**Structure and Guideline of the CourseWork**

**Data Mining and Web Analytics**

Introduction: The CW dataset is about **Car Price**. You are supposed to predict the car prices based on a set of predictors. In the original dataset, there are more than 20 predictors, but you have to choose **only ten** of them for the coursework. This is due the practicality, as more predictors means more plots, and more writings, and more features to work with, etc. Choose the predictors either by common sense, or after an initial modelling. I suggest that you choose the variables after an initial modelling on all predictors and finding the most influential variables.

The dataset and the metadata are provided in the coursework folder.

The structure of the report and the points to address are as follow:

**Introduction (5%)**

* What is the problem?
* Who are the stakeholders?
* Why does this problem matter to each stakeholder?

**Data Set and Visualization (40%)**

* Type of the dataset
* Dimensions of the dataset
* Variables, definitions, their types, and their roles
* Verbal presentation of at least two records
* Level of the data, what is each record about?
* Uni-variate visualization and commenting

o The measures of centrality and spread should be pointed to

* Bi-variate visualization and commenting (target vs predictor)

o Especially comment on the predictive power of the features

* Data Quality Assessment and treatment

o Outliers and extremes (definition and treatment)

o Missing values (definition and treatment)

**Model and Results (40%)**

* Predictive modelling formulation (what you want to predict and how)
* Type of the problem (classification? Regression?)
* Do we need data partitioning? Why?
* Underlining the performance metrics that you can use, and your metric choice for this problem (the definition of the metric, and the rationale behind picking it)
* Base-line model set-up and getting its performance on the testing set
* Feature engineering efforts
* Efforts to improve the model’s performance as summary table of the models you have used,

hyper-parameters of those models, any change in the features, and the performance on the

training and on the testing set

* A very brief introduction of every supervised/unsupervised method that has been used
* Error cost analysis from the point of view of stakeholder(s). Which error is the worst and

why?

**Conclusion (5%)**

* A non-technical summary of the project, the next steps that might be taken to improve the performance, and any new idea regarding relevant projects

o Non-technical means no jargon terms

o The audience of this report are decision-makers/stakeholders. They are domain savvy but not machine learning savvy

**Writing Style (10%)**

* Maintain a formal tone: Use a professional and objective tone throughout the assignment.
* Use headings and subheadings: Organize your report using headings and subheadings to guide the reader and create a clear structure.
* Use academic sources: When referencing or citing information, use credible academic sources such as scholarly journals, books, or reputable websites.

**Rules:**

* You work in a group, but you have to submit your reports separately.
* You must name all the group members on top of your report.
* The group reports should not be identical. You work on the problem with your groupmate, and you write down the report based on your own understanding, your own words, and your own perspective. Do not duplicate one report. You are supposed to write-down the report yourself, even though it would be similar to your groupmate’s, especially in the technical parts. The deviation is more expected in the intro or conclusion. [the variables, the plots, and the modelling parts are identical eventually, but your comments and notes should be in your own words.]
* Every plot or table must be placed in the text in the corresponding subsection.
* Follow the provided guide, use it to structure your report, and address the bulletpoints as many as you can.
* Use concise language, wandering or giving superfluous explanation undermines the credibility of your work.
* Use short paragraphs, with a clear relation to the previous paragraph, and expressive starting sentence.
* The word cap is 3000. I care about the quality, so verbose/prolix reports would be penalized, as it happens in the industry.
* Follow the ethics code. No copy-paste from anyone or anywhere (Turnitin will highlight any copied parts).