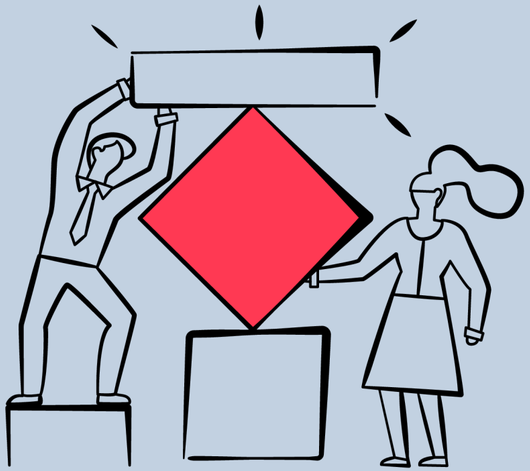

Car Dheko - Used Car Price Prediction

PRESENTATION

By: Mohamed yunus. T

Table Of Content

Executive Summary

1. The used car market has witnessed substantial growth, making accurate pricing increasingly crucial for both buyers and sellers. This report outlines the development of an advanced machine learning model at Car Dheko designed to predict used car prices through a thorough analysis of various features. The primary objective of the project was to enhance customer experience and optimize the pricing process via a user-friendly web application.

2. To achieve this, the project utilized a range of data science techniques. The process began with rigorous data preprocessing to ensure a clean and reliable dataset. This was followed by Exploratory Data Analysis (EDA) to uncover key insights and patterns that affect car prices. These preliminary steps were vital for the model training phase, where sophisticated machine learning algorithms were employed to build a highly accurate predictive model.

3. The project culminated in a Streamlit-based web application, providing users with an intuitive platform for accessing price predictions. This tool facilitates better decision-making and improves transaction efficiency. By integrating these data science methodologies, Car Dheko has significantly enhanced its capabilities in automotive pricing, establishing a new benchmark for precision and user experience in the used car market.

Introduction

A. Problem Statement:

In the automotive industry, accurately pricing used cars presents a significant challenge due to the multitude of factors that influence a car's value. Car Dheko aims to develop a machine learning model to precisely predict used car prices. This model will be integrated into an interactive web application, making it easily accessible for both customers and sales representatives.

B. Objective

The primary objective is to develop and deploy a machine learning model that can predict the prices of used cars based on input features such as make, model, year, fuel type, transmission, kilometers driven, and other relevant factors. The model will be integrated into a Streamlit application to deliver instant and accurate price predictions.

C. Scope

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

Data Collection And Preprocessing

Data Source

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

Data Cleaning And Preprocesing

Data preprocessing ensures a clean dataset for model training. Steps included converting price formats to numeric values, handling missing data by dropping columns with over 50% missing values or imputing with the median, and feature engineering, including encoding and scaling.

Data Preparation For Modeling

fter 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.

Exploratory Data Analysis (EDA)

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.

Key Insights

Correlation Matrix:
Revealed significant correlations between features like modelYear, km, and price.

Distribution Plots:
Identified skewness and outliers.

Outlier Detection:
Used IQR to detect price outliers.

Impact Of EDA On Model Development

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

Model Development

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.

Model Evaluation:

The models were evaluated using three key metrics: **Mean Squared Error (MSE)**, which measures the average squared differences between actual and predicted values; **Mean Absolute Error (MAE)**, which averages the absolute prediction errors; and the **R^2 Score**, which shows the variance explained by the model.

Model Used:

Linear Regression was used as a baseline model. Gradient Boosting and Decision Tree Regressors modeled non-linear relationships with hyperparameter tuning. Random Forest Regressor, an ensemble method, was chosen for robustness and accuracy, with parameters optimized using Randomized Search.

Result:

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

Model Deployment: Streamlit Application

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.

Features Of Application :
The application offers an intuitive user input interface with drop-down menus and sliders for entering car details like make, model, year, fuel type, and more, minimizing errors. It uses a trained Random Forest model for instant price predictions and provides visualizations to explain feature impacts.

Backend Implementation :
The application loads the trained Random Forest model using the joblib library, making it ready for predictions. User inputs are preprocessed similarly to the training data, ensuring consistency and accuracy in the predictions.

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.

Justification for Model Selection

The Random Forest model offers robustness due to its ensemble nature, making it less prone to overfitting compared to single decision trees. It consistently provided the most accurate predictions across all metrics (MSE, MAE, R^2) and effectively handles both numerical and categorical data, demonstrating versatility for diverse dataset features.

Conclusion

The deployment of the predictive model through the Streamlit application greatly enhances the customer experience at Car Dheko by providing quick and accurate price estimates, aiding both customers and sales representatives in decision-making. This tool not only streamlines the pricing process but also lays a foundation for future enhancements in predictive modeling. Future work could involve incorporating additional features like insurance details and seller ratings to refine predictions, developing city-specific models to account for regional price variations, and continuously updating the model with new data to maintain prediction accuracy over time.

Appendices

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

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

Thank You