QBUS2820
Predictive Analytics
Semester 2, 2017

Data Prediction Project (Assignment 1): Home Value Estimates

1. Key information

Required submissions: Written report (by Turnitin submission), predictions for the test data (through Kaggle), and Jupyter Notebook (through Ed).

Key deadlines: Friday September 22nd at 5pm (report deadline), Monday September 25th (Jupyter notebook and final submission deadline for the Kaggle competition).

Weight: 20 out of 100 marks in your final grade.

Groups: You can complete the assignment in groups of up to three students. There are no exceptions to this: if there are more than three you need to split the group.

Length: The main text of your report should have a maximum of 20 pages (15 is ideal). If you wish to include additional material, you can do so by creating an appendix. There is no page limit for the appendix. Keep in mind that making good use of your audience's time is an essential business skill. Every sentence, table and figure has to count. Extraneous and/or wrong material will reduce your mark no matter the quality of the assignment otherwise.

Marking and key rules:

- Carefully read the requirements for each part of the assignment.
- A separate rubric will indicate the marking criteria for the report.
- You must use Python for the assignment. It is fine to use Excel for data manipulation, though this is neither efficient nor recommended.
- Reproducibility is fundamental in data analysis, so that you will be required to submit a Jupyter Notebook that generates the results that appear in the main text of your report. Unfortunately, Turnitin does not accept multiple files, so that you will do this through Ed instead. Not submitting your code will lead to a loss of 50% of the assignment marks.
- Failure to read information and follow instructions, including Business School rules and guidelines, is subject to loss of marks.

2. Getting the data

You can download the data for the assignment and competition, as well as the variable descriptions, via the link below. You need to create a Kaggle account based on your university e-mail address in order to have access and make submissions.

https://inclass.kaggle.com/c/qbus2820-17

3. Problem description

Assessing property prices is fundamental for city governments, who need up-to-date and accurate information on market values to set taxes (such as property taxes and stamp duties), to make public policy decisions, and to negotiate with property developers. However, houses are illiquid assets: only a small fraction of properties go on the market in a given year, so that current prices are not directly available for most houses.

As a consultant working for a data analytics company, the city government approaches you to develop a model to predict house sale prices based on state-of-art techniques from predictive analytics. To enable this task, you are provided with a dataset containing highly detailed information on recent house sales. The response, sale price, is the last column in the dataset.

As part of the contract, you need to write a report for the client describing your analysis and an initial assessment of how your candidate models are likely to perform in practice based on a validation set. The client will then use a test set to evaluate your work. The client is trained in statistics.

4. Understanding the data

Information about the data are on the Kaggle page for the assignment. There are two data files, the training set and a validation-test set (simply called test on Kaggle), which omits the response values. Kaggle randomly splits the observations in the validation-test data into validation and test cases, but you do not know which ones end up in each set.

When you make a submission during the competition, you get a score based on the validation set. The validation scores are visible to everyone and provide an ongoing ranking of groups. At the end of the competition, Kaggle will rank the groups based on the test data only. Be careful not to overfit the validation set in attempt to improve your position, as this may lead to a disappointing result for the test data.

5. Written report

The purpose of the report is to describe, explain, and justify your solution to the client. Your solution is:

1. How you process the data.
2. Your basic understanding of the data (EDA).
3. Feature engineering (constructing relevant predictors from the raw data).
4. Description and justification of the two final models that you will submit for evaluation on the test data.

The client's time is important. The client has a limited attention span. The client is not interested in minor details. Be objective and concise. Find ways to say more with less. When it doubt, move material to the appendix.

Requirement:

Your report must include the validation scores for at least five different sets of predictions, including your two final models. You need to make a submission on Kaggle to get each validation score. The other three sets of predictions should be from different models or substantive variations thereof. You do not need to describe these three additional methods in detail.

Suggested outline:

1. Introduction: write a few paragraphs stating the business problem and summarising your final solution and results. Use plain English and avoid technical language as much as possible in this section (it should be for a wide audience).
2. Data processing and exploratory data analysis: provide key information about the data, discuss potential issues, and highlight interesting facts that are useful for the rest of your analysis.
3. Feature engineering.
4. Methodology (your two models, your rationale, how you fit them, some interpretation, etc).
5. Validation set results and comparison with other approaches.
6. Final remarks (non-technical).

6. Kaggle Competition

The purpose of the Kaggle competition is to incorporate feedback by allowing you to compare the performance of your model with the submissions from other groups.

Participation in the competition is part of the assessment and you must make sure that your final submission is correct. Your ranking in the competition will not affect your marks (apart from bonus marks, as explained below), however we will assess whether your participation represents a genuine effort to make good predictions and improve them.

Bonus marks:

The three teams with most accurate predictions for the test data will receive bonus marks. The first place will get 5 bonus marks for the unit, the second place 3 marks, and third place 1 mark. In order to qualify for the full bonus marks, the groups ranked first and second would need to give a short presentation (5-10 minutes) to share their successful approach with the rest of the class. If the group does not wish to do so, the bonus is 1 mark.

6. Tips

- Only a few predictors account for a large part of the variation in the response. The main challenge is to achieve an edge in predictive performance by making efficient use of the large amount of information available.
- Because the number of observations in the training data is reasonably large, you may only see an improvement from using methods such as regularisation once you have a sufficient number of regressors in the model.
- Several predictors have missing values. In many cases, the missing values mean that the field does not apply to that house. This is information, rather than lack of information as is traditionally the case with missing values.
- Please use Ed for troubleshooting.