Car Price Prediction – Project Report

Prepared By: Bhoomija
Tool Used: Jupyter Notebook
Model Used: Linear Regression
Dataset: car_price_prediction.csv

---

1. Introduction

Predicting the selling price of used cars is crucial for both buyers and sellers. It helps estimate a fair market value based on key vehicle features such as year, fuel type, kilometers driven, transmission type, and more.

This project uses a dataset of used cars to build a Linear Regression model that predicts car prices. The process involves cleaning data, exploring it visually (EDA), and training a machine learning model to estimate selling prices.

---

2. Exploratory Data Analysis (EDA)

The dataset contains various attributes like:

name, year, km_driven, fuel, seller_type, transmission, owner, mileage, engine, max_power, seats, and selling_price.

Key Insights:

Most cars run on Petrol followed by Diesel.

A large number of cars are between 5–10 years old.

Manual transmission is more common than automatic.

Outliers were found in columns like selling_price, km_driven, and engine.

Actions Taken:

- Converted year to car age.
- Removed missing values.
- Applied label encoding to categorical variables.

- Outliers removed using z-score thresholding.
- Features selected based on correlation and domain relevance.

---

3. Model Building

Model Used:

Linear Regression

Train-Test Split:

80% Training

20% Testing

```
X = df.drop('selling_price', axis=1)
y = df['selling_price']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

---

4. Model Performance

Metric	Value
R^2 Score	~0.78
MAE (Error)	~₹87,000
RMSE	~₹1.3 lakhs

The R^2 score suggests that about 78% of the variability in car prices is explained by the model. The Mean Absolute Error indicates that our predictions are off by roughly ₹87,000 on average.

---

5. Conclusion

The linear regression model performs fairly well given the simplicity and transparency of the approach.

There is room to improve accuracy by experimenting with advanced models like Random Forest or Gradient Boosting.

Additional data cleaning (especially standardizing mileage, engine, and power) can further boost model quality.

This project gives a practical foundation in regression modeling and real-world data preparation tasks.