CO7093/CO3093 - Big Data & Predictive Analytics
COURSEWORK 1
Simulation-based & Regression Models

Department of Informatics
University of Leicester

Assessment Information

Assessment Number	1
Contribution to Overall Mark	30%
Submission deadline	23 February 2018

Assessed Learning Outcomes

This second assessment aims at testing your ability to

- build up a predictive model based on probability distributions
- analyse a dataset to gain understanding of the data from within Python
- build up a regression model and evaluate it
- communicate your findings on your predictive model

How to submit

For this assignment, you need to submit the followings:

1. A short report (in .pdf) on your findings in exploring the given datasets, a description of your models and their evaluations, as well as any decisions or actions that may be taken following your analyses.
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.
3. A signed coursework cover

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

1 Predicting the Winner

Consider the Premier League dataset, which records the results of the Premier league matches thus far during the current season 2017/18. The data include Full Time Result, Full Time Home/ Away Team Goals, Half Time Home/Away Team Goals and other types of variables. Please have a look at the given notes along with the dataset in order to understand the abbreviations used in the dataset.

Objective: Using the given dataset, we would like to build up a model that can predict the winning team of the next premier league match between Manchester United and Manchester City by using simulation and a historical dataset.

Manchester City vs Manchester United

Analyse the given dataset in order to show the followings:

1. Check for missing values in the dataset, drop columns that may be irrelevant to the problem of predicting the winning team and provide a descriptive analysis of the dataset.
2. Extract all home and away matches played by both teams as well as the number of goals scored or conceded. You may present the results in the form of data frames.
3. To get a better picture of how Manchester United and Manchester City stack up against each other, juxtapose the teams' offensive and defensive performance data. For that purpose, plot the goal scores frequency of the Manchester City's away offense against Manchester United's home defense and the Manchester City's away offense against Manchester City's home defense.

Simulation

To predict the winning team, we can create fantasy games with an objective to estimate the probability that one team will beat another.

1. Use empirical distributions of goals scored by the two teams to predict the winning team by simulating a large enough fantasy games between both teams. Handle possible draws appropriately.
2. A balanced simulation should consider both the offensive and defensive performance of each team. Perform a balanced simulation of the match between both teams in order to predict the winning team. Handle possible draws appropriately.
3. Use and justify theoretical probability distributions to simulate Manchester City-Manchester United's games as paired random drawings from a theoretical probability distribution. Then, execute a balanced simulation of offensive and defensive performance with your probability models for goals scored. Handle possible draws appropriately.

2 Predicting a Stock price

Consider the AAPL stock, which records the daily AAPL stock prices from 1980s to date. The data include the Open – price when the market opens, High – the highest price on the day, Low – the lowest price of the day, Volume – the amount of stocks traded, Close –the price when the market closes, and the Adjusted close – the adjusted stock price to account for stock splits that could have occurred.

Objective: Using the given dataset, we would like to build up a model that can predict the price of the stock for the next *five* days.

1. Create the time series of the given stock prices. You should consider the adjusted close prices. Comment on the graph obtained spotting trends or possible sharp price changes.
2. Construct a predictive model of stock prices with any predictors you feel are relevant. You may introduce additional attributes into the dataset e.g. moving averages, see `https://en.wikipedia.org/wiki/Moving_average`, Bollinger bands, see `https://en.wikipedia.org/wiki/Bollinger_Bands`, etc. Justify why your model is appropriate to use.
3. Write down the mathematical equation of your fitted model and evaluate your model. Make sure to withhold a subset of the data for testing. You should aim for a model with a higher accuracy.
4. Include in your report a discussion if you could make any money with your predictive model.

Mark Scheme

The following areas are assessed:

1. Man. City vs. Man. U. model + justification + evaluation [30 marks]
2. AAPL Stock prices prediction model + justification + evaluation [30 marks]
3. Quality of coding [20 marks]
4. writing a report (up to 5 pages including graphs) interpreting the results [20 marks]

Indicative weights on the assessed learning outcomes are given above. The following is a guide for the marking:

- First ($\geq$ 70 to 100 marks): A complete coverage of data science techniques exploring the dataset; both predictive models are detailed and well justified along with the evaluation of the regression model and perhaps an attempt to evaluate how good your model for finding the winning team is; 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 science techniques exploring the dataset; both predictive models are justified with an appreciable accuracy for the regression model; and a well structured narrative on the results obtained from the dataset and any decisions that may be recommended.

- Second Lower ($\geq$ 50 to 59 marks): Some techniques used for model building and evaluation are overlooked; at least one predictive model partially justified with an appreciable accuracy is given; and a good narrative of the findings about the dataset with few deficiencies.
- Third ($\geq$ 40 to 49 marks): Essential data science techniques are covered; at least one 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.

Last Updated January 28, 2018 by Emmanuel Tadjouddine