Praxis Business School

CELEBRATE YOUR WORTH

PREDICTING CAR PRICES WITH MONGODB & SPARK

SANGITA KUNDU

EXECUTIVE SUMMARY

In summary, the objective of this project was to develop a predictive model using the Spark distributed data processing framework to estimate the selling prices of cars based on predictors such as transmission type, fuel type, car name, owner, and manufacturing year. However, our initial model's performance fell short, as indicated by an R^2 score of 0.45, suggesting room for improvement.

To refine the model, we conducted in-depth analysis, including residual analysis and feature importance assessment, to gain insights into the model's shortcomings and identify influential predictors. We recommended exploring alternative regression algorithms and incorporating advanced feature engineering techniques to enhance the model's predictive accuracy. Additionally, regularization methods were considered to address overfitting and improve generalizability.

By iteratively improving the model and incorporating these advanced techniques, we aim to develop a robust and accurate tool for estimating car selling prices. The insights gained from this project have the potential to assist car sellers, buyers, and dealerships in making informed decisions in the dynamic automotive market, empowering them to price their cars more accurately and optimize their strategies.

EXPLORATORY DATA ANALYSIS

EDA involved visualizing variables like transmission, fuel type, car name, owner, and manufacturing year using charts and plots. This helped uncover patterns, outliers, and relationships in the data, providing valuable insights.

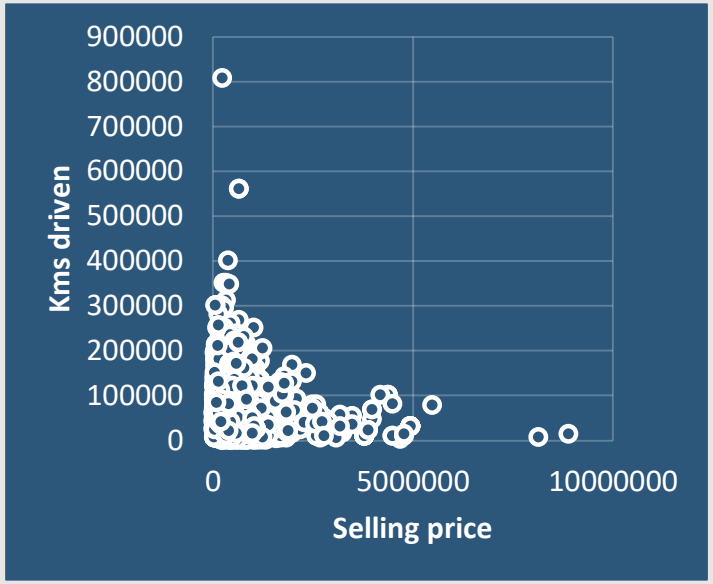


Chart 1:
Scatter plot for selling price vs kms driven

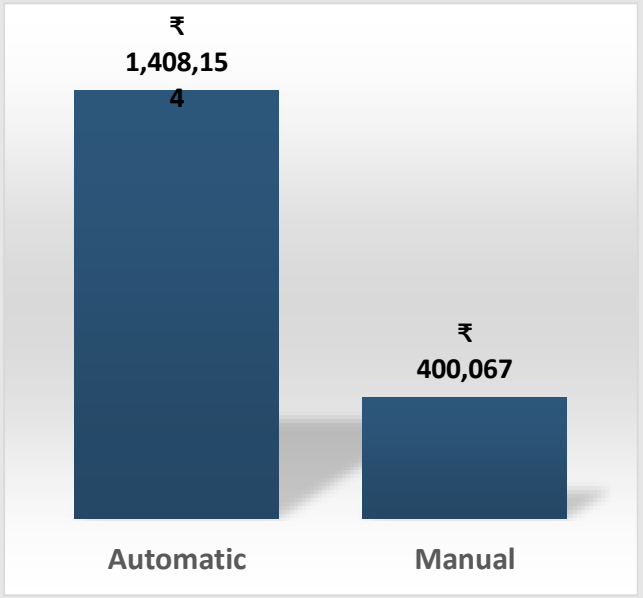


Chart 2:
Categorization based on fuel type vs average of selling price

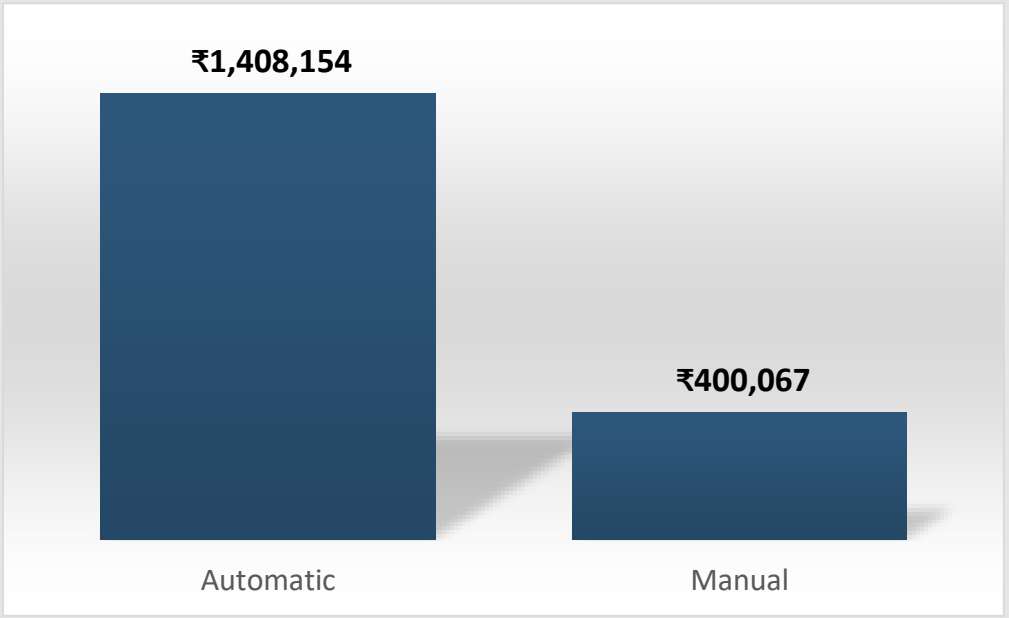
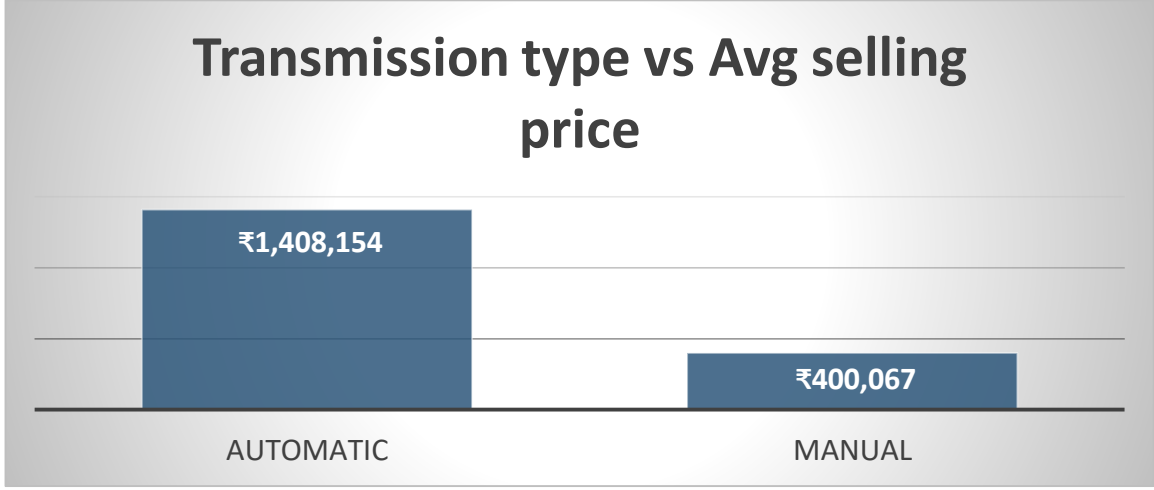


Chart 3:
Categorization based on ownership

Praxis Business School
CELEBRATE YOUR WORTH

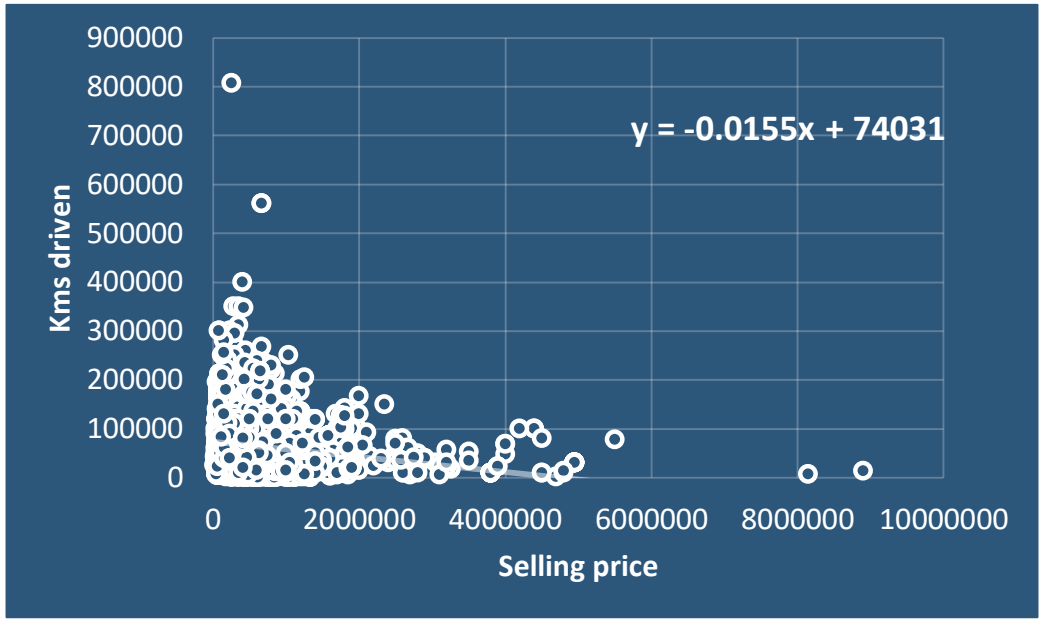
ANALYSIS INTERPRETATION



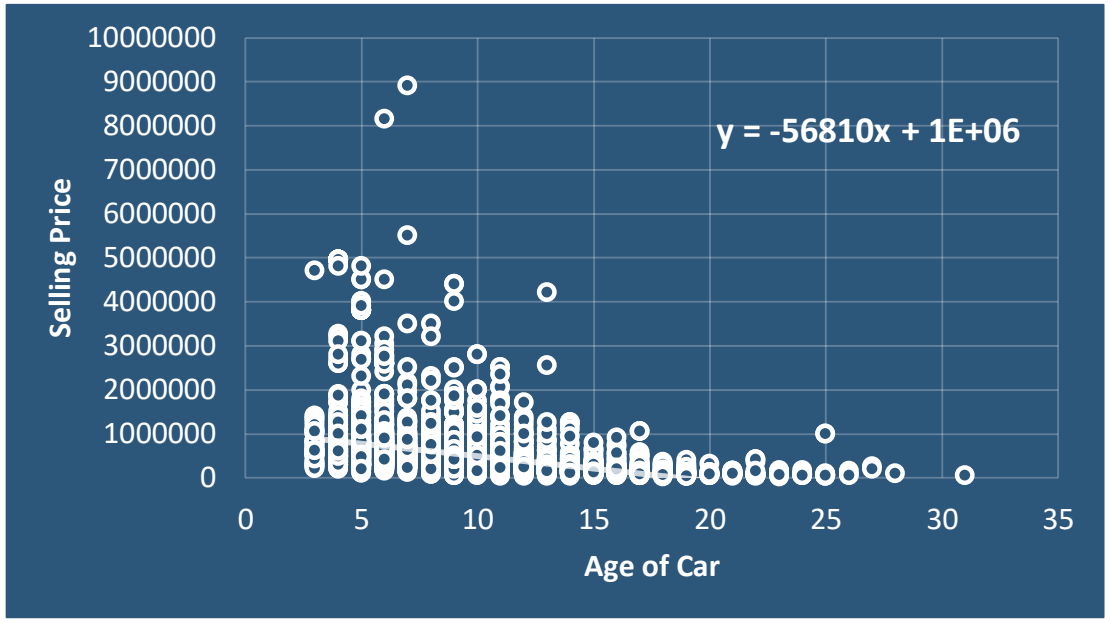
- Bivariate analysis revealed a negative correlation between kilometers driven and selling price, suggesting that as distance increases, the selling price tends to decrease.
- Diesel cars exhibited the highest selling prices, followed by petrol, CNG, and LPG.
- Test cars have the highest selling price, followed by first owners, second owners, third owners, and fourth owners.
- Automatic cars generally have 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.

Praxis Business School
CELEBRATE YOUR WORTH

DATA PREPARATION

In creating dummy variables, we followed the following steps:

Conversion of categorical columns to numerical:

Categorical columns were transformed using One-Hot Encoder to represent the categorical data in a numerical format for analysis.

String Indexer Usage:

We utilized the String Indexer class to encode categorical variables, such as the "Seller_Type" column. This process assigned unique numerical indices to each distinct category within the column. The resulting indexed values were stored in a new column called "seller_type_indexer."

Removal of unnecessary columns:

To optimize the dataset, redundant categorical columns were dropped as they were replaced with indexed and vector columns. This step eliminated duplicate information and enhanced the efficiency of subsequent analysis.

Praxis Business School
CELEBRATE YOUR WORTH

PIPELINE CREATION & NORMALIZING THE DATA

Pipeline stages creation:

The pipeline consists of two stages: type indexer and type encoder. These stages utilize transformers, namely Type_Indexer and Type_Encoder, for dataset preprocessing. By applying the fit() method to the pipeline object with the new_data dataset, the pipeline is trained and generates a fitted pipeline (pipeline_model) for transforming new data.

Standard Scaler Usage:

To ensure fair comparisons and reduce the impact of varying feature magnitudes, a Standard Scaler was applied to scale the features within a consistent range. This scaler transformed each value to a range between 0 and 1, enabling standardized comparisons across features.

Praxis Business School
CELEBRATE YOUR WORTH

MODEL BUILDING

The train-test split involves:

- Dividing the scaled_df dataset into two separate datasets: the training dataset and the test dataset.

- Utilizing the randomSplit() method, which takes two parameters: weights and seed, to achieve the split.

- The weights parameter determines the relative sizes of the resulting datasets.

- The seed parameter, although optional, is 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 and reproducible results.

After the split, the training dataset consists of 3098 records, while the test dataset consists of 1236 records.

OUTPUT

The training data is used to fit the linear regression model, and the test data is transformed using the same model for prediction.

The model's coefficients and intercept are calculated to interpret the model, make predictions, and evaluate its performance. However, the obtained R-squared value of 0.45 suggests 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

Praxis Business School

CELEBRATE YOUR WORTH

THANK YOU!