Used Car Price Prediction

Internship Project Report

Name: CHRISTY E

Date: 03.10.20025

1. Introduction

The Used Car Price Prediction project focuses on predicting the resale value of used cars based on various features such as year of manufacture, fuel type, kms driven, city, body type, and transmission type. The main objective of this project is to help users and dealers estimate the fair market value of a car.

2. Dataset Details

The dataset contains 7,400 records and 29 columns. After cleaning, 7,396 records remained.

✅ Dataset Loaded Successfully
Original Shape: (7400, 29)

✅ Data Cleaning Completed
Shape after cleaning: (7396, 29)

Preview of dataset (first 5 rows):
        car_name yr_mfr fuel_type kms_run sale_price   city \
0   maruti swift   2015   petrol   8063   386399 noida
1   maruti alto 800   2016   petrol   23104   265499 noida
2 hyundai grand i10   2017   petrol   23402   477699 noida
3   maruti swift   2013   diesel   39124   307999 noida
4 hyundai grand i10   2015   petrol   22116   361499 noida

   times_viewed body_type transmission   variant ... total_owners \
0   18715 hatchback   manual   lxi opt ...   2
1   2676 hatchback   manual   lxi ...   1
2   609 hatchback   manual sports 1.2 vtvt ...   1
3   6511 hatchback   manual   vdi ...   1
4   3225 hatchback   manual   magna 1.2 vtvt ...   1

   broker_quote original_price car_rating   ad_created_on \
0   397677   404177.0   great 2021-04-04 07:09:18.583
1   272935   354313.0   great 2021-03-22 14:07:32.833
2   469605   467480.0   great 2021-03-20 05:36:31.311
3   294262   374326.0   great 2021-01-21 12:59:19.299
4   360716   367216.0   great 2021-04-01 13:33:40.733

   fitness_certificate emi_starts_from booking_down_pymnt reserved \
0   True   8975   57960   False
1   True   6167   39825   False
2   True   11096   71655   False
3   True   7154   46200   False
4   True   8397   54225   False

   warranty_avail
0   False
1   False
2   False
3   False
4   False

[5 rows x 29 columns]

3. Exploratory Data Analysis (EDA)

To better understand the dataset, several exploratory analyses were performed.

3.1 Correlation Heatmap

➢ The heatmap shows strong correlations between sale_price and features like original_price, broker_quote, and emi_starts_from.

➢ Negative correlation is seen between yr_mfr and kms_run, which makes sense as older cars usually have more kilometers driven.


Correlation Heatmap
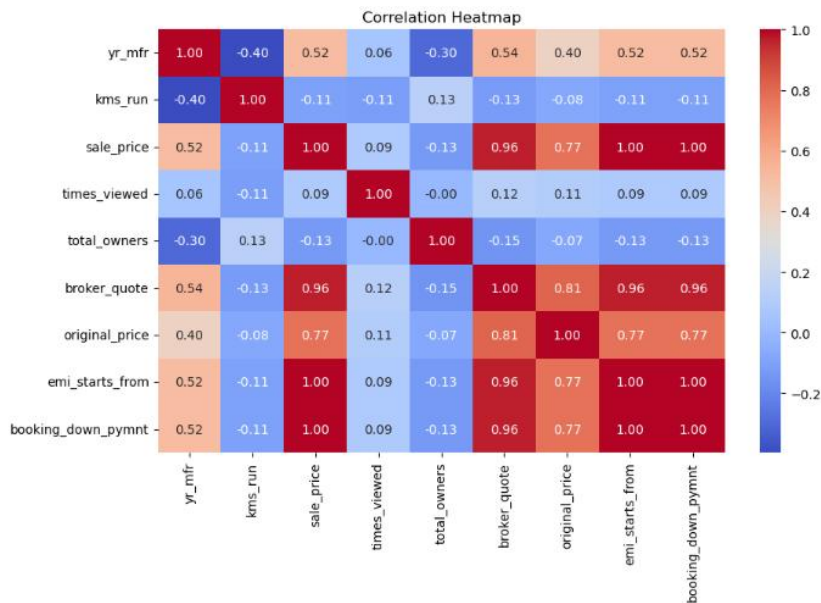
3.2 Scatterplot Sale Price vs Kms Run

➢ As expected, higher kilometers driven generally lead to lower resale prices.

➢ However, some outliers exist, showing that well-maintained older cars can still fetch high prices

Sale Price vs Kms Run (log scale)

3.3 Average Price by Body Type

➤ Luxury SUVs and Luxury Sedans command the highest resale prices.

➤ Hatchbacks and Sedans fall in the lower price range, which matches market trends.
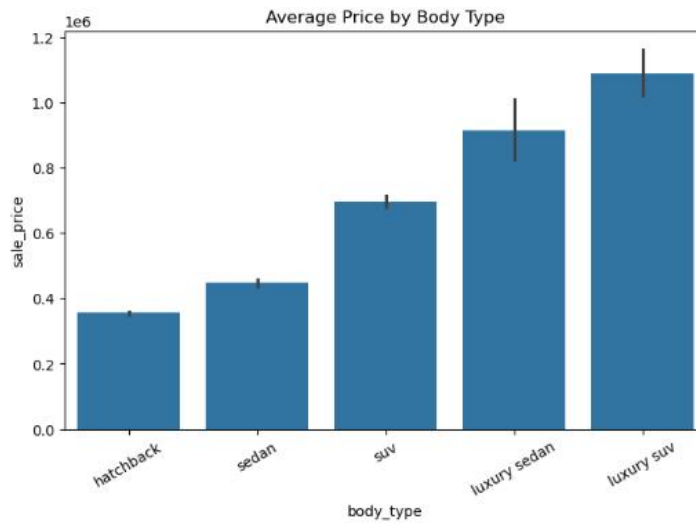
3.4 Average Price by Fuel Type

➤ Diesel cars generally sell at higher prices than petrol cars.

➤ Electric cars show wide price variation due to limited data and newer market trends.
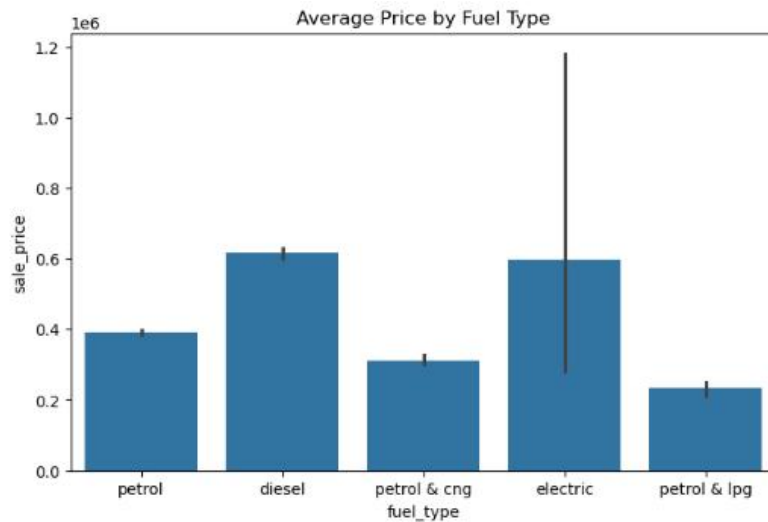
3.5 Year of Manufacture vs Sale Price

➢ Newer cars (2015 onwards) clearly have higher average resale values.

➢ Older cars (before 2005) mostly sell at very low prices, with some exceptions for premium models.
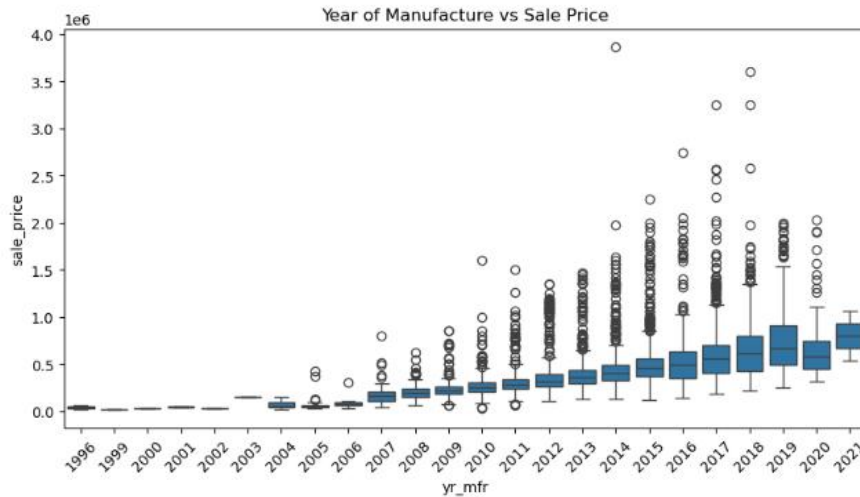
3.6 Top 10 Cities by Price

➢ Metro cities like Mumbai, Delhi, and Bangalore show higher average resale prices.

➢ Smaller cities show relatively lower resale values.

4. Model Building & Results

Three models were trained and evaluated:

➢ Linear Regression

➢ Random Forest

➢ XGBoost

Among these, XGBoost performed the best with the highest R² score and lowest RMSE/MAE.

4.1 Model Results and Comparison

```
📊 Model Performance on Test Data:
R² Score : 0.9356
RMSE      : 73,795
MAE       : 42,921

🔥 Top 5 Features Influencing Car Price:
body_type      0.513551
yr_mfr         0.168378
fuel_type      0.127482
transmission   0.080769
car_name       0.066269
dtype: float32

✅ Project Summary:
- Dataset Shape: (7396, 29)
- Total Cars Analyzed: 7396
- Model Used: XGBoost Regressor (trained on 8 features)
- Best Features: body_type, yr_mfr, fuel_type, transmission, car_name
- Achieved R² Score of 0.94 → model explains 93.6% of price variation.
```
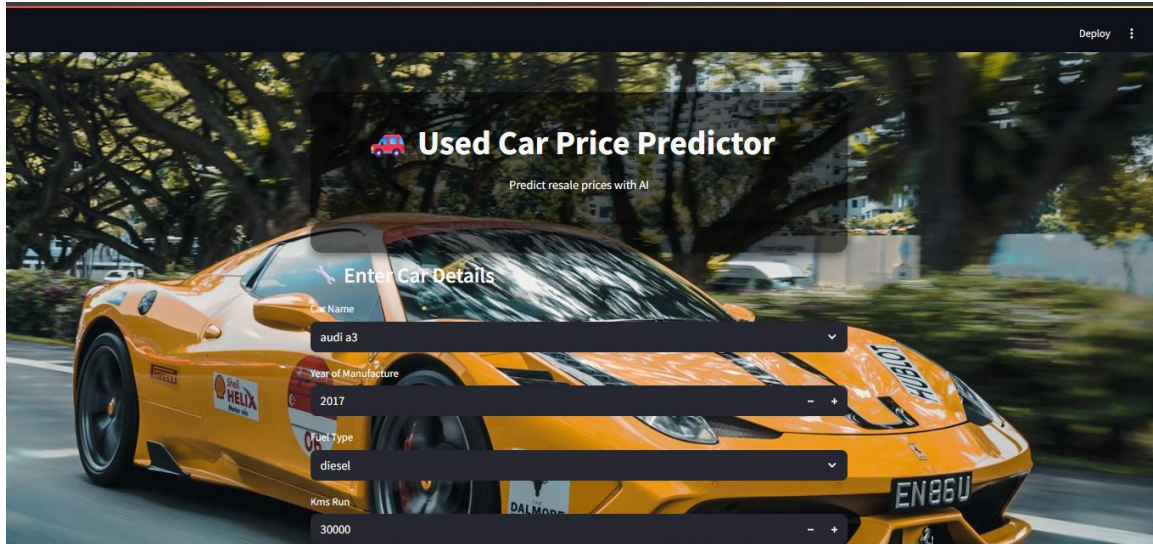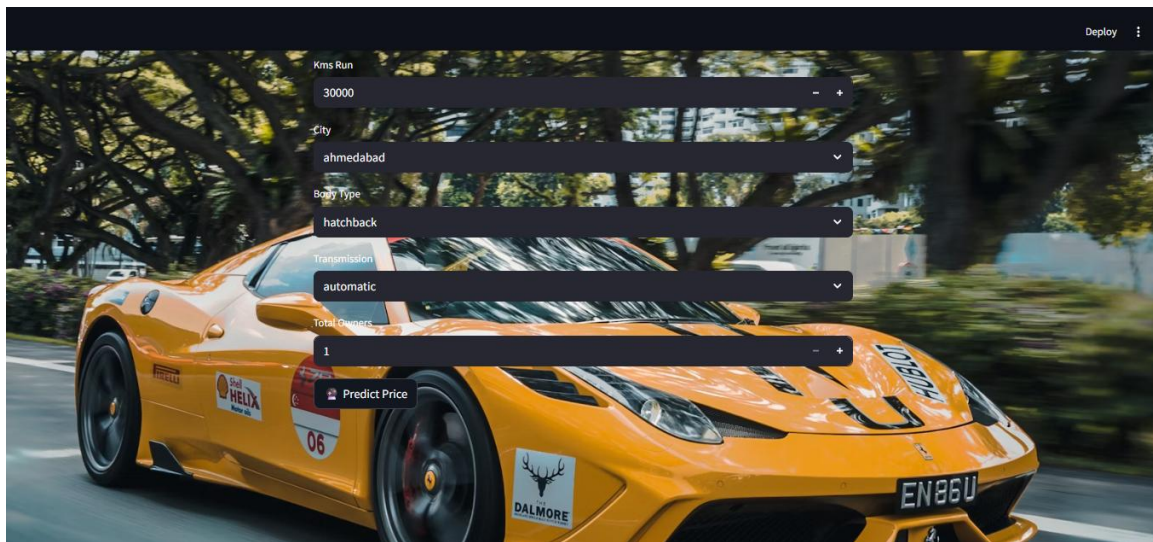
5. Streamlit Application

A Streamlit application is a Python program that creates interactive web apps for data visualization and machine learning with minimal code.

5.1. Streamlit Before

5.2 Streamlit After

6. Conclusion

The project successfully demonstrated the process of cleaning data, performing exploratory analysis, building predictive models, and creating a user-friendly application. Factors such as body type, fuel type, and year of manufacture significantly influence the resale value of cars. The XGBoost model was chosen as the best model for prediction tasks.