Index

Introduction

The project focuses on analysing and predicting car purchase amounts based on various features such as gender, age, annual salary, credit card debt, and net worth. The initial exploration involves visualizing the distribution of these features using Seaborn boxplots. To enhance the accuracy of the predictive model, outliers are identified and removed using the Interquartile Range (IQR) method.

Subsequently, a linear regression model is implemented using gradient descent. The goal is to establish a relationship between the input features and the target variable, i.e., the car purchase amount. The data is pre-processed by scaling to ensure uniformity, and the dataset is split into training and testing sets.

The gradient descent algorithm iteratively optimizes the model coefficients, and the progress is monitored by printing the updated coefficients at each iteration. Finally, the trained model is applied to the test set to make predictions.

This project aims to provide insights into the factors influencing car purchase amounts and to develop a predictive model that can assist in understanding and estimating the potential car expenditure based on individual characteristics.

Problem statement

In a world where individual financial decisions play a crucial role in shaping economic dynamics, there exists a need to understand and predict the factors influencing car purchase amounts. The challenge lies in developing a predictive model that can effectively analyze key individual characteristics, such as gender, age, annual salary, credit card debt, and net worth, to estimate the potential car expenditure.

Dataset

Main Context:-

As a vehicle salesperson, you would like to create a model that can estimate the overall amount that consumers would spend given the following characteristics:
customer name, customer email, country, gender, age, annual salary, credit card debt, and net worth

The model should anticipate the following (Problem Statement):

Amount Paid for a Car

Task type:

Regression

Algorithm:

The given problem statement can be solved using Machine Learning or Deep Learning Techniques

Note:

While reading csv you will face an error **UnicodeDecodeError**
Just do the following step while reading csv file:-

```
data = pd.read_csv("/kaggle/input/ann-car-sales-price-prediction/car_purchasing.csv",encoding='ISO-8859-1')
```

Dataset Link- https://www.kaggle.com/datasets/yashpaloswal/ann-car-sales-price-prediction

Methodology

1. **Data Visualization:**
    - The project begins by visualizing the distribution of various features such as gender, age, annual salary, credit card debt, net worth, and the target variable, car purchase amount, using Seaborn boxplots. This step helps in identifying the general trends and potential outliers in the dataset.
2. **Outlier Removal:**
    - Outliers, which can significantly impact the performance of a predictive model, are detected and removed using the Interquartile Range (IQR) method. This ensures that the subsequent linear regression model is trained on a more robust and representative dataset.
3. **Linear Regression Implementation:**
    - The linear regression model is implemented from scratch using gradient descent. The goal of linear regression is to establish a linear relationship between the input features and the target variable (car purchase amount in this case). The model is represented as:
      $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$
      where y is the target variable, $x_1, x_2, \ldots, x_n$ are the input features, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients to be optimized.

4. **Data Preprocessing:**
    - The input features and target variable are preprocessed. Features are scaled using Min-Max scaling to ensure uniformity and prevent certain features from dominating the others. Additionally, a bias term (intercept) is added to the feature matrix to account for the constant term in the linear regression equation.
5. **Training and Testing Split:**
    - The dataset is split into training and testing sets using the `train_test_split` function from `sklearn`. This division ensures that the model is trained on a subset of the data and tested on unseen data to evaluate its generalization performance.
6. **Gradient Descent Optimization:**
    - The core of the project involves iterating through the gradient descent algorithm to optimize the coefficients ($\beta_0, \beta_1, \ldots, \beta_n$). The algorithm minimizes the error between the predicted car purchase amount and the actual values in the training set. The gradients are computed for each feature, and the coefficients are updated iteratively to minimize the mean squared error.
7. **Prediction:**
    - The trained linear regression model is then applied to the test set to make predictions. The model's performance can be evaluated by comparing its predictions with the actual car purchase amounts in the test set.

LIBRARIES

The code utilizes several Python libraries for data visualization, data manipulation, and machine learning. Here's a brief description of the main libraries used in the provided code:-

1. Seaborn:-

Purpose:- Seaborn is a statistical data visualization library based on Matplotlib. It provides a high level interface for drawing attractive and informative statistical graphics.

Usage in Code:- Seaborn is used to create boxplots for visualizing the distribution of various features in the dataset.

2. Matplotlib:-

Purpose:- Matplotlib is a comprehensive plotting library. It is often used for creating static, interactive, and animated visualizations in Python.

Usage in Code:- Matplotlib is used to create subplots and display the boxplots.

3. NumPy:-

Purpose:- NumPy is a fundamental package for scientific computing with Python. It provides support for large, multidimensional arrays and matrices, along with mathematical functions to operate on these arrays.

Usage in Code:- NumPy is used for array operations, calculations, and manipulation of data.

4. Pandas:-

Purpose:- Pandas is a data manipulation and analysis library. It provides data structures like DataFrame for efficient data manipulation and analysis.

Usage in Code:- Pandas is used to handle and manipulate the dataset, including removing outliers.

5. ScikitLearn (sklearn):-

Purpose:- ScikitLearn is a machine learning library that provides simple and efficient tools for data analysis and modeling. It includes various tools for classification, regression, clustering, and more.

Usage in Code:- ScikitLearn is used for preprocessing (MinMaxScaler) and splitting the dataset into training and testing sets.

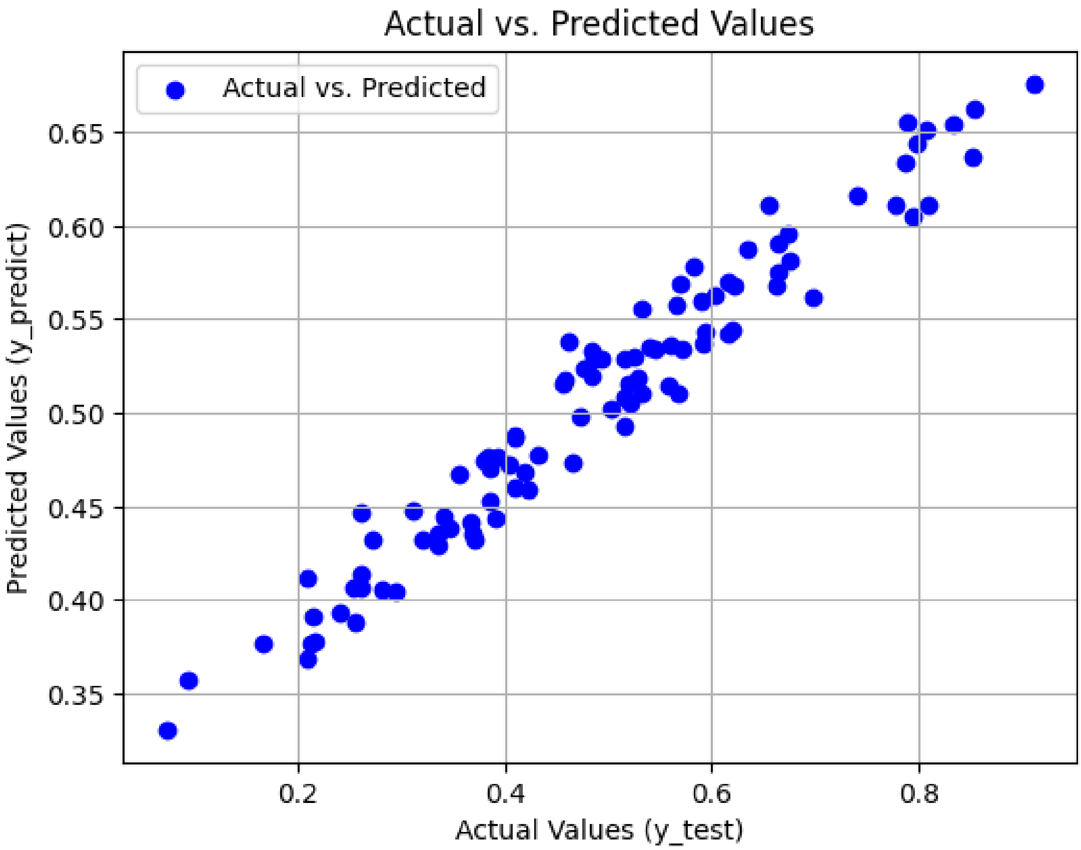
Results

The mean square error obtained is 0.0122 .

The root mean square error is 0.110 .

The r2 score is 0.640 .

The graph plotted between Actual and Predicted values is :



Conclusion

From above we can conclude that this model is good fit for predicting car sales as

Practical application:-

Forecasting car sales can help dealerships and manufacturers optimimsing inventory levels. By using this model they can prevent overstocking or understocking, reducing carry costs and minimizing risk of unsold inventory.

It can help to guide marketing efforts by identifying high demand waiting periods and customer preferences.

Manufacturers can use sales prediction to plan production schedules more efficiently . By aligning production with expected demand, companies can reduce lead times, improve resource allocation and avoid cost associated with excessive production .

GitHub link

GitHub project link- https://github.com/Harman4404/ml_project