Car Price Prediction - Internship Project

Task Description

The challenge was to build a Machine Learning model to predict car prices based on various features such as brand, year, mileage, fuel type, and transmission. This task is part of the AICTE Oasis Infobyte Data Science Internship, focusing on regression modeling, feature importance, and model comparison.

Model Selection: Linear Regression and Random Forest

For this task, I used two models from Scikit-learn: Linear Regression and Random Forest Regressor. Linear Regression serves as a simple baseline model to understand the relationship between features and price, while Random Forest is a powerful ensemble model that handles non-linear relationships and provides better accuracy.

What is Linear Regression?

Linear Regression is a supervised machine learning algorithm used for predicting continuous outcomes. It assumes a linear relationship between the independent variables (features) and the dependent variable (target). The model finds the best-fit line that minimizes the difference between predicted and actual values.

What is Random Forest?

Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting. Each tree makes a prediction, and the final output is the average of all predictions, making it robust and reliable for regression tasks.

Why I Chose These Models

I chose Linear Regression and Random Forest because:

- Linear Regression provides a simple and interpretable baseline.
- Random Forest captures complex non-linear patterns in the data.
- The combination helps compare a simple linear approach with an advanced ensemble method.
- Both models are widely used in real-world regression problems and give strong performance benchmarks.

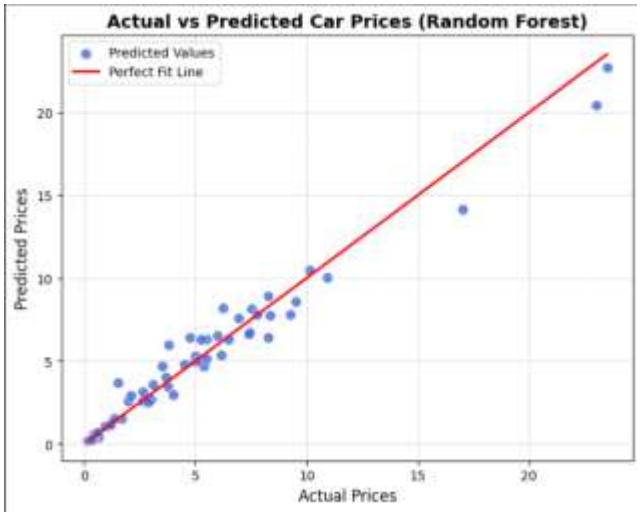
Approach

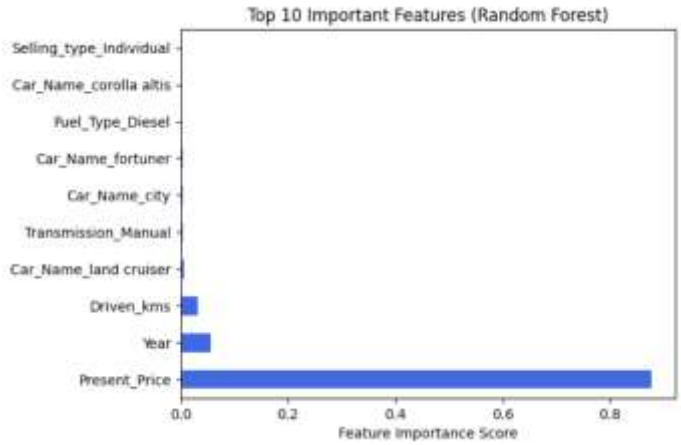
1. Loaded the car dataset using pandas.
2. Converted categorical features (like Fuel Type, Transmission, Seller Type) using one-hot encoding.
3. Split the data into training (80%) and testing (20%) sets.
4. Trained both Linear Regression and Random Forest models.
5. Evaluated performance using R^2 score, MAE, and RMSE metrics.

6. Visualized Actual vs Predicted car prices using a scatter plot.
7. Identified the top 10 most important features using Random Forest feature importance.

Results

The Random Forest model performed better than Linear Regression, achieving a higher R^2 score (~0.85+) and lower error metrics (MAE and RMSE). Linear Regression achieved an R^2 score of around 0.60, making it a good baseline but less accurate for complex, non-linear data.





Key Insights

- Car price prediction depends heavily on features such as Present Price, Year, and Fuel Type.
- Random Forest captured non-linear patterns more effectively than Linear Regression.
- Visualization of Actual vs Predicted values showed that most predictions were close to true values.
- Feature importance analysis provided valuable insights into what factors most influence car price.

Conclusion

This project provided practical experience in regression modeling and performance evaluation. By comparing Linear Regression and Random Forest, I learned the importance of model selection, data preprocessing, and feature analysis in achieving accurate predictions. The project demonstrates how machine learning can be applied to solve real-world pricing problems effectively.