Prediction & Classification - Used Vehicles Prices in Australia

Meet Our Team

Mathini Kanagaratnam
Business Expert

Anjali Patel
Data Analyst

Sairohit Chowdhary
Data Analyst

Man Chung Chan
Machine Learning Analyst

Yan Pui Siu
Machine Learning Analyst

Business Problem

The prices of pre-owned vehicles have skyrocketed, doubling since the onset of the pandemic.

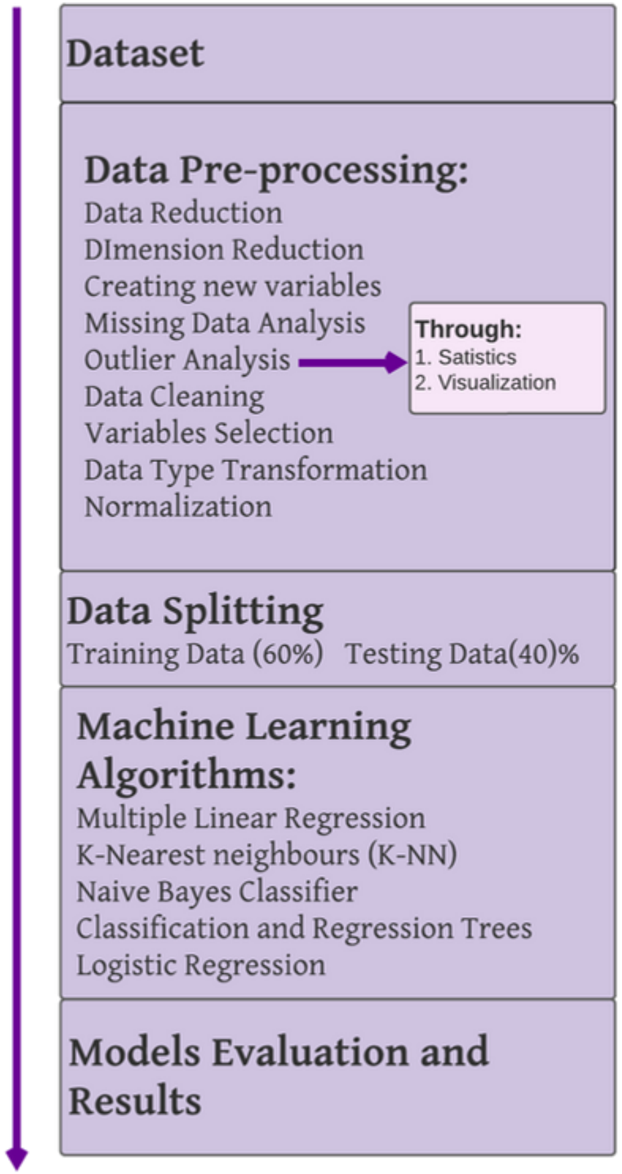
Although post-pandemic, the prices of used cars have decreased, they have not yet reached pre-pandemic levels.

Sellers and Buyers are unable to make a judgment on the price due to its volatility

Objective

- To predict the price of pre-owned vehicles, which can help customers to determine a fair price in a market where prices are volatile.
- Buyers can avoid paying an excessively high price for a used vehicle, and sellers can set a suitable price based on their urgency to sell
- Enabling customers to make decisions about which model to purchase if they plan to sell it in the future.
- Help car manufacturers to determine which models they should focus on producing in the used car market to stay competitive.

Workflow & Data Description



Variable Name	Description
ID	Vehicle ID
Name	Vehicle Name
Price	Price of Vehicle in AUD
Brand	Brand of Vehicle
Model	Model of Vehicle
Variant	Variant of Vehicle
Series	Series of Vehicle
Year	Manufacturing Year
Kilometers	Total Distance Driven
Type	Type of Vehicle
Gearbox	Type of Gearbox
Fuel	Type of Fuel Used
Status	New, Used, or Demo
CC	Size of Engine in Cubic Centimeters
Color	Color of Vehicle
Seating Capacity	Number of Seats

Data Pre-processing

01

Data Reduction

We concentrate only on predicting prices and classifying used cars by removing new and demo car records from the 'Status' variable

Number of Rows Before Data Reduction	Number of Rows After Data Reduction
17,048	16,304

02

Dimension Reduction

To simplify the dataset, we removed the 'Status' variable since it only contained records related to used vehicles.

03

Creating new Variables

A new variable called 'Age' was introduced to calculate the age of vehicles in the dataset by subtracting the manufacturing year from the current year (2023)

New Variable Name	Data Type	Description
Age	Numerical	Age of Vehicle
Price_Classification	Nominal	Fair or Not Fair
Price_Classificatio_ Numerical	Numerical	0 = Fair, 1 = Not Fair

04

Missing values analysis

Upon evaluation of the dataset with the newly created variables, we discovered that there were no missing values in the dataset.

Data Pre-processing

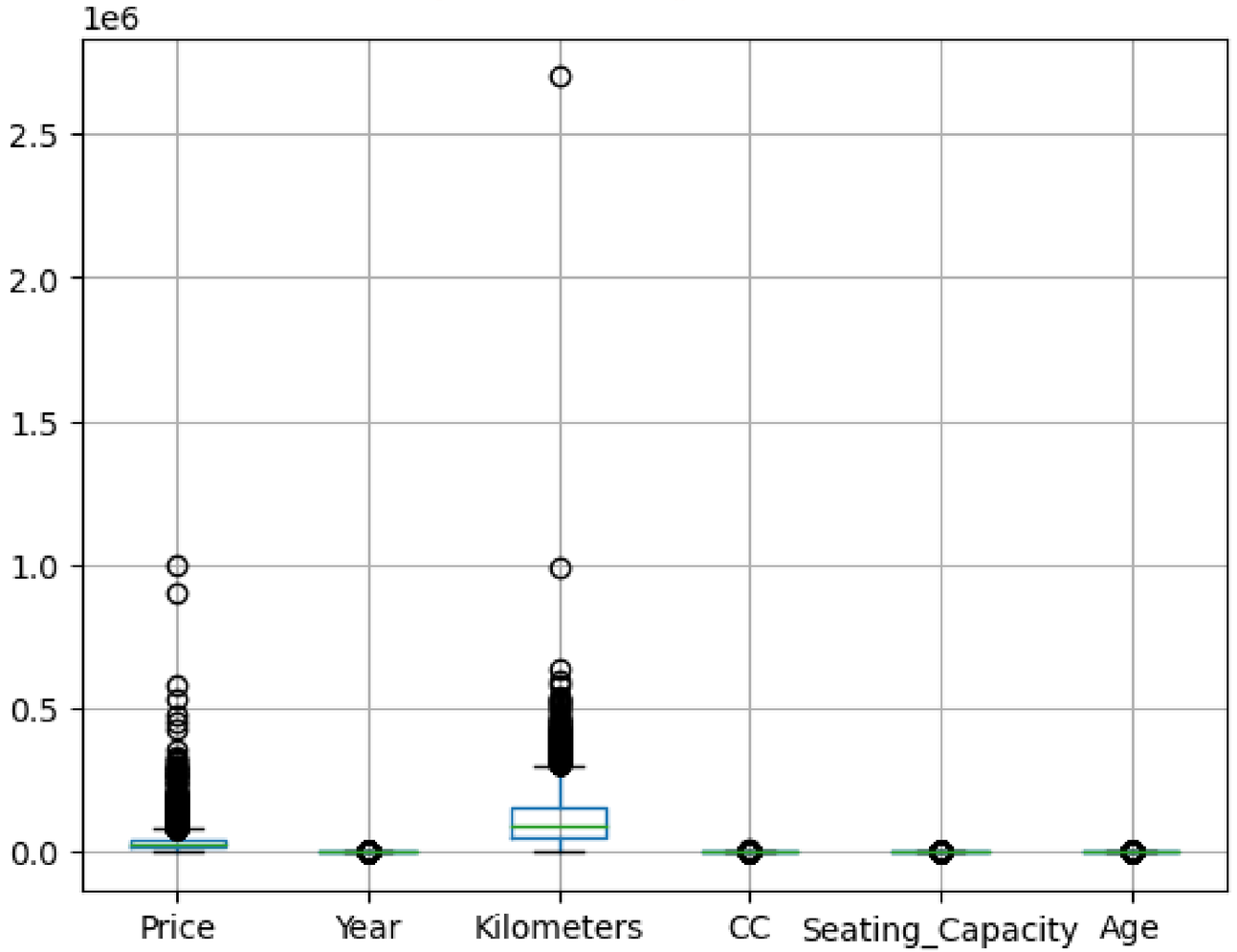
05

Data Cleaning

- Values exceeding the upper bound or falling below the lower bound of the "Price", "Kilometers", and "CC" variables were considered outliers.
- We eliminated the rows containing these outliers from our dataset by omission.

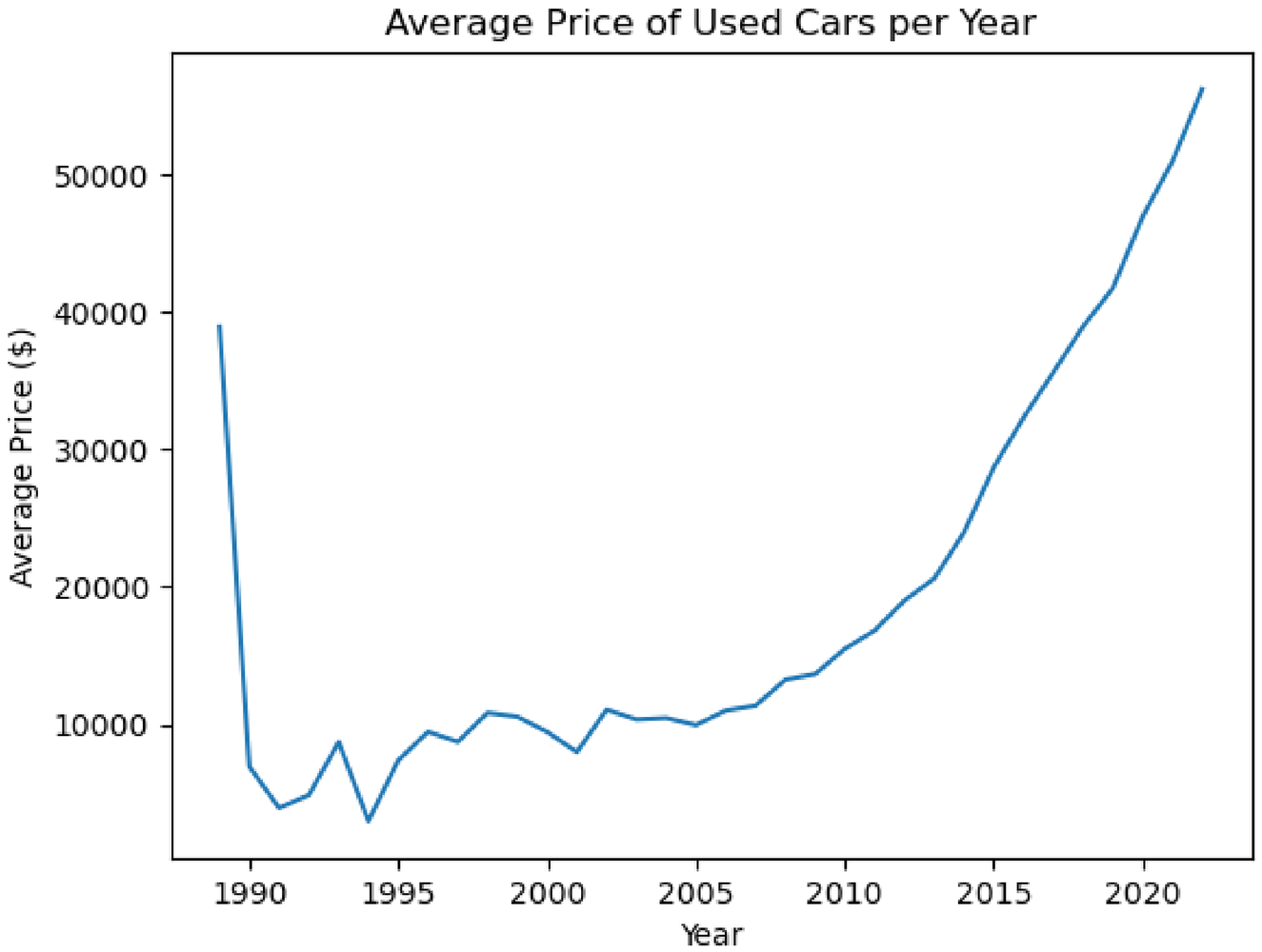
Number of Rows Before Omission	Number of Rows After Omission
16,304	14,894

Graph 1: Finding outliers



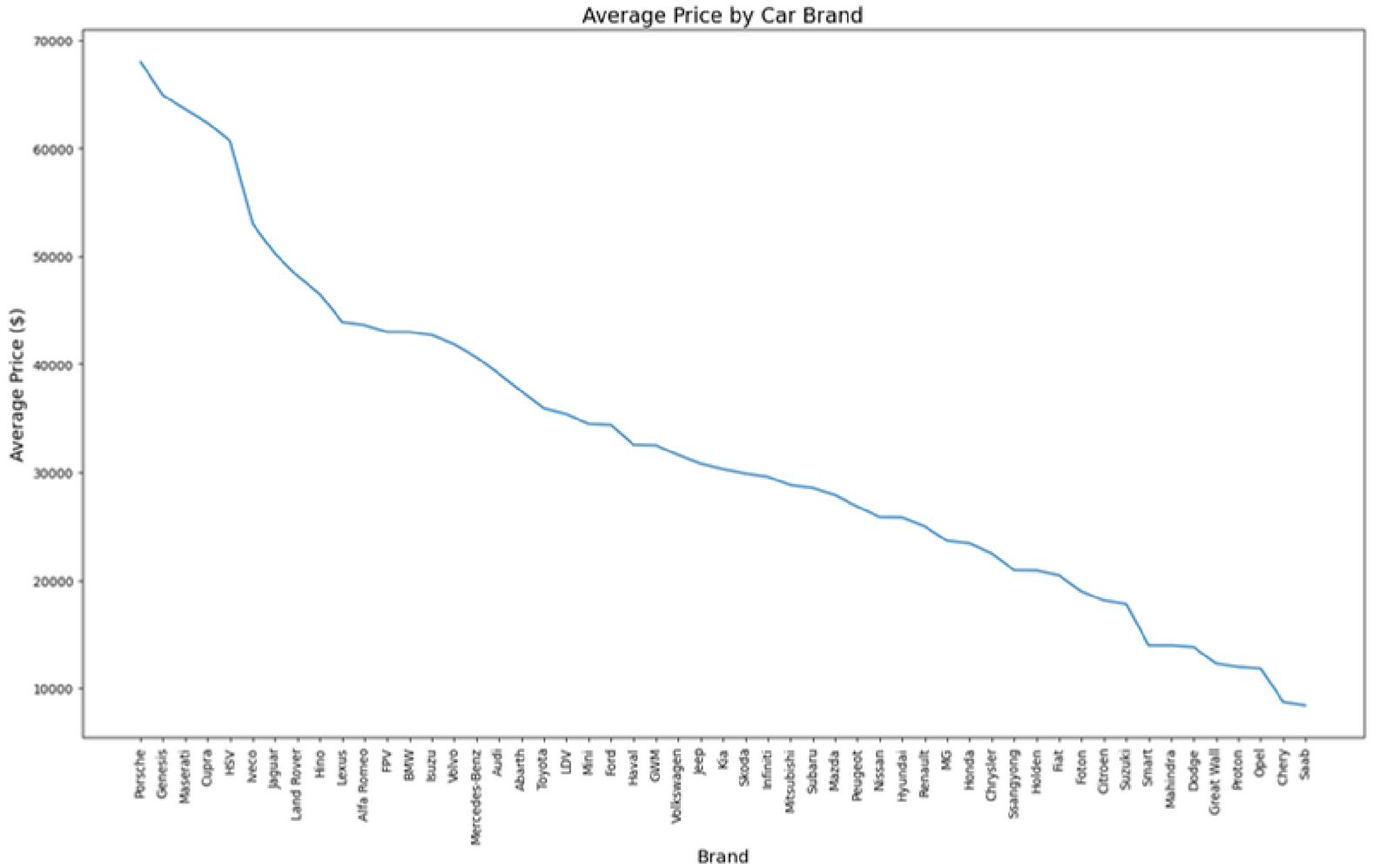
Data Visualization

Graph 2: Average price of used cars per year



Data Visualization

Graph 3: Average used car price for different brands



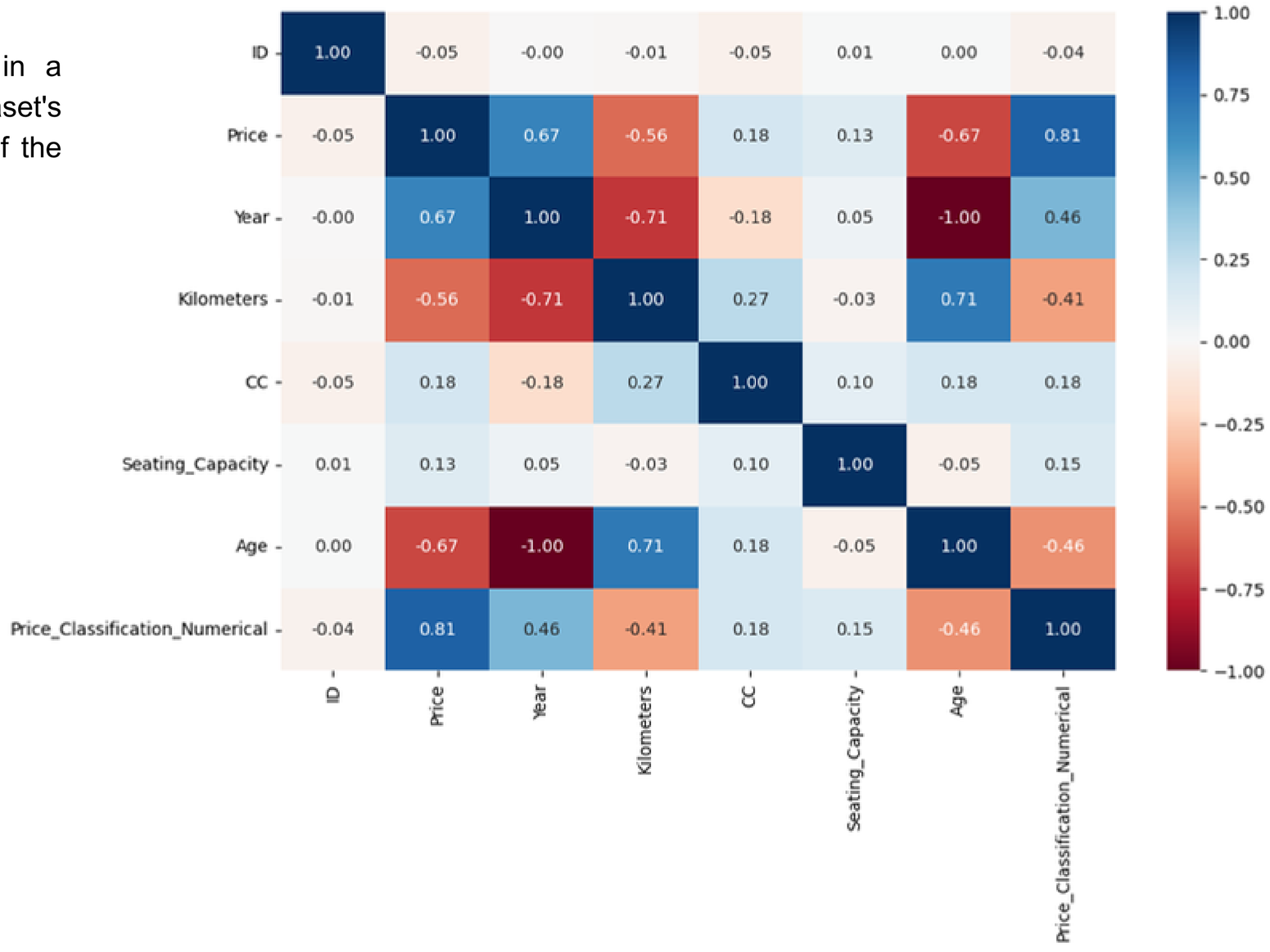
Car Feature Selection

Exhaustive Search

Aims to identify the most important features in a dataset, with the goal of reducing the dataset's dimensionality and enhancing the performance of the model.

Predictors
Brand
Kilometers
Type
Gearbox
Fuel
CC
Seating_Capacit
Age

Graph 4:Correlation heatmap

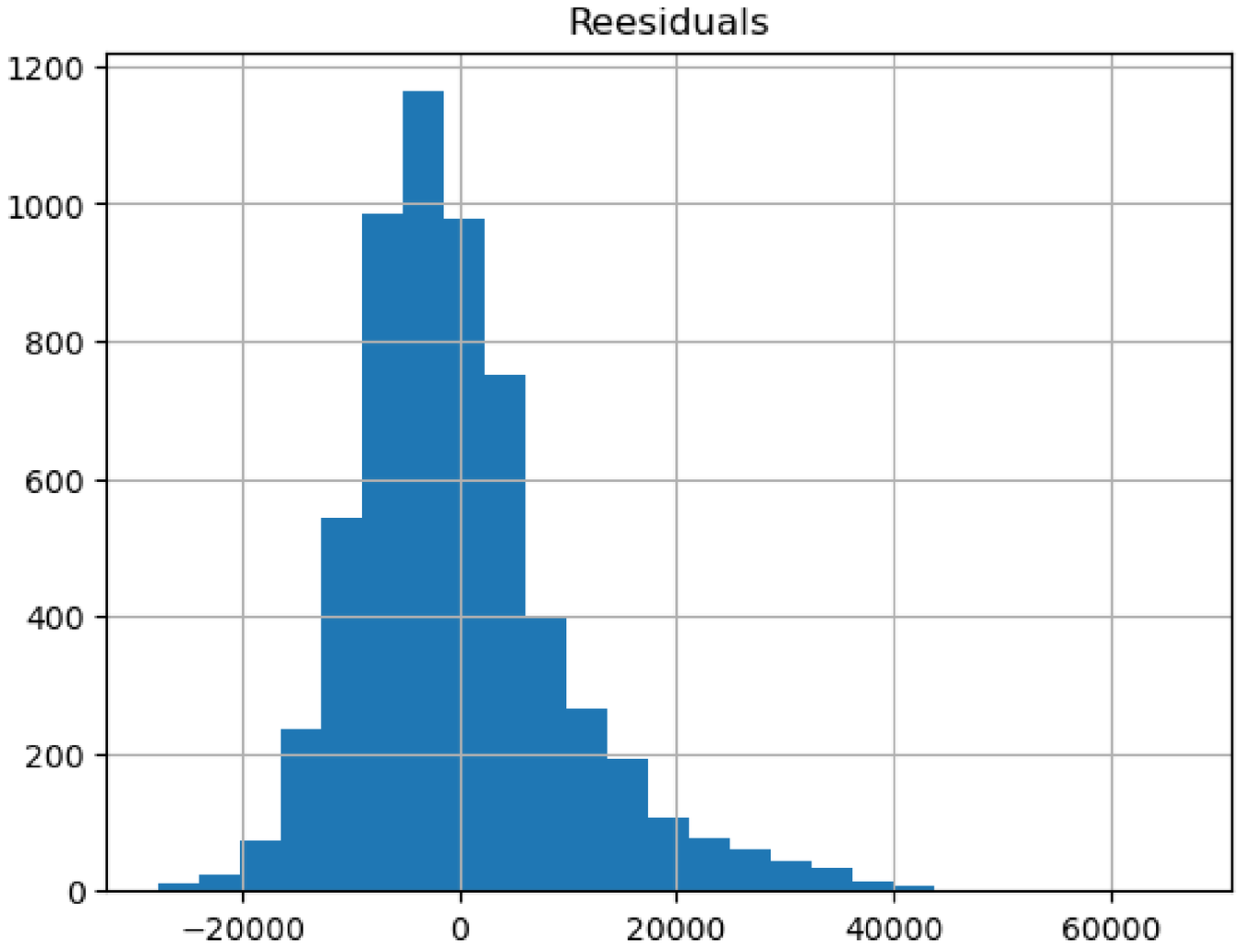


Predictive and Classification Models

	Predictive Model	**Classification Model**
Objective	Predicting Car Prices $	Classifying Car Prices
Result	AUD$	Fair / Not Fair
Threshold	/	Based on average price of used vehicles in 2022

Predictive Model Performance

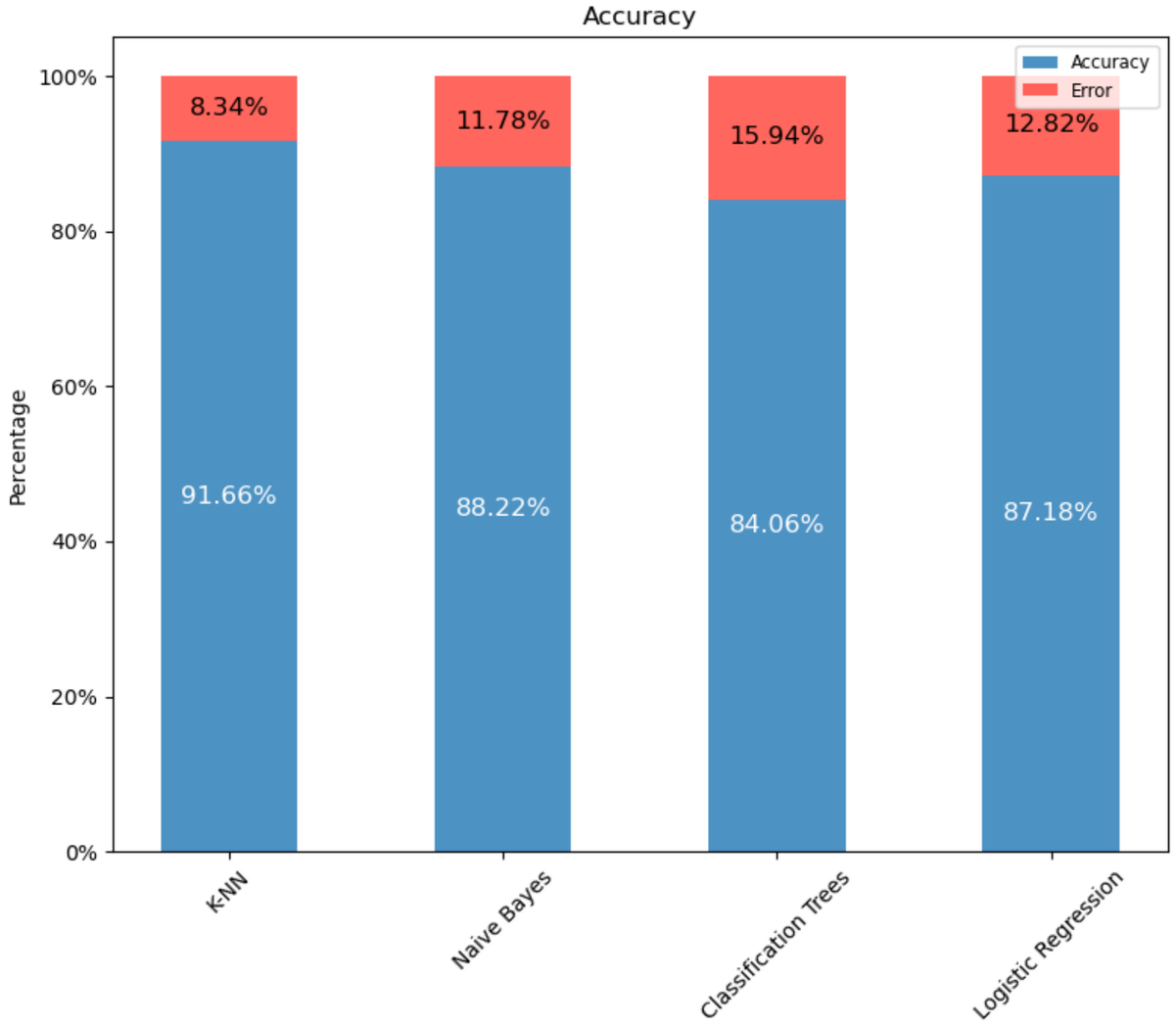
Graph 5: Distribution of errors



- Errors = Actual Values - Predicted Values
- Most concentrated towards lower end
- Indicates the predicted car prices are too high
- Error Percentage: 32%, (29% in other research paper)
- Conclusion: High Error Rate

Classification Models Performance

Graph 6: Models comparison

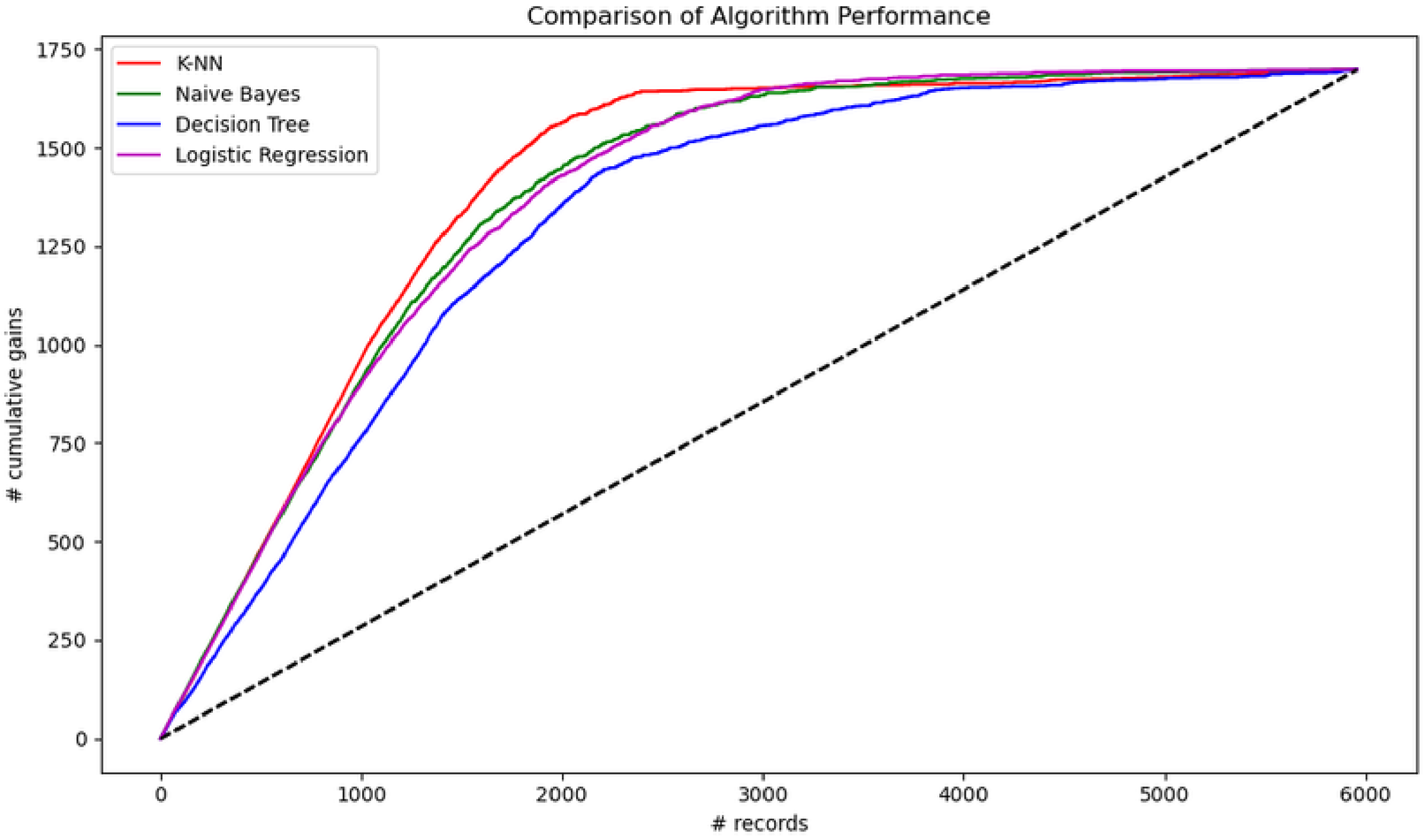


Error Rate from Lowest to Highest:

- **1. k-NN - 8%**
- 2. Naïve Bayes Classifier - 12%
- 3. Logistic Regression - 13%
- 4. Classification Trees - 16%

Classification Models Performance

Graph 7: Cumulative gains chart



- Measures model's ability to correctly identify fair car price
- X-axis: dataset size
- Y-axis: proportion of correctly identified fair car prices
- Higher and steeper curve: more effective model in achieving classification goal
- Larger area under curve: better overall performance in distinguishing between fair and unfair car prices accurately
- Best model: **k-NN**

Conclusions

Best Model:

K-NN algorithm

Compared to other similar research:

"Used Car Price Prediction using K-Nearest Neighbor Based Model"

91.7% Our report **vs** **85%** Compared report

- Outperforming the other algorithms
- Non-parametric approach - not rely on any assumptions about the data
- Steepest curve + largest area (Cumulative Gains Chart)

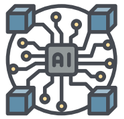
How K-NN Help to Improve Business Problem:

- **Potential Buyers** - guidance on whether a vehicle's price is reasonable + negotiations with dealers for a more favorable price
- **Sellers/ Dealerships** - determine optimal prices that maximize their profits
- **Manufacturer** - determine which models they should focus on producing in the used car market to stay competitive

Limitations

- Australia Dataset
  - not be generalizable to other regions

- Focus on 8 features of the car
  - other external factors: inflation rate, supply chain issues, level of maintenance



- Assumptions and limitations in each algorithm
  - affect the accuracy of the model predictions

Challenge

- Hard to apply all algorithms in a data set
  - Each algorithm has its data type requirements + data pre-processing steps

- Hard to find the price threshold in used car market based on vehile type --> take the overall average

Thank You