Praxis
Business School
CELEBRATE YOUR WORTH

REVVING UP SALES: PREDICTING CAR PRICES WITH MONGODB AND SPARK

Rupesh Kumar Sahu-C23024

CONTENT

01 EXECUTIVE SUMMARY

02 EDA

03 ANALYSIS INTERPRETATION

04 CORRELATION ANALYSIS

05 DATA PREPARATION

06 PIPELINE CREATION AND DATA NORMALIZATION

07 MODEL BUILDING

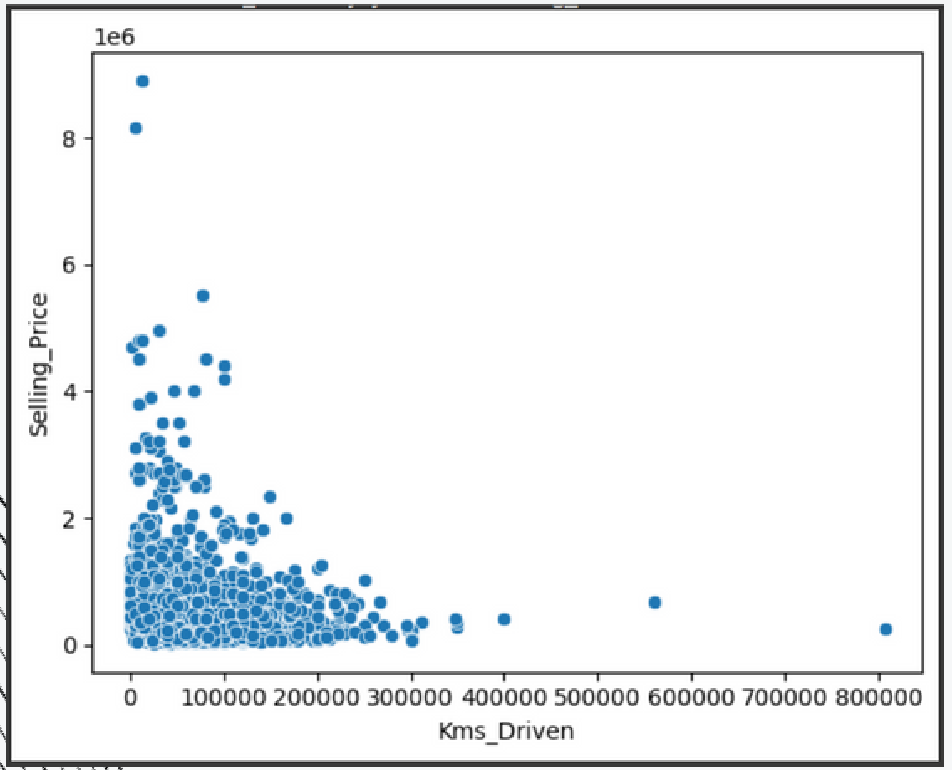
08 OUTPUT

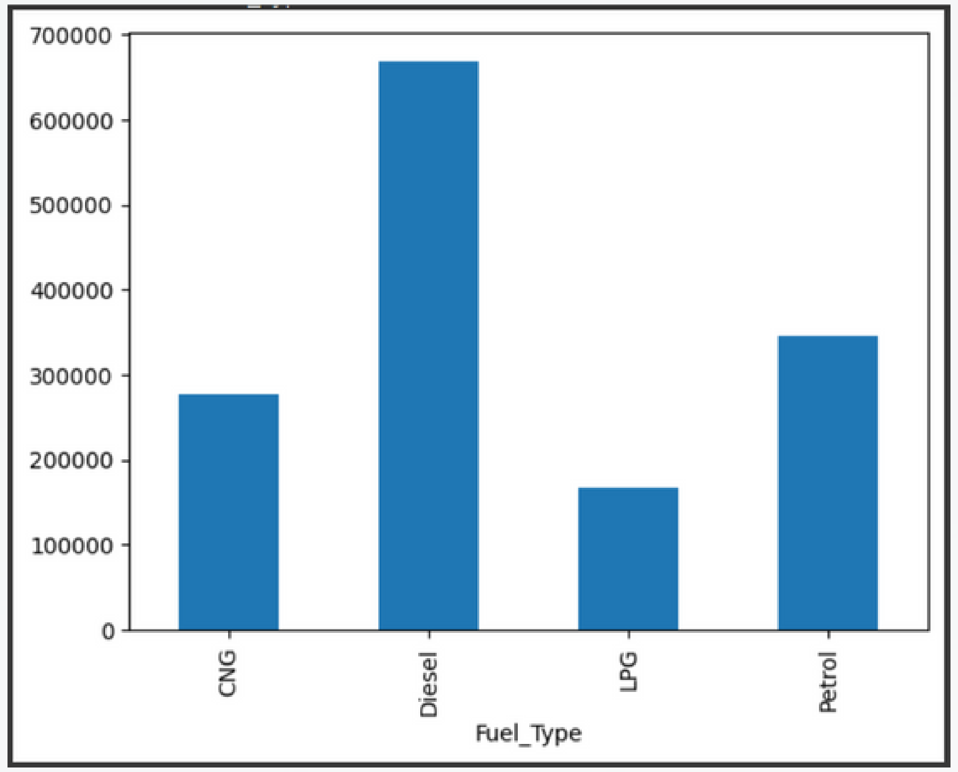
EXECUTIVE SUMMARY

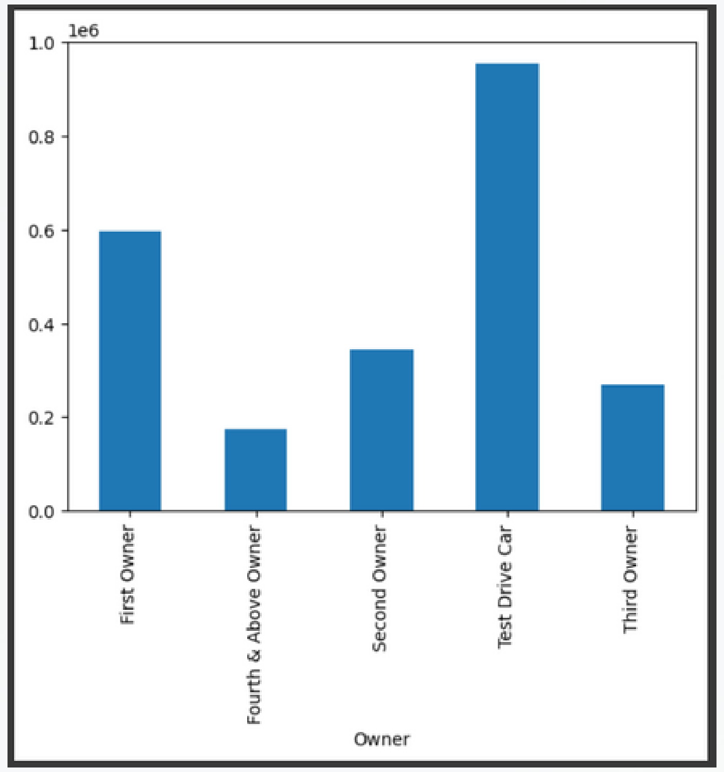
In this Linear Regression project, our objective was to develop a model using the Spark distributed data processing framework to predict the selling price of new cars based on predictors such as transmission type, fuel type, car name, owner, and manufacturing year. The integration of MongoDB with Spark provided efficient data management and analysis capabilities. However, the initial model achieved an R2 score of 0.45, indicating room for improvement. Further investigations, including residual analysis, feature importance assessment, and evaluation of assumptions, were recommended to refine the model. Exploring alternative regression algorithms, feature engineering techniques, and regularization methods, as well as employing cross-validation, were suggested to enhance the model's performance. Despite the current limitations, this project holds the potential to aid car sellers, buyers, and dealerships in estimating car selling prices, contributing to better decision-making in the automotive market.

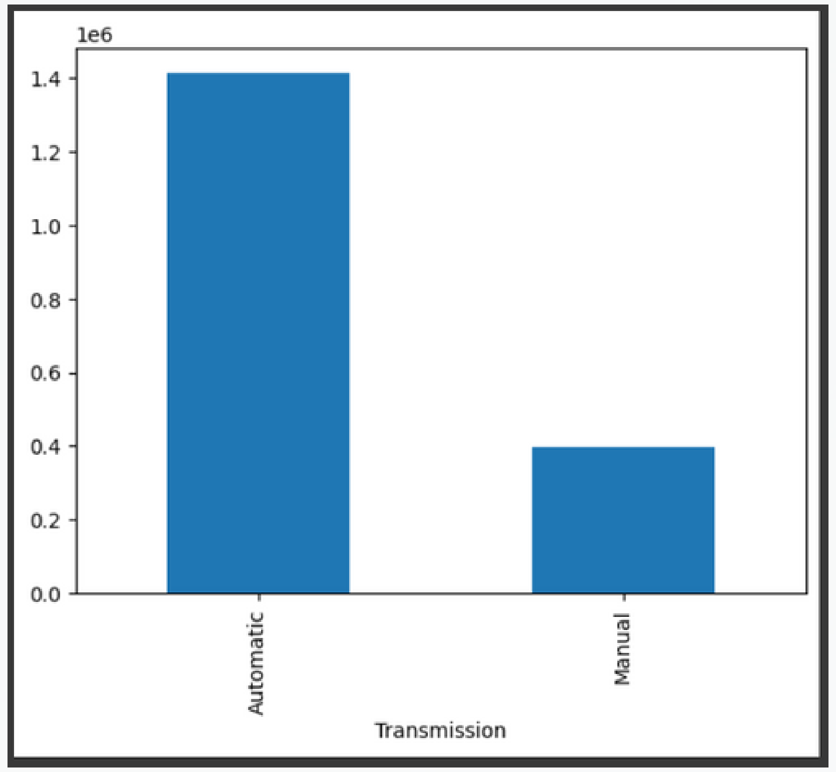
EXPLORATORY DATA ANALYSIS

EDA involved utilizing data visualization charts and plots for univariate and multivariate analysis. By visually examining variables such as transmission type, fuel type, car name, owner, and manufacturing year, we sought to uncover patterns, outliers, and relationships within the data, gaining valuable insights to guide further steps in the project.









ANALYSIS INTERPRETATION

- Based on the bivariate analysis, it can be observed that there is a negative correlation between the number of kilometers driven and the selling price of cars, indicating that as the distance increases, the selling price tends to decrease.
- Diesel cars have the highest selling price, followed by petrol, CNG, and LPG.

- In terms of ownership, test cars have the highest selling price, followed by first owners, second owners, third owners, and fourth owners.

- Automatic cars generally command a higher selling price compared to manual cars.

Praxis Business School
CELEBRATE YOUR WORTH

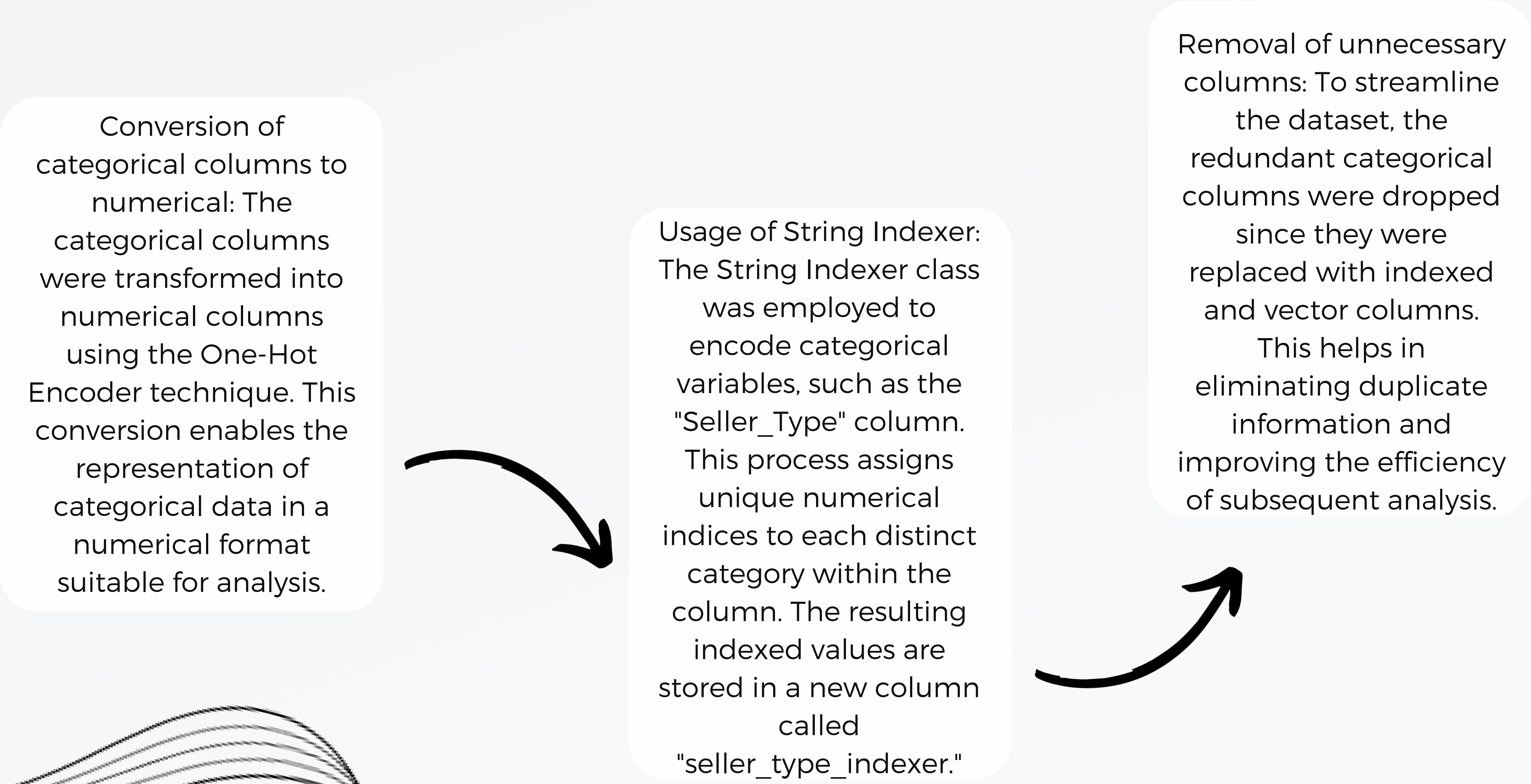
CORRELATION ANALYSIS

There is a negative correlation of -0.19 between the number of kilometers driven and the selling price. This suggests that as the kilometers driven increase, the selling price tends to decrease.

There is a negative correlation of -0.41 between the year of the car and the selling price. This indicates that as the car gets older (higher number of years), the selling price tends to decrease.

DATA PREPARATION

In the process of creating dummy variables, the following steps were performed:



Conversion of categorical columns to numerical: The categorical columns were transformed into numerical columns using the One-Hot Encoder technique. This conversion enables the representation of categorical data in a numerical format suitable for analysis.

Usage of String Indexer: The String Indexer class was employed to encode categorical variables, such as the "Seller_Type" column. This process assigns unique numerical indices to each distinct category within the column. The resulting indexed values are stored in a new column called "seller_type_indexer."

Removal of unnecessary columns: To streamline the dataset, the redundant categorical columns were dropped since they were replaced with indexed and vector columns. This helps in eliminating duplicate information and improving the efficiency of subsequent analysis.

Praxis Business School
CELEBRATE YOUR WORTH

PIPELINE CREATION AND NORMALIZING THE DATA

The creation of pipeline stages involves setting up a pipeline comprising two stages: a type indexer and a type encoder. These stages use transformers, namely the Type_Indexer and Type_Encoder, to preprocess the dataset. Once the pipeline is defined, the fit() method is applied to the pipeline object using the new_data dataset as input. This trains the pipeline and produces a fitted pipeline (pipeline_model) that can be utilized to transform new data.

In addition, a Standard Scaler is employed to scale the features within a consistent range. The Standard Scaler ensures that each value is scaled to a range between 0 and 1, enabling fair comparisons and reducing the impact of varying feature magnitudes.

MODEL BUILDING

- The train-test split involves dividing the scaled_df dataset into two separate datasets: the training dataset and the test dataset.
- This split is achieved using the randomSplit() method, which takes two parameters: weights and seed.

- The weights parameter determines the relative sizes of the resulting datasets, while the seed parameter is optional and used for reproducibility purposes.
- In this case, the training dataset is allocated 70% of the data, while the test dataset receives 30% of the data.

- The seed is set to 1234 to ensure consistent results. After the split, the training dataset contains 3098 records, while the test dataset contains 1236 records.

OUTPUT

- The training data is fitted into the linear regression model, and the test data is transformed using the same model for prediction
- The linear model coefficients and intercept are calculated to interpret the model, make predictions, and evaluate its performance.
- However, the obtained R-squared value of 0.45 indicates that the model's fit is not satisfactory. Further investigation is required to thoroughly evaluate the model and explore potential improvements.

Praxis
Business School
CELEBRATE YOUR WORTH

THANK YOU