# **Car Data Analysis Project Report**

**Prepared by:** Abdelrahman badawi  
**Date:** August 2024

# **Introduction**

The goal of this project is to predict car selling prices based on various features such as the year of manufacture, present price, kilometers driven, and more. The dataset consists of 301 entries with 9 features. The primary objective is to develop a model that accurately predicts the selling price, aiding in understanding the key factors influencing car prices.

# Exploratory Data Analysis

The dataset includes several features related to car characteristics. Below is a brief overview of the dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CarName** | **Year** | **Selling Price** | **Present Price** | **Kms Driven** | **Fuel Type** | **Seller Type** | **Transmission** |
| ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual |
| sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual |
| ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual |

Initial exploration showed no missing values in the dataset, and the data types were appropriate for analysis. The following visualizations provide insights into the distribution of key variables:

# **Distribution of Year**

The dataset includes cars from the years 2003 to 2018, with most cars being from recent years.

A graph of a graph

Description automatically generated with medium confidence

# Distribution of Selling Price

* Selling prices range from 0.1 to 35 lakhs, with a majority of cars priced under 10 lakhs.

A graph of a graph

Description automatically generated with medium confidence

# Distribution of Present Price

* Present prices range from 0.32 to 92.6 lakhs, showing a wide spread and indicating various car models with different price brackets.

A graph of a line graph

Description automatically generated with medium confidence

# Distribution of Kilometers Driven

* Kilometers driven varies from 500 to 500,000, with a significant number of cars having driven less than 50,000 kilometers.

A graph of a line graph

Description automatically generated with medium confidence

# Selling Price vs Fuel Type

A box plot was created to analyze the relationship between fuel type and selling price. The analysis revealed that diesel cars tend to have a higher selling price compared to petrol cars. This trend may be influenced by the higher initial cost of diesel vehicles and their better fuel efficiency.

A graph of a price comparison

Description automatically generated with medium confidence

# Selling Price vs Transmission

A graph of blue and black boxes

Description automatically generated with medium confidenceAnother box plot showed that cars with automatic transmission generally have a higher selling price compared to those with manual transmission. This reflects the market's preference for the convenience of automatic cars, especially in urban areas.

# Selling Price vs Seller Type

The box plot for seller type indicated that cars sold by dealers usually fetch higher prices than those sold by individuals. Dealers often refurbish and certify their vehicles, adding value that translates into higher selling prices.

A graph with a bar chart and a bar chart

Description automatically generated with medium confidence

# Selling Price vs Owner

**A box plot comparing selling prices based on the number of previous owners showed that cars with fewer owners are generally priced higher. Cars with multiple previous owners tend to have lower resale values due to concerns about their maintenance history.**

**A graph with blue and black boxes

Description automatically generated with medium confidence**

# Distribution of Transmission Types

**A count plot revealed that manual transmission is more common in the dataset compared to automatic transmission. This reflects the market's composition, where manual cars have traditionally been more prevalent.**

**A graph of a distribution of transmission

Description automatically generated**

# Distribution of Owners Types

**A count plot showed that most cars in the dataset are first-owner vehicles. This is consistent with the trend that first-owner cars tend to have higher resale values and are more sought after in the used car market.**

**A chart with a green rectangular bar

Description automatically generated**

# Distribution of Seller Types

**A count plot indicated that a significant portion of the cars is sold by dealers, with a smaller proportion being sold by individuals. Dealers dominate the used car market, which is reflected in this dataset.**

**A graph showing a number of types

Description automatically generated with medium confidence**

# Pairplot of Numerical Variables

**A pairplot was generated to show the relationships between the numerical variables, highlighting any potential correlations or trends.**

**A graph of different sizes and shapes

Description automatically generated with medium confidence**

# Data Preprocessing

**In the preprocessing step, categorical variables such as Fuel Type, Seller Type, and Transmission were encoded into numerical values. Additionally, new features like Car\_Age, Depreciation, and Price\_Ratio were created to enhance the model's predictive power. These new features provide additional insights into the car's characteristics, helping to improve the model's performance.**

# Model Development

**The Random Forest Regressor was chosen for this project due to its ability to handle non-linear relationships and interactions between features. The model was trained on an 80/20 split of the data and evaluated using various metrics.**

* **Training Data: 80% of the data was used for training the model.**
* **Test Data: 20% of the data was used to evaluate the model’s performance.**
* **Hyperparameters: Default hyperparameters were used for the Random Forest model.**

**A prediction function was created to estimate the selling price of new cars based on the trained model.**

# 5. Model Evaluation

**Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²). The results are as follows:**

* **Mean Absolute Error (MAE): 0.87**
* **Mean Squared Error (MSE): 1.93**
* **R-squared (R²): 0.94**

**These metrics indicate that the model performs reasonably well, with a good fit to the data. The R-squared value of 0.91 suggests that the model explains 91% of the variance in the selling price.**

# Effect of Features on Predicted Price

**For a car named 'Ciaz' manufactured in 2017, with 6,900 kilometers driven, manual transmission, and petrol as the fuel type, the model predicted a selling price of approximately 6.91lakhs.**

**A screenshot of a graph

Description automatically generated**

# Actual vs. Predicted Selling Price

**A scatter plot was generated to compare the actual selling prices with the predicted values. The plot shows that the model's predictions closely follow the actual prices, with most points lying near the diagonal line, indicating accurate predictions.**

**A graph with a line going up

Description automatically generated**