Table of Contents

[1. Introduction 2](#_Toc197131719)

[2. Data Overview and Preprocessing 2](#_Toc197131720)

[2.1. Dataset Description 2](#_Toc197131721)

[2.2. Data Quality Assessment 2](#_Toc197131722)

[3. Analysis and Findings (Exploratory Data Analysis) 3](#_Toc197131723)

[3.1. Overview of Key Car Features 3](#_Toc197131724)

[3.2. Car Price Analysis 5](#_Toc197131725)

[3.3. Other Insights: Feature Relationships 7](#_Toc197131726)

[3.4. Summary of EDA Findings & Business Implications 7](#_Toc197131727)

[4. Machine Learning Approach and Findings 8](#_Toc197131728)

[4.1. Machine Learning Task Definition (Supervised Regression) and Data Preparation for Modeling 8](#_Toc197131729)

[4.2. Model Training and Performance Evaluation 8](#_Toc197131730)

[5. Recommendations 9](#_Toc197131731)

[6. References 10](#_Toc197131732)

[7. Appendices 11](#_Toc197131733)

[Appendix A: Detailed Model Parameters/Code Snippets 11](#_Toc197131734)

# Introduction

One prominent participant in the used car market in Australia is Eureka Motors. Despite having a successful business, they understand that adopting new technology is essential to staying ahead of the competition. In today's market, pricing decisions based solely on traditional gut instinct can be erratic and may have an effect on profits and the speed at which cars sell. To get a sharper edge, Eureka is exploring how Artificial Intelligence (AI) can help them make smarter, data-backed decisions.

Our goal was to help Eureka solve this challenge. We deep dive into a dataset full of Australian used car listings to truly understand what influences prices – beyond just the obvious make and model. Think mileage, age, specific features, and market trends. Using this information, we've been developing and testing smart computer models (using machine learning) that can learn these patterns and help estimate.

Though out this work will provide Eureka Motors with practical tools and insights. It's not just about crunching numbers; it's about turning data into actionable advice. We aim to offer clear recommendations on how they can refine their pricing strategies, manage their inventory more effectively, and possibly even spot new opportunities within the competitive Australian used car landscape, all based on what the data tells us.

# Data Overview and Preprocessing

## Dataset Description

The dataset we will use is "EM140\_A2\_data.csv," which was made available by Eureka Motors especially for studying the used automobile market in Australia. Information from 5000 used car listings is included in the dataset, which is categorized using 19 variables. These variables detail the vehicle's identification (id), brand, model, year, condition, mileage, powertrain (drive\_type, fuel\_type, cylinder, engine\_size, fuel\_consumption), physical characteristics (body\_type, exterior\_color, doors, seats), and location (suburb, state). The target variable for predictive modeling and analysis is the listed price, denominated in Australian dollars. Using a machine learning model, this dataset will serve as the foundation for pricing and market insight estimation, cleaning, and insight investigation.

## Data Quality Assessment

In Figure 9 (see Appendix A), we address the need for model evaluation on unseen data by splitting the dataset into training and testing portions using train\_test\_split. We use ColumnTransformer to define a thorough preprocessing pipeline in coincidentally. In order to prepare numerical (like scaling) and categorical (like encoding) features for machine learning algorithms and so this can lead to a result that can improve model performance, this crucial tool to manage their different needs.

In Figure 10, data integrity is then enhanced by type correction and cleaning (refer to Appendix A). The 'price' column must be processed, including commas removed, conversion to numeric, and error management. Additionally, the 'year', 'kilometres', and 'cylinder' columns must have the appropriate integer data types. At this point, missing value checks are completed, and it is confirmed that all columns contain the intended data types, which is essential for further analysis.

The final preparation phase focuses on data quality refinement on the Figure 11 (see Appendix A). Outliers in 'price' and 'kilometres' are identified, and 'price' values are capped at $150,000 to reduce skewness and for the categorical columns like 'transmission' and 'condition' undergo standardization to ensure consistent representation. This can improve data reliability and uniformity, yielding a dataset optimized for effective modeling.

# Analysis and Findings (Exploratory Data Analysis)

## Overview of Key Car Features

A graph of a car price

AI-generated content may be incorrect.

Figure 1: Price Distribution Graph

The histogram above depicts a right-skewed distribution of used car prices in AUD, indicating that most automobiles are priced cheaper, peaking between $20,000 and $30,000. As prices rise, the frequency of cars declines dramatically, with a modest secondary high recorded around the $150,000 threshold.

A group of blue bars

AI-generated content may be incorrect.

Figure 2: Vehicle Characteristics (Kilometres, Age) Graph

The most common car type is the SUV, and the vast majority of cars are listed as used. Furthermore, automatic transmissions, unleaded fuel, and front-wheel drive are the prevalent features, with the highest concentration of these cars found in NSW (New South Wales).

A graph of a graph of a graph

AI-generated content may be incorrect.

Figure 3: Usage Characteristics (Kilometres, Age) Graph

The provided graph displays two histograms illustrating the distribution of kilometres driven and the age of cars. Both distributions appear in right-skewed shape, suggesting that a majority of the cars observed have relatively low mileage and are comparatively young.

## Car Price Analysis

A group of blue and black rectangles

AI-generated content may be incorrect.

Figure 4: Price Variation Across Key Features Graphs

The figure 4 above illustrates that vehicle prices in AUD exhibit significant variability depending on various features like body type, condition, fuel, transmission, drive type, and brand. Generally, newer vehicles, those with electric/hybrid fuel types, automatic transmissions, AWD, and luxury brands command higher median prices and often show wider price ranges.

Used automobiles, manual transmissions, and more common body shapes, such as hatchbacks, typically have lower median pricing. Across practically every category, there are multiple high-priced outliers, indicating a broad potential price spectrum outside the average interquartile ranges.

A group of graphs showing different sizes of data

AI-generated content may be incorrect.

Figure 5: Identification of Price Influencers Graphs

Analysis of the graphical data indicates strong negative correlations between vehicle price and both age and kilometers accumulated; price exhibits a clear downward trend as these variables increase.

In contrast, the relationships between price and engine size or fuel consumption demonstrate considerably higher variability and less defined trends. Because the data points for engine size and fuel consumption are so widely scattered, accurately predicting price based on these factors alone proves difficult.

This suggests that a vehicle's age and mileage are considerably stronger indicators of its lower market value than its engine specifications or fuel efficiency. In essence, these visualizations reveal that various vehicle characteristics affect pricing differently.

## Other Insights: Feature Relationships

A screenshot of a graph

AI-generated content may be incorrect.

Figure 6: Heatmap For Numeric Features

Significant interdependencies between the vehicle attributes are shown by examining the correlation heatmap. Notably, there is a strong positive correlation between engine size and cylinder count, and both variables also strongly correlate with higher fuel consumption. As expected, the year of manufacturing has a negative correlation with the age of the vehicle, whereas age has a positive correlation with the number of kilometers driven. Accordingly, a decrease in the listed vehicle price is typically associated with attributes like high mileage and advanced vehicle age.

## Summary of EDA Findings & Business Implications

Key findings:

* Price is right-skewed, influenced by age, kilometres, brand, condition, body type.
* SUVs, Utes, and high-end brands tend to have higher prices.
* There is a noticeable deterioration pattern as age and kilometers increase.
* There are feature relationships that are pertinent to modeling, such as age/kilometers and engine/cylinders.

Implications: Pricing strategy needs to account for these key drivers. Inventory mix could focus on high-value segments.

# Machine Learning Approach and Findings

## Machine Learning Task Definition (Supervised Regression) and Data Preparation for Modeling

Our aim was to create a model that could estimate the price of used vehicles based on past data provided. We chose a supervised learning approach—specifically, a regression model—because we were dealing with numerical data on a large range data.

After analyzing and plotting out the dataset above, we selected criteria that appeared to have the big influence on pricing, such as the car's branding, condition, mileage, and age.

For step preprocessing, we used One-Hot Encoding to convert categorical input (such brand and transmission) into a format that the model could easily understand. We also standardized numerical components to ensure that everything was on a consistent scale this play as a step that enables models like linear regression to avoid bias towards larger-range data in the dataset.

## Model Training and Performance Evaluation

Two distinct regression models were selected for this analysis is linear regression and random forest regressor:

1. A graph with a line and a red line

   AI-generated content may be incorrect.Linear regression: Chosen as a fundamental baseline model. It is computationally efficient and highly interpretable, providing clear coefficients that indicate the linear relationship between each feature and the price.

Figure 7: Linear Regression Actual vs. Predicted Prices Graph

* Performance:
  + R-squared (R²): 0.7423
  + Mean Absolute Error (MAE): $7,913.95 AUD
  + Root Mean Squared Error (RMSE): $12,261.77 AUD

1. A graph with a line and a dotted line

   AI-generated content may be incorrect.Random forest regressor: Selected as a more complicated, non-linear model. As an ensemble method based on decision trees, it can capture complex interactions between features and non-linear trends (like depreciation curves) more effectively than linear models.

* Performance:
  + R-squared (R²): 0.9046
  + Mean Absolute Error (MAE): $4,266.03 AUD
  + Root Mean Squared Error (RMSE): $7,459.86 AUD

Figure 8: Random Forest Actual vs. Predicted Prices Graph

Random Forest outperformed Linear Regression based on R².

The average prediction error (MAE) for Random Forest is approximately $4,266 AUD.

Model Strengths/Limitations:

1. Linear Regression: Interpretable coefficients but may underfit complex relationships.
2. Random Forest: Captures non-linearities, robust, but less interpretable ('black box'). Performance depends on data quality.

# Recommendations

Based on the findings from the data analysis process and the performance of the machine learning models built above, the recommendations that we going to suggest can usefully help Eureka Motors improve its business:

1. Implement model-assisted pricing guidance: Utilize the developed Random Forest regression model as a decision-support tool for vehicle appraisal and pricing. With a demonstrated Mean Absolute Error (MAE) of approximately $4,266 AUD on the test data, the model can provide a consistent, data-driven baseline estimate. This estimate should serve as a starting point then apply their expertise to account for specific vehicle nuances (e.g., unique features, exceptional condition, local market demand) not fully captured by the model features or reflected in the average error.
2. Optimize the inventory acquisition strategy: The data showing that there are some car categories that can be attracted with a higher pricing. Taking advantage of this, Eureka Motors should prioritize the acquisition of automobiles with great price resilience or better market worth.
3. Improve data gathering for important predictors: To increase the accuracy of future price estimates and market insights, prioritize improving the quality and consistency of data obtained during assessments. Standardize the inputs for categorical variables such as condition and drive\_type. Given their established influence on pricing, maintaining high accuracy for kilometers, year (and consequently car\_age), and brand is critical for dependable model performance.

**Future Improvements:**

* Regularly retrain the model with new sales and inventory data (monthly will have a best practice) to adapt to changing market dynamics.
* Explore incorporating features initially excluded, such as model (using techniques like target encoding or embedding layers if more advanced models are used or using deep learning for advanced) and potentially location (suburb or postcode-level data if available) to capture finer-grained market variations.
* Investigate more difficult outlier detection and treatment methods during preprocessing.
* Consider developing separate models for distinct vehicle segments (e.g., luxury vs. budget, commercial vs. passenger) if performance analysis indicates significant variation across segments.
* Explore incorporating external data sources, such as broader economic indicators or real-time auction data, to further enhance market context and prediction accuracy.

# References

# Appendices

## Appendix A: Detailed Model Parameters/Code Snippets

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 9: Handling Missing Data Code

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 10: Data Type Correction Code

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 11: Outlier Identification and Treatment