Detailed Report for "Car Dheko - Used Car Price Prediction"

1. Project Overview

Title: Car Dheko - Used Car Price Prediction
Category: Machine Learning, Web Application
Tools and Technologies: Python, Streamlit, Pandas, Scikit-learn, Pickle, Matplotlib, Seaborn

The **Car Dheko - Used Car Price Prediction** project is a machine learning-driven application designed to predict the resale value of used cars based on various features such as brand, model, fuel type, mileage, transmission, and other car specifications. This project uses a trained machine learning model to predict prices with a user-friendly web interface built with **Streamlit**.

2. Problem Statement

Used car buyers and sellers often struggle to determine the correct market price for a vehicle due to the numerous factors that can affect the price, such as brand, car condition, mileage, and location. The goal of the **Car Dheko - Used Car Price Prediction** project is to create an automated system that can predict the resale price of used cars based on these factors. This can help users make informed decisions when buying or selling a used vehicle.

3. Data Collection

The dataset for this project is gathered from a **used car sales platform**, which contains a variety of car attributes, such as:

- **Fuel type** (Petrol, Diesel, Electric, CNG, etc.)
- **Body type** (SUV, Sedan, Hatchback, etc.)
- **Transmission** (Manual, Automatic)
- **Owner count** (Number of previous owners)
- **Brand** (Car brand like Maruti Suzuki, Hyundai, etc.)
- **Model** (Specific car model)
- **Model Year**
- **Insurance validity**
- **Kilometers driven**
- **Mileage** (km)
- **Seats**
- **Color**
- **City**

4. Data Preprocessing and Feature Engineering

The data was processed to handle missing values, duplicate rows, and outliers. Following preprocessing, the dataset was cleaned, transformed, and encoded to make it suitable for machine learning algorithms. Some of the key steps include:

- **Handling Missing Values**: If any attribute had missing values, they were either dropped or filled with appropriate values based on the data type.
- **Encoding Categorical Variables**: Categorical variables like **Fuel Type**, **Body Type**, **Transmission**, and **Brand** were encoded into numerical representations using **Label Encoding** or **One-Hot Encoding**.
- **Feature Scaling**: Features such as **Mileage**, **Kilometers Driven**, and **Model Year** were scaled to ensure the model was not biased toward any particular feature with higher numerical values.

5. Model Selection and Training

For predicting the price of used cars, multiple machine learning algorithms were tested. These include:

- **Linear Regression**: A basic regression algorithm, used to predict continuous values.
- **Random Forest Regressor**: A more advanced ensemble model that performs well with structured data and captures non-linear relationships.
- **Gradient Boosting Regressor**: A boosting algorithm that iteratively improves predictions by focusing on misclassified instances.

The **Random Forest Regressor** was chosen as the final model due to its robustness, ability to handle non-linear data, and superior performance based on evaluation metrics.

Evaluation Metrics:

- **R-squared (R^2)**: Measures the proportion of the variance in the dependent variable (car price) explained by the model.
- **Mean Absolute Error (MAE)**: The average of the absolute errors between predicted and actual prices.
- **Root Mean Squared Error (RMSE)**: Measures the square root of the mean of squared errors, giving more weight to larger errors.

6. Model Deployment and Web Interface

Once the model was trained, it was serialized using **Pickle** to save the model and load it later for predictions. A simple web application was developed using **Streamlit**, where users can input the car specifications via a form in the sidebar.

Key Features of the Web Application:

- **Car Selection**: Users can select the car's **Fuel Type**, **Body Type**, **Transmission**, **Owner Count**, **Brand**, **Model**, **Model Year**, **Insurance Validity**, **Kilometers Driven**, **Mileage**, **Seats**, **Color**, and **City**.
- **Predicted Price**: After entering the car details, the user clicks on a button to trigger the **Price Prediction**, which uses the trained model to estimate the resale price.
- **Responsive UI**: The UI is responsive, providing instant feedback to users. The car details are displayed in a card format for better readability.

7. User Interface (UI) Design

The **Streamlit** app has a simple yet elegant user interface with:

- A **sidebar** where the user selects the car specifications.
- A **main panel** displaying the car's details and the predicted price.
- A **color scheme** with a dark background and contrasting fonts, making the app visually appealing.

8. Challenges Faced and Solutions

- **Handling Missing Data**: Some entries in the dataset had missing values, especially in the **Mileage** and **Insurance Validity** columns. These were handled by imputation or removal of rows with critical missing data.
- **Data Imbalance**: Certain car models or brands had fewer records, which could have impacted the model’s performance
- **Feature Selection**: Some features were initially irrelevant or had a high correlation with others, leading to multicollinearity. Feature selection techniques helped identify and keep only the most impactful features.

9. Future Improvements

While the model and application work well for most cases, there are areas for further improvement:

- **Incorporating More Data**: More diverse datasets, including more regional and brand-specific data, would improve the accuracy of predictions.

- **Deep Learning Models**: Exploring deep learning models like **Neural Networks** for better handling of complex patterns in the data.
- **Price Adjustment for Market Trends**: Including data related to market trends and demand/supply for specific car models could help fine-tune price predictions.
- **User Authentication**: Adding user authentication could allow users to save their car details and predictions.

10. Conclusion

The **Car Dheko - Used Car Price Prediction** serves as a useful tool for anyone looking to buy or sell a used car. By leveraging machine learning techniques, the app provides an accurate prediction of the car's resale value based on various factors. This can save users both time and money by helping them make better-informed decisions.

11. References

- **Pandas**: For data manipulation and preprocessing.
- **Scikit-learn**: For machine learning models and evaluation.
- **Streamlit**: For creating the interactive web application.
- **Pickle**: For saving and loading the trained model.