MA5011: Assignment

2025-04-28

Introduction to Predictive Analytics

- Modelling Challenge

Instructions

1. The assignment is worth 50% of the module grade.
2. The project details are on the next two pages.
3. The project analysis should be carried out in R.
4. Your project report must be submitted in a **PDF format** (you can easily do this by writing your answers in a Microsoft Word document and then printing to pdf).
5. The project report should have as its first page your name and student number.
6. As a guideline this report should be ~ 2500 words and 10-15 pages including figures.
7. You must also submit the code file (either .R or .Rmd) you used for analysis. Ensure you comment your code.
8. Your code file should be called studentname.R, i.e. "philippawilkes.r" or "philippawilkes.rmd"
9. In summary, submit 3 files to the assignment page on Brightspace:

- PDF report document
- code file
- dataset

10. The deadline is 23:59 Tuesday 20/05/2024.

Assignment

1. Dataset selection

- Choose one of the datasets provided below. Alternatively, you can use an internal work dataset, or other dataset of interest, that has the possibility of being modelled - it must contain a response (dependent) variable of interest, as well as a number of explanatory (independent) variables. You may discuss this with me.

(a) Dataset 1 - Seattle house sales dataset.
Link: [https://www.kaggle.com/datasets/harlfoxem/housesalesprediction]

(b) Dataset 2 - Ames house sales dataset.
Link: [https://www.kaggle.com/datasets/prevek18/ames-housing-dataset]

(c) Dataset 3 - Ecological dataset.
Link: [https://www.kaggle.com/datasets/abrambeyer/openintro-possum]

(d) Dataset 4 - Transport dataset.
Link: [https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma]

2. Introduction

- Briefly describe the context and source of your dataset.
- Pose a research question that you wish to answer - you may refine this as your analysis progresses. At a minimum for the research question, identify your response (dependent) variable; you will then be examining how one or more independent variables affect it.

3. Exploratory descriptive analytics

- Summarise and explore each variable in the dataset, providing the appropriate summary statistics, visualisations and confidence intervals.
- If you omit any variables from later modeling justify those omissions here.

4. Diagnostic analytics

- Formulate and carry out hypothesis tests to explore relationships you identified as interesting in the exploratory data analysis.
- Report p-values, test statistics, and confidence intervals, and interpret each in context.

5. Correlation and relationship exploration

- Use scatterplots to explore relationships between continuous variables.
- Calculate and discuss correlation coefficients for continuous variables, particularly between response variable and predictor variables.
- If appropriate for the data, use chi-squared tests to check associations among categorical predictors.

6. Model fitting

- Based on analysis in the previous steps, select one continuous predictor and fit a simple linear regression to the response (you may do this more than once, but one properly explained model is what is required).
- Ensure to test all assumptions.

7. Conclusion

- Summarise your key findings in plain language.
- Discuss implications for the original business/research question.
- Suggest any next steps or further analyses.

General notes

- Don't include code in the report. Instead, include relevant code outputs of interest (e.g. t.test outputs, simple linear regression outputs etc.), you can copy and paste or screen shot/ snip these for example.
- Ensure all figures and tables correctly labelled and cited in the text.
- Analysis is iterative, you do not have to report all the analysis that you do (you can include it in the code script if you wish). You want the analysis to have a flow so that it can be easily followed.
- The numbered steps are a guide, depending on your data you may want to concentrate more on some steps than others.