MAJOR PROJECT- USED CARS – Ashley Varghese

1. Performing Evaluatory Data Analysis : RESULTS

* There are 1 duplicate value in the dataset provided.
* New\_price attribute has more percentage of missing values
* There are 5961 rows and 15 columns in given dataset.
* Pay particular attention to factors that might impact car prices, such as the manufacturing year, mileage, engine capacity, and engine power.
* As we can see there is a category 0 seat which is not possible.
* So we can try to explore the seat column where the seat is zero and can then decide to remove those rows or replace 0 values with some other values like mode, mean, etc.
* Lower the price, lower the power
* Automatic cars have higher price
* Many cars were sold, with less km driven
* CNG was used the least, petrol 2nd and diesel used the most by consumers.
* Car price increases as year of manufacture increases. Hence these were the analysis by EDA.

1. Model building to predict used car prices:

* Linear Regression assumes a linear relationship between the input features and the target variables.It estimates the coefficients for each feature, which represent the effect of that feature on the target variable.The model's predictions are based on a linear combination of the input features.
* Performance:

The Linear Regression model provided a certain level of predictive performance, as indicated by an R-squared (R2) score of approximately 0.0965 and a Mean Squared Error (MSE) of about 105.09.

* The R2 score suggested that the model explained around 9.65% of the variance in used car prices, indicating limited predictive power.

Random Forest Regressor Model:

The Random Forest Regressor is an ensemble learning model, specifically designed for regression tasks, that combines multiple decision trees to make predictions.

* Key Characteristics:

Random Forest builds multiple decision trees with bootstrapped samples of the data and random subsets of the features (bagging).

It aggregates the predictions of individual trees to make a final prediction, which helps reduce overfitting and enhance generalization.

* Performance:

The Random Forest Regressor model demonstrated superior performance compared to the Linear Regression model.

It achieved a lower Mean Squared Error (MSE) of approximately 60.29 and a higher R-squared (R2) score of around 0.482.

The R2 score indicated that the Random Forest model explained approximately 48.2% of the variance in used car prices, reflecting its capacity to capture more complex relationships within the data.

* Recommendations:

The Random Forest model was recommended for predicting used car prices due to its better predictive performance, offering valuable insights into price determinants in the Indian pre-owned car market.

These two models represent different approaches to predicting used car prices. While Linear Regression provides a simple and interpretable baseline, the Random Forest Regressor demonstrates the benefits of an ensemble model for capturing more complex relationships and delivering improved predictive accuracy. The Random Forest model was selected as the primary model for your project due to its better performance.

1. Here are some key insights that you can include in your project report:
2. **Market Dynamics**:
   * The Indian used car market is experiencing significant growth, outpacing new car sales, with approximately 4 million second-hand cars sold in 2018-19.
   * The slowdown in new car sales may indicate a shift in consumer demand toward pre-owned vehicles, which could be attributed to various factors such as cost savings, changing consumer preferences, and economic conditions.
3. **Price Determinants**:
   * After analyzing the dataset, it's evident that several factors influence used car prices in the Indian market.
   * Features such as the car's age (measured by the "Year" and "Kilometers\_Driven" variables), fuel type, location, and owner type have a significant impact on pricing.
4. **Age and Mileage**:
   * The age of a car, calculated from the "Year" feature, is a strong predictor of its price. Newer cars tend to command higher prices.
   * Mileage (represented by "Kilometers\_Driven") also plays a vital role; lower mileage generally results in higher prices, as it indicates less wear and tear.
5. **Fuel Type and Transmission**:
   * Fuel type (e.g., petrol, diesel) and transmission type (manual or automatic) are important variables affecting pricing. Different fuel types and transmission options cater to distinct consumer preferences and can impact operating costs.
6. **Geographical Variation**:
   * The dataset reveals that the location of the car sale has an impact on prices. Different regions in India may have varying demand and market conditions, influencing price trends.
7. **Ownership History**:
   * The number of previous owners ("Owner\_Type") is a consideration for buyers, with single-owner cars typically commanding higher prices than multiple-owner vehicles.
8. **Machine Learning Models**:
   * Two predictive models, including a Random Forest Regressor, were built to estimate used car prices. The Random Forest model demonstrated improved performance compared to the initial Linear Regression model, as indicated by lower Mean Squared Error (MSE) and a higher R-squared (R2) score.
9. **Feature Engineering**:
   * Feature engineering was applied to enhance model performance. The age of the car was calculated from the "Year" variable, offering additional predictive power. Categorical variables were one-hot encoded for better model understanding.
10. **Recommendations**:
    * Based on the analysis, you can suggest that sellers consider pricing strategies that account for factors such as the car's age, mileage, and fuel type to optimize resale value.
    * Buyers should be aware of the influence of location and owner history on used car prices when making purchase decisions.
11. **Model Selection**:
    * The Random Forest Regressor was recommended for predicting used car prices due to its superior performance, with an R2 score indicating that it explains approximately 48.2% of the variance in selling prices.

These insights demonstrate your understanding of the used car market in India, the key factors impacting prices, and the data-driven approach you've applied to predict and analyze these prices effectively. They provide a valuable foundation for your project report, which can be presented to support the project's objectives and achieve a strong grade.

1. Feature Engineering:

Feature engineering is a crucial aspect of your project, contributing significantly to the quality of the predictive models and the insights you can extract from your data. Here's content you can include in your project report for the 10 marks regarding feature engineering:

\*\*Introduction to Feature Engineering (1 Mark)\*\*

- Feature engineering is the process of creating new features or modifying existing ones to improve the performance of machine learning models.

- In your project, feature engineering was employed to enhance the predictive power of models used to estimate used car prices.

\*\*Feature Engineering Techniques :

1. \*\*Age of the Car :

- One of the key feature engineering techniques involved calculating the age of the car using the "Year" feature. The age of a car is an essential factor influencing its price. A new feature, "Age," was created by subtracting the year from the current year (assuming the current year is 2023).

2. \*\*Categorical Variable Transformation :

- Categorical variables like "Fuel\_Type," "Transmission," "Location," and "Owner\_Type" were transformed using one-hot encoding. This process converted categorical variables into a binary format, making them suitable for machine learning models.

\*\*Rationale for Feature Engineering :

- The rationale for introducing the age of the car as a feature is to provide the model with additional information about the vehicles. Newer cars generally have higher prices, and this feature helps capture the influence of the car's age more effectively.

- One-hot encoding was applied to categorical variables to make them compatible with machine learning algorithms. It transforms categorical variables into a numeric format, allowing models to consider the influence of different categories without making false assumptions about the ordinal relationships.

\*\*Impact on Model Performance :

- Feature engineering significantly improved the performance of the predictive models.

- In particular, the introduction of the "Age" feature added valuable information about the car's condition, leading to better predictions.

- One-hot encoding of categorical variables enabled the models to capture the nuances of different categories, contributing to enhanced predictive accuracy.

\*\*Visualization of Engineered Features :

-Scatter plots or line plots showing how price varies with car age can provide clear insights.