QBUS6810 S2 2022 Marking Notes

This document is based on feedback given to students from S2 2022 on where they could have improved their marks. Please use this as a checklist to ensure that you are submitting the best report you can. Be warned that if you follow this blindly without thinking you might find yourself losing marks because you have a different dataset!

1 Introduction

In general the *Introduction* was done well, with groups generally receiving a mark of 4 or 5 for this section. Most groups did a good job at describing Airbnb and outlining the purpose of the report, but fell short when it came to summarising the contents of the report.

Reasons groups lost marks

- □ **Not describing the purpose of the report in enough detail.**
  The purpose of the report is to analyse factors that affect listing price and to build a model that predicts listing price.
- □ **Not summarising the contents of the report in enough detail.**
  It's nice to briefly describe the sections that are contained in the report and to also briefly list out all of the models you attempted to build.
- □ **Not describing the results of the report in enough detail.**
  A report isn't a movie or a book, it's okay to spoil the ending! You should have provided a brief summary of which variables affected listing price and summarised the performance of the models you built. This includes providing the validation root mean squared error (RMSE) of your best model.

2 Data Processing

In general the *Data Processing* section was done well, with groups generally receiving a mark of 4 for this section.

Reasons groups lost marks

- □ **Not describing the amount of data that was available.**
- □ **Not describing the kinds of data that was available.**
- □ **Not describing how the data was processed.**
- □ **Not describing how missing values were handled.**
- □ **Not *justifying* why missing values were handled the way they were.**
- □ **Missing values were not handled appropriately.**
- □ **Not identifying that `Price` was the response variable given in units of Autralian dollars (AUD).**

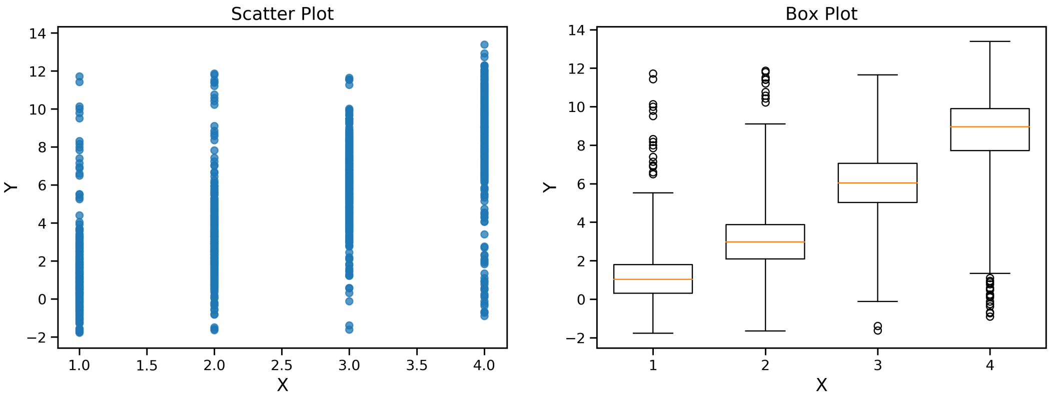

Figure 1: Both of these figures are constructed using the same data. Which do you think is the most effective in communicating a relationship between X and Y?

3 Exploratory Data Analysis (EDA)

In general this section could have been improved, with groups generally receiving a mark between 4 and 7. The goal is to tell a 'story' about your data and highlight interesting relationships that are relevant to the business context i.e. relevant to the prediction of Airbnb listing prices. You then need to clearly describe how your EDA findings will affect your model building decisions. It is useful to also frame your findings by providing insights based on domain knowledge. Most groups could have improved their EDA by structuring their EDA more clearly. For example, starting with univariate analysis before moving onto multivariate analysis.

Reasons groups lost marks

- □ **The overall structure of the EDA was hard to follow.**
  A lot of groups could have improved by dividing up the analysis into different sections (e.g. univariate analysis follow by multivariate analysis). Additionally, EDA is about telling a 'story'. For example, you might start by looking at how `Price` is correlated with other variables and then investigating the variables that appear to have the largest correlation with `Price`.
- □ **The EDA wasn't grounded by the business context.**
  It's important to provide insight and context when performing EDA. For example, you should go beyond saying, *There is a positive correlation between longitude and listing price.* Instead you could say, *Our EDA shows that there is a positive correlation between longitude and listing price. This is unsurprising as lines of longitude increase as you move towards Sydney's coast and indicates that coastal listings are more expensive than inland listings. A majority of Airbnb users are holiday-goers and Australian beaches are extremely popular among tourists.*
- □ **Not investigating the response variable**.
  The response variable is probably the most important variable in your data, so you need to inspect it to check its distribution and whether there are extreme values that may affect your predictions.
- □ **Not identifying which variables would be useful for model building.**
  Remember that one of the key goals of the project is to build models to predict listing price. This means, one of the key goals of EDA is to identify which variables will make good predictors. It would have been nice to include a summary at the end of your EDA clearly identifying which predictors you were going to use for your model building.
- □ **Not identifying which variables would be useful for feature engineering.**
  It is important that throughout your EDA you comment on which variables may need to be feature engineered before they can be used as model predictors and also identify or speculating on potential new features that could be constructed from existing variables.
- □ **Not discussing outliers or extreme values.**
  A lot of groups didn't mention outliers or extreme values at all. The groups that did mention them often treated them quite superficially. It's good to identify outliers/extreme values and to discuss how they would be handled (you may choose not to do anything about them, but you still need to say that) and whether they would have any affect on your models.

□ **Not reporting on the relationship between location and price.**
This was an important relationship to investigate because you were later required to construct a location-based feature, which must be well *justified*. This means that in your *Feature Engineering* section you should have referred to your analysis in your EDA.

□ **Ignoring variables without any explanation.**
There were numerous variables that were made available in the dataset and it is okay if you don't use all of the variables, but you at least need a sentence explaining why you were ignoring them. For example, you might say, *We have chosen not to use `host_location` as we do not believe that it would affect listing `price` since the whereabouts of the host doesn't affect the characteristics of the listed property.*

□ **Inappropriate visualisations.**
The *quality* of the visualisations were assessed in *Writing and Presentation II*, but the *appropriateness* of the visualisations was assessed in *EDA*. Groups need to consider how effectively their visualisations communicate information and consider whether another visualisation would be more appropriate (e.g. Figure 1).

□ **Putting every figure in the appendix.**
Figures are a key part of the EDA and are very effective at communicating relationships. These should have been placed in the main body of the report. The appendix should be used for supplementary material and should include details that are beyond the scope of the assignment. The reason why the appendix is not included in the page limit is that it is not considered part of the body of the report and so was not assessed as part of the report. It makes it also very difficult to read your report if you have to look up every single figure in the appendix.

4 Feature Engineering

In general this section could have been improved, with groups generally receiving a mark between 5 and 8. The main reason groups lost marks was because they didn't justify their design choices, they didn't provide enough detail for a reader to replicate their feature engineering process or because they performed very little feature engineering.

Reasons groups lost marks

□ **The feature engineering was not substantive.**
If your group only create a location-based variable and no other variables, you lost marks. Groups needed to perform substantial feature engineering to achieve high marks in this section. For example, some groups performed natural language processing, or reduced the number of categorical variables by merging categories.

□ **Choices made during the feature engineering process were not well explained.**
You needed to provide enough detail that a someone else would be able to replicate your feature engineering process. For example, it was not enough to say that you reduced `Neighbourhood` into 3 groups based on the suburb's average listing price. You need to explicitly define each group. Did you decide that the most expensive group corresponded to... the top 10% of suburbs? The top 5 most expensive suburbs? Suburbs where the average listing price was over $1000 AUD? etc.

□ **Choices made during the feature engineering process were not well justified.**
Your group needed to explain their design decisions. For example, if you engineered a feature which was the `distance` of a listing to the Opera House, you need to explain why you selected the Opera House, and how you think your `distance` variable would affect listing price.

□ **A new location-based variable was not created.**

□ **A feature engineering technique has been misapplied.**

5 Methodology General

In general this section could have been improved, with groups generally receiving a mark between 3 and 5.

Reasons groups lost marks

□ **The choice of models was not well justified.**
Your group needed to include some introductory sentences for each model explaining why that particular model was chosen. For example, you could say that your linear model served as a baseline model as it is a very simple model and could be used to benchmark more complex models against. Additionally, linear models are easy to interpret and can be used to identify predictors that are strongly related to listing prices. In contrast, a stacked model is difficult to interpret, but is expected to offer better prediction accuracy, and hence could be used by Airbnb to recommend listing prices to hosts.

□ **Assumptions were not discussed.**
Your group needed to discuss model assumptions. For example, you could have mentioned that a linear model assumes a linear relationship between the predictors and the response. It was not enough to simply list assumptions. You needed to discuss assumptions in relation to the task. For example, you could refer to your EDA to discuss whether your assumption of linearity was likely to hold or not.

□ **The overall structure of the methodology section was difficult to follow.**

□ **Did not include a linear model, regression tree or advanced tree-based method or stacked model.**

6 Models

In general this section was done poorly, with most groups receiving between 3 and 6 for each model. In the model description, groups needed to provide enough detail that a data scientist would be able to read their report and reproduce each model.

Reasons groups lost marks on the model description

□ **Not clearly identifying the response and the predictors.**
Often it was unclear whether the model was predicting `price` or `log(price)`. Additionally, it was generally not always clear which predictors were being used. Groups that were successful gave a summary of the predictors (e.g. listed the number of predictors used) and then provided more comprehensive details in the appendix.

□ **Not clarifying whether predictors were standardised/normalised.**
Predictors should be standardised/normalised for ridge and lasso regression and it was not always clear whether this happened or not. It was also not clear whether the predictors were standardised/normalised for other models, especially if standardising/normalising was discussed in *Feature Engineering*, but were not mentioned in the *Methodology.* It is important to know whether standardisation/normalisation has occurred because it affects the model interpretation.

□ **Not including a relevant equation.**
Often it's nice to include an equation of the final model where appropriate.

□ **Not describing relevant tuning parameters.**
It is important to describe some of the important tuning parameters and how they affect the final model. For example, for ridge regression, this would be the penalty parameter λ in $y = \beta_0 + \sum_{i=1}^{P} \beta_i X_i + \lambda \sum_{i=1}^{P} \beta^2$, where P is the number of predictors and in regression trees could be the pruning parameter α, which controls the size of the tree.

□ **Not describing or justifying how the tuning parameters were selected.**
Your group should have specified how you selected each of their tuning parameters. For example, you might have used 5 fold cross-validation and performed a random search. Alternatively, you may have used a grid search, in which case you should also report on which values you tested.

□ **Not reporting on the final tuning parameters were used.**
Your group needed to report the value of all tuning parameters used in your final model.

□ **Not reporting on the final model parameters used.**
Model parameters are different from tuning parameters. Model parameters include β coefficients, or the weights of the level 1 models in a stacked model.

□ **Not including a relevant diagram.**
Visualisations are often a great form of communication. It was often useful to include diagrams for your models where appropriate, e.g. a graph showing the magnitude of the β coefficients, a graph of your regression tree or a diagram of your stacked model. Note that sometimes if the figure is quite large, e.g. you had 50 β coefficients, it is better to provide a smaller diagram in the main report and a larger more comprehensive figure in the appendix.

Reasons groups lost marks on the model interpretation

□ **Not attempting to interpret their model.**
The model interpretation was worth 3 marks for each model. Groups that did not attempt to interpret their model automatically lost 3 marks.

□ **Model interpretation was not detailed enough.**
A lot of groups attempted to interpret their models, but did not provide enough detail. Remember that while you might provide more technical explanations in this section, you still need to make sure you interpret your models for a non-technical audience. For example, you would need to explain what it means if a β coefficient is large in magnitude or if a β coefficient is close to 0 or if a β coefficient is negative. You also need to relate your findings to the business context. For example, if `bedrooms` is associated with a large β coefficient, it means that listings with more bedrooms are rented at a higher price. You could then comment on whether your findings were consistent with your EDA.

□ **Model interpretations were incorrect.**
Groups often forgot to account for a log transformation of the response or the standardisation/normalisation of the predictors in their analysis.

□ **Model interpretation was not converted back to meaningful units.**
Often groups predicting `log(price)` did not convert their interpretation back into meaningful units. For example, discussing `log(price)`, which had values such as 5.42 or 4.8 is not intuitive. It is much more meaningful to the audience if you had discussed `price` in AUD.

□ **Model interpretation was not related back to the business context.**
Groups need to provide some useful analysis and recommendations based on their findings and summarise their interpretations with some general statements.

How to interpret your model

Here are some suggestions for interpreting your models.

Linear models

- Analysing the β coefficients:
  - Identify which predictors are associated with the large β coefficients (in magnitude).
  - Identifying which predictors have β coefficients close to 0.
  - Identifying which predictors were associated with a negative β coefficient.
  - Explaining how a 1 unit increase in a predictor affected the response (see Lecture 2 slide 33).
- Each of the items listed above needs to still be *interpreted.* For example you need to explain that if a β coefficient is close to 0, it means that thee is a very weak (or non-existent) linear relationship between that predictor and the response.

Regression trees

- Analysis the splits that are high up in the tree and the value the splits occur at.
- Identifying which predictors (if any) were not used in the tree and discussing how regression trees perform variable selection.
- Analysing variable importance.

- Note that simply describing how to read a tree is not meaningful analysis. For example, saying *if you have more than 2, but less than 3 bedrooms and a longitude less than 153, then your listing price is expected to be $150 AUD* does not provide any intuition as to how bedrooms and location affect listing `price`. You still need to discuss general relationships between your predictors and the response.

Random forest and boosting for regression trees

- Analysing variable importance.
- Note that groups needed to provide analysis beyond just including a figure. This means that groups needed to explain what variable importance measures i.e. what does a variable importance of 100% vs 0% mean?
- Findings could have been related back to simpler models and the EDA and you could discuss whether the models were consistent in identifying the same important/unimportant predictors.

Model stacking

- Analysing the contribution of each level 1 model to the final model.
- Model stacking allows you to combine both linear models and non-linear models and evaluate the contributions. If a stacked model relies solely on the linear model, then it's an indication that the relationship between the predictors and the response is linear. Otherwise, if the stacked model relies heavily on non-linear models, it indicates that the relationships are non-linear.
- If a particular level 1 model dominates the stacked model, then you could try to provide some interpretation based on that level 1 model.
- A lot of groups didn't attempt to interpret their stacked model, and didn't comment on why stacked models are difficult to interpret.

Validation and Comparisons

In general this section could have been improved, with groups generally receiving a mark between 2 and 5.

Reasons groups lost marks

□ **It was unclear whether the RMSE was for the training data or the validation data.**

□ **The validation RMSE clearly didn't correspond to the Kaggle leaderboard**.
For example, the validation RMSE was calculated for `log(price)` instead of `price`.

□ **The value of the training/validation RMSE were clearly wrong**.
The training and validation RMSE was expected to be ∼$100-250 AUD.

□ **Not enough detail was given in the model comparison.**
For example, a lot of groups did not describe the relationship between model complexity, validation RMSE and interpretability.

□ **Model comparisons were not made in reference to the business context.**
There were two main goals of the assignment. The first was to understand what factors affect listing price and the second was to build a model to predict listing price. You need to identify and discuss whether particular models were more suited for a particular purpose.

□ **Not identifying model similarities.**
Comparisons should involve identifying both differences and similarities. For example, you could have commented on whether all of their models identified the same variables as having a large, or no effect on listing price.

□ **Not enough detail was given on model limitations (in reference to the business context).**
Groups often listed limitations for models without providing an explanation. For example, instead of just saying *regression trees have a low prediction accuracy*, you should explain what this means and make a judgement as to whether regression trees are a preferred model for predicting listing `price` or not.

7 Conclusion

In general this section was done well, with most group receiving between 2 and 3.

Reasons groups lost marks

□ **Not summarising which variables affected listing price.**
Most groups listed which variables affected listing price, but didn't describe the relationship, or provide context. For example, instead of saying, *The number of bedrooms affects price* you could have said, *Listings with more bedrooms were generally rented for a higher price*, that way you are describing *how* `bedrooms` affects `price`. You can then provide some insight into why this trend might occur.

□ **Not reporting the performance of their best model.**
One of the outcomes of the project is that you have create a model that Airbnb hosts can used to help them evaluate their property and predict a sensible listing price. It's important that you describe your model and state its performance i.e. the validation RMSE. For example, a model with a validation RMSE of $10 AUD is vastly different to a model with a validation RMSE of $1000 AUD. *Note: in practice you would quote the test RMSE, but due to time constraints we only expected you to quote the validation RMSE.*

8 Writing and Presentation I

Most groups received a mark between 2 and 4 for this section. In future, groups could improve their marks by spending some time polishing their reports.

Reasons groups lost marks

□ **Excessive grammatical and spelling errors.**

□ **Divisions between sections were unclear.**

□ **Reporting too many significant figures.**

□ **Abbreviations were used without being defined.**

□ **Equations were given as screenshots.**
Reports where equations were typed using an equation editor appeared much more professional.

□ **Symbols and letters given in equations were not defined.**

□ **Units were not mentioned throughout the report.**
This was particularly true in the *Validation and Comparison* section. Any time you reported on listing `price`, you should give the units. Similary, units should be provided in all tables.

□ **The report contained Python code.**
Groups that included screenshots of a line (or more) of Python code were heavily penalised. Groups were also penalised if there were references to to Python code, e.g. mentioning a function such as `isna()`. Code references are not appropriate in a report.

9 Writing and Presentation II

Most groups received a mark between 2 and 4 for this section. In future, groups could improve their marks by spending extra time checking their tables and figures.

Reasons groups lost marks

□ **Figure labels were too small to read.**
A majority of groups had axes labels that were too small to read. Reports where almost all figure labels were unreadable were more heavily penalised.

□ **Figures were missing labels or legends.**

□ **Figures were missing units or had the wrong units.**
For example, *Price* (AUD not specified) or *log(Price) (AUD).*

□ **Figures/tables were provided as low quality screenshots**.

□ **Figures appeared in the report and were not accompanied by text analysis**.

□ **Figures were convoluted and did not convey much meaning**.
A common example is that students would include a figure of their full regression tree, but the nodes overlapped and were impossible to read. When a figure is impossible to read, the figure is no longer useful. A much more successful approach taken by some groups was to show the first few rows of the regression tree in the main report and attach a figure of the full regression tree in the appendix.

□ **Figures were not well selected**.
In the EDA, a lot of groups included too many figures. For example, they included 20 different scatter plots to demonstrate the relationship of various variables with the response. Often these were not discussed in the text or were only superficially mentioned. A much more successful approach was to select a few of these plots and discuss them in detail and include the remaining figures in the appendix.

10 References

Most groups received a mark of 1 for this section.

Reasons groups lost marks

□ **Not including a reference list.**

□ **Inconsistent formatting of the reference list.**

□ **In-text referencing using the author's first name rather than their surname.**

11 Jupyter Notebook

Most groups received a mark of 4 for this section. Groups received full marks only if their notebook was exceptionally well presented.

Reasons groups lost marks

□ **Not enough inline comments accompanying the code.**

□ **Excessively long cell outputs.**
For example, looping through the entire dataset and printing each row.

□ **Sections in the notebook were unclear.**
Some notebooks could have benefited from more headings in markdown cells.

□ **Installing a Python package inside their Jupyter notebook.**

□ **Using a Python package that wasn't on the list of allowed packages.**