Assignment A1: Classification

Student Name	Yash Mistry		Student No	218612723
Problem attempted	Complex Model 80-100%	Simple Model 40-79%	Student Id	Ydmistry
Place "Yes" in one only	?	yes	*Do not attempt a complex model unless you can complete a simple model first!*	

Partial Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Problem	5	4	3	2	1	0
Data Exploration	10	8	6	4	2	0

Final Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Solution	5	4	3	2	1	0
Data Preparation	20	16	12	8	4	0
Model Development	30	24	18	12	6	0
Model Evaluation	30	24	18	12	6	0

Brief Comments	(see notes below)	Total 0 to 100

Read these notes

These and the following notes are trying to help you!
Read the rubric on how the report content is going to be assessed!
Your partial submission may not be perfect but has to reflect a genuine effort.
We expect your partial submission sections to be improved for the final submission.
We will not look at your partial submission until we mark the final submission.
We will assess the final submission and its mark stands.
However, we will deduct marks if the quality of the partial submission is poor.

Note: We will severely penalise the final submission when
the partial submission is late or missing.

Do not attempt a complex model unless you can complete a simple model first!
If you cannot formulate a complex problem, you will not get extra points for other complex criteria.
Use the font already used in the template, i.e. Arial 10 (and not MyTiniestFont 2).
If any submission aspects could only be determined by running the process, the marks will be severely reduced.

Note: If it is not in this report, it does not exist and does not get marked!

So, we will not check your RapidMiner scripts to check anything that was missing from the report.
Any part which carries points but is missing in the report gets zero marks.
We expect consistency between the report and RapidMiner scripts, so...

Note: Anything reported that cannot be substantiated
by RapidMiner scripts will be marked as zero.

It means that we will check the RapidMiner scripts when in doubt or even just curious.

Executive problem statement (one page)

Aim

In this Assignment we have to figure out, which of the Airbnb rentals are good to their neighbourhood. Also speculate how/if the reviews have any relation with the attractiveness with the property.

Expectation

To identify which of the following attributes should be taken into consideration when it comes to attractiveness of the property and which is the attractive rentals of Airbnb come under which neighbourhood.

Working

We must work our way into to figuring out what kind of problems and what kind of attributes we will face and clean up during the process.

Factors such as neighbourhood, the price of the rentals, the even the reviews will come under consideration when it comes to figuring out the solutions of our problems

With this process, we will benefit with out understanding what kind of neighbourhood are most attractive. We will even find out what kind of rental attracts majority of reviews.

Data exploration (one page)

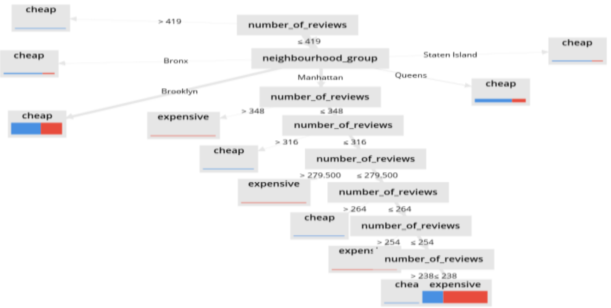

Fig1.1

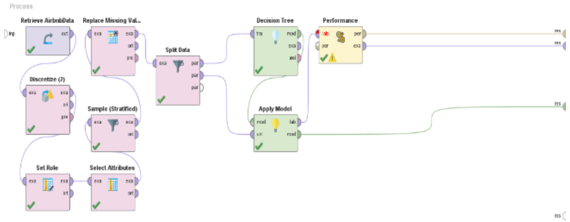

Fig1.2

To look at what price point the rentals at Airbnb are well worth it , we need to check that which of the rentals are in attractive range to the people with price point and reviews point. To figure out what type of rental attracts more in the data we have to do more process. Places such as Brooklyn and Manhattan have the most reviews. But with more processing we can figure out the rental situation of the property and its attractiveness.

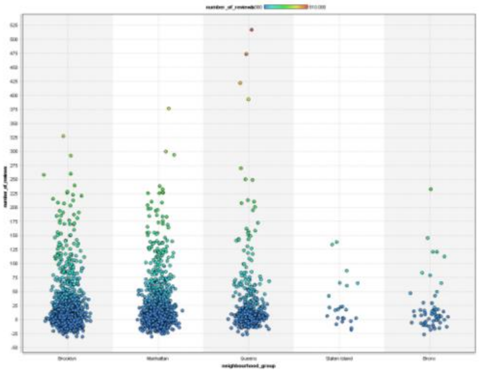

Fig1.3

from the Fig 1.1 ,1.2 and 1.3 we can see that most of the reviews are in the neighbourhood location of Brooklyn, Manhattan, Queens.

From the process I can see that most of the rentals who are from Brooklyn,Manhattan,Queens are much more worth the price given their room type.

Executive solution statement (one page)

Aim
To clearly articulate your understanding of the business solution to management.

Simple Model
The business solution is succinctly described for executives and justified.
Cross-references with the technical sections of the report provided for support, e.g. to tables, charts and plots.
Answer to business question (C) is given and justified.

Complex Model
In addition, business decisions and actions that are supported by the complex solution are well explained.
Cross-refs with technical sections support exec summary.
Answer to business question (C) is given and justified.

Hints
Ensure that whatever problem you describe can be solved using the provided data.
Make sure the exec summary describes the solution from the business perspective and not a technical perspective.
Use business language and not computer / mathematical / statistical / data science language.
The solution statement should describe the high level benefit and not the methods of their delivery.
Think and state who the company clients are and what the likely benefits of this project are for them.
Ensure that your solution clearly matches the problem statement.
Ensure that the solution is formulated in terms of achieving the high-level aim.
Do not include any charts or tables in the solution statement section.
However, cross-reference your problem statement with tables or charts from the following section, e.g. you can refer to them as "… (see Figure 1)" or "As shown in Table 4…".

If you need to support your statements / analysis / argument with references to any published materials, use Harvard citation style as described in: http://www.deakin.edu.au/students/studying/study-support/referencing/harvard. As the executive summary should not take even one page, we suggest to include your bibliographic references at the bottom of this page, immediately below the executive summary (or problem description).

All comments, such as this, which are not part of your submission should be deleted to save space.

Data Preparation (one page)

Aim
To demonstrate your understanding of data by describing complex relationships between attributes.
Depending on the selected model some attributes may need to be transformed or new attributes created.

Simple Model
Relationships between attributes, are explored and visualised.
Selection of labels and predictors recommended and justified.
Attributes are generated and transformed as needed.
All charts annotated (with text and arrows) to highlight important insights.

Complex Model
In addition, in a complex solution all attribute weights are used to select the most useful attributes, which is tested and justified.
All missing values and data errors handled adequately.

Hints
Many hints are identical to those in the section on “Data Exploration” so read them!
Some preliminary data exploration has already been conducted in the previous sections.
Focus on depicting attributes relationships and not their individual characteristics,
Include here the text of your analysis with tables and charts.

Your analysis and description could include:

- What relationships exist between numerical attributes (e.g. using correlation tables or scatter plots)
- What relationships exist between nominal attributes (e.g. using stacked bars, block, heat/tree maps)
- What is the weight between predictor attributes and labels and what does it mean
- Any other more creative visualisations / tabulations, possibly with some value aggregation

Avoid indiscriminate “dumping” of tables, charts or code into this section – all content must have some purpose.
All included charts, tables or RM processes (or their parts) have to be described or used in the discussion.
Make sure that all charts, tables and important results are labelled for cross-referencing, e.g. “Figure 1 - Histogram of Overall Rating” or “Table 4 – Comparison of model performance”.

All comments, such as this, which are not part of your submission should be deleted to save space.

Model Development (one page limit)

Aim
To explain details of developed classification models and selected methods for data preparation and reporting.

Simple Model
k-NN classification model included.
The analytic process, its operators and their parameters described and annotated (with text and arrows).
The values of the model parameters are justified (such as k).
Operators annotated (with text and arrows) to highlight important insights.

Complex Model
In addition, when working on the complex solution, a Decision Tree is included as the second classification model.
The model parameters must be selected and justified (e.g. the values for “depth” or “pruning”).
Class imbalance is investigated, dealt with and justified.

Hints
Your textbook will be extremely helpful in this task.
Include here screenshots of all or parts of the RM process.
If your process is very large, consider splitting it into sub-processes or separate processes.
If your process does not fit into this page, include only the most important parts.
By including arrows and text boxes (e.g. with numbers to refer to) annotate each operator and its properties.
Note that some of your justifications may utilise cross-referencing with tables or charts from other sections.
Avoid indiscriminate “dumping” of RM processes/models into this section – all content must have some purpose.
You may include a brief description of the operators and what they did but this is NOT the aim of this section.
Do not include definition of terms or a “textbook” description of operations – we already know this!

All comments, such as this, which are not part of your submission should be deleted to save space.

Model Evaluation (one page)

Aim
To report and explain the performance of developed classification models.

Simple Model
The model is hold-out validated using accuracy and kappa.
Validation results are analysed, interpreted and reported.
A statement is included with justification on to what degree the model advice can actually be trusted.

Complex Model
In addition, all models of a complex solution are cross-validated (in which case no hold-out validation is needed) and independently tested, performance tabulated and compared - the best model identified.
In addition to accuracy and kappa measures / charts such as AUC and ROC are also used.

Hints
Your textbook will be extremely helpful in this task.
If you have few results to report, include here screenshots of your results, e.g. confusion table or ROC charts.
If you have many results to report, include here a table of all results.
You need to describe and explain your results.
It is the most important that you include here the detailed analysis of your results –
explain the impact of the obtained results on the future use of the model to support decision making.
Avoid indiscriminate "dumping" of performance results – all content must have some purpose.

All comments, such as this, which are not part of your submission should be deleted to save space.

Any materials, analysis or reports that do not fit into 7 (seven pages in total, including the front page) will not be looked at or marked.