School of Mathematics and Statistics

MAST30034: Applied Data Science

Assignment 2

Due date: No later than 5:00pm on Friday 9th October 2020

Weight: 20%, Maximum marks: 20

Project Overview

The aim of this project is to make a qualitative analysis of the New York City Taxi and Limousine Service Trip Record Data. The data set covers trips taken in various different types of licensed taxi and limousine services in the New York City area. The data is freely available to download from `www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`.

You are free to choose the tools and techniques you use to perform the analysis. You will be required to prepare a self-contained report of up to 15 pages detailing the steps taken in performing your attributes analysis and the output of modelling and analysis.

Project Details

You are free to select a period of time to analyse, as well as the type of licensed taxi you wish to focus on, it is mandatory for you to work on a large scale of dataset ($n \geq 100000$). You are also free to select attributes you want to study. You are required to analyse at least **FIVE** attributes (before feature selection) for this assignment. These attributes are to be used as candidate features for model selection and/or parameter tuning. Your report should explain and justify your selection decision. The first stage of the project is to access and report the target data via descriptive statistics for a group of selected attributes to characterise the data and make a clear research goal. Following that, you should build at least **ONE** appropriate statistical model to explain the relation between your attributes. You are expected to refine your model (e.g. feature selection for supervised learning models or a suitable criterion for optimal number of clusters), and evaluate the performance of your model (e.g. classification error, MSE, SSE). You are also expected to highlight key findings based on your results and note findings that you believe are important or unanticipated.

Report

Your report should be a maximum of 15 pages and cover at least the following items:

- Identify the research problem and attributes you want to study.
- Choose appropriate data and describe the procedures for processing and analysing the data.
- Interpretation of results: Description of trends, comparison of groups, or relationships among your chosen attributes.
- Identify the most important attributes based on certain criterion and your chosen response.
- Evaluate the performance of your model with an appropriate procedure.
- Make recommendations or prediction based on your results, or actions to be taken in practice to further improve the performance.

Citation style

You are free to use any citation style such as APA, Harvard etc. Please ensure that the name, year and title of publication is clearly stated in the reference page.

Assessment

Your report will be assessed according to the following checklist:

Research problem, quality and clarity of report (4 marks)	□ Lists appropriate research goals succinctly (1m) □ Quality writing, spell-checked, correct grammar, and comprehensible sentence structures (1m) □ Identifies potential stakeholders, and explain how research is relevant to stakeholders (1m) □ Conclusion: provides recommendations for potential stakeholders based on analysis of findings (1m)
Data and Attribute Selection (2 marks)	□ Clearly states and justify data period (1m) □ Clearly states and justify choice of five (or more) attributes to be analysed (1m) □ Use of an appropriate external dataset (Bonus: 2m)
Pre-processing and Cleansing (3 marks)	□ Clearly states pre-processing and/or feature engineering steps (1m) □ Clearly states data cleansing steps (1m) □ Appropriate justification for pre-processing steps, as well as steps for handling missing data (1m)

Descriptive analysis (3 marks)	☐ Appropriate choice of summary statistics and suitable graphical tool for presenting for each attribute (1m) ☐ Investigate pairwise relationship between attributes (1m) ☐ Clear description of each attribute based on summary statistics and appropriate plots (1m)
Modelling (6 marks)	No marks possible without any statistical modelling ☐ Clearly specificies the statistical model, with appropriate use of equations (1m) ☐ State and check all model assumptions (1m) ☐ Succinctly justify choice of model, including how it helps to address research goal (1m) ☐ Fit model to training data with all attributes before model refinement (1m). ☐ Refine model and find optimal values of tuning parameters using an appropriate procedure (1m). ☐ Evaluate model performance on testing/validation data with appropriate metrics and procedure (1m)
Analysis of result(s) (2 marks)	☐ Fit final model after refinement and interpret model parameters where appropriate (1m) ☐ Make recommendation on how to use final model (1m)

Submission details

Submissions should be made via Turnitin on the LMS.

- A penalty of 10% of the available marks will be deducted for each day or part-thereof that the submission is late.

Extension policy: More details on the process of applying for a penalty waiver can be found on `https://ask.unimelb.edu.au/app/answers/detail/a_id/5667/~/applying-for-an-extension`

Plagiarism policy: You are reminded that all submitted project work in this subject is to be your own individual work. Automated similarity checking software will be used to compare submissions against each other and known public source code. It is University policy that cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student concerned.

Tips on Getting Started

If you're unsure of how to start this project, try going through some of the models you have used in the previous subjects. Depending on the choice of model(s) and attribute(s), you may need to perform some creative feature engineering or transformation on the dataset. Following this, you should then discuss any issues or any interesting aspects that appeared during your experimentation.

For example, consider the scenario where your data is linearly separable through the use of a transformation:

- Consider performing a descriptive analysis before model fitting to identify issues with your data such as linear separability, missing values, outliers etc.
- For supervised learning models, consider the linear separability of your data. When there is linear separability, some models perform well (e.g. SVM), whereas some models (e.g. logistic regression) fail to converge. The kernel trick may be used to induce linear separability if it is desired.
- Penalised regressions (e.g. ridge, LASSO) tend to perform poorly if the number of features is much lesser than the sample size.
- Consider performing feature engineering to generate more useful features. Do not perform it excessively though as it will lead to overfitting.

Your report should justify any feature engineering or transformation, as well as the choice of model. Additionally, aim to discuss the expected vs actual performance of the model and report on any notable finding you came across.

Further Hints

- Sub-sampling may help you to increase the scope of data you can cover.
- Explain your handling of missing/unreasonable data and why any missing data does not undermine the validity of your analysis. You should report and justify the size of data that has been removed.

- When you are trying to make comparisons, make sure your measurement is of the same scale.
- You may want to try different methods for your analysis.
- Always tell the reader what to look for in tables and figures. Be as factual and concise as possible in reporting your findings.
- If necessary, define unfamiliar concepts and provide the appropriate background information to aid your finding.