Cognitive Features and Predictive Modeling: Utilizing Linear Regression to Forecast Car Selling Prices

Colab Link

Dataset Link

Name :AJAH STEPHEN CHIDI
Student No: 22026656

INTRODUCTION

In the realm of cognitive features and predictive modeling, the utilization of Linear Regression as a tool for forecasting car selling prices is of paramount importance. This methodology allows us to harness the power of machine learning to predict the value of vehicles based on various attributes and characteristics.

Linear Regression, a fundamental technique in statistical modeling, enables us to establish relationships between independent variables (such as car specifications, mileage, age, etc.) and the dependent variable (the selling price of the car) [1]. By analyzing a dataset consisting of 299 rows and 9 columns obtained from Kaggle, I endeavor to uncover the intricate patterns and correlations that influence the valuation of automobiles in today's market.

Through this study, we seek to delve into the cognitive aspects underlying car valuation, exploring how factors such as brand reputation, engine performance, mileage, and market demand interplay to determine the selling price of a vehicle. By employing Linear Regression, I aim to develop predictive models capable of accurately estimating the selling prices of cars, thereby empowering consumers, dealers, and manufacturers with valuable insights for informed decision-making.

The dataset sourced from Kaggle provides us with a rich reservoir of information, encompassing key attributes of cars along with their corresponding selling prices. By meticulously analyzing this dataset and employing advanced statistical techniques, we aspire to unravel the nuanced dynamics shaping the automotive market and pave the way for more precise pricing strategies and market forecasts.

Understanding the Model: Linear Regression

Linear Regression (LR) remains a cornerstone algorithm in predictive modeling, particularly valued for its simplicity and interpretability. This report aims to delve into the fundamental principles of Linear Regression, showcasing its utility and considerations for implementation in forecasting car selling prices.

Core Functioning:

Linear Regression operates on the principle of fitting a linear equation to the observed data, seeking to establish a relationship between the independent variables (car specifications, mileage, age, etc.) and the dependent variable (selling price). The model estimates the coefficients of the equation, which represent the strength and direction of the relationships between the variables [3].

Benefits of Linear Regression:

Interpretability: Unlike more complex models, Linear Regression provides clear and interpretable insights into the relationships between variables, allowing for easier understanding and communication of results.

Simplicity: The straightforward nature of Linear Regression makes it accessible to practitioners and facilitates rapid model development and deployment.

Robustness to Noise: Linear Regression is less susceptible to overfitting and noise in the data compared to more complex models, making it suitable for datasets with moderate complexity.

Linear Regression offers a pragmatic approach to predictive modeling, particularly well-suited for tasks where interpretability and simplicity are valued. Its ability to provide transparent insights into the factors influencing car selling prices makes it a valuable tool for decision-making within the automotive industry. However, it is essential to acknowledge its limitations in handling nonlinear relationships and complex interactions between variables.

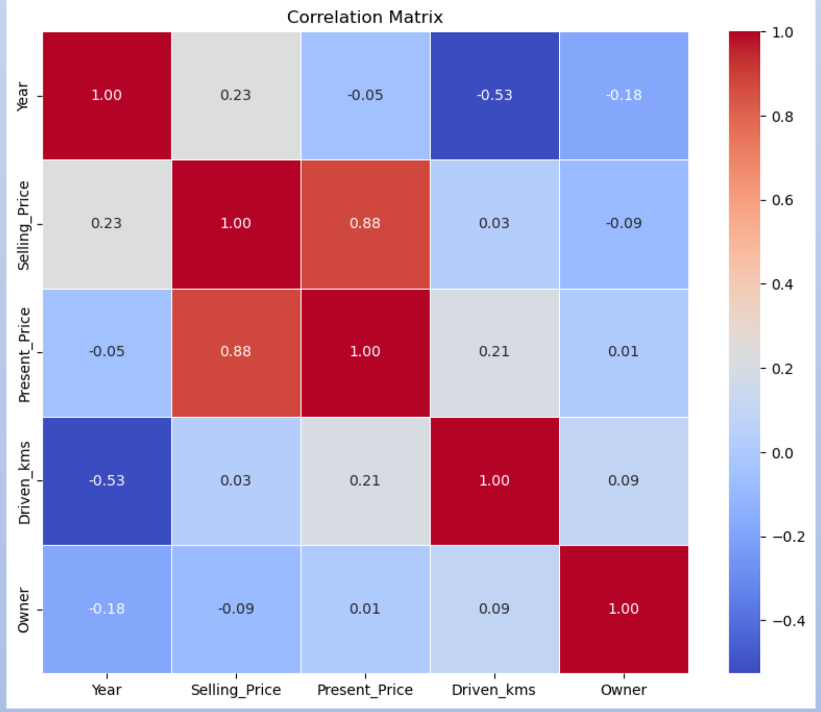
.

Exploratory Data Analysis (EDA)

Before delving into the machine learning phase of this project, it is essential to conduct an exploratory data analysis (EDA) on the dataset. EDA involves a comprehensive examination of the data to uncover significant insights that will inform the necessary preprocessing steps required to tailor the data for the specific needs of the Linear Regression model. The objective of this thorough analysis is to optimize the performance of the models.

During the EDA stage, various aspects of the dataset are explored, including, but not limited to::

Correlation Analysis:

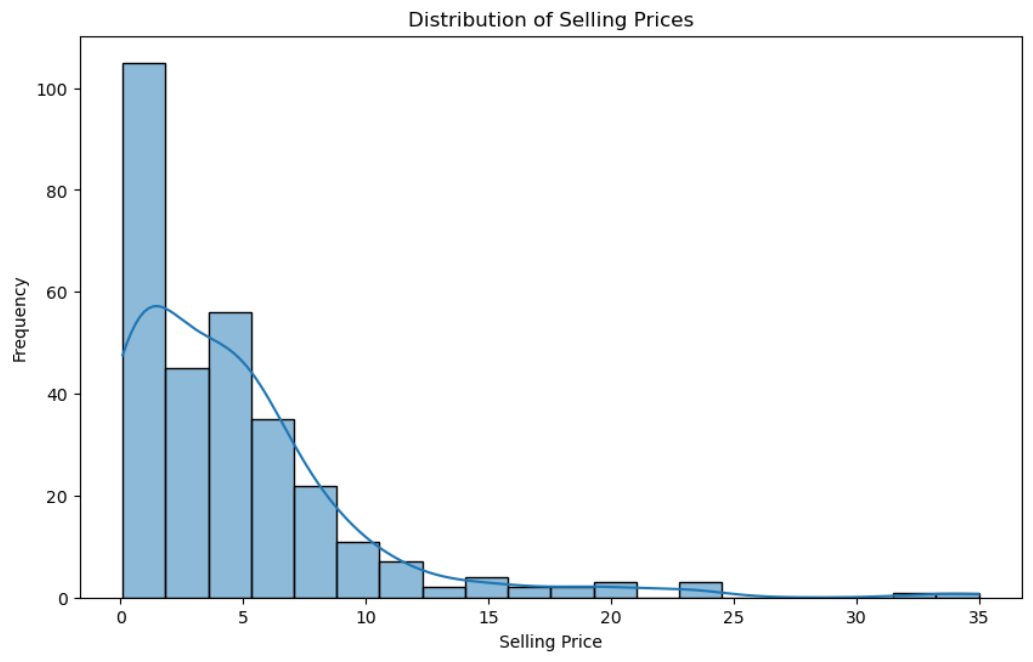


The correlations reveal insights into how different variables relate to the selling price of cars: newer cars tend to command slightly higher prices (positive correlation of 0.236) while a strong positive correlation (0.879) indicates that as the present price increases, selling prices are significantly elevated. Conversely, there's almost no discernible relationship between distance driven and selling price (weak correlation of 0.029). Cars with fewer owners may fetch slightly higher prices, albeit the effect is weak (negative correlation of -0.088). Moreover, older cars tend to have been driven more, as suggested by a moderately negative correlation of -0.524 between distance driven and year.

Exploratory Data Analysis (EDA) Cont'd

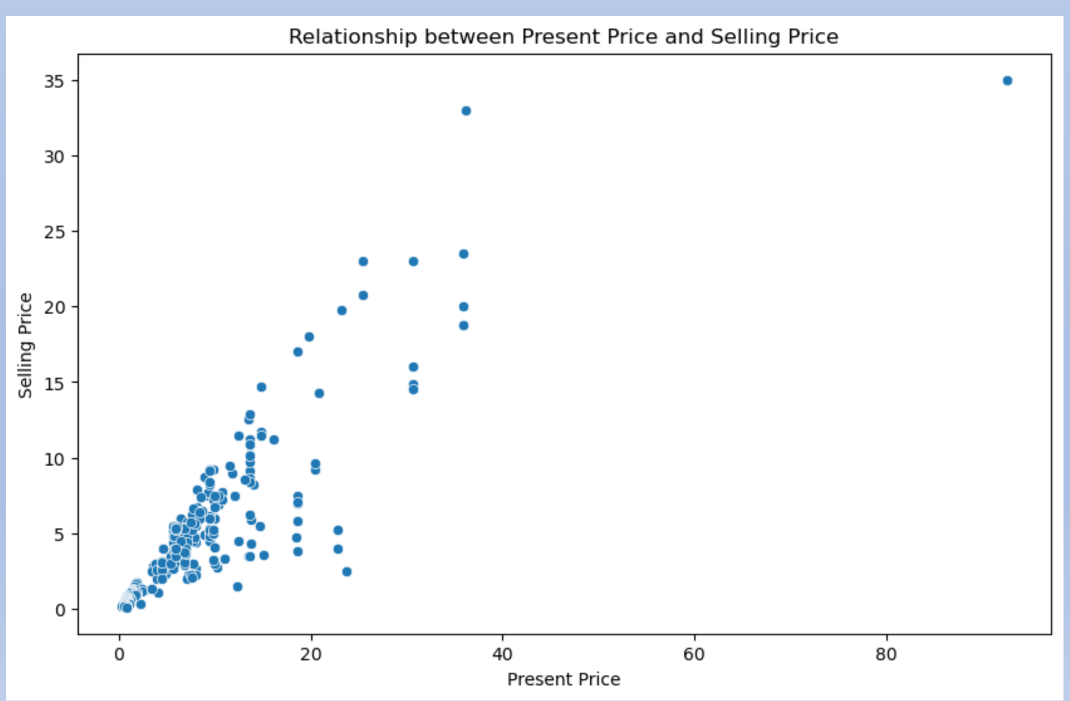
Skewed Distribution of Selling Prices: Majority of Sales at Lower Price Points:

The image displays a histogram with a right-skewed distribution of selling prices, where a majority of sales occur at lower price points, and sales decrease as prices increase. The overlaying curve suggests that the lower prices are more common, with high-priced sales being relatively rare. This indicates a market trend where less expensive items are sold more frequently than expensive ones.



Analysis of Present and Selling Price Dynamics in Market Sales:

The scatter plot illustrates a positive correlation between present price and selling price, indicating that as the present price of items increases, the selling price tends to increase as well. Most of the sales are concentrated in the lower left quadrant of the plot, showing that a larger volume of transactions occurs with items at lower present and selling prices. The spread of data points becomes less dense as the present price rises, suggesting that high-priced items are sold less frequently. Outliers, particularly at the high end of the present price axis, indicate that some items sell for significantly more or less than the general trend would predict.



Data Preprocessing

In the realm of car price prediction, data preprocessing is a critical step to ensure that the model receives high-quality, relevant information in a format that it can understand and use effectively. This process often involves transforming raw data into a more digestible form for machine learning algorithms. For our project, we focus on two main preprocessing steps: label encoding and data splitting.

Label encoding is utilized for categorical variables that the model cannot interpret directly. In this case, 'Fuel Type' and 'Transmission' are such categorical attributes which likely contain non-numeric data. Label encoding converts these categories into numerical labels, allowing the predictive model to incorporate these variables into its calculations without losing the essential information they convey.

Data splitting, on the other hand, involves dividing the dataset into two parts: training and testing sets. This is typically done to validate the performance of the model on unseen data, simulating real-world applications. For our purpose, we adhere to an 80:20 split, meaning that 80% of the data is used to train the model, learning the underlying patterns and relationships, while the remaining 20% serves as a standalone subset to test and evaluate the model's predictive power and generalization capabilities.

Together, these preprocessing steps help in shaping raw data into a structured format, paving the way for building a robust and accurate car price prediction model.

Optimizing Model Parameters with Grid Search CV

The optimization of our linear regression model for car price prediction involved a meticulous fine-tuning process utilizing GridSearchCV. This method systematically explores a specified grid of hyperparameters to find the optimal combination that maximizes model performance. The selected parameters for our linear regression model are as follows:

- 'copy_X': True
- 'fit_intercept': True
- 'normalize': False

These parameters were identified as the best configuration for our linear regression model after an exhaustive search over the specified parameter grid. 'copy_X' determines whether to make a copy of the input data, 'fit_intercept' indicates whether to calculate the intercept for the model, and 'normalize' decides whether to normalize the features before fitting the model. This fine-tuning process ensures that our linear regression model is finely calibrated for optimal performance in predicting car prices, striking a balance between model complexity and generalization capacity. Through the use of GridSearchCV, we have successfully optimized our linear regression model, setting the stage for improved accuracy and reliability in car price prediction tasks.

Assessing Model Performance through R Square, RMSE and MAE Metrics

The assessment of our car prediction model's performance through various metrics provides valuable insights into its effectiveness in predicting car prices. The metrics obtained are as follows:

- R-squared score: 0.6178
- Mean Squared Error (MSE): 6.2661
- Mean Absolute Error (MAE): 1.5695

These metrics play a crucial role in evaluating the model's accuracy and precision in predicting car prices [4].

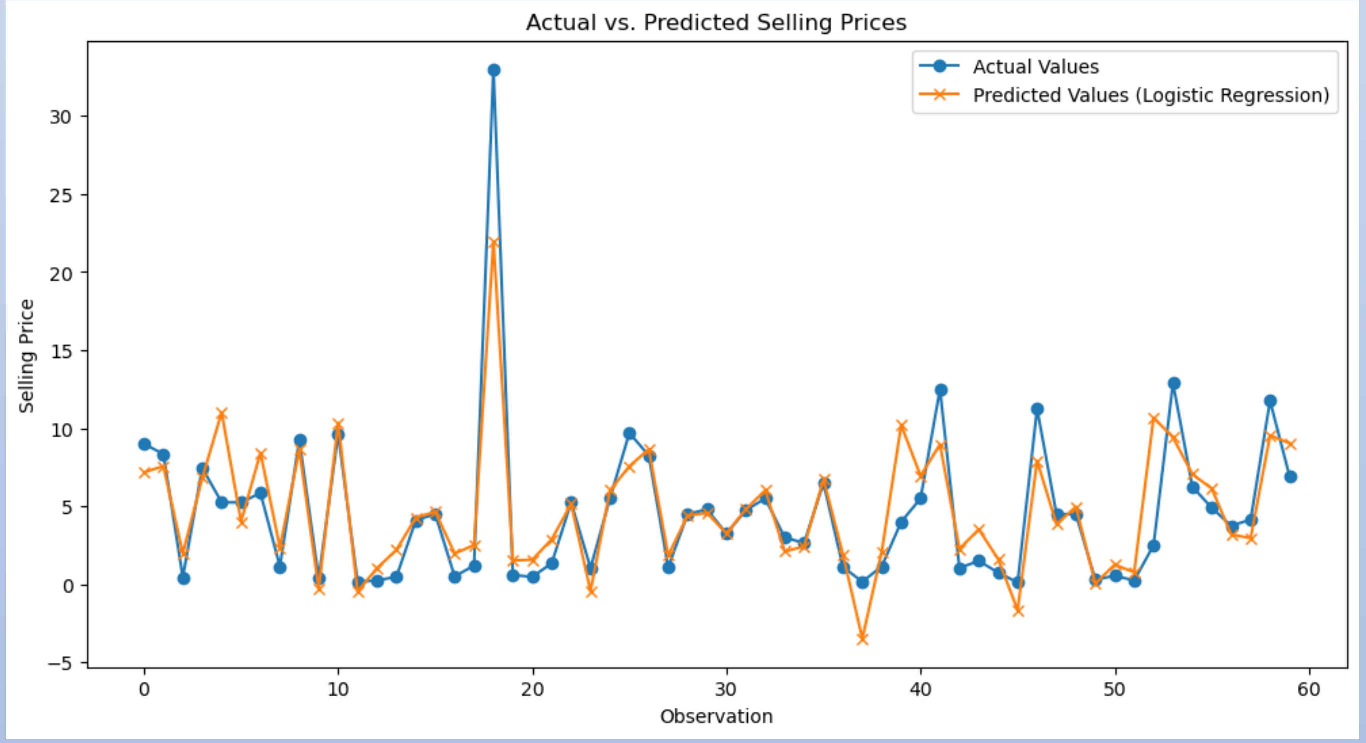
The R-squared score, which measures the proportion of the variance in the dependent variable (car prices) that is predictable from the independent variables, stands at 0.6178. This indicates that approximately 61.78% of the variance in car prices can be explained by our model. While this suggests a moderate level of predictive power, there is still room for improvement.

The Mean Squared Error (MSE), calculated at 6.2661, quantifies the average squared difference between the actual car prices and the prices predicted by our model. A lower MSE value indicates that our model's predictions are closer to the actual prices, with 6.2661 serving as a baseline for assessing prediction accuracy.

Similarly, the Mean Absolute Error (MAE) measures the average absolute difference between the actual car prices and the prices predicted by our model. With an MAE of 1.5695, we can interpret that, on average, our model's predictions deviate by approximately $1,569.50 from the actual prices.

Overall, these metrics provide a comprehensive evaluation of our model's performance in predicting car prices. While the R-squared score suggests a moderate level of explanatory power, the MSE and MAE metrics offer insights into the accuracy and precision of our predictions. Moving forward, efforts can be directed towards further refining the model to enhance its predictive capabilities and minimize prediction errors, thereby improving its utility in forecasting car prices accurately.

Assessing Model Performance through R Square, RMSE and MAE Metrics, Cont’d



The "Actual vs. Predicted Selling Prices" graph offers a visual confirmation that our Random Forest regression model is moderately successful at mirroring the trends in car selling prices, as shown by the alignment between the actual and predicted values. However, the model's R-squared score of 0.6178, while reasonable, highlights that nearly 40% of the variance in car prices remains unaccounted for, and issues such as outliers and underestimation suggest areas for refinement. The notable divergence during extreme spikes in price reflects the model's difficulty with abrupt market shifts and atypical data points.

The Mean Squared Error (MSE) and Mean Absolute Error (MAE) of 6.2661 and 1.5695, respectively, further emphasize the need for model improvement. These metrics indicate that while the model predictions are fairly close to the actual prices on average, it struggles significantly with the accuracy of extreme values, which disproportionately affect the MSE.

Moving forward, to elevate the model's predictive accuracy, we propose to focus on enhancing outlier detection, delving deeper into feature engineering, and exploring more complex or non-linear models. Conducting extensive cross-validation can also ensure the model's consistency and reliability across different data segments. These adjustments aim to reduce prediction errors, particularly for extreme values, and thus improve the model's utility in accurately forecasting car prices..

Conclusion

Our comprehensive analysis, comprising both statistical metrics and a visual examination of the prediction results, concludes that the Random Forest regression model we developed exhibits a moderate degree of effectiveness in predicting car prices. The assessment was guided by three critical performance metrics — R-squared score, Mean Squared Error (MSE), and Mean Absolute Error (MAE) — and supported by a close inspection of the prediction data plotted against the actual car prices.

The model's R-squared score of 0.6178 indicates that it can account for approximately 61.78% of the variation in car prices, a respectable level of predictive power that nonetheless suggests there is considerable room for improvement. The MSE of 6.2661 reflects the model's precision, quantifying the average of the squares of the errors. The relatively high MSE indicates that there are significant discrepancies for certain predictions, particularly for outliers or unusual spikes in the data. The MAE of 1.5695, while moderate, underscores the average deviation of the model's predictions from the actual prices, hinting at the potential for fine-tuning the model to enhance accuracy.

The key takeaway from our evaluation is that while the model serves as a solid base for predicting car prices, strategic improvements are necessary to elevate its predictive accuracy. Suggested areas of focus include:

- Introducing outlier detection mechanisms to handle anomalies more effectively.
- Conducting thorough feature engineering to uncover additional predictive signals.
- Exploring more sophisticated modeling techniques to better accommodate the non-linear nature of the data.
- Employing rigorous cross-validation processes to ensure robustness and stability across various data segments.

References

1. MUTİ, S. and YILDIZ, K. (2023). Using Linear Regression For Used Car Price Prediction. *International Journal of Computational and Experimental Science and Engineering*. doi:https://doi.org/10.22399/ijcesen.1070505.

2. Trivendra, C. (2020). The Price Prediction for used Cars using Multiple Linear Regression Model. *International Journal for Research in Applied Science and Engineering Technology*, 8(5), pp.1801–1804. doi:https://doi.org/10.22214/ijraset.2020.5290.

3. Alhakamy, A., Alhowaity, A., Alatawi, A.A. and Alsaadi, H. (2023). Are Used Cars More Sustainable? Price Prediction Based on Linear Regression. *Sustainability*, 15(2), p.911. doi:https://doi.org/10.3390/su15020911.

4. COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS AND REGRESSIONS FOR CAR PRICE PREDICTION. (2019). *Bulletin of V. N. Karazin Kharkiv National University Economic Series*, (97). doi:https://doi.org/10.26565/2311-2379-2019-97-04.