Integrative Predictive Model for Automobile Sales Price Forecasting

Abstract— In the context of automotive sales, this study addresses the imperative for a predictive model to estimate consumer expenditure based on various customer attributes such as name, email, country, gender, age, annual salary, credit card debt, and net worth. The focus of the model is the anticipation of the amount paid for a car, constituting a regression task. The comparative analysis explores the efficacy of three distinct algorithms—Artificial Neural Networks (ANN), Multiple Linear Regression, Lasso and Random Forest—in predicting car sales prices. The research aims to discern the strengths and weaknesses of each algorithm, shedding light on their respective performances in the given context. By delving into this integrative approach, the study seeks to provide valuable insights for automobile salespersons, offering a nuanced understanding of predictive modeling techniques for enhanced decision-making in the dynamic realm of automotive sales.

I. INTRODUCTION

In the ever-evolving landscape of automobile sales, the ability to predict and understand consumer behavior is paramount for success. The Integrative Predictive Modeling for Automobile Sales Price Forecasting aims to address this crucial need through a comparative analysis of advanced algorithms—Artificial Neural Networks (ANN), Multiple Linear Regression, Lasso and Random Forest. As a vehicle salesperson, the objective is to create a robust regression model capable of estimating the overall amount consumers would spend on a car, leveraging key customer attributes such as name, email, country, gender, age, annual salary, credit card debt, and net worth. The problem at hand revolves around predicting the "Amount Paid for a Car," constituting a regression task. Accurate forecasting in this domain is essential for sales professionals to tailor their approach, optimize inventory, and enhance customer satisfaction. The chosen algorithms—ANN, Multiple Linear Regression, Lasso and Random Forest—represent a diverse set, each offering unique strengths in handling complex relationships within the dataset. This study recognizes the significance of historical data in developing precise predictive models. By analyzing past demand patterns, market dynamics, and consumer trends, the model aims to provide insights that empower sales teams to make informed decisions. The comprehensive approach involves not only visualization but also the identification of subtle trends and shifts in demand dynamics. The ultimate goal is to equip the automotive industry with proactive forecasting tools, enabling a deeper understanding of customer preferences and optimizing production schedules. In the pursuit of enhancing forecasting accuracy, this research endeavors to combine the strengths of different algorithms, offering a comparative analysis that sheds light on their performance in the context of automobile sales price forecasting. The integration of advanced predictive modeling techniques serves as a beacon for sales professionals, guiding them toward improved decision-making, operational efficiency, and ultimately, greater success in the competitive automotive market.

II. LITERATURE REVIEW

The car purchasing dataset contains five hundred rows with columns as customer name, customer email, gender, age, country, annual salary, credit card debt, net worth and car purchase amount. Below Table 1 shows the variables and the detailed explanation of each.

Variables in the dataset	Explanation

Customer name	Unique identifier for each customer.
Customer email	Email address of the customer
Country	Country of residence
Gender	Gender of the customer (0: Male, 1: Female)
Age	Age of the customer
Annual Salary	Annual income of the customer
Credit Card Debt	Amount of credit card debt
Net Worth	Total net worth of the customer
Car Purchase Amount	Monetary value customers are likely to spend on a vehicle

Table 1: Variables in the dataset and description

In the realm of automobile sales, the significance of predictive modeling for estimating consumer spending on car purchases cannot be overstated. The dataset under consideration encompasses crucial variables such as customer name, email, gender, age, country, annual salary, credit card debt, net worth, and car purchase amount. This literature review focuses on integrative predictive modeling for forecasting automobile sales prices, emphasizing the task's regression nature. To address the problem statement of predicting the amount paid for a car, three distinct algorithms are proposed: Artificial Neural Networks (ANN), multiple linear regression, Lasso, and random forest. The comparative analysis aims to discern the most effective approach. The relevance of such models for salespersons lies in their ability to leverage customer information, ultimately aiding in optimizing sales strategies. Previous studies have demonstrated the efficacy of visualization techniques in demand forecasting, showcasing the potential synergy between visualization and predictive modeling. In [3] the authors explore the development and evaluation of supervised learning-based models for used car price prediction. Specifically, the study focuses on the application of Artificial Neural Network (ANN) models, employing the Keras Regression algorithm, and compares their performance with other machine learning algorithms such as Random Forest, Lasso, Ridge, and Linear Regression. Experimental results have shown that the Random Forest model with a Mean Absolute Error value of 1.0970472 and R2 error value of 0.772584 has given the less error among all the other algorithms.The investigation aims to contribute insights into enhancing decision-making for automobile sales through a comprehensive analysis of diverse modeling techniques.

III. Methodology

1. Data Preparation

Preprocessing is a crucial step in the execution of any machine learning model as it involves preparing the raw data to make it suitable for the model. The quality of the data directly impacts the model's learning ability and the information that can be derived from it. For instance, the dataset obtained from Kaggle comprises 500 rows and 9 columns and may contain noise, missing values, outliers, or be in an unusable format. Therefore, a comprehensive data analysis and preprocessing were necessary before applying any machine learning model. The initial task involved checking for missing values in the dataset after loading it into Jupyter Notebook, and no missing values were found. Subsequently, a detailed analysis for duplicate values was conducted, and no duplicate customer details were identified.

2. Exploratory Analysis

After completing data preprocessing, which involved handling missing values and eliminating duplicates, an exploratory data analysis was conducted. As part of this analysis, a box plot was used to detect potential outliers in the dataset. Additionally, scatter plots were created to visualize the relationship between the car purchase amount and other variables, aiming to identify suitable features for prediction. Figure 1, Figure 2, Figure 3 shows the scatterplots of Net

worth, Annual salary and credit card debt v/s car purchase amount.

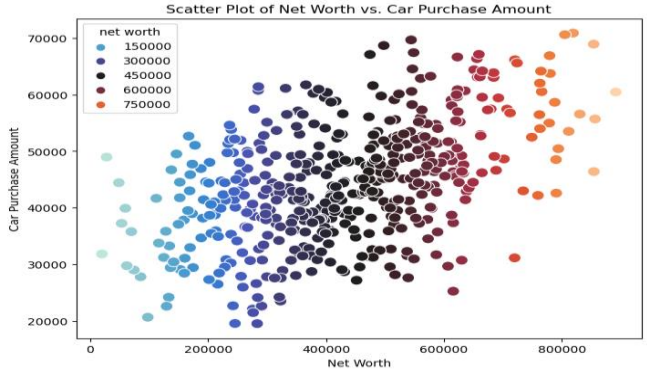

Figure 1. Scatter plot of Net worth v/s Car purchase amount

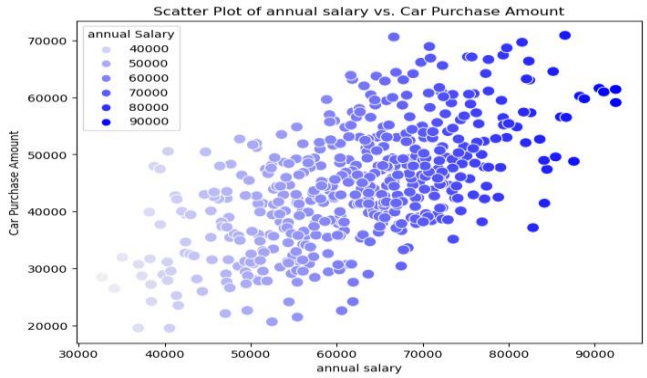

Figure 2. Scatter plot of annual salary v/s Car purchase amount

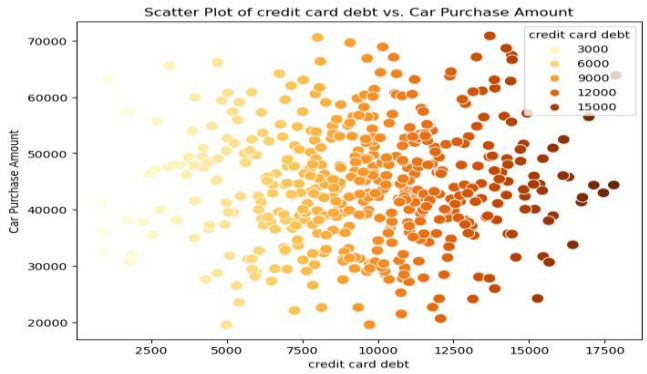

Figure 3. Scatter plot of credit card debt v/s Car purchase amount

The relationship between annual salary and net worth is positive, while the relationship between credit card debt and the car purchase amount is negative. Figure 4 also shows the heatmap to find the correlation between variables.

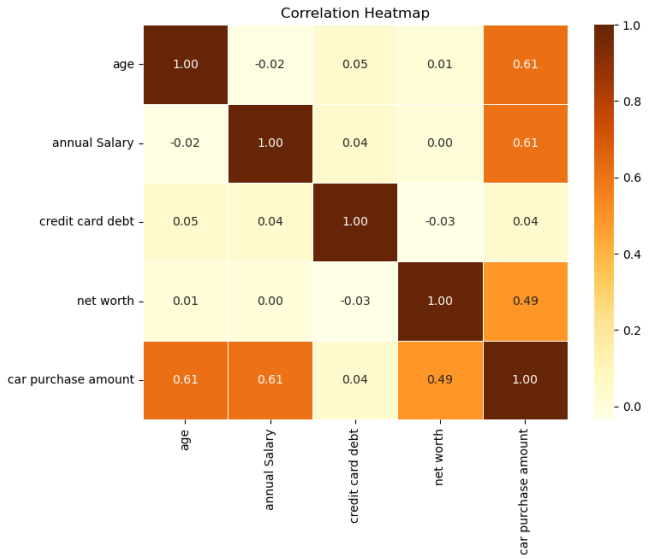

Figure 4. Shows the heatmap to find the correlation.

3. Finding and Removing outliers

After detailed analysis of finding missing values and duplicate rows, the next task was to find out the outliers in the given data. The goal of spotting outliers is usually to draw attention to values that require further examination. Outliers are values that are significantly different from the majority of the data. Here the attributes which need further attention are 'age', 'annual salary', 'credit card debt', 'net worth', which is used as the predictor to predict the car purchase amount. So, it is needed to find the outliers in these attributes. After detailed analysis and considering the values, a box plot was plotted which is shown in Figure 5 below.

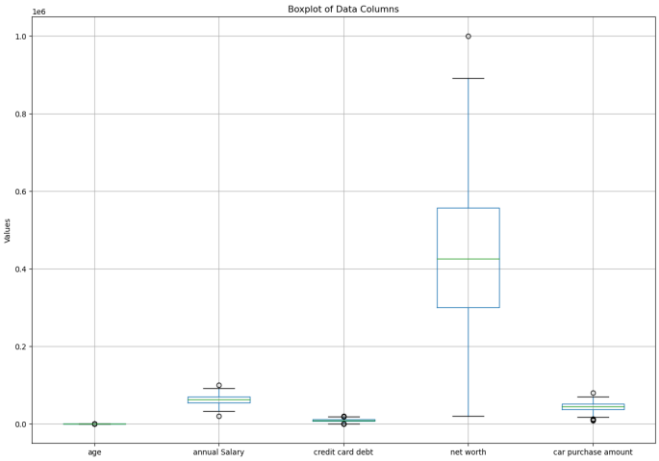

Figure 5. Shows the boxplot to find the outliers.

Outliers were found in every variable plotted which was eliminated using the Inter quartile range method. Now the data is ready for transformation and scaling.

4. Feature scaling the data

Feature scaling is a crucial step in the data preprocessing process before constructing a machine learning model. It involves transforming the features in a dataset so that their values are all on the same scale. This is necessary because real-world datasets often contain features that vary in magnitude, range, and units. As a result, machine learning models require feature scaling to interpret these features on the same scale. In the context of neural networks, they perform best when the predictors and outcome variable are on a scale of when employing a logistic activation function. Therefore, before entering any variables into the network, they should be scaled to an interval. In a specific example, feature scaling was performed on variables such as 'age', 'annual salary', 'credit card debt', 'net worth', and 'car purchase amount' to predict the car purchase amount.

5. Backward elimination and Forward selection to find suitable predictors.

Backward and forward selection are two popular procedures used when analyzing predictors in a dataset. Backward selection begins with all predictor variables in the model. It then iteratively removes the least significant predictor variable at each step until none of the remaining predictors fulfill the stated requirement. Forward selection begins with a null model that has no predictors. It then adds the most important predictor variable at each stage until no more predictors fulfill the required condition (Shmueli et al., 2020, p.172). After careful consideration of both age, annual salary, credit card debt, net worth were found as suitable predictors.

IV. Model Development

1. Artificial Neural Networks

A neural network was utilized with various configurations, including experimenting with a single hidden layer containing 3, 6, 9, and 12 nodes. The model's performance was evaluated based on its capacity to learn and generalize from the training data.

2. Multiple Linear Regression

A fundamental multiple linear regression model was implemented to comprehend the linear relationships between the independent variables and the target variable. This served as a benchmark for assessing the performance of more complex models.

3. Lasso Regression

Lasso regression, a regularization technique, was applied to the multiple linear regression model to prevent overfitting and potentially enhance generalization on unseen data.

4. Random Forest

Random Forest, an ensemble learning method, was employed to capture complex relationships and interactions within the data. This model is capable of handling non-linearities and provides insights into feature importance.

V. Model Evaluation Metrics

The following metrics were used to assess the performance of the models:

Mean Squared Error (MSE): This metric quantifies the average squared difference between the predicted and actual values. The MSE provides a measure of the model's accuracy, with lower values indicating better performance. R-squared (R2): R-squared measures the proportion of the variance in the target variable that is predictable from the independent variables. A higher R-squared value indicates a better fit of the model to the data.

VI. RESULTS AND DISCUSSION

The predictive model estimates the overall amount consumers are likely to spend on purchasing a vehicle. The model aims to provide valuable insights into customer spending behavior, enabling more informed sales strategies.

1. Artificial Neural Network

A Multilayer Perceptron Regressor (MLPRegressor) is utilized to create a predictive model for car purchase amounts based on customer characteristics. The dataset is split into

training and testing sets, and the model is trained using the training data. Scenario 1 is defined with default settings, including a hidden layer with 3 neurons, logistic activation function, a maximum of 500 iterations, and a random state for reproducibility. The model is then evaluated on the test set, and metrics such as Mean Squared Error (MSE) and R2 Score are calculated to assess its performance. Additionally, the regressionSummary function and a DataFrame are used to provide a detailed summary of the regression results, including predicted values, actual values, residuals, and percentage errors. This allows for a comprehensive analysis of how well the model fits the data and the accuracy of its predictions. Subsequently, the model's architecture was fine-tuned in Scenario 2 (6 neurons), Scenario 3 (9 neurons), and Scenario 4 (12 neurons). Comprehensive evaluations were conducted, including Mean Squared Error (MSE) and R2 Score calculations for each scenario.

Notably, Scenario 4 with 12 neurons exhibited the least error, since it has the least MSE. To facilitate a holistic comparison, bar plots visualizing MSE and R Squared across scenarios were generated, highlighting the superior performance of the model with 12 neurons. Figures 6 and 7 show the R-squared and MSE for different numbers of nodes in the hidden layer. This comprehensive analysis underscores the efficacy of the MLPRegressor in predicting car purchase amounts, with the identified optimal configuration as 12 number of nodes in the hidden layer.

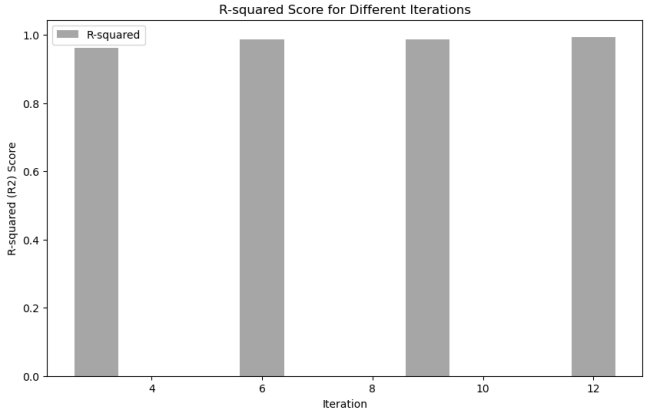

Figure 6: R-squared Score for different number of neurons

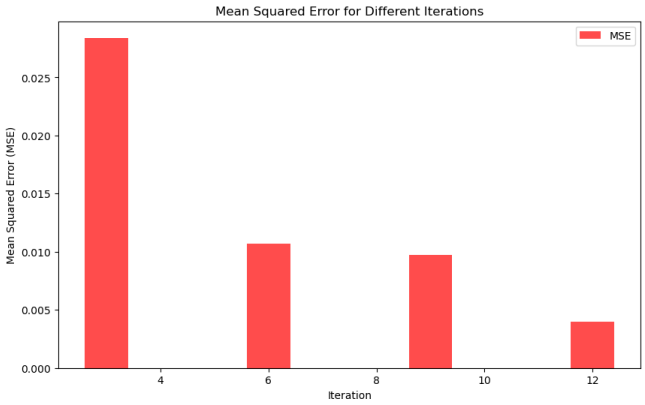

Figure 7: Mean Squared error for different number of neurons.

This comprehensive approach to analysis and prediction contributes not only to assessing the model's efficacy but also to uncovering patterns and relationships within the dataset, enabling informed decision-making and potentially uncovering valuable insights for real-world applications.

2. Multiple Linear Regression

In this analysis, a Multiple Linear Regression (MLR) model was constructed to predict the target variable based on a set of features. The dataset was split into training and testing sets, and the MLR model was trained on the training set. The model's performance was evaluated on the test set using standard regression metrics, including Mean Squared Error (MSE) and R-squared (R2) score, which quantify the accuracy and goodness of fit. The calculated MSE reflects the average squared difference between predicted and actual values, while the R2 score indicates the proportion of variance in the target variable explained by the model. After evaluation, the obtained Mean Squared Error (MSE) approximately close to 0 indicates an exceptionally low level of prediction error on the test set and an R2 Score of nearly 1.0 (0.9999) might suggest the possibility of overfitting. Overfitting occurs when a model learns the training data too well, capturing noise and random fluctuations that are specific to the training set but do not generalize to new, unseen data. Regularization techniques are employed to mitigate overfitting.

3. Lasso Regularization

Lasso regularization is introduced to enhance model accuracy, reduce overfitting, and improve overall

performance. Lasso adds a regularization term to the linear regression objective function, penalizing the model for having too many features with non-zero coefficients. This encourages the model to select a subset of the most important features, effectively performing feature selection.

By penalizing overly complex models, Lasso helps prevent overfitting and enhances the model's ability to generalize well to new data.

In this analysis, Lasso Regression was employed with an alpha value of 1.0 to model the relationship between the predictor variables and the target variable. The Mean Squared Error (MSE) was calculated to assess the accuracy of the model's predictions on the test data. The obtained MSE serves as a key indicator of the model's performance, providing insight into its ability to generalize to new, unseen data. The results contribute crucial information for understanding the predictive capabilities of the Lasso Regression model and guide potential adjustments for optimal performance in future analyses.

4. Random Forest

In this analysis, a Random Forest Regressor was employed to model the relationship between predictor variables and the target variable, representing the car purchase amount. The Random Forest model, consisting of 100 decision trees, was trained on the scaled training data. Subsequently, the model was evaluated on the test set, and the performance metrics were calculated. The Mean Squared Error (MSE) and R-squared (R2) score were employed to assess the accuracy and explanatory power of the model respectively.

Random forest regression sees the dataset from several perspectives compared to a decision tree that keeps digging the dataset in one direction. This feature eases the central decision tree problem, overfitting the training set. The relationship between trees in the forest is not too strong to affect other trees, but strong enough to make the ensemble outperform any standalone tree. This mechanism makes trees in the forest prevent themselves from individual errors.

After fitting the training data to each model, we have the following results.

ML Models	MSE	R-squared
ANN	0.0040	0.9940
Multiple Linear Regression	0.000001	0.9999
Lasso	0.7843	0.9999
Random Forest	0.0370	0.9515

Figure 7: Comparative Analysis

Figure 7 shows comparative analysis of the models. The ANN model stands out for its remarkable accuracy, presenting a very low Mean Squared Error (MSE) and an impressive R-squared (R2) score of 0.9940. These metrics collectively signify an excellent fit to the data, highlighting the model's capacity to capture intricate patterns within the dataset. Multiple linear regression tends to experience overfitting issues, and lasso regression, despite its advantages, demonstrates a higher mean squared error (MSE) when compared to alternative models.

On the other hand, Lasso Regression, while offering advantages, exhibits a higher mean squared error compared to alternative models, underscoring potential limitations in its predictive accuracy. Finally, the Random Forest model, a potent ensemble method, demonstrates robust predictive capabilities with a moderate MSE and a strong R2 score of 0.9515. It means that the artificial neural network model suits the data structure the best and makes the best prediction. Figure 8 shows the plot for R-squared for all the models created.

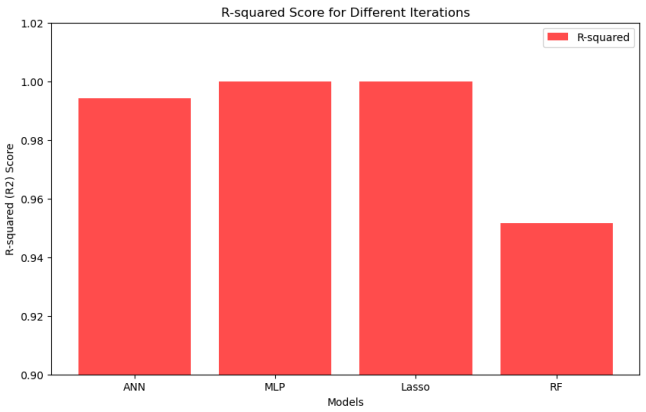

Figure 8: Bar plot of R-squared for all the models.

VII. CONCLUSION AND FUTURE WORK

In conclusion, this study undertook a comprehensive exploration of predictive modeling techniques in the domain

of automotive sales, with a specific focus on estimating consumer expenditure for car purchases. The comparative analysis encompassed four distinct algorithms: Artificial Neural Networks (ANN), Multiple Linear Regression, Lasso, and Random Forest. Results revealed the remarkable accuracy of the ANN model, showcasing its ability to capture intricate patterns within the dataset. However, the study also identified nuances and limitations in other models, such as potential overfitting in Multiple Linear Regression, higher Mean Squared Error (MSE) in Lasso Regression, and the balanced performance of Random Forest. However, there is still much room to improve. Other models like Naive Bayes, LSTM, or Gradient Boosting algorithms can be applied to determine whether better results can be obtained.

Additionally, considering the dynamic nature of the automotive market, the incorporation of external factors, such as economic indicators, industry trends, and marketing campaigns, could enhance the model's predictive power. Exploring the potential of reinforcement learning techniques to optimize sales strategies and adapt to changing market conditions is another promising direction. Moreover, deploying the developed predictive models in a real-world setting and gathering feedback from sales professionals could provide valuable insights for model refinement and practical application. By addressing these avenues, future research can contribute to the refinement and advancement of predictive modeling techniques in the context of automotive sales, ultimately empowering sales professionals with more accurate and actionable insights.

VIII. REFERENCES

[1] Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for Business Analytics: Concepts, techniques and applications in Python. Wiley-Blackwell.*

[2] Varshitha, J., Jahnavi, K., & Lakshmi, C. (2022). Prediction of used car prices using artificial neural networks and machine learning. 2022 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/iccci54379.2022.9740817

[3] ANN - Car sales price prediction. (2022, September 30). Kaggle. https://www.kaggle.com/datasets/yashpaloswal/ann-car-sales-price-prediction/data