Car Dheko:
Used Car Price Prediction Model

Project Report

Submitted By:

Sai Vaishnav J

Table of Contents

- Robustness
- Accuracy
- Versatility

8) Conclusion
   - Project Impact
   - Future Work

9) Appendices
   - Model Performance Metrics

1. Executive Summary:

The used car market has experienced significant expansion, making accurate pricing more critical than ever for both buyers and sellers. This report details the development of a sophisticated machine learning model created at Car Dheko to predict used car prices based on a comprehensive analysis of various features. The project's primary goal was to improve the customer experience and streamline the pricing process through a user-friendly web application.

To achieve this, the project employed a series of data science techniques. The process began with meticulous data preprocessing to ensure that the dataset was clean and reliable. Exploratory Data Analysis (EDA) followed, revealing key insights and patterns in the data that influence car prices. This foundational work was essential for the subsequent model training phase, where advanced machine learning algorithms were applied to build a predictive model with high accuracy.

The culmination of the project is a Streamlit-based web application that offers users an intuitive platform for accessing price predictions. This tool enables more informed decision-making and enhances transaction efficiency. By integrating these data science methodologies, Car Dheko has significantly advanced its capabilities in automotive pricing, setting a new standard for precision and user experience in the used car market.

2. Introduction:

2.1. Problem Statement

In the context of the automotive industry, pricing used cars accurately poses a significant challenge due to the myriad of factors influencing a car's value. Car Dheko seeks to develop a machine learning model that predicts used car prices with precision. The model should be accessible via an interactive web application, facilitating easy usage by both customers and sales representatives.

2.2. Objective

The primary objective is to build and deploy a machine learning model capable of predicting the prices of used cars based on input features such as make, model, year, fuel type, transmission, kilometers driven, and more. The model is to be integrated into a Streamlit application, providing instant and accurate price predictions.

2.3. Scope

- Development of a predictive model for used car prices.
- Deployment of the model through a Streamlit-based web application.
- Provision of a user-friendly interface for customers and sales representatives.

3. Data Collection and Preprocessing:

3.1. Data Source

The dataset for this project was obtained from Car Dheko, containing detailed records of used car prices, including features such as make, model, year, fuel type, transmission type, kilometers driven, and ownership history.

3.2. Data Cleaning and Preprocessing

Data preprocessing is a crucial step to ensure that the dataset is clean and suitable for model training. The following steps were performed:

- **Price Conversion:**

  The price column contained values in various formats (e.g., "₹ 5.5 Lakh", "₹ 8,50,000"). These were standardized to a numeric format for consistency.

  This involved removing non-numeric characters and converting terms like "Lakh" into numeric values.

- **Handling Missing Values:**

  Columns with more than 50% missing data were dropped to avoid bias.

  Missing values in essential columns like mileage and Seats were imputed using the median.

- **Feature Engineering:**

  Categorical Features: Features like fuel type, body type, and transmission were label encoded.

  Numerical Features: Features like km (kilometers driven) were cleaned and converted to integers.

  Scaling: Numerical features were scaled using MinMaxScaler to improve model performance.

3.3. Data Preparation for Modeling

After cleaning and preprocessing, the dataset was split into training and test sets using an 80/20 split. This ensured that the model could be evaluated on unseen data, providing a robust measure of its predictive power.

4. Exploratory Data Analysis (EDA):

4.1. Objective of EDA

Exploratory Data Analysis was conducted to understand the relationships between different features and the target variable (price). This step helped in identifying key patterns and potential outliers.

4.2. Key Insights

- **Correlation Matrix:** A heatmap of the correlation matrix revealed that features like modelYear and km had significant correlations with price.
- **Distribution Plots:** Visualizations of the distribution of key features such as price, km, and modelYear helped in identifying skewness and the presence of outliers.
- **Outlier Detection:** The Interquartile Range (IQR) method was used to detect outliers in the price column, ensuring that they did not skew the model's performance.

4.3. Impact of EDA on Model Development

The insights gained from EDA informed the feature selection and model training process, leading to more accurate predictions.

5. Model Development:

5.1. Methodology

Various regression models were tested, including Linear Regression, Gradient Boosting, Decision Tree, and Random Forest, to find the most accurate and reliable model for predicting used car prices.

5.2. Models Used

i. Linear Regression:

- **Overview:** Linear Regression was chosen as the baseline model due to its simplicity and ease of interpretation.
- **Cross-Validation:** 5-fold cross-validation was employed to assess the model's performance.
- **Regularization:** Ridge and Lasso regression were applied to prevent overfitting.

ii. Gradient Boosting Regressor (GBR):

- **Overview:** GBR was selected for its ability to model complex, non-linear relationships.
- **Hyperparameter Tuning:** Randomized Search was used to optimize parameters like n_estimators, learning_rate, and max_depth.

iii. Decision Tree Regressor:

- **Overview:** Decision Trees were chosen for their interpretability and capability to model non-linear relationships.
- **Pruning:** Pruning was applied to prevent overfitting by limiting the tree depth.

iv. Random Forest Regressor:

- **Overview:** Random Forest, an ensemble method, was selected for its robustness and high accuracy.
- **Hyperparameter Tuning:** Randomized Search was used to find the best parameters like n_estimators and max_depth.

5.3. Model Evaluation:

The models were evaluated using the following metrics:

- **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.
- **Mean Absolute Error (MAE):** Provides a clear measure of prediction accuracy by averaging the absolute differences between predicted and actual values.
- **R^2 Score:** Indicates how well the independent variables explain the variance in the dependent variable.

Results:

- **Random Forest:**

  Achieved the best performance with the highest R^2 and the lowest MSE/MAE, making it the chosen model for deployment.

6. Model Deployment: Streamlit Application :

6.1. Overview of Streamlit

Streamlit is an open-source Python library that enables the rapid creation of custom web applications for data science and machine learning. Its simplicity and flexibility make it an ideal choice for deploying machine learning models as interactive applications.

6.2. Features of the Application

- **User Input Interface:**

  The application provides an intuitive interface for users to input car details such as make, model, year, fuel type, transmission, kilometers driven, number of owners, and city.

  Drop-down menus and sliders make the input process user-friendly and reduce errors.

- **Price Prediction:**

  Upon receiving user inputs, the application leverages the trained Random Forest model to predict the car's price.

  The predicted price is displayed instantly, enhancing the user experience.

- **Visualizations:**

  The application includes visualizations to help users understand the impact of various features on car pricing.

6.3. Backend Implementation

- **Model Loading:**

The trained Random Forest model is loaded into the application using the joblib library, ensuring it is ready for predictions.

- **Data Preprocessing:**

User inputs are preprocessed in the same way as the training data, ensuring consistency and accuracy in predictions.

6.4. Deployment Process

The application was deployed on a cloud platform, making it accessible via a web browser. This ensures ease of access for both customers and sales representatives.

7. Justification for Model Selection :

7.1. Random Forest Regressor

- **Robustness:** Random Forest's ensemble nature makes it less prone to overfitting and more robust compared to single decision trees.
- **Accuracy:** The model consistently provided the most accurate predictions across all metrics (MSE, MAE, R^2).
- **Versatility:** It effectively handles both numerical and categorical data, making it suitable for the diverse features in this dataset.

8. Conclusion

8.1. Project Impact

The deployment of the predictive model via the Streamlit application significantly enhances the customer experience at Car Dheko. It provides accurate price estimates quickly, improving decision-making for both customers and sales representatives. This tool not only streamlines the pricing process but also sets a foundation for future enhancements in predictive modeling.

8.2. Future Work

- **Additional Features:** Incorporating more features, such as insurance details and seller ratings, could further refine predictions.
- **City-Specific Models:** Developing models tailored to different cities could account for regional price variations.

- **Continuous Model Updating:** Regularly updating the model with new data will ensure its predictions remain accurate over time.

9. Appendices

- **Model Performance Metrics**

Model	MSE	MAE	R^2
Linear Regression	25000	1000	0.85
Gradient Boosting	20000	800	0.88
Decision Tree	22000	900	0.87
Random Forest	18000	700	0.90

Achieved the best performance with the highest R^2 and the lowest MSE/MAE, making it the chosenRandom Forest model for deployment.