CS5228 Final Project Report
GROUP NAME

CHAN CHEAH CHA
A0189006A
E0324590

HONG JIAYONG

PHANG DAO YI JOSEPH

***Abstract*—**

I. Introduction

Introduction.

A. Motivation

Motivate and outline the goals and questions you address. Why is this work important, what are the challenges, will benefits from the results. In short, what problem are you trying to solve? For example, simply aiming for a top rank on Kaggle is not a sufficient motivation :).

II. Data Exploration

In this section, we explain the data preparation process, including checking data quality, exploring the data, and pre-processing it for modeling.

A. Understanding the Data

The dataset contains both numerical and categorical attributes, which represent different aspects of the cars listed for resale. The training dataset consists of 25,000 records with 30 attributes, while the test dataset consists of 10,000 records with 29 attributes (`price` target variable excluded).

To better understand the data and its context, we conducted background research focusing on key attributes that influence resale car prices. This analysis will aid in the interpretation of our dataset and guide our preprocessing steps.

1) Price-Related Attributes: The price of a car in the resale market is influenced by several key attributes and and understanding them is essential for accurate price prediction.

- **Deregistration Value** The car's value at the end of its COE tenure. For PARF cars (less than 10 years old), the deregistration value includes both the PARF rebate and COE rebate. This suggests that PARF cars have higher overall deregistration value compared to COE cars, which may positively impact the resale price.
- **Depreciation** Depreciation value is directly computed from the resale price and deregistration value over the remaining years of COE left.

2) Market-Related Attributes: Market dynamics also play a significant role in determining resale prices. Understanding these attributes is essential for a comprehensive analysis of the resale market.

- **Demand and Supply** The oversupply of similar models can lead to lower prices, particularly for cars commonly used in ride-sharing services. Attributes such as model, make, and manufactured (year) are related to this aspect.
- **Popularity or Rarity of Car Model** Rare or exotic cars may fetch higher prices, making this a potential feature for model classification.
- **Consumer Preferences** Shifting trends towards eco-friendly vehicles can increase demand for hybrid or electric cars, which is linked to the fuel_type attribute.
- **Regulatory Changes** Fluctuations in COE prices affect different vehicle classes, necessitating potential feature construction from engine_cap, power, and registration dates.

These insights from market-related attributes will enhance our understanding of the pricing dynamics in the resale car market.

Describe the dataset in detail, including the types of attributes (categorical, numerical, etc.), and provide an overview of how these attributes might influence car prices.

B. Data Quality Check

In this dataset, significant missing values were identified across various features, as summarized below:

- **Indicative Price** This feature is completely missing for all records, which is a significant concern as it represents a key aspect of resale price analysis.
- **OPC Scheme** and **Original Registration Date** Both of these features have missing data for nearly all records, with over 98% missing. This could severely limit our ability to analyze vehicle history and ownership.
- **Lifespan** Approximately 90% of the records lack this information, making it challenging to determine the expected value depreciation over time.
- **Fuel Type** Missing for about 76% of the records, this attribute is essential for understanding consumer preferences and market trends towards eco-friendly vehicles.
- **Mileage** With around 21% of the records missing mileage data, this could impact the resale price estimation as mileage is a significant determinant of car value.

- **Other Attributes**: Several other fields exhibit varying degrees of missing data, ranging from 0.03% to 15.25%.

Basic sanity check conducted on date type columns. There are no future dates found in year, original_reg_date and reg_date. Dates for reg_date are consistent as well with no reg_date being earlier than original_reg_date. Evaluate the quality of the data by checking for missing values, inconsistencies, duplicates, and outliers. Discuss how you addressed any issues related to data completeness and correctness.

C. Initial Feature Engineering for EDA

1) Binary Variable Creation:

- **OPC Scheme** Most of the values are missing. Assuming that NA values means that the listed car is not under the OPC scheme, this column is transformed to a binary form where 1 indicates the presence of an OPC scheme and 0 indicates its absence.
- **Parf** Based on the background context, we know that PARF cars usually have a higher resale price than COE cars. This information can be extracted from the category column under the 'parf cars' and 'coe cars' labels. 1 signifies that the car is cateogrised as PARF and 0 otherwise.
- **Rare** Similar to PARF, rarity of the car can be extracted from the category column under the label 'rare & exotic'. 1 signifies that the car is labelled as rare and 0 otherwise.
- **Vintage** Similar to PARF, the vintage nature of the car can be extracted from the category column under the label 'vintage cars'. 1 signifies that the car is labelled as vintage and 0 otherwise.

2) Remaining Age: One of the key features we derived during the feature engineering process is the "AGE-remaining" column, which quantifies the remaining lifespan of a vehicle's Certificate of Entitlement (COE). This feature is particularly important in the context of the Singapore car resale market, as the COE directly impacts the vehicle's resale value.

COE information is extracted from the title column. Specifically, we looked for instances where the title contains the term "COE" and extracted the relevant date from the title. For vehicles without explicit COE dates in the title, we utilized the maximum of the original_reg_date and reg_date columns to estimate the end date of the COE. By adding 10 years to this maximum date, we derived the COE end date for those vehicles. Using the derived COE end date, we computed the "AGE-remaining" column. This was accomplished by finding the difference in years between the current date and the COE end date.

D. Exploratory Data Analysis (EDA)

Perform an initial exploratory analysis to understand the distribution of data and relationships between key variables. This may include descriptive statistics, visualizations (e.g., histograms, scatter plots), and identifying trends or patterns.

Type of Vehicle SUVs, luxury sedans, and sports cars are the most common vehicle types in the dataset. SUVs are highly represented, with relatively lower price variability compared to luxury sedans and sports cars.

Fuel Type The most common fuel type is diesel, followed by petrol-electric and petrol. Petrol-electric vehicles also show a significant range in pricing, with some outliers that could represent premium models. This suggests that hybrid vehicles are valued well in the resale market. Electric vehicles have a narrower price range compared to diesel and petrol-electric, suggesting that the market for used electric vehicles may not be as established yet, potentially affecting their resale value.

PARF cars have a slightly higher range compared to coe cars. Rare cars can command much higher resale prices.

1) Feature Correlation: The correlation matrix reveals several significant relationships among the various features in the dataset.

Firstly, there are strong positive correlations observed between power and engine capacity, with a correlation coefficient of **0.87**. This indicates that vehicles with larger engine capacities tend to exhibit higher power outputs. To leverage this relationship, future analyses could benefit from creating a power-to-weight ratio feature, which may provide additional insights into vehicle performance.

A high correlation of **0.95** between road tax and engine capacity suggests that vehicles with larger engines incur higher road tax. This relationship implies that including engine capacity as a feature in price prediction models is crucial, especially for regulatory considerations.

There is a strong positive correlation of **0.81** between depreciation and price, indicating that vehicles with higher prices typically experience lower depreciation rates. This relationship suggests that enhancing the model with features capturing vehicle condition or age could improve price predictions.

A very high correlation of **0.92** between deregistration value and price reinforces our research that higher-priced vehicles often possess higher deregistration values. Incorporating deregistration value as a predictor variable will likely enhance the model's predictive capability.

On the other hand, a negative correlation of **-0.52** between the number of owners and depreciation suggests that vehicles with more previous owners tend to depreciate more. This finding highlights the importance of considering ownership history as a feature in the model, as it could provide valuable information on expected depreciation.

Additionally, the correlation of **-0.39** between mileage and price indicates that higher mileage vehicles are generally associated with lower resale prices. Thus, mileage should be treated as a critical variable in the price prediction model, as it could significantly influence resale values.

Moderate correlations also exist, such as the **0.70** correlation between power and price, which further supports the inclusion of power as a predictor variable in the model. Moreover, the correlation of **0.60** between curb weight and road tax suggests that heavier vehicles typically incur higher road taxes. This relationship may warrant further investigation into how weight influences pricing and regulatory costs.

The "AGE-remaining" column has a positive correlation of **0.40** with "price," indicating that vehicles with more time left on their COE tend to maintain higher resale values. This is crucial for potential buyers as a longer COE often equates to less immediate financial obligation.

In contrast, vintage and OPC scheme exhibit low correlations with other numerical features, indicating a lack of strong linear relationships.

Given these insights, future steps should include feature engineering to create interaction terms based on these correlations, as well as investigating multicollinearity among the highly correlated features.

III. Data Preparation

A. Creating Interaction Terms

power_engine is the interaction term between power and engine_cap. As mentioned in Feature Correlation subsection, there is strong positive correlations between these two features.

$$\text{power\_engine\_interaction} = \text{power} \times \text{engine\_cap} \quad (1)$$

B. Data Cleaning

Handle missing values, outliers, and any inconsistencies in the dataset. Discuss the strategies used, such as imputing missing values or removing invalid records.

1) Handling Missing Values: KNNImpute

2) Handling Outliers: We utilize Principal Component Analysis (PCA) to reduce dimensionality, followed by DBSCAN to identify outliers in the dataset. The dataset is standardized to ensure all features contribute equally to distance calculations. PCA is applied to the standardised dataset to transform the high-dimendional data to a lower-dimensional representation while preserving as much variance as possible. This is to facilitate the clustering process by reducing complexity. NearestNeighbours used to find the k-nearest ditances. The knee point value is used to determine the epsilon for DBSCAN. After DBSCAN was run, the data points that are not assigned to any cluster were classified as anomalies.

There are 57 anomalies found. 57% of the cars are rare sports cars like ROLLS-ROYCE or FERRARI. Since high performance sports cars have difference price dynamics due to their rarity, brand prestige and demand, it is expected for these data points to be outliers. Interaction terms like power_engine ratio and the rare should be able to capture these attributes which will help the model understand the price dynamics and recognise their impact on resale prices effetively. These outliers labelled as rare will not be removed since it will lead to information loss on rare premium cars with lower frequency counts.

The other outliers consist of points with extremely high values of depreciation above $500,000 or extremely low prices at around $38000 for a car with 8 years of AGE-remaining. Since the number of outliers (24) is low, we can safely remove these outlier without losing too much information.

We also remove data points where AGE-remaining is less than 0. This is logically inconsistent because a vehicle cannot be sold if it is no longer operational or has exceeded its regulatory lifespan. Hence, The presence of such outliers can negatively affect model accuracy and generalizability.

C. Data Reduction

Reduce the dimensionality of the data by removing redundant or irrelevant features. This could include dropping highly correlated variables or irrelevant fields, as well as feature selection techniques. Use AGE_remaining instead of AGE_current since there is a a positive correlation with price.

D. Data Transformation

Transform the data as needed for modeling. This may involve scaling numerical features, normalizing the data, or applying logarithmic transformations to certain attributes.

E. Data Discretization

Discuss how you discretized continuous features (if necessary), grouping them into categories or bins for better interpretability or model performance.

F. One-Hot Encoding (OHE)

Explain how you converted categorical variables into numerical format using One-Hot Encoding (OHE) to ensure the machine learning models can interpret these variables. Provide details on which categorical variables were transformed and how.

IV. Data Mining Methods

A. Regression Techniques Applied

Detail the regression techniques you selected, such as Linear Regression, Random Forest, or XGBoost. Briefly mention how you chose these models.

B. Hyperparameter Tuning

Describe the process of hyperparameter tuning, including the methods (e.g., grid search, random search) used to optimize the models.

C. Feature Importance

Explain how you measured the importance of different features in your regression models. This could include methods like permutation importance or looking at feature coefficients in linear models.

D. Model Evaluation

Evaluate the models using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Compare the results between different models to determine which performed best.

E. Error Analysis

Discuss the error distribution, identifying where the models underperform (e.g., particular car types, extreme mileage values). Analyze the causes of errors and any data or model limitations.

F. Limitations & Future Work

Explain the principal limitations of your approach. This might include data constraints, model assumptions, or computational challenges. Suggest potential extensions or improvements for future work, such as more sophisticated feature engineering or applying additional machine learning techniques.

V. Conclusion

Summarize the findings of your project, including key insights on predicting car resale prices, feature importance, and the performance of various models.

VI. Instructions (TO BE REMOVED)

The final report will be a PDF document in the format of a scientific paper of at most 8 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. The appendix typically may include plots or tables that elaborate on the results of your EDA or your evaluation. For the layout and presentation of the report, we provide templates Download templatesfor Word and LaTeX.
Your report should include the name and student IDs of all team members as well as your team name – pick something cool :). Please also include a breakdown of your workload, i.e., some overview what team member was (mainly) responsible for each part 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 – although this might differ from your exact project task:

A. Goal

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 to come up with features for training a regressor. It is part of the project for you to justify, derive and evaluate different features. Besides predicting the 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.

B. Evaluation

The evaluation metric for this competition is Root Mean Squared Error (RSME). The RSME is a common metric to evaluate regression tasks. We use the RSME (instead of the Mean Squared Error) so that the error values have the correct unit, which is SGD for this task.

C. Submission Format

Submission files should contain two columns: Id and Predicted, separated by a comma – see example-submission.csv for an example. The order of the predictions have to match the order of the test data (test.csv). For example the line with Id=0 should contain the prediction for the first test sample, and so on.

D. Report Structure

Motivation. Motivate and outline the goals and questions you address. Why is this work important, what are the challenges, will benefits from the results. In short, what problem are you trying to solve? For example, simply aiming for a top rank on Kaggle is not a sufficient motivation :).

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. 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.

VII. ORIGINAL TEMPLATE FOR REFERENCE!!!

VIII. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

IX. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

X. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections X-A–X-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m^2” or “webers per square meter”, not “webers/m^2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm^3”, not “cc”.)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{2}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(2)", not "Eq. (2)" or "equation (2)", except at the beginning of a sentence: "Equation (2) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and,

conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.