CS5228 Final Project – Task Description (preprint)

2021/2022 Semester 1

1. Overview

The purpose of the final project is for you to show how you perform data mining tasks in a practical setting. Given a dataset and a set of tasks, you need to select appropriate techniques to solve these tasks, justify design and implementation issues, as well as interpret your results and assess any limitations of your approach. We tried to design the project to make the tasks both interesting and relevant, but also to provide you with enough flexibility for your approach addressing the tasks. As often emphasized in the lectures, there is rarely one single best way to solve a data mining task, and many steps will therefore benefit of even require your own creativity to come up with appropriate solutions. While this project will require a certain amount of effort, we also hope that you will have fun completing it, and that it will be a valuable learning experience.

In this project, we look into the market for used cars in Singapore. Car ownership in Singapore is rather expensive which includes the very high prices for new and used cars (compared to many other countries). There are many stakeholders in this market. Buyers and sellers want to find good prices, so they need to understand what affects the value of a car. Online platforms facilitating the sale of used cars, on the other hand, want to maximize the number of sales/transactions. But also the government is interested in the car market to evaluate and potentially adjust their laws and policies regarding car ownership. In short, data about car resale transactions can give you a wide range of interesting insights. To cover various such as aspects and to provide you with some flexibility, the final project is split into 3 subtasks:

- **Task 1: Prediction of Car Resale Prices** – given the information about a used car, your task is to predict its price based on past resale transactions. This regression task is implemented as a Kaggle InClass Competition.
- **Task 2: Car Recommendation** – given the information about a car a user is interested in, your task is to find "meaningful" alternatives that can be shown to the users in form of recommendations.
- **Task 3: Open Task** – given the car resale dataset, you explore own ideas to gain interesting insights into the data. There are no specific rules, apart from your insights being *non-trivial* and *useful* (cf. our definition of data mining in Lecture 1).

In the following, after giving a brief overview to the core dataset, we detail on the 3 subtasks. If you have any questions, please do not hesitate to post your question on the LumiNUS forum, or send me an email (chris@comp.nus.edu.sg)

2. The Dataset

The core dataset of past car resale transactions has has been collected from sgCarMart. Figure 1 shows an example listing for a used car for sale on sgCarMart, giving an overview over the different features for a transaction (e.g., the mileage, road tax, coe, number of previous owners, power, and so on). This dataset provides the basis for all 3 subtasks. However, you are free to explored and collect additional data to further improve your results. This might be particularly interesting for Task 3 where you can define your own goal and task you want to address.

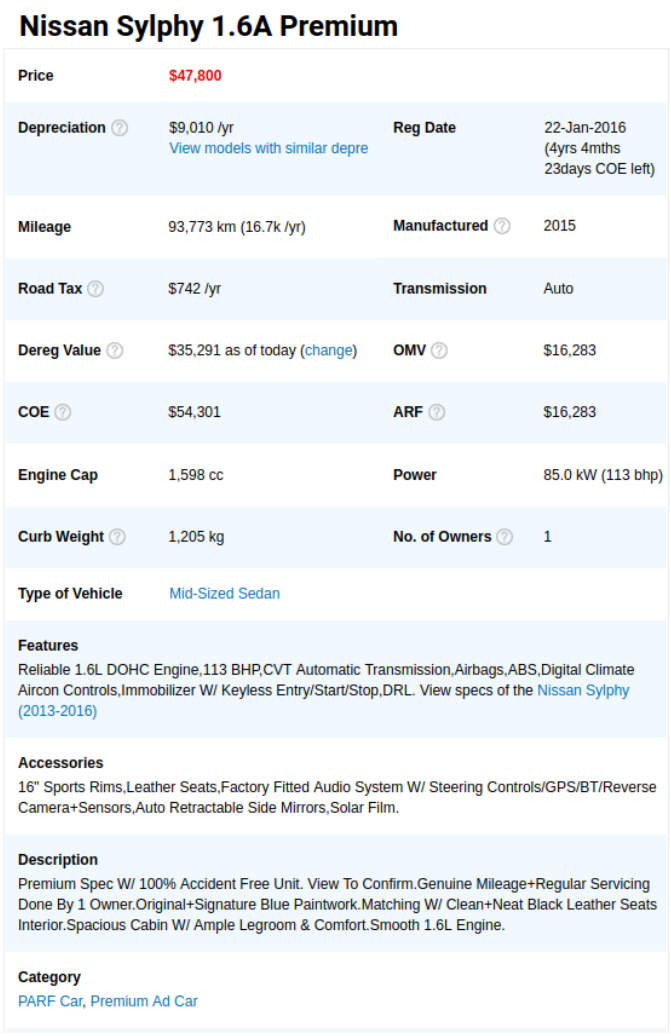
We believe that the meanings of all attributes are rather self-explanatory. However, we do provide a brief description of each attribute on the Kaggle page for our InClass competition. If you still have questions, you can ask your question on the forum or in an email.

3. The Tasks

The following subsections give an overview to the 3 subtasks and outline the requirements for the final submission. As a general note, when tackling the different subtasks, you are not limited by the methods or algorithms covered in the lectures. You are also free to use any packages (NumPy, pandas, matplotlib, scikit-learn, NetworkX, etc.) to implement your solutions.

3.1. Task 1: Prediction of Car Resale Prices

The goal of this task is to predict the resale price of a car based on its properties (e.g., make, model, mileage, age, power, etc). It is therefore first and foremost a **regression task**. These different types of information allow you come up with features for training a regressor. It is part of the project for you to justify, derive and evaluate different features. Besides to prediction outcome in terms of a dollar value, other useful results include the importance of different attributes, the evaluation and comparison of different regression techniques, an error analysis and discussion about limitations and potential extensions, etc.



Nissan Sylphy 1.6A Premium

Price	$47,800		
Depreciation	$9,010 /yr View models with similar depre	**Reg Date**	22-Jan-2016 (4yrs 4mths 23days COE left)
Mileage	93,773 km (16.7k /yr)	**Manufactured**	2015
Road Tax	$742 /yr	**Transmission**	Auto
Dereg Value	$35,291 as of today (change)	**OMV**	$16,283
COE	$54,301	**ARF**	$16,283
Engine Cap	1,598 cc	**Power**	85.0 kW (113 bhp)
Curb Weight	1,205 kg	**No. of Owners**	1
Type of Vehicle	Mid-Sized Sedan		

Features
Reliable 1.6L DOHC Engine,113 BHP,CVT Automatic Transmission,Airbags,ABS,Digital Climate Aircon Controls,Immobilizer W/ Keyless Entry/Start/Stop,DRL. View specs of the Nissan Sylphy (2013-2016)

Accessories
16" Sports Rims,Leather Seats,Factory Fitted Audio System W/ Steering Controls/GPS/BT/Reverse Camera+Sensors,Auto Retractable Side Mirrors,Solar Film.

Description
Premium Spec W/ 100% Accident Free Unit. View To Confirm.Genuine Mileage+Regular Servicing Done By 1 Owner.Original+Signature Blue Paintwork.Matching W/ Clean+Neat Black Leather Seats Interior.Spacious Cabin W/ Ample Legroom & Comfort.Smooth 1.6L Engine.

Category
PARF Car, Premium Ad Car

Figure 1. Example of a listing for a used car for sale on sgCarMart. The dataset will provide all attributes contain in the listing as attributes.

Task 1 will be implemented as **Kaggle InClass competition**. On the competition page on Kaggle, you can download various files. `train.csv` and `test.csv` split the dataset into the training and test set. Naturally, `training.csv` will contain the numerical attribute `price` for each car; this column is missing in `test.csv`. The predictions you submit should be via a `csv` file with a single column that contains the predicted resale price for each row in the test dataset; we provide the file ‘`example-submission.csv`‘ to show you an example of a submission. To prevent overfitting to the leaderboard, you are allowed to make 3 submissions per day to the leaderboard.

3.2. Task 2: Car Recommendation

Many websites recommend users similar or related content while browsing their content. For example, when watching a video on YouTube, you have quick access to other videos that YouTube thinks you might also be interested in. However, sgCarMart does not provide such a feature when browsing a car listing. The goal of this task is to design and implement a **recommendation system** to support such a feature. Similar, to Task 1, there are different approaches to solve this. It is therefore important – apart from the implementation of your solution itself – the discuss alternative approaches and justify your design decisions. If you make any important assumptions (e.g., users can specify their preferences, or you utilize a user’s browsing history of previous cat listings), you should also make those explicit and justify why those are realistic and meaningful assumptions.

Since ”good” recommendations are very subjective, Task 2 will require you to submit a **self-contained Jupyter notebook** incl. any code or data used in this notebook (note that your source code will be part of your submission anyway). The notebook should motivate your approach and important design decision (like in the final report) and should provide a method `get_top_recommendations()` that computes your, say, top-10 recommendations given a row from the dataset (i.e., given the listing if a car a user is currently browsing on). Of course, `get_top_recommendations()`, may use any additional information you want to utilize beyond the provided dataset. We will provide more details regarding the submission of your solution for Task 2 towards the end of the project.

3.3. Task 3: Open Task

Task 1 & 2 describe two goals to gain insights into the dataset. Now, in Task 3, we expect you to explore your own goals to extract further **non-trivial and useful** information from the data. You are welcome to utilize your results from Task 1 & 2. This task can be addressed from the perspective of the different stakeholders, e.g.:

- Buyers might benefit from an advanced search engine that ranks results based on whether a car listing seems to be a bargain or not (bargain = true price is much lower than the predicted price).
- sgCarMight might benefit from an in-depth analysis of past transactions find the most (un-)profitable car models.
- Car manufacturers might want to compare used cars’ prices with the prices of new models to asses the long-term value of makes and models.
- A governmental organization might want to assess if Singaporean car buyers value eco-friendly car models or not.
- ...

As you can see from these previous example, tasks may require to collection of additional data for the analysis (e.g., collecting the prices of new car models). It’s up you to design your task around available data. Your report should contain a clear motivation for your task of choice and why the results are of practical interest for any stakeholder. If

your task computes individual results for a given input (like in Task 2), we also strongly encourage to organize your solution as a self-contained Jupyter notebook.

4. Deliverables

4.1. Progress Report

The progress report will be a simple slide deck as PDF document of approx. 10-15 slides. The purpose of the progress report is two-fold: (a) to give us a chance to check if your project goes into the right direction, and (b) to provide you with a little incentive to better start early than too late. There is no official layout or structure. As the name suggests, it should outline your progress with your project work (e.g., goals and questions, EDA results, first design decisions or results, but also with issues/challenges/obstacles that you are facing). The last 1-2 slides should outline the next steps until the end of the project.

- Deadline for the progress report: TBD

4.2. Final Report

The final report will be a PDF document of at most **10 pages** including tables, plots and figures, but excluding references and the appendix. The appendix may contain supplementary content but should be used sparingly. As a rule of thumb, the report should be readable and completely comprehensible without the appendix. This typically may includes plots or tables that elaborate in the results of your EDA or your evaluation. For the layout and presentation in the report, we will provide a Word and LaTeX template.

4.3. Structure & Content

Your report should include the name and student IDs of all team members as well as your team name. Please also include a breakdown of your workload, i.e., some overview what team member was (mainly) responsible for each parts of the project. This can be a table, Gantt chart, etc. to be added to the appendix.

While the overall structure of the report is up to you, it should cover the following aspects:

- **Motivation.** Motivate and outline the goals and questions you address. Note that this is also relevant for Task 1 & 2 as different teams may focus on different aspects for those tasks.
- **Exploratory Data Analysis & Preprocessing.** Explain and justify your approach to understand the data, and how it informed your data preprocessing steps (e.g., data reduction, data transformation, outlier removal, feature generation).
- **Data Mining Methods.** Describe how you chose and applied appropriate data mining techniques (e.g., regression and classification models, recommendation methods). This description should include which techniques you used, how you chose their hyperparameters, etc. Note that you do not need to explain the techniques themselves. However, in case of more advanced methods or models, you should add relevant references.
- **Evaluation & Interpretation.** Evaluate and compare the performance of different methods. Discuss which method(s) performed best and why. Understand in what cases your methods perform bad, and discuss principle limitations and potential future steps for improvement.

The structure of your report should, of course, reflect the 3 different subtasks you need to address in this project. While *EDA & Preprocessing* might be only one section in your report, *Data Mining Methods* and *Evaluation & Interpretation* will arguably require their own instances for each subtask.

4.4. Submission

The final submission contains both the report as PDF document as well as your source code, uploaded to LumiNUS in a zipped folder. Instead of the source code, you can also add a link to a GitHub repository. Note that the reproducibility of your approach is part of the grading (cf. Section 5) which includes the organization and readability of your code.

- Deadline for the progress report: TBD

5. Grading

In a nutshell, a good grade requires that your approaches are well motivated and are methodologically sound, and that the outcome – mainly the report but also your source code – is of a high quality. In more detail, we weigh the core criteria for the grading as follows:

Methodological Quality (60%). While the exact distribution may depend on your chosen project and approach, methodological quality generally covers the following aspects:

- **Preprocessing:** appropriate preprocessing methods are chosen (informed by the results of the EDA) and correctly implemented; missing values, categorical attributes, etc. are handled correctly.
- **Visualization:** appropriate plots, figures and tables are used to visualize results, architectures and work flows.
- **Methods:** applied methods are well motivated and correctly implemented; alternatives are discussed and design decision are justified.
- **Evaluation:** different methods are compared or evaluated using appropriate metrics and experimental setups (e.g., cross-validation); common errors and principle limitations are evaluated and discussed.

Quality of Report (30%). The report describes your methodology and explains your results in a clear, concise

and comprehensible manner. Related work should be appropriately referenced; the limit of 10 pages should not be exceeded (excluding references and appendix!).

Reproducibility (10%). The code you submit is complete, well-organized and readable. Simply put, it should be easy for an outsider to use and understand your code to retrace your steps and reproduce results. This is particularly relevant for self-contained Jupyter notebook required for Task 2 (and maybe Task 3 depending on your goals).

Important: In case of the Kaggle InClass competition for Task 1, your position on the public and private leaderboard will only be used as part the bigger picture, primarily as part of the methodological quality. Getting a good grade does not require a top position on the leaderboards as long as the overall approach is sound of and high quality. Of course, a sound approach and good results typically go hand in hand, and results (significantly) below the average indicate problems with the methodology. The main purpose of implementing Task 1 as a Kaggle InClass competition is to provide you with incentives for solving this task, to give you a way to compare your solutions with the ones of others, and to hand out bragging rights to the top competitors.

6. FAQ

Please use this Google Doc to submit any immediate questions and comments about the project you have. Based on your feedback there, we will extend this FAQ section and publish an updated version of this project description.