UNIVERSITY OF ST ANDREWS

ID5059 Knowledge Discovery & Data Mining
School of Computer Science
School of Mathematics and Statistics

Tom Kelsey
Mar 2019

ID5059 GROUP PROJECT DESCRIPTION

Important dates

The due date for your individual and group project reports is 9pm on Friday the 19th April (end of week 10). Upload a zipfile containing your individual report, your group report and any supplementary information (code, spreadsheets, etc.) to Moodle. Each group member should submit the group project, thereby avoiding a single point of failure.

The group presentations will take place during the 11am lecture slots of week 11. Eleven groups at 10+ minutes per group will fill these slots.

1 Overview

This year groups have a choice of two projects.

1. A Kaggle competition, or
2. an online Machine Learning Course.

You can identify your fellow group members on MMS.

2 The Kaggle Project

The problem will be a completed competition on Kaggle. For those not familiar, Kaggle is a competition platform for predictive modelling problems, often with cash prizes. The data and description of the problem are available via Kaggle.com - you will have to register.

You will all be working on the same classifcation problem, which can be found on Kaggle as the 'Porto Seguro's Safe Driver Prediction' competition[1]. A description of this is paraphrased below:

> Nothing ruins the thrill of buying a brand new car more quickly than seeing your new insurance bill. The sting's even more painful when you know you're a good driver. It doesn't seem fair that you have to pay so much if you've been cautious on the road for years. Porto Seguro, one of Brazil's largest auto and homeowner insurance companies, completely agrees. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. In this competition, you're challenged to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. While Porto Seguro has used machine learning for the past 20 years, they're looking to Kaggle's machine learning community to explore new, more powerful methods. A more accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

In short this this is a classification problem. You will work in the teams you have been assigned to and submit models to the competition collectively.

2.1 Data

The data is available via Kaggle upon registering. There are 3:

1. *train.csv* This contains the data that you will be using to train your models. It has a full complement of reponse data (the target class) and covariates for some 600,000 observations.
2. *test.csv* This is the test data that you will apply your model to for competitive assessment. In short it has covariates but no response that you can see. You will make predictions from your model to this data, then upload the results to see where you are on the leader-board.
3. *sample_submission.csv* An example of the format that you submission must conform to for uploading to Kaggle.

[1] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/

2.2 Resources

- You can obtain the data from the Kaggle website.
- There is an amount of online discussion around all such competitions - you can get tips there.
- Google scholar - research has never been so easy!
- Use social networks to post questions and/or generally discuss the modelling with your group and class.

3 The Open Machine Learning Course

There are three Spring 2019 assignments[2] to go with the tutorials and resources supplied[3].

Your group will work on the first (Exploratory Data Analysis of US flights) and third (Decision trees, Random Forest, and gradient boosting) of these. (The second assignment involves previous assignments that we haven't been part of.)

The demo assignments on the same web page should provide useful experience and insights for the deliverable assignments, but you do not have to complete these or report on them.

3.1 Resources

- Any of the material supplied by the Open Machine Learning Course.
- There are other groups and individuals working on this course at the moment, and there will be online discussions - you can get tips there.
- Google scholar - research has never been so easy!
- Use social networks to post questions and/or generally discuss the modelling with your group and class.

4 What to do

For either project you need to produce:

- a group report,
- individual contribution statements,
- a group presentation.

In exchange for these services you will be awarded marks that can be redeemed against a degree. Specifically as a proportion of the grades in this module:

- group report and presentation (10%),
- individual report (10%).

Hence this project represents one fifth of the total grades available for this module. You will not be assessed solely on the performance of your models.

[2] https://mlcourse.ai/assignments

[3] https://mlcourse.ai/

4.1 The group components

Clearly there is a group component - you will have to decide collectively which project to tackle and how to go about analysing it. You may like to divide some tasks amongst your group members e.g.

1. data cleaning
2. checking of cleaning and summarising the data
3. research of particular analysis method

Although your individual reports are independent works, the analysis itself should be a group effort. You will have to describe the different aspects of the analysis in your independent report, so you'll have to share with your team mates and understand what others have done. If you personally focussed on a modelling method, feel free to make that more detailed in the report, with other people's given more light treatment.

4.1.1 Group report

This is intended as a report for your (imagined) client. Assume that they are not analysts and are primarily interested in what your modelling means in real-terms for their company. So there should be no esoteric terminology or modelling details to interfere with your message. They are interested in deploying your model for use - it will form the backbone of some decision support software i.e. they will make good or bad decisions based on your recommendations. This should be brief, no more than 4 pages in total. There should be a very succinct executive summary at the start selling your method/results. What does your approach offer this type of business? Why should they part with a large consulting fee or licence your model? If you can argue a good Return On Investment (ROI) then you're in a good position. Feel free to speculate about costs associated with the different decisions implied. Also bear in mind this document is something of a sales pitch - it should look the part, not some dry analyst's report (which will be your individual technical reports).

You should also attempt to give some insight into what is driving the response. You have many variables at your disposal and the client would appreciate some insight into what characterises the targets. Note this is where the client can give your model the "sniff-test" for anything suspicious. They may not be analysts, but they know typically know their market - if you are producing nonsense it may be obvious, even if the methods are complex (known in the trade as 'client-visible' problems).

4.1.2 Group presentation

Your group will provide a 10 minute presentation (plus time for questions) in week 11 of semester 2. This presentation is for a more technical audience (your classmates and myself) so I will expect details of how you implemented your methods and relevant performance measures. I expect that you will divide the presentation amongst yourselves e.g. 1 person describes the problem and data, another person describes the analysis method, the other describes the findings.

4.2 Individual component

You must each provide a report which you can imagine is the technical appendix to your non-technical group report. This report should:

1. outline the data, the problem and the methods you used to analyse the data.
2. give summaries of your model's performance, both as measured during the model development e.g. CV measures, and the model's ultimate performance against the test data.

3. *proper referencing* where appropriate. This includes references within the text and a proper reference section as in an academic journal e.g. see a *biometrics* article. If you do not know how to do this ask me - it is essential that you do this properly at this level.
4. you must indicate what software and computer specifications were used in the fitting of your models.
5. it should be no more than 3000 words.
6. do not include raw computer output or code in the body of the text - you can put these sort of things in an appendix or as supplementary files in your zipped submission if you wish.

Although the analysis will have been spread over the group, you should demonstrate some understanding of all the methods applied. If you were in charge of a particular component of the analysis in your group, this may be more detailed.

5 Tips

- You can use whatever software you like. The Open Machine Learning Course assumes use of python (and python provides very good tool support for projects of this type), but if you prefer – and can figure out how to – do the work in another language, then that is fine.
- Make sure you have familiarised yourself fully with the data from the outset. Do a lot of exploratory work to identify anything unusual - note that this is a good place to divide up some work across your group. There may be no data-cleaning *per se* required depending on the data, but you should still examine the data carefully (indeed the difference between the winners and losers might come down to treatment of odd cases).
- You can pre-process your data as you see fit e.g. drop variables, condense categories etc.
- Start small (particularly in the case of a large dataset). Consider developing baseline model based on an initial sample of a workable size.
- Use the group to your advantage - you can divide up the research and reading and share notes on what you find.
- You can emulate the test datasets by creating appropriate validation sets.
- Do not become obsessed with finding the best model in the world. Aim to produce a reasonable model to base your project on as a minimum attainable goal. Once you have something concrete you can look to improve things.
- The group report describes your “best" model, and presents your evidence for why it is “good". This is because your imagined client doesn't care about any other models that were considered but rejected. The target for your individual report is not imagined – it is me, my colleagues and the formal examination board. Therefore in this document you should report on all the models considered, how they were evaluated, and why they were not chosen for presentation to your client.
- You can use any document preparation system you wish (e.g. MS Word or LaTeX). The quality of your writing and presentation will be taken into account during assessment (although minor spelling and grammatical errors will be ignored, since I will not assume when grading that the author is a native speaker of English). All tables should have captions, and all figures should have written legends. Like the title of the paper itself, each caption and legend should convey as much information as possible about what the table or figure tells the reader. For example

Figure 3. Comparison of serum AMH concentrations with NGF population and with NGF recruitment The red line is the log-unadjusted validated AMH model, peaking at 24.5 years. The blue line denotes the decline in NGF population [15], with peak population at 18-22 weeks gestation. The green line denotes the numbers of NGFs recruited towards maturation [22], with peak numbers lost at age 14.2 years on average. Each quantity has been normalised so that the peak occurs at 100%. Correlation coefficients (r) are given for AMH concentrations against the other two curves for birth to 24.5 years and for 24.5 to 51 years.

- Use appendices to show completeness of analysis. If you experimented with multiple choices of model tuning parameter (say) you can list all the outputs in an Appendix to your report.

Have fun!