



Capstone Project

Navneet kaur

Problem Statement:

- The used car market in India is a dynamic and ever-changing landscape. Prices can fluctuate wildly based on a variety of factors including the make and model of the car, its mileage, its condition and the current market conditions. As a result, it can be difficult for sellers to accurately price their cars.
- I have the sales data of all the cars sold during the time frame of 1983 to 2020. I am going to analyse this data set in order to help them expand their business, gain and retain customers, and stand out the competitions they face.

The data set has 8128 data points with 12 features in it related to :

Car Details - *Car name, transmission, fuel type, number of seats, year of manufacturing.*

- **Engine Details** - *Mileage, Engine type, Maximum power in BHP.*
- **Sale Details** - *Selling price, kilometers driven by the car.*



Approach:

We propose to develop a machine learning model that can predict the price of a used car based on its features. The model will be trained on a dataset of used cars that have been sold on Cardekho.com in India.

Benefits:

- The benefits of this solution include:
- Sellers will be able to more accurately price their cars which will help them to sell their cars faster and for a higher price.
- Buyers will be able to find cars that are priced more competitively.
- The overall used car market in India will become more efficient.



Numerical features:

- ✓ Selling_price
- ✓ Year
- ✓ Km_driven
- ✓ Mileage
- ✓ Engine
- ✓ Max_power

Categorical features

- ✓ Name
- ✓ Fuel
- ✓ Seller_type
- ✓ Transmission
- ✓ Owner
- ✓ seats

1.Data Cleaning and Preparation:

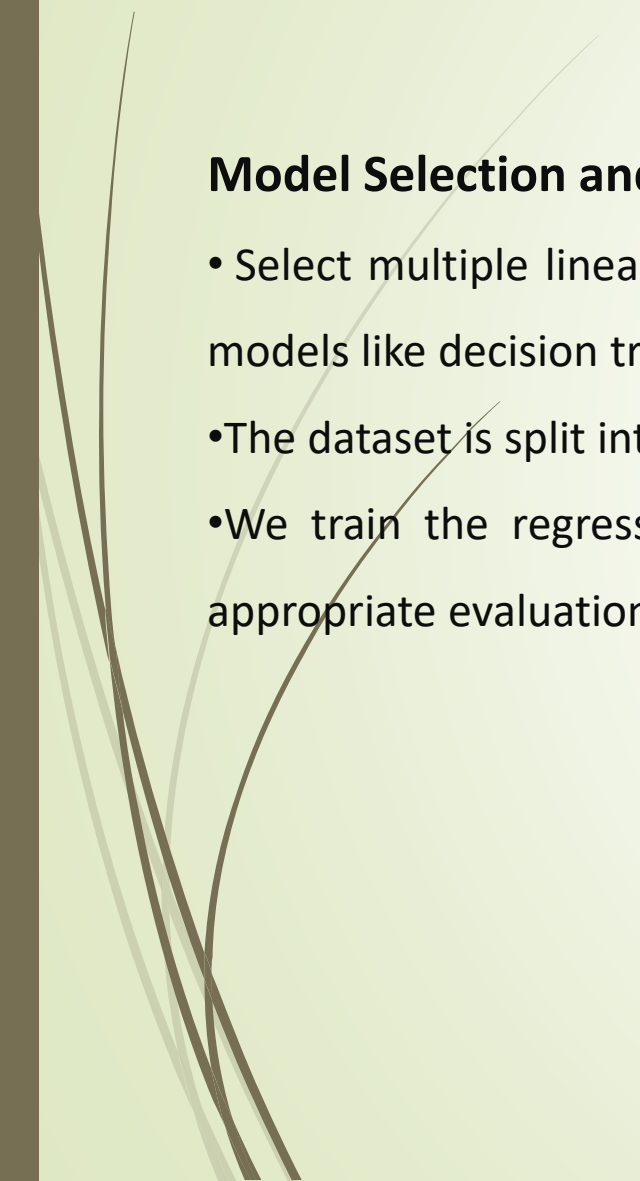
1. Begin by inspecting the dataset for missing values, outliers, and inconsistencies. Missing values can be imputed or dropped depending on their impact. Outliers may need to be handled cautiously; they could represent genuine data points or errors.
2. Categorical variables like transmission, seller-type, owner and fuel type are converted into numerical format using techniques like one-hot encoding.
3. Numerical features might need normalization or scaling, especially if they have different scales. This ensures that all features contribute equally to the model.

2.Exploratory Data Analysis (EDA):

1. Explore the distribution of features and their relationships with the target variable (selling price). Scatter plots and pairplot are useful for visualizing these relationships.
2. Identify correlations between features using correlation matrices or pairwise plots. Strong correlations between features could indicate multicollinearity, which may affect the model's performance.




Model Selection and Training:

- Select multiple linear regression as the baseline model due to its simplicity and interpretability. Other regression models like decision trees, random forests, and gradient boosting are also considered for comparison.
 - The dataset is split into training and testing sets. Cross-validation may also be used for robust evaluation.
 - We train the regression models on the training set and evaluate their performance on the testing set using appropriate evaluation metrics such as mean squared error (MSE), and R-squared.
- 

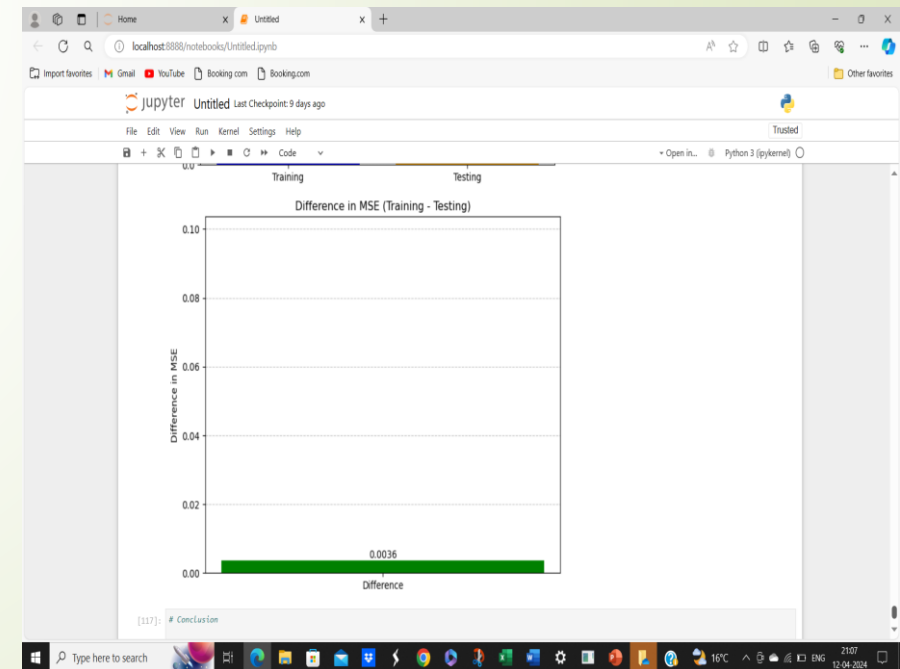
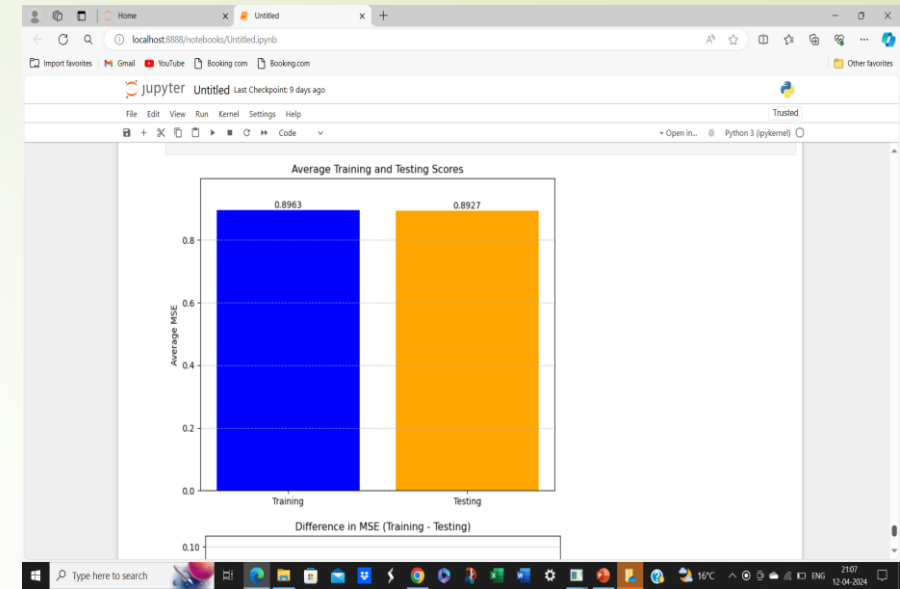
Findings of the Model


- ✓ The OLS regression results indicate that the model has an R-squared value of 0.896, indicating that approximately 89.6% of the variance in the target variable is explained by the independent variables included in the model.
- ✓ The coefficients for the independent variables represent the change in the target variable for a one-unit change in the respective independent variable, holding all other variables constant.
- ✓ The p-values associated with each coefficient indicate the statistical significance of the corresponding independent variable. Variables with p-values less than the chosen significance level (usually 0.05) are considered statistically significant
- ✓ The confidence intervals provide a range within which the true population parameter is likely to fall.
- ✓ The coefficients for `year_log`, `km_driven_log`, `engine_log`, and `max_power_log` are statistically significant at the 0.05 level, suggesting that these variables have a significant impact on the target variable.


- 
- ✓ The coefficients for the categorical variables (e.g., name_2, fuel_2, seller_type_2, transmission_2, owner_2, etc.) represent the change in the target variable relative to the reference category (usually the category with the lowest index)
 - ✓.Some categorical variables have coefficients with p-values less than 0.05, indicating that they are statistically significant predictors of the target variable
 - ✓.The Omnibus test, Durbin-Watson statistic, Jarque-Bera test, and Kurtosis provide information about the normality and autocorrelation of residuals
 - ✓ Based on these results, you can conclude that the model provides a good fit to the data and that several independent variables are statistically significant predictors of the target variable. However, further analysis and evaluation may be necessary to assess the model's predictive performance and identify any potential issues or areas for improvement.

Small Difference Between Training and Testing Scores:

- ❖ Based on the training score of 0.8963, the testing score of 0.8927, it seems that model is performing well and is likely achieving a balanced fit.
- ❖ The small difference between the training and testing scores indicates that the model's performance doesn't degrade significantly when applied to new data. This suggests that the model is not excessively complex, which could lead to overfitting, nor is it too simplistic, which could result in underfitting.



- 
- ❖ Based on the training score of 0.8963, the testing score of 0.8927,
 - ❖ and the small difference between them, it seems that your model is performing well and is likely achieving a balanced fit.
 - ❖ **Small Difference Between Training and Testing Scores:** The small difference between the training and testing scores indicates that the model's performance doesn't degrade significantly when applied to new data. This suggests that the model is not excessively complex, which could lead to overfitting, nor is it too simplistic, which could result in underfitting.




In conclusion, based on the evaluation of the model's performance, it appears that the multiple linear regression model is performing well and achieving a balanced fit. With both the training and testing scores being high and exhibiting a small difference between them, the model demonstrates the ability to capture underlying patterns in the data while generalizing well to new, unseen data.

Next steps could include:

1.Further Analysis and Interpretation: Dive deeper into the model's coefficients to understand the specific impact of each feature on the selling price of used cars. This analysis can provide valuable insights into factors driving pricing decisions in the market.

2.Model Refinement: Consider refining the model further by experimenting with different feature engineering techniques, exploring interactions between features, or trying more complex regression algorithms to potentially improve performance.



3. Validation and Sensitivity Analysis: Validate the model's performance on additional datasets or conduct sensitivity analysis to assess its robustness to variations in data and potential changes in market conditions.

4. Integration and Deployment: Integrate the trained model into your business operations, allowing it to assist in pricing decisions for used cars. Deploy the model into production, either as a standalone application or integrated into existing systems, to streamline pricing processes and enhance decision-making.