CS5228 Knowledge Discovery and Data Mining

Final Report

Group 42: Top Gear

Dong Suhong
School of Computing
National University of Singapore
A0290555Y

Wu Tong
School of Computing
National University of Singapore
A0255954R

Yang Haohong
School of Computing
National University of Singapore
A0286177N

Zhang Haochen
School of Computing
National University of Singapore
A0290607A

***Abstract*—This project addresses the challenges of predicting used car prices in Singapore, where ownership costs are notably high. Motivated by insights into the unique market dynamics, we developed models to forecast resale prices using attributes like make, model, mileage, age, and engine power. Our approach involved extensive data preprocessing and exploratory analysis to uncover key patterns. We implemented a hierarchical Random Forest model structured at multiple levels (make, model, global), alongside linear regression and XGBoost models. While the hierarchical approach added depth, it did not significantly improve accuracy, highlighting complexities in capturing price-influencing factors. The study offers a framework for stakeholders to better understand and navigate the used car market in Singapore.**

***Index Terms*—component, formatting, style, styling, insert**

I. Motivation

Upon arriving in Singapore for my graduate studies, I had my first introduction to the expensive car ownership here through a conversation with my landlord. During our discussion about housing arrangements, he asked if I needed parking, and, being unfamiliar with Singapore's automotive landscape, I took the opportunity to learn more about the local situation. To my surprise, I discovered that car prices in Singapore are staggeringly high—usually twice or even higher than that in China. My landlord suggested considering a used car instead, emphasizing the affordability and flexibility it offers amid Singapore's strict car ownership regulations and pricing policies. This personal experience stayed with me and highlighted the importance of the used car market for residents and expatriates alike.

In conducting this project, we aim to address the challenges and opportunities in understanding Singapore's used car market. While we now have access to valuable datasets, predicting resale prices for cars is a complex task due to various influencing factors, such as make, model, mileage, age, and engine power, each contributing differently to the car's value. Additionally, Singapore's unique regulatory environment, including policies like the Certificate of Entitlement (COE), makes it difficult to directly apply insights from international studies, necessitating a localized approach to price prediction.

This study aims to achieve a few specific goals. Primarily, we aim to build a predictive model for used car prices, while also examining how different features contribute to a car's valuation. By providing an accurate assessment of what factors most affect resale price, this research can help newcomers like students or expatriates make informed decisions when purchasing a vehicle, potentially avoiding overpriced offers. Additionally, this research may benefit other stakeholders, including online car marketplaces, financial institutions, and insurance companies, by providing insights into market pricing dynamics and helping them better serve their clients.

II. Exploratory Data Analysis

This dataset has 25,000 entries, 30 columns, and multiple missing values across features.

For numerical features, `manufactured` (year), `curb_weight`, `power`, `engine_cap`, and cost-related attributes (`coe`, `road_tax`, `omv`) show wide distributions and potential outliers, with `indicative_price` entirely missing. `Price` ranges from 700 to 2.9 million, with significant high-end outliers. For accurate modeling, we will handle outliers and focus on relevant features identified in this analysis.

For classification characteristics, `make`, `model`, `type_of_vehicle`, and `category` show high variability, with `eco_category` containing only a single unique value, making it uninformative. High-cardinality features (`title`, `description`) may require special handling, while frequently occurring values in features such as `make` (e.g., Toyota) and `type_of_vehicle` (e.g., SUV) provide insights into common categories in the dataset.

A. Numerical Characteristics

Some visualization methods help us to get an overview of the distribution and outlier characteristics of numerical features in the dataset:

- Most features, including `curb_weight`, `power`, `engine_cap`, `dereg_value`, `mileage`, `omv`, `arf`, and `price`, show right-skewed distributions with a concentration of lower values and long tails extending toward higher values.
- `no_of_owners` has a discrete distribution, with the majority of cars having 1-2 previous owners.
- `manufactured` shows a cluster around recent years, suggesting that the dataset mostly includes newer vehicles.
- Boxplots confirm the presence of significant outliers in several features, such as `depreciation`, `coe`, `road_tax`, and `price`.
- Outliers are particularly evident in cost-related variables (`price`, `depreciation`, `coe`, `road_tax`), which may represent high-end or rare vehicles.
- For features like `no_of_owners`, the boxplot reflects limited variation due to the discrete nature of the data.

Most features are right-skewed, with newer vehicles and common lower values. Significant outliers in cost-related features, especially `price` and `depreciation`. Limited variation in `no_of_owners`.

B. Classification Features

For the top 20 categories for each categorical feature in the dataset:

- **Popular Categories:**
  - `make`: Dominated by brands like Toyota, Honda, and Mercedes-Benz.
  - `model` and `type_of_vehicle`: Popular models and types show high diversity.
  - `category`: Mainly consists of categories like ”parf car” and ”premium ad car”.
  - `transmission`: Predominantly ”auto” with fewer ”manual” vehicles.
  - `fuel_type`: Mostly ”diesel” and ”petrol”, though many values are missing.
- **High Prevalence of Missing Values:**
  - `opc_scheme`, `lifespan`, `eco_category`, and other features have many ”Missing” values, indicating incomplete data.
- **Low Variability:**
  - Some features like `eco_category` and `opc_scheme` lack diversity, limiting their potential use in analysis.

III. Relationship Analysis

A. Numerical Features and Target Variables

For the relationship between the target variable `price` and various numerical features:

- **Strong Correlations:**
  - Features like `dereg_value` and `arf` show positive correlations with `price`, indicating higher values are associated with more expensive cars.
  - `depreciation` also shows a strong positive trend with `price`.
- **Weaker Correlations:**
  - Features like `power`, `engine_cap`, and `coe` display some correlation with `price`, but with more scattered patterns.
- **Unique Patterns:**
  - `manufactured` and `curb_weight` show clustering around newer and lighter vehicles with higher prices.
  - `mileage` has an inverse relationship, with lower mileage generally correlating with higher prices.

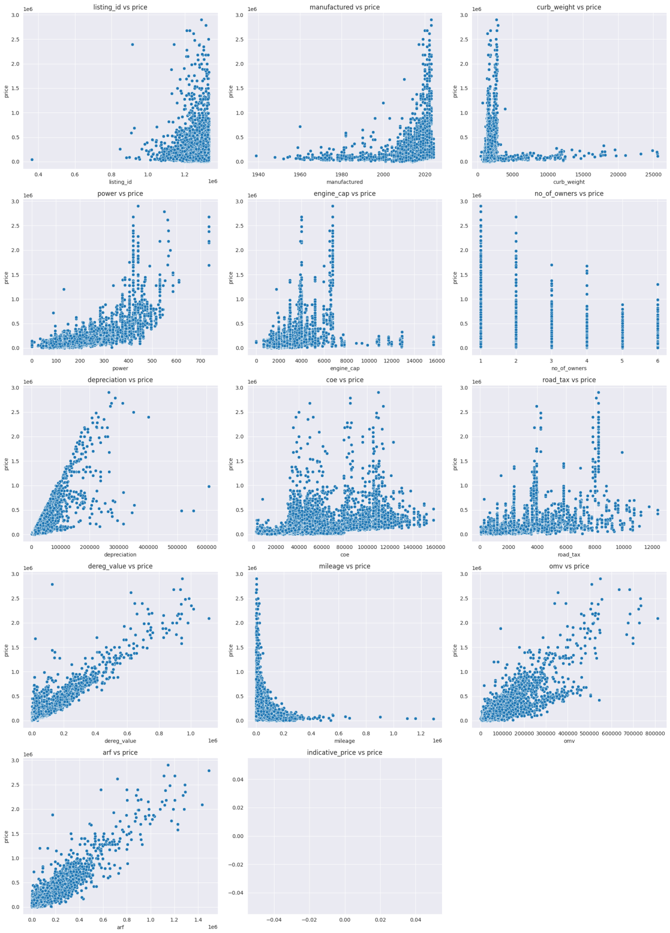

Fig. 1. Relationship between numerical features and target variables

B. Categorical Features and Target Variables

For the relationship between the top 10 categories of each categorical feature and the target variable, `price`. Here are the main takeaways:

- **Make and Model:** High-end brands like Mercedes-Benz and models such as AMG have higher median prices, showing the impact of brand and model on vehicle pricing.

- **Vehicle Type and Category:** Luxury and premium vehicle categories show higher price distributions compared to others, reflecting the influence of vehicle type.
- **Fuel Type and Transmission:** Diesel and automatic transmission vehicles have a wider range of prices, indicating variability based on fuel and transmission types.
- **Other Features:** Fields like `opc_scheme`, `lifespan`, and `eco_category` exhibit less variation in `price`, suggesting minimal influence from these categories on pricing.

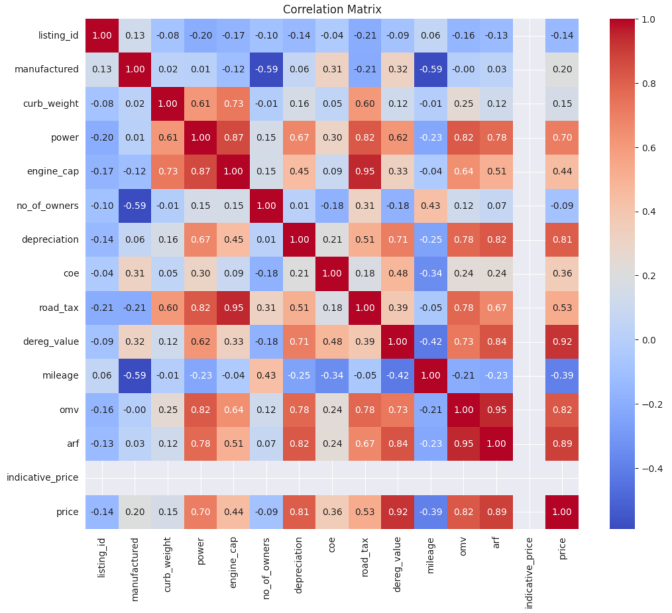

Fig. 2. Correlation matrix heat map

C. Correlation Analysis

This correlation matrix provides insights into the relationships between numerical features in the dataset. Key points are as follows:

- **Strong Positive Correlations:**
  - `price` shows strong correlations with `dereg_value` (0.92), `arf` (0.89), `depreciation` (0.81), and `omv` (0.82), suggesting that these cost-related features significantly influence vehicle price.
  - `engine_cap` and `power` are highly correlated (0.87), indicating that larger engine capacity often aligns with higher power output.
  - `dereg_value`, `omv`, and `arf` also exhibit strong intercorrelations, indicating overlapping aspects in vehicle valuation metrics.
- **Moderate Positive Correlations:**
  - `road_tax` has a moderate correlation with `price` (0.53) and `engine_cap` (0.95), which aligns with larger engine sizes often incurring higher road tax.
- **Weak and Negative Correlations:**
  - `mileage` shows a moderate negative correlation with `price` (-0.39), reflecting that higher mileage might decrease vehicle value.
  - Features like `no_of_owners` and `listing_id` have weak correlations with most other variables, suggesting limited impact on `price`.

IV. Data Preprocessing

To optimize the dataset's compatibility with various machine learning algorithms, we conducted comprehensive data preprocessing and cleaning. The key steps are detailed as follows:

A. Handling Missing Values

- **Removing Features with High Missing Rates:** Several features were identified with over 75% missing values, including `original_reg_date` (98.99%), `fuel_type` (76.49%), `opc_scheme` (99.40%), `lifespan` (90.56%), and `indicative_price` (100.00%). Due to the insufficient information provided by these features, they were excluded from the dataset.
- **Removing Single-Value Features:** Certain features, such as `eco_category`, contained only one unique value ("uncategorized") across all entries. Due to their lack of variability, these features do not contribute to predictive modeling and were therefore removed.
- **Hierarchical Imputation for Remaining Missing Values:** For numerical features with missing values, a hierarchical imputation approach was applied:
  - Group by `make` and `model` and perform iterative imputation within each group for groups with sufficient data points.
  - For remaining data points, group by `make` and apply iterative imputation at this broader level.
  - Finally, apply iterative imputation across the entire dataset to fill any remaining missing values.

  This hierarchical approach respects natural groupings within the data, ensuring that missing values are filled with contextually relevant information and improving the reliability of the imputed values.

B. Feature Extraction and Transformation

- **Removing Unstructured Text Features:** Features such as `title`, `description`, `features`, and `accessories` contain unstructured text data with a high percentage of missing values. Due to the complexity of processing such data in this modeling phase, these features were excluded to reduce noise.
- **Processing Year Features:** The features `reg_date` and `manufactured` were found to be highly correlated (correlation coefficient of 0.887). To handle missing values in `manufactured`, the year from `reg_date` was used as a substitute, creating a new feature `reg_year`, and the original date features were removed.

- **Processing Categorical Features:** To enable categorical features to be used in numerical models, one-hot encoding was applied to type_of_vehicle, category, and transmission. For instance:
  - type_of_vehicle was transformed into 11 binary columns.
  - category expanded into 16 binary columns.
  - transmission represented by 2 binary columns.

  This transformation enables models to process categorical information numerically, with each unique category represented by a separate binary column.

C. Outlier Removal

Data points with exceptionally high annual mileage were identified as outliers and removed to prevent training bias. Additionally, significant outliers in features such as curb_weight, power, depreciation, coe, road_tax, dereg_value, mileage, omv, and price were analyzed through boxplots and appropriately handled to ensure data consistency before training.

D. Final Processed Dataset Suitability

The cleaned and processed dataset, now containing both continuous and dummy variables, is compatible with a variety of regression models, including:

- **Linear Models:** The simplified structure following dummy variable conversion makes the dataset suitable for linear regression.
- **Ensemble Methods (Random Forest):** The dataset's structure allows complex models to capture non-linear relationships and feature interactions effectively.
- **Regularized Models (Lasso):** The dataset supports regularization techniques, aiding in feature selection and reducing overfitting.

V. Data Mining and model training

A. Linear Regression

- **Model Selection and Training:** A linear regression model [1] was trained to establish a baseline for predictive accuracy. Linear regression was chosen for its simplicity and interpretability, allowing us to understand the primary linear relationships within the data.
- **Evaluation:** The model was trained on an 80-20 train-test split, and we recorded the Root Mean Square Error (RMSE) as the primary performance metric. The linear regression model achieved an RMSE of approximately 43,720, which indicates the average prediction error and serves as a benchmark for further model improvements.

B. Random Forest Regressor

- **Model Selection and Training:** The Random Forest Regressor was chosen for its ability to capture complex, non-linear relationships within the dataset. Configured with 1,000 trees (n_estimators=1000) and parallel processing (n_jobs=-1), this model aggregates predictions from multiple decision trees, enhancing accuracy and reducing overfitting.
- **Evaluation:** The model achieved an RMSE of approximately 26,713, a significant improvement over simpler models. This low RMSE indicates that the Random Forest effectively captures the data's complex patterns, making it well-suited for accurate price prediction.

C. XGBoost Regressor

- **Model Selection and Training:** The XGBoost Regressor [2] was employed for its strong performance with tabular data and ability to handle non-linear relationships effectively. This model was configured with the following hyperparameters to balance model complexity and performance:
  - n_estimators=2000: A higher number of trees allows for greater model complexity.
  - learning_rate=0.01: A low learning rate ensures gradual learning to prevent overfitting.
  - max_depth=6: Limits tree depth, controlling overfitting.
  - subsample=0.8 and colsample_bytree=0.8: These sampling techniques help regularize the model by training each tree on only 80% of the data and features, respectively.
  - random_state=42: Ensures reproducibility.
  - n_jobs=-1: Utilizes all available CPU cores for efficient training.
- **Evaluation:** The XGBoost model achieved an RMSE of approximately 25,700. This result indicates a good fit to the data, with the model capturing complex relationships while being tuned to reduce overfitting. The lower RMSE compared to simpler models underscores the model's predictive strength.

D. Hierarchical Random Forest

The Hierarchical Random Forest model was designed to exploit hierarchical relationships within categorical features (make and model) by building separate models at different levels. This multi-layered structure aimed to capture patterns within finer-grained groups and improve predictive accuracy. The model training approach was as follows:

- **Level 1 (make, model):** For each unique (make, model) pair with at least 10 records, a Random Forest model was trained, aiming to capture specific patterns within this subset.
- **Level 2 (make):** When (make, model) combinations had insufficient data, a broader model was trained for each make category. This level generalizes patterns across each make group.
- **Level 3 (Global):** A global Random Forest model was trained on the entire dataset, providing a fallback for cases without sufficient data at the previous levels.

Each model used default Random Forest parameters, focusing on computational efficiency while capturing hierarchical information.

The hierarchical approach, however, did not yield the expected performance improvement. While the layered structure allowed the model to adapt to different data levels, the overall predictive accuracy was limited, and the model's hierarchical complexity did not translate into significantly better results. The approach highlights the challenges of leveraging hierarchical structures effectively in cases where data distribution or sample size constraints limit the model's ability to generalize. Further tuning or alternate approaches may be necessary to improve performance.

VI. Evaluation and Interpretation

After evaluating multiple models, including Linear Regression, Random Forest, XGBoost, and Hierarchical Random Forest, we selected the standard Random Forest Regressor as the final model. Linear Regression, though interpretable, resulted in higher errors and served as a benchmark. XGBoost demonstrated strong performance with a lower RMSE, but its complexity and tuning requirements limited its practical efficiency in this context. The Hierarchical Random Forest approach aimed to capture layered relationships within make and model levels; however, it did not yield the desired accuracy improvements, likely due to sample size constraints in certain groups. The Random Forest Regressor achieved a balance between simplicity and predictive strength, capturing non-linear patterns effectively with an RMSE significantly lower than baseline models. Its robust performance and ease of implementation made it well-suited for our needs, making it the optimal choice for accurate, reliable predictions.

References

[1] Singh Saini, Prabaljeet, and Lekha Rani. ”Performance Evaluation of Popular Machine Learning Models for Used Car Price Prediction.” International Conference on Data Analytics and Insights. Singapore: Springer Nature Singapore, 2023.

[2] Cui, Baoyang, et al. ”Used car price prediction based on the iterative framework of XGBoost+ LightGBM.” Electronics 11.18 (2022): 2932.