To-do list

1. choose variables (rates/speed ..., external data preferred)
2. data pre-processing & cleaning
3. basic analysis (Descriptive Statistics) Lec 3

☆ 4. Models : find 1 response variable & predictors Lec 4

Apply linear/glm/... ⟹ perform gradually a model selection

(improve model. e.g. remove useless predictors)

(model comparison technique)

5. Analysis of the final model (e.g. goodness of fit,
test of significance,
analysis based on reality)

School of Computing and Information Systems
MAST30034: Applied Data Science
Assignment 2
Due date: No later than 11:59pm on Tuesday 10th Sep 2019
Weight: 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 https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. The whole data set is large, covering many years, you are not expected to analyse it all, only a subset that you are free to choose.

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, however, large scale of data is preferred. You are also free to select attributes you want to study. Your report should explain and justify your selection decision. better external source / transformation of original

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 a statistical model to explain the relation between your input variables and response variables (both types of variable may be chosen by you from the data). Transformation of the data or using external data set may result in higher marks, if it is clearly justified. You should then refine your model. For example, you may improve your model by investigating the correlation of your selected attributes, and ranking the importance of your input variables based on clearly self-defined criterion. Justified your selection of the final model. 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.
- Make recommendations or prediction based on your results, or actions to be taken in practice to further improve the performance.

Assessment

Your report will be assessed across a number of areas, including:

- Quality of your research problem
- Justification of data and attribute selection
- Quality of your model and attribute relations
- Quality and clarity of interpretation of results
- Quality and clarity of report

Submission details

Submissions should be made via Turnitin on the LMS.

- Late submissions will incur a deduction of 2 marks per day (or part thereof).
- If you submit late, you MUST email the subject co-ordinator, Chris Culnane, cculnane@unimelb.edu.au.

Extension policy: If you believe you have a valid reason to require an extension you must contact the subject co-ordinator, Chris Culnane cculnane@unimelb.edu.au at the earliest opportunity, which in most instances should be well before the submission deadline.

Requests for extensions are not automatic and are considered on a case by case basis. You will be required to supply supporting evidence such as a medical certificate. In addition, your git log file should illustrate the progress made on the project up to the date of your request.

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.

Further Hints

- Using external data set may result in higher marks.
- 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 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.