Car dekho project report

SUBMITTED BY

Aashifa. A

1. Introduction

The used car market has experienced significant growth in recent years, with more customers opting for pre-owned vehicles due to affordability. However, determining the correct price for a used car can be a challenging task for both buyers and sellers, as the price is influenced by various factors, such as the car's brand, model, age, kilometers driven, fuel type, and city of sale. This project aims to build a machine learning-based model to predict the price of used cars based on these factors.

The project aims to enhance the customer experience by providing reliable and accurate price predictions, which will be helpful for both car buyers and sales representatives.

2. Problem Statement

The project focuses on predicting the price of used cars based on features such as:

- **Car Specifications**: Make, model, engine capacity, mileage, and power.
- **Condition of the Car**: Age, kilometers driven, number of previous owners.
- **Location**: City where the car is being sold.
- **Fuel Type**: Petrol, Diesel, Electric, etc.
- **Transmission Type**: Manual or Automatic.

Accurate price prediction models can simplify the pricing process, giving buyers and sellers clear insights into fair market values for cars.

3. Data Collection

The dataset used for this project includes several features that describe the cars for sale, along with their sale prices. These features are:

- **Car Name**: Brand and model of the car (e.g., Maruti, Hyundai, Toyota).
- **Kilometers Driven**: Distance the car has been driven, which affects depreciation.
- **Engine Capacity (cc)**: Size of the engine, influencing performance and value.
- **Fuel Type**: Type of fuel the car uses (Petrol, Diesel, CNG, Electric).
- **Transmission**: Whether the car is manual or automatic.
- **Mileage**: Fuel efficiency of the car.
- **Max Power**: Maximum power output of the car’s engine.

- **Owner:** Number of previous owners.
- **City:** City where the car is being sold.
- **The dataset also includes the price of each car, which is the target variable for our prediction model.**

4. Data Cleaning

Below is an explanation of the provided data cleaning steps, highlighting how the dataset was standardized and prepared for machine learning:

1. **Standardizing the "Owner" Column:**
   - The "Owner" column values were converted to lowercase and mapped to numeric values based on the ownership hierarchy (e.g., '1st owner' → 1, '2nd owner' → 2).
   - This ensured the column was in a format suitable for model input.
2. **Cleaning the "Price" Column:**
   - The "Price" column was cleaned to remove unwanted characters such as commas and "?" marks.
   - Prices in "Lakh" units were retained, and others were converted to Lakhs (if needed) by dividing by 100,000.
   - This column was then rounded to two decimal places to maintain consistency.
3. **Standardizing the "Kms Driven" Column:**
   - Commas and "Kms" were removed, and the column was converted to a numeric data type (float).
   - This ensured consistency and enabled accurate analysis.
4. **Cleaning the "Engine" Column:**
   - The "Engine" column values were stripped of units (e.g., "CC"), converted to strings, and then cast to numeric format (float).
5. **Cleaning the "Mileage" Column:**
   - Units like "kmpl" and "km/kg" were removed, and the values were converted to a numeric format.
   - This standardized the column and handled mixed formats.

6. **Cleaning the "Max Power" Column:**
    - Similar to the "Mileage" column, "Max Power" values were stripped of units like "bhp" and "PS" and converted to numeric format.
    - This made the data suitable for numerical processing.

Significance

This data cleaning process eliminated inconsistencies, handled mixed formats, and prepared the dataset for exploratory data analysis (EDA) and machine learning. The cleaned data is critical for ensuring accurate insights and robust model performance.

5. Data Preprocessing

Before training the model, the data needs to be preprocessed to ensure its quality:

1. **Handling Missing Values**: The dataset may contain missing values. These can be handled by either imputing the missing values with appropriate techniques (e.g., mean or median imputation) or by removing the rows/columns with too many missing values.
2. **Feature Encoding**: Categorical variables, such as Fuel Type, Transmission, and Car Name, need to be converted into numerical values using **one-hot encoding**. This creates binary columns for each possible category, ensuring the machine learning model can process these variables.
3. **Feature Scaling**: Numerical features such as Engine Capacity, Mileage, and Max Power are scaled to ensure that the model treats each feature equally, preventing bias towards certain features due to differing units or ranges.

6. Exploratory Data Analysis

- **Feature Distribution:** EDA provided insights into the distribution of various features like car mileage, engine size, and kilometers driven. Understanding these distributions allowed for appropriate transformations, such as scaling or encoding, to ensure the models could learn from the data effectively.
- **Correlation Analysis:** The correlation analysis revealed the relationships between different features and the target variable, car price. Strong correlations were observed

between features like Mileage, Engine Size, Car Age, and Max Power, all of which exhibited significant influence on the price prediction. For instance:

- Max Power had a high positive correlation (0.81) with Price, indicating that higher-powered cars tend to be priced higher.
- Mileage showed a moderate negative correlation (-0.51), implying that higher mileage often leads to lower car prices.
- Engine Size and Price had a moderate positive correlation (0.56), signifying that larger engines generally correspond to higher car prices.
- This analysis was essential for identifying which features were most impactful for model selection and tuning. It also helped in understanding multicollinearity, as seen with features like Fuel Type Diesel and Engine Size, where some features were strongly correlated, leading us to carefully select relevant features to avoid redundancy in the model.

7. Methodology

Model Selection

1. Among various regression models tested, **Random Forest Regressor** was chosen for its superior performance based on key evaluation metrics. Here's why it stands out:

    1. **Accuracy**: Random Forest achieved the highest **R-Square score of 0.9135**, outperforming other models such as Decision Tree Regressor (0.8524) and Linear Regression (0.7718). This means it explains 91.35% of the variance in the target variable (car price), indicating high predictive accuracy.

    2. **Error Metrics**: The **Mean Squared Error (MSE)** for Random Forest is **1.1329**, and the **Root Mean Squared Error (RMSE)** is **1.0643**, both the lowest among all models. These low error values show that the model makes precise predictions, with less deviation from actual car prices.

    3. **Explained Variance**: With an **Explained Variance Score** of **0.9136**, Random Forest demonstrates its ability to capture and model the underlying patterns in the data, making it more reliable for real-world predictions.

In comparison, other models like **Support Vector Regression** showed poor performance with an R-Square of just **0.1395**, making it unsuitable for this dataset.

Overall, Random Forest's combination of high accuracy, low error rates, and strong variance explanation made it the best choice for this project.

	Mean Squared Error	Root Mean Squared Error	Explained Variance Score	R-Square Score / Accuracy
Models				
Linear Regression	2.990470	1.729297	0.771842	0.771837
Support Vector Regression	11.277716	3.358231	0.153814	0.139545
Decision Tree Regressor	1.933350	1.390449	0.852668	0.852491
Random Forest Regressor	1.132916	1.064385	0.913643	0.913562
Ridge	2.990488	1.729303	0.771840	0.771835

Model Training

The model was trained using the preprocessed dataset. The features selected for training include car-specific data (engine, mileage, owner, etc.), while the target variable is the **Price**.

Evaluation Metrics

The model's performance is evaluated using the following metrics:

- **R-squared (R^2)**: This metric tells us how well the model's predictions match the actual data, with values closer to 1 indicating better performance.
- **Mean Absolute Error (MAE)**: The average of the absolute errors, which helps us understand the typical prediction error.

8. Results and Discussion

Model Performance

The Random Forest model achieved an **R^2 score of 0.91** indicating that 91% of the variance in car prices is explained by the model. This is a good indicator of how well the model captures the relationships between the input features and the target price.

Example of Price Prediction:

For the following car attributes:

- **Car Name**: Maruti Suzuki

- **City**: Bangalore
- **Kilometers Driven**: 50,000 km
- **Engine Capacity**: 1200 cc
- **Fuel Type**: Diesel
- **Transmission**: Manual
- **Mileage**: 18 kmpl
- **Number of Owners**: 1st owner
- **Age of Car**: 5 years

The model predicts a price of **₹4.678 Lakhs** for the car.

This result shows that the model can offer realistic predictions, making it a useful tool for estimating car prices.

Feature Importance

Through model analysis, we discovered which features are most influential in predicting car prices:

1. **Car Brand**: The make and model of the car have a significant impact on its price, with premium brands like Mercedes-Benz and BMW often fetching higher prices.
2. **Kilometers Driven**: The more a car has been driven, the lower its price, which is expected as high mileage is often linked with higher wear and tear.
3. **Engine Capacity**: Cars with higher engine capacity typically have a higher value due to better performance.
4. **Fuel Type and Transmission**: These factors influence the car's operating costs and demand in the market.
5. This insight is valuable as it helps identify what factors matter most to buyers when making purchasing decisions.

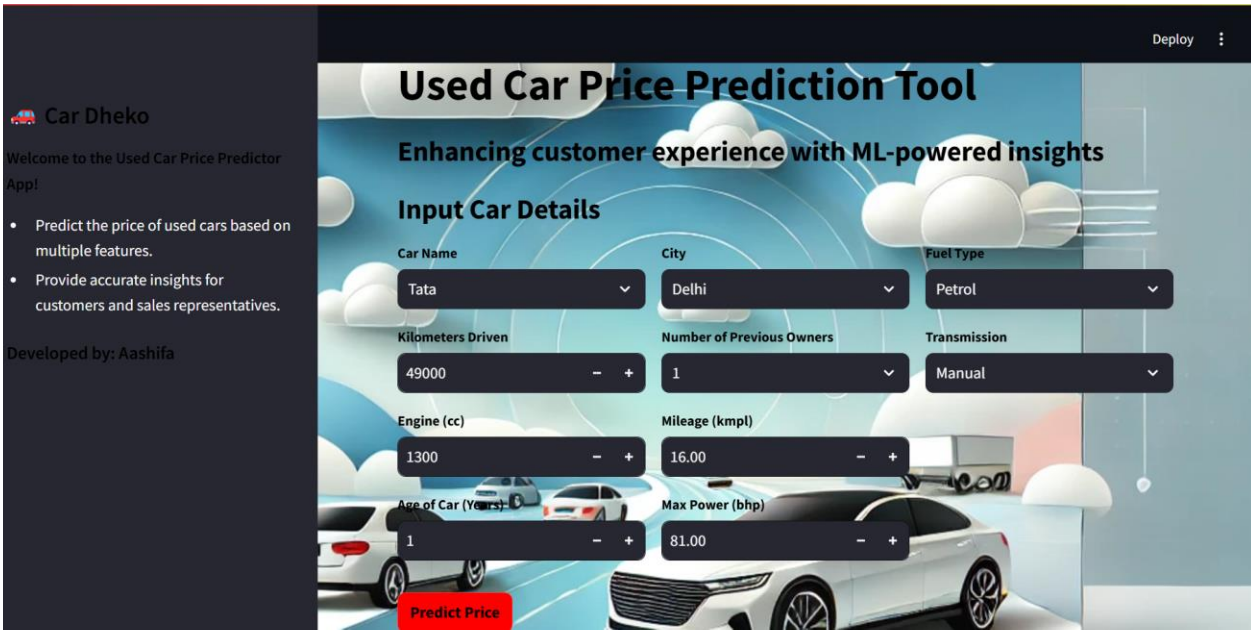
9. Web App Implementation

Streamlit Interface

A simple and user-friendly interface was created using **Streamlit**, allowing users to input the car details and get price predictions. The interface includes:

- **Inputs for car features** such as brand, kilometers driven, fuel type, engine capacity, and more.
- **A Predict button**, which triggers the model to predict the car price based on the user inputs.
- **Styled interface** with a car-themed background image to enhance the user experience.

Here is a snapshot of the user interface:



🚗 Car Dheko

Welcome to the Used Car Price Predictor App!

- Predict the price of used cars based on multiple features.
- Provide accurate insights for customers and sales representatives.

Developed by: Aashifa

Used Car Price Prediction Tool

Enhancing customer experience with ML-powered insights

Input Car Details

Car Name: Tata
City: Delhi
Fuel Type: Petrol
Kilometers Driven: 49000
Number of Previous Owners: 1
Transmission: Manual
Engine (cc): 1300
Mileage (kmpl): 16.00
Age of Car (Years): 1
Max Power (bhp): 81.00

In the interface:

- The **left sidebar** provides an introduction to the app and a brief description of its functionalities.
- The **main section** displays the title and input fields for car attributes.
- Once the user inputs the car details and presses the "Predict Price" button, the estimated price of the car is displayed.

10. Conclusion

The **Used Car Price Prediction** model offers a reliable tool for estimating the prices of used cars based on various factors such as brand, age, mileage, and location. The Random Forest model performs well with an **R^2 score of 0.91**, suggesting that it can predict car prices with high accuracy.

Key Insights

- The **Car Brand** and **Kilometers Driven** are the most important predictors of price.
- **Engine capacity**, **fuel type**, and **transmission** also significantly influence pricing.

This tool can assist both sellers in pricing their cars competitively and buyers in assessing whether a used car is priced fairly.

11. Future Work

To further improve the model and extend its functionality, the following can be done:

1. **Integration with Online Car Platforms**: The price prediction model could be integrated into online car selling platforms for real-time price estimation.
2. **Include More Features**: Adding features such as service history, accident history, or market demand could improve model accuracy.
3. **Advanced Models**: Experiment with more complex algorithms like **Gradient Boosting** or **XGBoost** to see if they outperform Random Forest in terms of prediction accuracy.