Group Project

This year the group project 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 classification problem, which can be found on Kaggle as the “Porto Seguro's Safe Driver Prediction” competition. Their description is paraphrased as follows:

> 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.

This is a classification problem. Work in the teams you have been assigned to.

Data

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

1. `train.csv`: This contains the data that you will be using to train your models. It has a full complement of response 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 your submission must conform to for uploading to Kaggle.

Resources

- There is online discussion around all such competitions - you can get tips there.

- Use our Piazza.com question and answer forum, or other social networks, etc to post questions or generally discuss the modelling with your group, and with the ID5059 community as a whole.

Deliverables

Kaggle submission

1. Submit a formal entry to the competition via the “late submission” link so that your group appears on the official Kaggle leaderboard for this competition. Submit your entry as a group, but each student in the group must appear as a member of the group.

Each individual student must upload the following things **via moodle**:

2. A group report (in PDF form, details below).
3. An individual contribution statement (in PDF form, details below)

Optionally, you may also submit any code that you believe may prove useful in assessing your work (ideally in the form of an annotated Jupyter notebook).

Important: Do not upload any of the data, or your model’s predictions.
Important: Upload materials to moodle only. Do not upload any materials to MMS.

These will contribute to your overall grade in the module as follows:

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

Hence this project represents one fifth of the total grades available for this module. Note that you will not be assessed solely on the performance of your models, or rather, your position on the leaderboard.

The group components

Clearly there is a group component - you will have to decide collectively how to go about your work. You may like to divide some tasks amongst your group members, for example

- data cleaning
- checking of cleaning and summarising the data
- 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 teammates 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.

Group report

- 1 page (single sided) in total, submitted as a PDF file.

This is intended as a report for your imagined client. Assume that they are not analysts, and that they 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.

It should be as succinct as possible. 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.

Consider giving 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 may look for anything suspicious about your work. They may not be analysts, but they 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).

Individual report

- 2 pages (single sided) in total, submitted as a PDF file.

You must each provide a single-authored report that is the technical counterpart to your non-technical group report. This report should:

- outline the data, the problem and the methods you used to analyse the data.
- give summaries of your group’s model performance, both as measured during the model development e.g. CV measures, and the model's ultimate performance against the test data.
- It should be no more than 2 pages (single sided).
- Do not include raw computer output or code. You can upload such things optionally as supplementary material 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.

Tips

- You can use whatever software you like.
- Make sure you have familiarised yourself fully with the data from the outset. Do 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 required depending on the data, but you should still examine the data carefully (performance differences 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.
- 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, or my colleagues marking the project. Therefore in this document you can 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), but the submitted reports must be in PDF form. The quality of your writing and presentation will be taken into account during assessment.

Deadline

Deliverables must be **submitted via moodle by Friday 17th of April** (in week 10). Starting on the 18th of April, marks M as a function of time t decay exponentially, $M(t) \propto \exp(-t/\tau)$, with $\tau = 4$ days.