reported to the **four-decimal point**.
10. The report should be **NOT more than 10 pages** including everything like text, figure, tables and small sections of inserted codes, etc, but excluding the appendix.
    Especially for Task B, you should write a complete report including sections such as business context, problem formulation, data processing, EDA, feature engineering, methodology, analysis, conclusions and limitations, etc.

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.
- Use "***random_state= 0***" 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 code file 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)

Tasks

Task A. User Comments Classification (25 Marks)

Your goal is to build a Random Forest (RF) classifier that classifies whether a **user comment is spam or not**.

Use the user comments dataset "*User_Comments.csv*" which contains the user comments of Youtube videos. Use "*train_test_split*" function to split 80% of the data as your training data, and the remaining 20% as your testing data.

General instructions:

1. "*CLASS*" in the data is the target variable.
2. 5-fold cross validation if needed.
3. Make sure set your random number generator seed to 0 for this question: "*np.random.seed(0)*".

(a) Use the following Python package:

from sklearn.feature_extraction.text import TfidfVectorizer

Build a bag of words representation of the "*CONTENT*" column with:

- Max 500 features
- Remove the top 3% of frequently occurring words
- A word must occur at least twice to be included as a feature
- Remove common English words

(b) Build a random forest classifier and use cross validation to optimise the parameters of the random forest. You need to at least optimise the number of trees in the random forest and can explore and optimise other parameters as well.

Use the following Python packages:

from sklearn import ensemble

from sklearn.model_selection import GridSearchCV

With your CV selected optimal parameters' values, re-train the RF on the full training set and produce your best performing model. Report your random forest settings that achieve the best classification.

Test your best performing model on the test set, and you must achieve an average score ("*avg / total*") of **at least 0.97** for precision, recall and f1-score of "*sklearn classification_report*". Report your "*sklearn classification_report*" output.

(c) Based on your cross validation results from *GridSearchCV*, plot the "*mean_test_score*" and "*mean_train_score*" vs number of trees on the same Figure.

If you optimised other parameters, then fix these parameters to their optimal values.

(d) Produce a histogram of the depths of the trees of your best performing model.

(e) Report the top 10 most important text features of your best performing model.

Task B. Moneyball (35 Marks)

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. 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.

1. Problem description

Based on the models we have learnt from QBUS6850 unit, select 2 models to predict NBA player salary from performance statistics. Note: you may try models that are not covered in the lecture, while **at least one of the presented models** must be the model that we have covered in QBUS6850 unit.

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 machine learning techniques. To enable this task, you were provided with a dataset containing highly detailed performance of the NBA players.

As part of the contract, you need to **write a report** according to the details below. You can use the given test set to evaluate the performance of your work.

The performance/scoring metric is: **Root Mean Squared Error (RMSE).**

2. Understanding the data

You can download the "*NBA_Train.csv*" and "*NBA_test.csv*" data for the Canvas. The response is the **SALARY($Millions)** column in the dataset. The target variable "SALARY" is omitted in the test set.

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

3. 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.

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 selected 2 models, your rationale, how you fit them, some interpretation, etc).
5. Report and interpret the test set performance.
6. Final analysis, conclusion, limitations and future work suggestions.

See the NBA glossary Table on the next page.

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.