CIT Institiúid Teicneolaíochta Chorcaí
Cork Institute of Technology

Data Analytics and Visualisation

Data9005

Assignment 2

Due: 11th May (Monday) 2020, 11.59pm (GMT)

Please submit your assignment via Canvas.

This assignment is worth 50% of your module.

In your submission include a declaration page, further details will be given around this, if you want to submit before I give further details use the declaration text as given in Research Methods for assignment 2.

Standard late penalties apply.

Please do not share this dataset with anyone or publish it online.

Question 1

Daniel, the manager of a telecoms company wants to apply analytical techniques so that he can predict if a RF radio mast, used in mobile phone communication, is susceptible to outages due to adverse weather conditions. To this end he would like you to use appropriate algorithms and/or machine learning techniques to aid him with this task.

Daniel's company has a large nationwide network of RF radio masts and these are affected by 'outages' due to adverse weather conditions, mainly very heavy rain and heavy mist. Outages mean that there is no available mobile phone service.

To help you Daniel has commissioned an engineer to classify 2,186 radio masts are either *okay* – i.e. not susceptible to adverse weather conditions and *under*, i.e. under engineered and so susceptible to adverse weather conditions.

He has also supplied you with a scoring dataset of 936 radio masts for you to classify as whether *okay* or *under*.

The data sets are up on Canvas, in the assignment folder. The training data is called **RF_TrainingDatasetA_Final.xlsx,** and the scoring data **RF_ScoringDatasetA_Final.xlsx.** The data may need to be cleaned before you can use it for analysis/modelling purposes.

Data details

Both data sets contain the follows attributes; note the scoring data does not have the label **Eng_Class** attribute.

RFDBiD: This is a unique ID number for each radio mast.

Eng_Class: This is the label variable, either okay or under (engineered).

Antennafilename1, 2: These are indicator variables (i.e. not to be used in classification).

Outcome: This was used by the engineer as a double checking mechanism for Eng_Class, and the last 105 were not completed by the engineer due to time constraints.

AntennagaindBd1: etc these are 74 variables that define each radio link.

Further details (from the engineer): The suffix for many variables typically ends with the unit and a code 1 or 2 to represent each side of the RF link. Example "Antennagaindbd1", "Antennagaindbd2". these example attributes represent the antenna gain associated with each side of the RF link with units in dBd. Antenna gain is expressed as a dipole / isotropic.

Note for the most part links are engineered with symmetrical parameters for both sides of the RF link. For example the Antenna type may be the same at both sides. For other attributes however there may be differences in the configuration values associated with both sides of the RF link.

The attribute FrequencyMhz represents the frequency of operation for the link. From this it may be possible to create a model for each individual frequency band.

It may be possible to group and build models based on other attributes.

The goal is to try to create a model based on the training set, apply the model to the scoring set to predict what links are well engineered and what links are not so well engineered.

The following slideshare is quite detailed, but it will give you an idea of the type of elements involved in the link design.

http://www.slideshare.net/SAIFUUU/microwave-link-design [slide 10 & 21] - many of the details discussed match the attribute names in the dataset.

Note this is a data mining/machine learning task so this technical knowledge is given mainly to give you a context for the data set.

Please answer the following questions, referencing your work appropriately:

Using R answer the following questions, giving your code where appropriate. The use of RMarkdown is optional but recommended; you do not need to use RMarkdown for part g) below, you copy and paste this into your output or give it as a separate document. Make sure that the code you submit runs correctly, it is recommended to check this just before you submit. It is suggested to use the last 3 digits of your ID no. as a set.seed where necessary so your results are reproducible.

a) Investigate the data by carrying out some exploratory data analysis (EDA). Perform the necessary data cleaning/data reduction tasks and outline how you do this in R. Word count 750.

20 marks

b) Set up a training/testing methodology. Using a least 2 models, tune these models and compare your results. Give your best model and comment.

10 marks

c) Perform feature selection on your model in c). Explain how you do this, giving a rational and comment on your results.

10 marks

d) For your best model explain to Daniel how this model works. Give and explain the cost/loss function used in your modelling. Word count 750.

20 marks

e) Daniel is primarily concerned with finding the under engineered masts as these are the ones that cause outages, so incorrectly 'scoring' a mast as *under* when is it *okay* is not as bad as incorrectly 'scoring' a mast as *okay* when it is *under*; you can take the ratio here of misclassification 'costs' as 1:h, where h = {8, 16, 24}, i.e. h can take a value of 8, 16 or 24. Redo your modelling using your best model above and comment on your new results.

15 marks

f) Using the scoring data set provided predict whether these radio masts will be okay or under engineered using your best model to part d) and comment.

5 marks

g) Explain in your own words and using appropriate referencing **ANY TWO** of the following (400 – 750 words per topic), use diagrams where appropriate. Reference your work.

(i) How does convolution layers and max pooling actually work in CNNs for an image classifier with 3 dimensions, i.e. height, length and colour channels? Give also a broad overview understanding of what both layers do in deep CNN for image classification.

(ii) Explain what a 3D tensor is. For a ConvNet which has a 3D tensor as input, give the generalised mathematical equations, with full indexing, for a mid network convolutional layer, explaining what each term is, the input to the layer and how the output of this layer is connected to the input of the next layer.

(iii) Describe in detail a deep learning methodology that are used in either (you may not use the same methodology as you use in another assignment, e.g. Research Methods):

(1) text analytics

or

(2) time series analysis.

20 marks

[Total 100 marks]