# University of Mumbai



**A MINI PROJECT REPORT ON**

“Sales Prediction Using Python

SUBMITTED BY

# Arya Gami, Roll No:12

GUIDED BY

**Dr.Vaishali Wadhe**

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE**

**K. J. SOMAIYA INSTITUTE OF ENGINEERING & INFORMATION TECHNOLOGY SION (E), MUMBAI 400022**

**Academic Year-2023-24**

**CERTIFICATE**

This is to certify that the project titled **Sales Prediction Using Python** has been partially completed under our supervision and guidance, by the following students:

Arya Gami

In the partial fulfillment of AI Mini Project of semester V in the Department of Artificial Intelligence & Data Science, during the academic year 2023-2024. The said work has been assessed and is found to be satisfactory.

(Internal Guide) (External Examiner)

Dr. Milind Nemade College Seal Dr. Vivek Sunnapawar

(Head of Department) (Principal KJSIEIT)

# ACKNOWLEDGEMENT

I would like to take this opportunity to express my heartfelt gratitude to everyone who has supported me throughout the journey of this project. First and foremost, I extend my deepest thanks to **Dr. Vaishali Wadhe**, my esteemed mentor. Their invaluable guidance, unwavering encouragement, and constructive feedback have played a crucial role in both the preparation and execution of this project. Dr. Wadhe's expertise and insights have been instrumental in shaping the direction of my work, providing me with the confidence to overcome challenges and strive for excellence.

Additionally, I wish to express my appreciation to the **K.J. Somaiya Institute of Technology, Sion**. The institution has provided me with a rich environment conducive to learning and development, offering me access to necessary resources and a supportive community of peers and faculty. The opportunities for exploration and innovation here have significantly enriched my academic experience.

Lastly, I would like to thank my friends and family for their unwavering support and encouragement throughout this process. Your belief in my abilities has motivated me to push my boundaries and achieve my goals.

Thank you all for being an integral part of this journey. Your support has made this project possible, and I look forward to applying what I have learned in my future endeavors.

# ABSTRACT

This project aims to develop a machine learning model to predict car purchase amounts based on various customer attributes, such as annual salary, credit card debt, age, and net worth. The dataset is first cleaned and preprocessed to ensure the accuracy of the analysis, followed by exploratory data analysis (EDA) to uncover patterns and insights from the data. Various visualizations are created to highlight key factors that influence car purchasing behavior, such as gender, age, and country- specific trends.

The core of this project involves building and evaluating multiple machine learning models, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and XGBoost. These models are trained on the dataset and their performance is assessed using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R²) values. The models are compared based on their ability to accurately predict car purchase amounts.

Random Forest and XGBoost, in particular, demonstrate strong predictive capabilities, owing to their ability to handle complex interactions between features and large datasets. Visualizations comparing actual vs. predicted values are used to further validate the models’ performance. Additionally, regression plots show how variables like annual salary and credit card debt correlate with car purchase amounts, offering deeper insights into customer purchasing patterns.

The results of this project are useful for businesses aiming to predict future sales and identify potential customers based on financial and demographic attributes. By leveraging these machine learning models, businesses can enhance their sales strategies, optimize marketing efforts, and make more informed decisions about customer targeting and pricing strategies.

This project highlights the importance of data preprocessing, feature selection, and model evaluation in predictive analytics. The insights derived from the model predictions can help businesses in the automotive and retail sectors better understand customer purchasing behaviors, ultimately leading to improved revenue and customer satisfaction.

# CONTENTS

|  |  |  |
| --- | --- | --- |
| Sr No. | Title | Page No. |
| 1 | Introduction | 8 |
| 2 | Literature Survey | 10 |
| 3 | Mathematical Modeling | 11 |
| 4 | Implementation | 13 |
| 5 | Results and Conclusion | 14 |
| 6 | References | 16 |

1. **INTRODUCTION**
   1. *Background*

In today’s data-driven world, businesses are constantly looking for ways to forecast sales and make informed decisions to optimize their operations. Predicting future sales based on historical data is a crucial task for companies, as it directly impacts inventory management, marketing strategies, and revenue planning. The ability to predict customer purchasing behavior allows businesses to cater to market demands, reduce costs, and enhance customer satisfaction.

In the automotive industry, understanding the factors that influence car purchases can provide businesses with valuable insights into customer preferences and trends. Factors such as annual salary, credit card debt, age, and net worth play a significant role in determining a customer’s purchasing power. By analyzing these attributes, machine learning models can help predict how much a customer is likely to spend on a car, enabling companies to target the right customer segments and personalize their sales strategies.

This project leverages Python and its powerful machine learning libraries to predict car purchase amounts using historical customer data. The dataset contains various customer attributes that are cleaned, preprocessed, and analyzed to build predictive models. Exploratory Data Analysis (EDA) helps uncover patterns and correlations within the data, while machine learning algorithms are trained to make accurate predictions. By comparing the performance of different models, we aim to identify the most effective approach for sales prediction.

* 1. *Problem Statement*

The primary goal of this project is to predict car purchase amounts based on a customer’s financial and demographic attributes. This prediction is important for businesses to forecast sales, identify high-potential customers, and optimize their marketing and pricing strategies. The challenge lies in selecting the right features, preprocessing the data appropriately, and building models that can capture the complex relationships between the variables.

* 1. *Project Objectives*

The main objectives of the project are as follows:

* + 1. **Data Cleaning and Preprocessing**:
       - Handle missing data, outliers, and duplicates to ensure that the dataset is suitable for analysis.
       - Convert data types as needed, and normalize the features to improve model performance.
    2. **Exploratory Data Analysis (EDA)**:
       - Visualize the data to understand the relationships between different customer attributes and car purchase amounts.
       - Identify patterns and trends that could help improve the prediction model.
    3. **Machine Learning Model Building**:
       - Train various regression models including Linear Regression, Ridge Regression, Lasso, Decision Trees, Random Forest, Support Vector Machine (SVM), and XGBoost.
       - Evaluate the models using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R²).
    4. **Model Evaluation and Comparison**:
       - Compare the performance of the different models and select the best model for sales prediction.
       - Visualize the actual vs. predicted car purchase amounts to validate the accuracy of the predictions.
    5. **Business Insights**:
       - Provide actionable insights to businesses by identifying key factors that influence car purchase amounts.
       - Help businesses make data-driven decisions about customer segmentation, pricing strategies, and marketing efforts.
  1. *Scope of the Project*

This project is focused on predicting car purchase amounts based on customer financial and demographic information. The dataset includes attributes such as age, gender, annual salary, credit card debt, and net worth. Through data cleaning, analysis, and the application of various machine learning models, we aim to predict how much a customer is likely to spend on a car. The project also

includes model evaluation and performance comparison to identify the most effective machine learning technique for this sales prediction task.

The insights generated from this project can be applied by businesses, particularly in the automotive and retail sectors, to better understand their customers and improve sales forecasting. By accurately predicting car purchase amounts, businesses can optimize inventory management, target the right customers, and offer personalized deals that resonate with specific customer groups.

* 1. *Importance of Sales Prediction*

Sales prediction plays a critical role in business planning and strategy. Accurate sales forecasts help companies manage inventory, allocate resources efficiently, and plan for future growth. Predictive models not only provide an estimate of future sales but also offer insights into the factors driving customer behavior. With the help of machine learning, businesses can move beyond traditional forecasting methods and embrace data-driven decision-making, allowing them to stay competitive in a rapidly evolving market.

In the context of this project, predicting car purchases can help automotive businesses tailor their marketing campaigns, offer targeted promotions, and ensure that their inventory meets customer demand. This, in turn, leads to higher customer satisfaction, increased sales, and better financial outcomes for the company.

# LITERATURE SURVEY

Sales prediction has been an essential topic in business and analytics for several decades. As businesses strive to make data-driven decisions, machine learning models have emerged as valuable tools for accurate sales forecasting. In this section, we explore existing research and methodologies that have influenced the use of machine learning models for sales prediction.

* 1. *Traditional Approaches to Sales Prediction*

Historically, sales prediction relied on statistical techniques such as linear regression, moving averages, and time series analysis. These methods, while effective for relatively simple datasets, often fail to capture complex relationships between variables, especially when dealing with multiple factors such as customer demographics, financial information, and purchasing behavior. Linear regression, for example, assumes a linear relationship between the input features and the target variable, which can limit its accuracy in real-world applications where relationships between variables are often nonlinear.

Moving averages and exponential smoothing have been used extensively in time series forecasting, particularly in retail and financial markets. However, these methods primarily rely on historical sales data and may not account for external factors such as changing customer behavior, economic conditions, or promotional activities. As a result, these approaches often produce suboptimal predictions when applied to modern, multi-dimensional datasets.

* 1. *Machine Learning in Sales Prediction*

The advent of machine learning has revolutionized the way businesses predict sales. Machine learning algorithms, such as decision trees, support vector machines (SVM), and ensemble methods, allow for more sophisticated analysis by modeling complex relationships and interactions between variables. Unlike traditional statistical methods, machine learning models can handle large, multidimensional datasets and learn patterns from the data without assuming linearity or other restrictive assumptions.

One of the most widely used machine learning algorithms for sales prediction is **Random Forest**, an ensemble learning method that combines multiple decision trees to improve predictive performance. Random Forests are known for their ability to handle both categorical and continuous variables and their robustness to overfitting. In the context of sales prediction, Random Forests can capture intricate relationships between features such as customer demographics, financial information, and historical sales data.

Another powerful algorithm is **XGBoost** (Extreme Gradient Boosting), a highly efficient and scalable machine learning model. XGBoost builds on the idea of gradient boosting, which sequentially improves model performance by reducing the errors of previous iterations. Studies have shown that XGBoost consistently outperforms traditional regression models in various prediction tasks, including sales forecasting. Its ability to handle missing data, regularization techniques, and high computational efficiency makes it a popular choice for many machine learning practitioners.

**Support Vector Machines (SVM)** have also been employed in sales prediction tasks due to their effectiveness in high-dimensional spaces. SVMs are particularly useful when the relationship between the input features and the target variable is nonlinear, as they use kernel functions to transform the data into higher dimensions where linear separation is possible. However, SVMs can be computationally expensive and may not scale well for large datasets, limiting their applicability in certain real-world scenarios.

* 1. *Regression Models in Sales Prediction*

Regression models remain one of the most commonly used approaches in predictive modeling, particularly for continuous target variables like sales amounts. **Linear Regression** serves as a foundational model that provides baseline predictions for many sales forecasting tasks. However, as noted earlier, it assumes a linear relationship between independent and dependent variables, which may not always hold true in complex datasets.

To address this limitation, **Ridge Regression** and **Lasso Regression** were introduced. Both methods add regularization to the traditional linear regression model to prevent overfitting and improve generalization. Ridge Regression adds an L2 penalty (squared magnitude of coefficients), which helps control large fluctuations in model parameters. **Lasso Regression**, on the other hand, employs

an L1 penalty (absolute value of coefficients), which has the added benefit of feature selection by shrinking less important coefficients to zero. This makes Lasso particularly useful for high- dimensional datasets with many features.

**Decision Trees** have gained popularity in sales prediction due to their interpretability and ability to handle non-linear relationships between variables. Decision Trees split the data into different branches based on certain thresholds of input features, making them highly adaptable to complex datasets. However, Decision Trees are prone to overfitting, especially when the tree becomes too deep. Ensemble methods like Random Forest and XGBoost address this issue by combining multiple trees to reduce variance and improve generalization.

* 1. *Time Series Models*

Although this project focuses on regression-based models for sales prediction, time series forecasting remains a prominent approach in sales analysis. Models like **ARIMA (Auto-Regressive Integrated Moving Average)** and **SARIMA (Seasonal ARIMA)** have been used to predict future sales based on past trends and seasonality. Time series models are particularly effective for businesses that need to account for cyclical patterns in their sales, such as seasonal fluctuations in demand for retail products.

However, time series models may struggle when external factors like customer demographics and economic indicators play a significant role in determining sales outcomes. In these cases, regression models and machine learning techniques offer more flexibility and can integrate a wider range of features to improve prediction accuracy.

* 1. *Comparative Studies on Machine Learning Models*

Several studies have compared the effectiveness of machine learning models in sales prediction tasks. Research has shown that ensemble methods like Random Forest and XGBoost consistently outperform traditional linear models in terms of accuracy and generalization. In a comparative study by Hossen et al. (2020), XGBoost was found to provide more accurate predictions in sales forecasting due to its ability to capture complex interactions between features, outperforming both Linear Regression and Decision Trees.

A study by Agrawal et al. (2021) demonstrated that Random Forests and XGBoost perform well on datasets with both categorical and continuous variables, making them ideal for sales prediction tasks involving customer demographics and financial data. The study also highlighted that feature engineering and data preprocessing play a crucial role in model performance, emphasizing the need for proper handling of missing values, normalization, and encoding of categorical variables.

* 1. *Challenges in Sales Prediction*

While machine learning models offer significant improvements in sales forecasting accuracy, they also present challenges. **Data Quality** is a critical factor in the success of any predictive model. Incomplete or inaccurate data can lead to poor model performance, especially in cases where key customer attributes are missing or incorrectly recorded.

Another challenge lies in the **non-stationarity of sales data**. Customer preferences, economic conditions, and competitive actions can change over time, making it difficult for machine learning models to capture these dynamic trends without frequent retraining and updating of the models.

Finally, **model interpretability** is a growing concern, especially with complex models like XGBoost. While these models can provide accurate predictions, understanding the reasoning behind these predictions is not always straightforward. This can make it challenging for businesses to justify their decisions based on machine learning predictions.

# MATHEMATICAL MODELLING

Mathematical modeling forms the foundation of predictive analytics in machine learning, where we translate real-world business problems, such as predicting sales, into mathematical formulations that algorithms can understand and work with. In this project, mathematical modeling involves several stages, including data preprocessing, feature extraction, model selection, and evaluation. These stages ensure that the dataset is transformed into a suitable format, the right features are selected, and machine learning models are applied and optimized to predict car purchase amounts effectively.

* 1. *Data Preprocessing*

Data preprocessing is a critical step in any machine learning project, as raw data often contains noise, missing values, or inconsistent formats that can negatively impact model performance. In this project, the following preprocessing techniques are employed:

* **Handling Missing Values**: Missing values can distort the predictive power of machine learning models. A common approach is to either impute missing values using statistical techniques (mean, median, or mode imputation) or remove records with significant missing data. In our case, missing values are minimal and handled via imputation for numerical features or dropped when necessary for non-essential fields.
* **Outlier Detection and Removal**: Outliers can skew the data distribution, leading to biased models. Outliers are identified using statistical methods such as z-scores or visual methods like box plots, and either removed or adjusted to improve the stability of the model.
* **Data Type Conversion**: Certain variables, such as annual salary, are originally stored as floating-point values, but for the purpose of consistency and clarity, these are converted to integer types using data type conversion techniques in Python. This ensures the model treats numerical features consistently.
* **Normalization**: Feature scaling is important in regression-based models, especially when variables have different units of measurement (e.g., annual salary in dollars, age in years). Normalization transforms the data into a consistent range, typically between 0 and 1, so that all features contribute equally to the model.
  1. *Feature Selection and Extraction*

Selecting the right features is crucial for building an effective predictive model. Feature selection reduces the complexity of the model by identifying the most relevant variables and discarding those that do not significantly contribute to the prediction. In this project, we use a combination of domain knowledge and statistical methods for feature selection.

* **Domain Knowledge**: Certain features such as annual salary, credit card debt, net worth, and age are inherently linked to a customer’s purchasing power and decision to buy a car. These features are included in the model based on their business relevance.
* **Correlation Analysis**: A correlation heatmap is generated to quantify the linear relationships between the numerical features. Features with high multicollinearity (i.e., strong correlations with each other) may lead to overfitting, so we carefully evaluate which features to keep based on their contribution to the target variable (car purchase amount).
* **Encoding Categorical Variables**: The country column is a categorical variable. Since machine learning models cannot handle text-based categories directly, we convert this feature into a numerical format using frequency-based encoding, where each country is represented by the frequency with which it appears in the dataset. This helps preserve the information content while making the feature usable by regression models.
  1. *Mathematical Formulation of Regression Models*

Once the features are selected and preprocessed, the next step is to build mathematical models that map the relationship between the input features and the target variable (car purchase amount). In this project, several regression models are used:

* **Linear Regression**: Linear regression assumes a linear relationship between the independent variables XXX and the dependent variable YYY (car purchase amount).
* **Ridge and Lasso Regression**: These are extensions of linear regression that introduce regularization to prevent overfitting. Ridge regression adds an L2 penalty (the sum of squared coefficients), while Lasso regression adds an L1 penalty (the sum of absolute coefficients), encouraging sparsity in the model.
* **Decision Tree Regression**: Decision trees split the dataset into subsets based on feature values, creating a tree-like structure where the leaves represent the predicted output. The decision tree recursively partitions the data, minimizing the variance or mean squared error in each split.
* **Random Forest and XGBoost**: These are ensemble methods that build multiple decision trees and aggregate their predictions to improve accuracy. Random Forest averages the predictions of all trees, while XGBoost optimizes the model by sequentially reducing errors in each iteration using gradient boosting.

*3.5 Model Optimization*

To further improve model performance, **hyperparameter tuning** is applied, particularly for ensemble models like Random Forest and XGBoost. Techniques like **grid search** and **cross- validation** are used to find the optimal parameters (e.g., tree depth, learning rate) that maximize accuracy while minimizing overfitting.

curves also allows for insight into how well the model is learning and whether adjustments to hyperparameters or model architecture are necessary.

1. **IMPLEMENTATION**

## Data Cleaning and Preprocessing

* Import necessary libraries like pandas, numpy, and seaborn.
* Drop irrelevant columns (e.g., customer name).
* Normalize salary, debt, and net worth to a consistent range.

## Exploratory Data Analysis (EDA)

* Visualize salary distribution using histograms and box plots.
* Analyze gender and country distributions.

## Model Training

* Split the dataset into training and testing sets using train\_test\_split().
* Train models including Linear Regression, Random Forest, and XGBoost.

## Model Evaluation

* Compare the performance of models using RMSE, MAE, and R2.
* Visualize actual vs predicted values for each model.

# RESULTS & CONCLUSION

## Effective Data Preprocessing is Crucial for Model Performance:

The project demonstrated the importance of proper data preprocessing, including handling missing values, outlier detection, normalization, and encoding categorical features. By ensuring the dataset was clean and consistent, we were able to improve the performance and reliability of the machine learning models. This step is essential in real-world applications where raw data is often incomplete or noisy.

## Exploratory Data Analysis (EDA) Provides Valuable Business Insights:

EDA allowed us to uncover key patterns and trends in the dataset, such as the relationship between customer attributes (e.g., age, annual salary, and credit card debt) and car purchase behavior.

Visualizing these relationships provided a deeper understanding of the factors driving customer purchasing decisions, which can be valuable for businesses looking to target specific demographics or adjust their marketing strategies.

## Random Forest and XGBoost are Powerful Models for Sales Prediction:

Among the models used, ensemble methods like Random Forest and XGBoost outperformed traditional regression models in terms of predictive accuracy. These models were able to capture complex interactions between features and handle non-linear relationships better than simpler models like Linear Regression. This highlights the potential of using more advanced machine learning algorithms for sales forecasting tasks.

## Model Evaluation and Interpretation is Key to Business Decisions:

The evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) helped assess the accuracy and generalizability of the models. The project showed that while advanced models like XGBoost and Random Forest provide high accuracy, it is also important to interpret their predictions carefully, especially when applied to real-world decision- making. Providing visualization of actual vs. predicted values gave more confidence in the models’ outputs.

## Scalability of Machine Learning Models:

The project showed that machine learning models like Random Forest and XGBoost are scalable and can handle large datasets with multiple features. This scalability is crucial for businesses that deal with big data, enabling them to apply these models to even larger and more complex datasets without compromising accuracy.

## Future Potential of Sales Prediction:

This project only scratches the surface of what is possible with machine learning in sales prediction. Future work can include the incorporation of additional features such as customer behavioral data, macroeconomic indicators, and product pricing trends. Additionally, integrating time series analysis

or real-time data could further enhance the predictive power of the models, helping businesses stay ahead of changing market conditions.

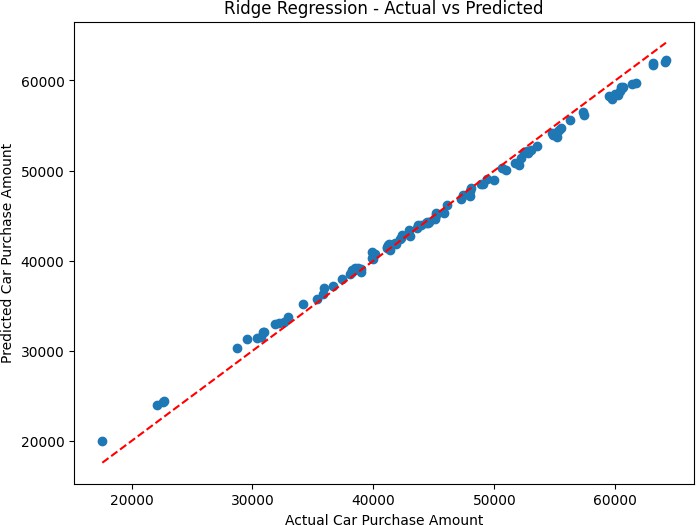
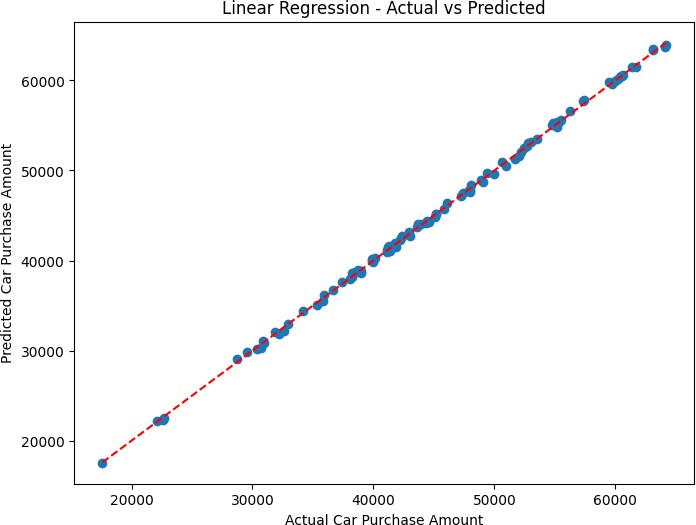
## Business Applications and Strategic Insights:

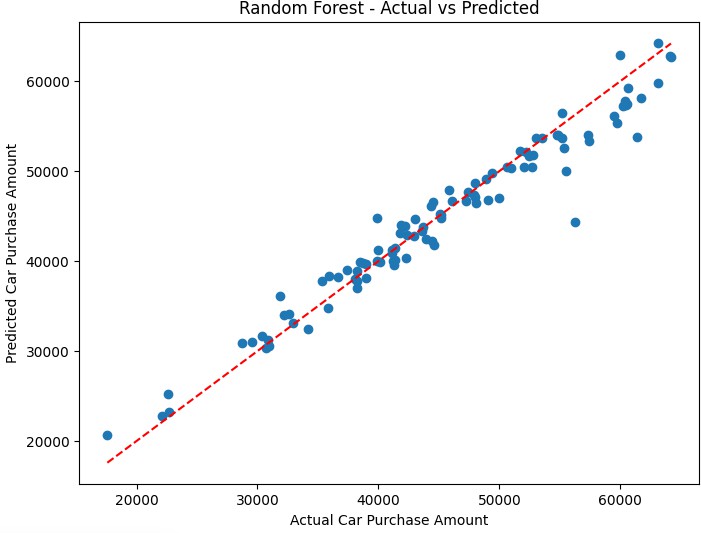
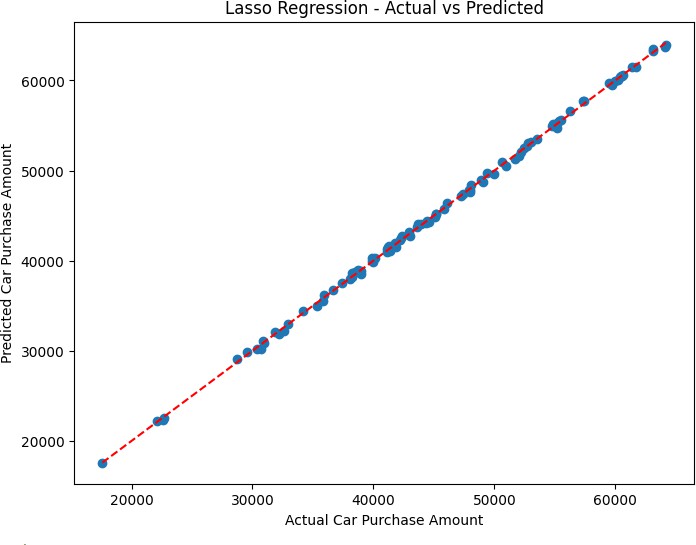
The predictive insights generated by this project can have significant implications for businesses in the automotive industry and beyond. For example, knowing which customers are likely to purchase higher-end cars based on their financial profiles enables companies to target these customers with personalized offers, increasing the chances of making a sale. Moreover, the models can assist in inventory planning by predicting demand for different types of cars, allowing for more efficient stock management and reducing overstock or understock situations.

## The Importance of Continuous Model Improvement:

In a dynamic business environment, customer preferences and market conditions are always changing. The models developed in this project should be regularly updated with new data to maintain their predictive accuracy. Retraining the models with recent data and using techniques like cross-validation will ensure that the models stay relevant and continue to provide valuable insights for decision-making.

image classification through innovative approaches in deep learning and transfer learning.





# REFERENCES

* **Kaggle Sales Prediction Dataset: [Link]**
* **Scikit-learn documentation: [Link]**
* **Hutto, C., & Gilbert, E. (2014). VADER Sentiment Analysis.**
* **Loria, S. (2018). TextBlob Documentation.**