

CDS6214

Data Science Fundamentals

Project (40%)

Group Number: G10 I

Title:

| **ID** | **Name** | **Contribution (in %)** |
| --- | --- | --- |
| 242UC24511 | Chan Jia Jun | 33 |
| 242UC24513 | Chen Yong Qi | 30 |
| 242UC24515 | Tong Jian Hao | 37 |
|  | **Total** | **100** |

Note: Please ensure all group members agree to the submitted contribution (in %).

| **YouTube Link** |  |
| --- | --- |
| **Dataset Link** |  |

**Table of Contents**

[**Introduction and Objectives 2**](#)

[**Dataset Description 3**](#)

[**Exploratory Questions and Answers 4**](#)

[**Modeling and Evaluation 5**](#)

[**Discussion and Insights 6**](#)

[**Conclusion 7**](#)

[**References 8**](#)

# Introduction and Objectives

In the competitive used car market, pricing decisions are crucial for dealerships, platforms, and individual sellers. This project aims to analyze a comprehensive dataset of used car listings to extract meaningful insights that can guide pricing strategies, inventory valuation, and risk assessment. Our objective is to explore patterns, relationships, and predictive factors influencing car prices using statistical analysis and machine learning techniques.

The scope of this project includes descriptive and predictive analysis focused on price, vehicle condition indicators, and brand-related trends. Through this, we aim to assist stakeholders in making data-driven decisions.

# Dataset Description

The dataset contains 10,000 entries of used cars with 13 features including make year, mileage, engine capacity, fuel type, ownership history, color, transmission, accidents reported, insurance validity, service history, and brand. All data entries were pre-cleaned to remove duplicates and missing values. Categorical variables such as fuel type and brand were label-encoded for modeling.

Target variable: price\_usd (continuous).

# Exploratory Questions and Answers

**Q1: How does service history impact used car prices?**

Analysis shows that vehicles with a complete service history command higher average prices compared to those with partial or no records. This indicates that buyers perceive well-documented maintenance as an assurance of quality and reliability.

**Q2: How has the trend in engine size evolved over the years?**

Average engine capacity has declined gradually in more recent manufacturing years. This suggests a shift toward more fuel-efficient or eco-friendly models, which can help manufacturers and resellers adjust their inventory strategies accordingly.

**Q3: Do certain car colors correlate with higher accident reports?**

Yes. The analysis reveals that darker-colored cars, such as black and grey, report slightly higher average accident counts. This insight could be useful for insurance companies in risk profiling or premium adjustments.

**Q4: What are the ownership trends among economy vs high-end car brands?**

Economy car brands tend to have more previous owners than high-end cars across the same make years. This insight helps resellers estimate depreciation patterns and resale value adjustments.

# Modeling and Evaluation

A Random Forest Regressor was trained using features such as brand, mileage, engine size, owner count, fuel type, and color. The dataset was split into 80% training and 20% testing. Categorical features were label-encoded. Model performance was evaluated using two metrics:

* Mean Squared Error (MSE): 1,486,082.94
* R² Score: 0.8132

The R² score of 0.8132 indicates strong explanatory power, meaning over 81% of the price variance can be explained by the input features.

# Discussion and Insights

**\*\*Implications:\*\***

The insights gained can be directly applied in strategic pricing, inventory planning, insurance underwriting, and marketing decisions within used car dealerships and online platforms.

**\*\*Limitations:\*\***

The dataset does not include location data, market trends over time, or real-time listings, which could further affect price prediction accuracy.

**\*\*Assumptions:\*\***

It is assumed that all recorded data fields are accurate and that label-encoded categories maintain meaningful relationships with price.

**\*\*Recommendations:\*\***

Future models could benefit from incorporating geolocation, time-on-market, and customer ratings. Additionally, integrating external data sources such as competitor listings could improve pricing strategies.

# Conclusion

This project demonstrated how data analysis and machine learning can provide actionable insights for the used car market. Through four targeted questions and a predictive model, we were able to identify key factors influencing car pricing and ownership trends. The findings suggest that structured data-driven decisions can enhance profitability and operational efficiency in the automotive resale industry.

# References

American Psychological Association. (2020). \*Publication manual of the American Psychological Association\* (7th ed.).

Scikit-learn Developers. (2024). \*Scikit-learn documentation\*. https://scikit-learn.org/

XGBoost Contributors. (2024). \*XGBoost documentation\*. https://xgboost.readthedocs.io/

Pandas Documentation. (2024). \*Pandas User Guide\*. https://pandas.pydata.org/docs/

Seaborn Documentation. (2024). \*Seaborn visualization library\*. https://seaborn.pydata.org/