Car_Dekho_Data Prediction

Objective:

The objective of this project is to develop a model that can predict the selling price of a new car. The target variable or dependent variable in this project is the "Selling Price," which represents the price at which a car is expected to be sold.

To achieve this objective, we will use several predictors or independent variables that can potentially influence the selling price of a car. These predictors include:

1. Transmission: This variable indicates the type of transmission system in the car, such as manual or automatic. The transmission type may have an impact on the car's selling price.
2. Fuel Type: This variable represents the type of fuel used by the car, such as petrol, diesel, or electric. The choice of fuel type can affect the car's price due to variations in fuel prices and demand.
3. Car Name: This variable identifies the specific make and model of the car. Different car brands and models have different price ranges based on factors such as brand reputation, features, and market demand.
4. Owner: This variable indicates the number of previous owners of the car. Generally, cars with fewer owners may have a higher selling price due to their perceived better condition and maintenance history.
5. Year: This variable represents the manufacturing year of the car. The age of the car can influence its selling price, with newer cars typically having higher prices.

By analyzing these predictors, we aim to build a Linear Regression model that can learn the relationship between these variables and the selling price of cars. The model will then be used to predict the selling price of a new car based on the given predictor values.

The project involves several steps, including data collection, data preprocessing, exploratory data analysis, feature engineering, model training and evaluation, and finally, using the trained model to predict the selling price of new cars.

The ultimate goal is to develop an accurate and reliable predictive model that can assist car sellers, buyers, and dealerships in estimating the selling price of new cars based on various influential factors, enabling better decision-making and pricing strategies in the automotive market.

Database Used:

In this project, we utilize the power of MongoDB as a NoSQL database. MongoDB is a highly flexible and scalable database solution that enables us to store and manage large volumes of data with dynamic schemas. It offers a document-based model, allowing us to work with data in a more natural and intuitive manner.

To enhance our data processing capabilities, we integrate MongoDB with Apache Spark. Apache Spark is a robust and distributed data processing framework that enables high-performance analytics on large datasets. It provides a wide range of functionalities for data manipulation, transformation, and analysis.

By leveraging the built-in connectors and libraries provided by Spark, we seamlessly integrate MongoDB into our data pipeline. These connectors facilitate the seamless exchange of data between Spark and MongoDB. We can effortlessly read data from MongoDB collections directly into Spark, enabling us to perform complex analytics and transformations on the data.
Furthermore, Spark's distributed computing architecture enables us to scale our analytics operations efficiently, making it suitable for handling large volumes of data. This scalability allows us to process and analyze the dataset effectively, even in environments with high data velocity and variety.
In this project, we also employ Spark for exploratory data analysis (EDA). EDA is a critical phase in data analysis, where we gain insights and understand the underlying patterns and relationships within the dataset. With Spark's advanced analytical capabilities, we can perform comprehensive univariate and bivariate analysis on the data. This includes various statistical measures, visualizations, and advanced techniques to uncover meaningful insights and patterns.
Overall, the integration of MongoDB with Apache Spark in this project empowers us to efficiently manage, process, and analyze large and dynamic datasets. By leveraging Spark's distributed computing capabilities and MongoDB's flexibility, we can perform advanced analytics, gain valuable insights, and make data-driven decisions effectively.

Model Used:

After applying the Linear Regression model using the Spark distributed data processing framework, we obtained an R2 score of 0.45. An R2 score of 0.45 indicates that our model explains only 45% of the variance in the target variable, which implies that the model is not a good fit for the data.

Conclusion:

In conclusion, our objective in this Linear Regression project was to develop a model that predicts the selling price of new cars based on various predictors such as transmission, fuel type, car name, owner, and year. However, our initial model, implemented using the Spark distributed data processing framework, yielded an R2 score of 0.45, indicating that it explains only 45% of the variance in the target variable.

While this R2 score falls short of our desired level of predictive accuracy, it presents an opportunity for further investigation and improvement. By delving deeper into the analysis of the regression results, conducting residual analysis, assessing feature importance, and evaluating assumptions, we can gain valuable insights into the limitations of our model and its potential areas of enhancement.

Additionally, we can explore alternative approaches to feature engineering, consider regularization techniques, and experiment with different regression algorithms to potentially improve the model's performance. Cross-validation can also be employed to assess the generalizability of the model and validate its predictive ability.

Ultimately, the goal is to develop a robust and accurate predictive model that accurately estimates the selling price of new cars. By iteratively refining our approach, addressing model limitations, and incorporating advanced techniques, we can strive to achieve a better fit for the data and provide valuable insights for car sellers, buyers, and dealerships in the automotive market.