ITNPBD6 Data Analytics Assignment Spring 2021

World Of Bargains

Store Performance

Project Objective:

Develop logistic regression, decision tree and neural network models that will identify whether stores will perform well or poorly.

Context:

Ivor Buquetlowd, the owner of a chain of over 100 shops in the UK, would like you to help grow his business effectively. The shops are all similar, but they make amounts of revenue varying from £1 million to nearly £5 million per year. He would like a computerised system to analyse the performance of his existing shops to help choose new locations for shops.

Your Task:

You must develop logistic regression, decision tree and neural network models that will identify whether stores will perform well or poorly. You can use Orange, Python, R, or any data mining package of your choice. The data for the assignment is in a file storedata.csv included with this document. The datasets contains the details of 136 stores. The data describes the following aspects of each store:

- Town
- Store ID
- Manager name
- Staff numbers
- Floor Space and Window Space
- Car park (yes or no)
- Demographic score
- Location (Shopping Centre, High Street, Retail Park)
- 40 min, 30 min, 20 min and 10 min drive time population size
- Store age
- Clearance space in store
- Competition number (how many competing stores are near ours)
- Competition score (from how good the competing stores are)
- Performance: Whether Ivor regards the store's profitability as "Good" or "Bad"

Feature 'Performance' is the response variable. You must construct a model that can accurately predict 'Performance' from the other features.

Submission Requirement:

You should hand in a report describing the modelling process you followed and your results. You should attempt to frame the problem in the form of CRISP-DM framework to better facilitate the discussion. Refer to the relevant CRISP-DM stages at reach stage of your report. You do not need to submit code or data. The report is worth 100 marks in total and must cover the following:

Introduction [10 marks]

Describe the task you were given: is it classification or regression? describe the data you received and the requirements of the finished system, including why data mining is suitable for this task. Define any terminology that you will use in the report (for example, model, variable, task, etc.).

Data Summary [10 marks]

List the variables that you found in the file provided by the company. For each one, say whether it should be treated as categorical or numeric; nominal, ordinal, continuous or discrete; and whether or not it is likely to be of use in building the solution. Explain your decisions: if you rule out any variables at this stage, you can justify your choice using summary statistics, or a histogram plot of its distribution.

Data Preparation [15 marks]

Describe what you did with the data prior to the modelling process. Show histograms of the data before and after any pre-processing that you carried out. (you do not need to give histograms of all variables, just the ones that need some cleaning) If you corrected any mis-typed or corrupted entries in the data, report what you changed, such as any rules you used, or examples of specific data points that were cleaned.

Modelling [40 marks]

You must use three different techniques and build models with each: these should include one tree-based model, one based on logistic regression, and one based on neural networks. Try to make each model perform as well as it can: if you varied the hyperparameters of a model, show which hyperparameters you varied and how this impacted on the results. Describe how you split the data for training, validation and testing purposes. Be methodical and record each result. This stage is a little like scientific research – you are carrying out experiments in your search for the best solution. Once you have a solution, show how you verified its robustness. For the three different techniques report on their comparative ability to predict store performance, but only select a single model for the final test.

Don't try to find a perfect or extremely accurate model - one does not exist! We are interested in the procedure you followed and the justification you give for choosing particular model types/parameters/features.

Results and Errors [15 marks]

Analyse and describe the level of accuracy the model achieves and the errors your model makes. Show a confusion matrix for each model. Are there any areas of the data where it performs worse than in others, and are there any types of error that World of Bargains would want to avoid more than others? Show a lift curve or a ROC curve for the decision as to whether or not a shop might be profitable and explain what it tells you.

10 marks will also be allocated for adopting a structured approach to the whole process and identifying the six CRISP-DM stages.

Submission:

The deadline for this assignment is 12 noon on Friday 16 April 2021. Please submit your report via the iStirling submission system.

You do not need to submit the models that you built, just the report.

As a guideline, your report should be around 3000 words. However, this is not a strict limit and no penalties will be issued for reports outside this range – just write what you need to provide the required information clearly and concisely. You can assume that the client has a good technical understanding of data mining and statistics, so do not shy away from technical terms in your report. Where you use them, however, explain what they mean in plain language too.

To maximise your mark, make sure you follow the instructions above and include everything that is asked for in the report.