Car Price Prediction Report

Name: K.GANESHA MOORTHY.

Course Name: DATA SCIENCE

Instructor Name: SIR DINESH

Date of Submission: 07/05/2025

1. Executive Summary

This report presents the analysis of car price prediction using machine learning techniques. A dataset from used cars was processed, including various features like car name, year of manufacture, fuel type, and kilometers driven. We performed data preprocessing, exploratory data analysis (EDA), and built a Linear Regression model to predict selling prices. The model showed reasonable performance, indicating potential for decision support in online car marketplaces and dealerships.

2. Introduction

Problem Statement

The used car market is growing rapidly, and pricing can be influenced by numerous features. Manually estimating car prices is prone to error and bias. Machine learning provides a scalable and data-driven approach to accurately estimate car prices.

Objective

- Clean and preprocess the car data.

- Perform exploratory data analysis (EDA).

- Build and evaluate regression models for accurate price prediction.

- Interpret the model outputs and identify key influencing features.

Dataset Description

Source: Dataset loaded from 'cars24data.csv'.
Target Variable: Selling_Price
Features: Name, Year, Kilometers Driven, Fuel Type, Transmission, Owner, Mileage, Engine, Power, Seats.

3. Data Preparation

Data Cleaning

- Loaded the dataset using pandas.
- Checked and handled missing values.
- Cleaned and converted non-numeric columns (e.g., Mileage, Engine, Power).
- Applied Label Encoding on categorical variables (Fuel Type, Transmission, Owner).

Data Transformation

- Converted relevant string values into numerical types for regression modeling.

Outliers Detection

- Used scatter plots and descriptive statistics to visually assess and handle outliers.

4. Exploratory Data Analysis (EDA)

- Generated heatmaps to study correlation between features.
- Visualized distributions of variables like Kilometers Driven, Selling Price.
- Observed strong correlation between car age, engine power, and price.

5. Modeling and Analysis

Model Selection

- Linear Regression was selected as the baseline model for this regression task.

Model Training and Testing

- Split the data into training and testing sets (80:20 ratio).
- Trained the model using `LinearRegression()` from scikit-learn.
- Evaluated predictions on the test set.

Performance Metrics

- R^2 Score
- Mean Absolute Error (MAE)
- Visualization of Actual vs Predicted Prices

6. Results and Interpretation

- The model achieved a reasonable R^2 score indicating decent predictive power.
- Top Influencing Features: Car Age, Engine Power, Kilometers Driven.
- Model showed better performance on newer and less driven cars.

Visualizations

- Correlation heatmap
- Predicted vs Actual Price Plot
- Feature Importance Graph

7. Conclusion

Summary of Analysis

- A regression pipeline was developed and tested.
- Achieved good results for predicting car prices with linear regression.
- Model can be improved by using ensemble models or adding external economic factors.

Limitations

- Limited dataset size and scope.
- Model performance may drop for outlier vehicles or rare configurations.

Suggested Next Steps

- Implement feature scaling and advanced models.
- Create a UI tool for price estimation.

8. Appendix

Code Snippets: (Add relevant cleaned code and function blocks here).

References:

- scikit-learn documentation
- pandas and seaborn libraries
- Online ML tutorials