OM 420: Predictive Business Analytics

Instruction for group project on Data Analytics

About data

These files provide detailed data about the circumstances of personal injury road accidents, the types of vehicles involved and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form. Information on damage-only accidents, with no human casualties or accidents on private roads or car parks are not included in this data.

A list of the variables contained in the files can be found at "variable-lookup.xls". As well as giving details of date, time and location, the accident file gives a summary of all reported vehicles and pedestrians involved in road accidents and the total number of casualties, by severity. Details in the casualty and vehicle files can be linked to the relevant accident by the "Accident_Index" field.

Objectives:

Objective 1: Explore the data and discover three interesting findings.

- You may use any approach we learned in exploratory data analysis lectures.

Objective 2: Build a meaningful prediction model.

1. Specify a target variable which you want to predict.
2. Build multiple prediction models for the target variable
    - ✓ Try at least three different types of models (KNN, LDA, Tree, Random Forest, etc).
    - ✓ For each model type, generate multiple models by using different sets of input variables.
    - ✓ Also, generate multiple models by changing parameters (if there is any).
3. Evaluate the generated models and make a comparison. Choose the best one, explain your reasoning, and report its final performance.
    - ✓ The process will involve trial and error. Do not expect to get an interesting finding or a good prediction model at your first try.
    - ✓ Do not forget that data preparation is the most important step in any data mining project.

Presenting Your Work

You will present your work in two forms, reflective presentation and a written report.

Reflective Presentation

- Discuss your overall approach in handling this project. For example, what are your findings? What are the procedures you had in deriving the findings? Your observations on the dataset? How many methods did you apply, and what are their corresponding results? What are the prediction models you created? Which model you think is the best? What additional information you may need to improve the prediction ability? What was the most difficult part in this project, and how did you overcome it? …… You do not need to mention all of the above – they are just examples.
- There is no need to show all of your detailed analyses/results in the presentation. You may present one part of your analyses as an example – it is up to your choice.
- The presentation will be evaluated on organization, presentation skills, evidence of in-depth analyses, and time management. These four criteria will be equally weighted.

Project Report

- Submit a report in PDF format, showing all your work done for this project.
- Do not exceed 7 pages (excluding appendix). Use font size 11. You should keep it as concise as possible, but at the same time include all the discussions, charts and findings you think are necessary.
- Appendix may include only the final form of your codes and screenshots of the results generated by the codes.
- You may also include screenshots of the results generated in the body text of the report, so choose where to include them as appropriate.
- Do not submit any original script file. Show all the necessary results in the report.
- Regarding Goal 1, present each of your findings as follows:
    - ✓ A well-written (and as short as possible) sentence that describes your finding.
    - ✓ A plot or a table that supports the validity of your finding. This plot or table must be easy to understand for someone who knows nothing about data mining.
    - ✓ A brief explanation of how you found it and why it is important and interesting.
- Regarding Goal 2, present your prediction analysis as follows:
    - ✓ The target variable must be a binary variable. For example you can define a new variable as “casualties_high” which is 1 if the number of casualties is more than or equal to 2, and 0 otherwise.

- ✓ A well-written (and as short as possible) sentence that describes your objective
- ✓ A brief explanation of your models and how you created them
- ✓ Evaluation of your models and comparison
- ✓ You can use one model which is not covered in the lectures (e.g., neural network, SVM, etc.). This is optional

- The report will be evaluated on the following. These four criteria will be equally weighted.
  - ✓ Interestingness: high if the potential impact of your finding or prediction is big
  - ✓ Validity: high if your evidence and analysis is convincing, low if it's not
  - ✓ Technique: high if it demonstrates your proficiency on data mining techniques
  - ✓ Writing: structure, logic, and writing skills.

Bonus mark – Each group can receive some bonus marks by raising questions for other groups