COMP 309 — *Machine Learning Tools and Techniques*

Assignment 3: Kaggle Competition

16% of Final Mark — Due: 11:59pm Friday 4th September 2020

1 Objectives

The goal of this assignment is to help you tie together all the concepts you have learnt in the first half of this course in the lectures and assignments. To aid you in completing this assignment, you should review the major aspects of the course that have been explored so far, such as:

- Data understanding, cleansing, and pre-processing,
- Machine learning concepts,
- CRISP-DM and pipelines in general,
- Feature manipulation, including feature selection, feature construction and imputation,
- Statistical design and analysis of results.

These topics are (to be) covered in lectures 01–12. Research into online resources for AI is encouraged, where the rabbit-hole[1] will provide useful jumping off points for further exploration.

2 Question Description

The price of power is a constant topic of discussion in New Zealand, with frequent reporting in the media[2]. You and your academic colleagues are fed-up and have decided to investigate whether the cost of power fairly aligns with how much power is generated around the country. Your newfound expertise in machine learning tools and techniques provides a strong foundation for your data mining.

The overall aim of this assignment is to develop the best possible machine learning system to **predict the price of electricity in Wellington**. The hope is that your model will give a better understanding of the relationship between electricity generation and cost[3].

We have set up a Kaggle InClass Competition[4] to facilitate finding the best machine learning system for officials to use. You are expected to analyse the provided data, design and improve your own machine learning pipeline, and consider the consequences of applying your pipeline to this data.

Note the data is real. Thus, you could attempt to find the original dataset and create a look-up table. This is not permitted as it misses the point of the course. We want to see the model produced with the understanding of the patterns, rather than see perfect results.

[1] `https://ecs.victoria.ac.nz/Courses/COMP309_2020T2/RabbitHole`
[2] `https://www.stuff.co.nz/business/117818574/genesis-price-rise-sparks-fears-generators-will-pocket-214m-in-consumer-savings`
[3] `https://www.em6.co.nz/`
[4] `https://www.kaggle.com/about/inclass/overview`

2.1 Preliminary: Accessing the Kaggle InClass Competition

To access the class competition, you **must** use the below url. **Please do not share this publicly** as it will allow anybody to access our competition, which will make the experience less enjoyable for your classmates. Deliberate cheating is a disciplinary matter, so please don't go there either.

Competition link: `https://www.kaggle.com/t/492c160cf6fb452cbc3dcbd9a5802868`

You will need to register a Kaggle account. It is perfectly fine (and expected) to use a pseudonym as your Kaggle username so your classmates do not know your real-life identity. However, **you will need to fill out the following form** so that the lecturers and tutors can link your Kaggle result to your ECS account. No other people will have access to this information! Each time you change your Username, please update the form (it would make sense to do this only once, but past experience suggests that students will change their Username a few times). We cannot give credit to top-scoring students if we cannot link a username to an actual/course name! Usernames must be suitable for work and respectful.

Please fill out the following form:
`https://docs.google.com/forms/d/1W5BiWbBrs-HCXUhJJ_w_UebFmVqPTvJeZcJn0XCddPs`
Please submit as part of your report.

Once you have completed the above steps, please verify that you can access the following page:
`https://www.kaggle.com/c/comp309-2020/overview` (when logged in).

Once successful, please accept the terms of the competition so that you may proceed to the rest of the assignment!

2.2 Core: Exploring and understanding the Data [40 marks]

We have created a data processed version of the power-price data. This is to be used in classification competition. We have split the data into training and test set.

In the Completion part the training set is to be used to create your model. You can use any machine learning tool, e.g. WEKA, SciKit Learn. Your model will need to be able to predict the class of 'unseen' test data (i.e. features, but not class, are provided).

It is often much more effective to first learn about the properties of a dataset (business and data understanding) before applying machine learning to it. You should begin by familiarising yourself with the dataset by reading the "Overview" and "Data" tabs of the Kaggle competition. Please download the dataset from the Data tab (in .csv format). You should now spend some time examining the data and taking notes of any interesting patterns you find.

Requirements

Using any tools you find useful, you should explore and analyse the dataset. You should draw upon your previous experiences and what you have learnt in this course to find a number of interesting patterns. You may wish to start by examining the quality, completeness and representation of individual features.

In your report (to be submitted electronically), you should spend approximately two pages describing the following regarding the Core part:

- (20 marks) Highlight the findings of your dataset exploration. You should identify important patterns (e.g. large correlation between variables) clearly using examples (say, five illustrative patterns), and discuss the potential consequence this may have on your results. To achieve a high mark, you should consider more complicated patterns, such as feature interactions. Use your judgement and justify what is an important pattern.

- (20 marks) Visualisation is an important aspect of this task. Please illustrate at least one important finding of your work.

2.3 Completion: Developing and testing your machine learning system [50 marks]

You may use any ML tools you wish, but a good solution will consider a number of factors, such as: pre-processing steps, the properties of the dataset and generalisation/over-fitting. Decisions around how to split your labelled data into training/testing/cross-validation set/s are your choice, which are important and should be explained.

Your file will need to consist of two columns. The Id and the predicted price only. The csv should include your predictions for all 1464 instances in the dataset, plus a header line (1465 lines). For example:

```
Id, Price
1464, 229.1
1465, 31.3
1466, 4.20
...
2926, 22.2
```

Note that the Ids for the test data (the unlabelled data) start at 1464, as the Ids from 0 to 1463 correspond to the training set (the labelled data).

Part of the test data is to be used by Kaggle to test your model in order to create the public leaderboard. The remaining part of the test set is used by Kaggle to test your model for the private leaderboard. You will submit your complete answer to the test set but will not know which instances have been used for the public or private leaderboard (this helps prevent over-fitting and gaming the system).

Once you are satisfied with your initial attempt (do not spend too long on it!), you should upload your output csv to the submissions page of the competition: `https://www.kaggle.com/c/comp309-2020/submissions`.
Once your submission has been processed, you will be able to see your classification accuracy on the public leaderboard. You should use this feedback to further improve your system. For example, if your leaderboard performance is much lower than on your own test set, you have over-fitted your model. You should use your judgment to decide how extensively to change your system. This may only be tweaking parameters, or you may decide to try a completely different algorithm. Note that you are limited to 4 submissions per day, but submitting this many may be to your detriment as you may "over-train" on the public leaderboard!

Do not be discouraged if your performance appears low on the leaderboard: we are interested in novel/interesting solutions even if they have lower performance, and you may come out on top on the private leaderboard anyway.

Requirements

You should refine your machine learning system a number of times (*at least 3*, including the initial system) based on the performance you achieve on the public leaderboard. Your submitted report should contain up to 4 pages regarding the Completion component:

- (15 marks) Discuss the initial design of your system, i.e. before you have submitted any predictions to the Kaggle competition. Justify each decision you made in its design, e.g. reference insight you gained in the Core part.
- (25 marks) Discuss the design of one or more of your intermediary systems. Justify the changes you made to the previous design based on its performance on the leaderboard, and from any other additional investigation you performed.
- (10 marks). Use your judgement to choose the best system you have developed — this may not necessarily be the most accurate system on the leaderboard. **Make sure you select this submission as your final one on the competition page before the deadline**. Explain why you chose this system, and note any

particularly novel/interesting parts of it. You should submit screen captures and/or the source & executable code required to run your chosen submission so that the tutors can verify its authenticity.

2.4 Challenge: Reflecting on your findings [10 marks]

Until now, we have been focusing on achieving the best performance possible — but consider whether this is all that ML tool users should consider?[5]

The dataset has a number of features that could be used to produce a machine learning system that, while accurate, may have a number of biases towards certain energy groups. Should the officials be worried that using your model in their analyses could be non-beneficial to sectors of society? Should we penalise with carbon tariffs coal-fired power stations (and its employees)? Should we build new hydro-power schemes (in areas of outstanding natural beauty/biodiversity)?

Requirements

You should consider the interpretability of your final chosen model from the Challenge part, and analyse any ethical concerns associated with it.
Your report (1 page on the Challenge component) should address the following questions:

- (10 marks) How easy is it to interpret your chosen machine learning model? Discuss any ethical/political/society consequences of how it uses the chosen features to make a prediction. For example, consider advising a politician on the effects/consequences of building a new energy generation facility?

[5] `https://towardsdatascience.com/can-a-machine-be-racist-5809b18e5a91`

3 Relevant Data Files and Kaggle Information

The datasets, and additional information about the Kaggle competition can be found online:

`https://www.kaggle.com/c/comp309-2020/`

4 Assessment

We will endeavour to mark your work and return it to you as soon as possible, however Covid restrictions could delay timings. Your position on the private leaderboard will help improve the final grade for the top twenty students in the Completion competition only, but will not be the main consideration. The tutor(s) will run a number of helpdesks to provide assistance in the assignment to answer any questions regarding what is required.

5 Submission Guidelines

5.1 Submission Requirements

1. A document that consisting of the report of all the individual parts. The document should mark each part clearly. The document can be written in PDF, text or the DOC format, but needs to be submitted in .pdf format of no longer than 12 pages excluding appendices.

5.2 Submission Method

The programs and the PDF version of the document should be submitted through the web submission system from the COMP309 course web site **by the due time**.

There is NO required hard copy of the documents.

KEEP a backup and receipt of submission.

Submission should be completed on School machines, i.e. problems with personal PCs, internet connections and lost files, which although eliciting sympathies, will not result in extensions for missed deadlines.

5.3 Late Penalties

The assignment must be handed in on time unless you have made a prior arrangement with the lecturer or have a valid medical excuse (for minor illnesses it is sufficient to discuss this with the lecturer). The penalty for assignments that are handed in late without prior arrangement is one grade reduction per day. Assignments that are more than one week late will not be marked.