Impact of the Architecture Attributes on Linear Regression (A Tutorial For Analyzing Customer Spending w.r.t Dependent & Independent Features)

Name: __Zeeshan Ali_____

ID:___ 23036973_____

1. Introduction:

1.1 Overview of Linear Regression

Linear regression is a statistical method use to predict the value of a dependent variable base on one or more independent variables. It examines the relationship between these variables to model and estimate outcomes. Fit a linear equation to the observe data linear regression establishes the strength and direction of the relationship between the variables. This technique is widely apply in various fields like economics, finance and social sciences for forecast and analysis. Simplest form, simple linear regression involves one independent variable while multiple linear regression uses more than one. The goal is to find the best-fit line or regression line that minimizes the difference between the predictive and actual values. Linear regression is said to be one of the most commonly use and powerful tools in statistics due to its simplicity and effectiveness to understand and predict relationships between variables [1].

Linear regression architecture is a framework that shows the relationship between a dependent variable and one or more independent variables. It uses a straight-line pattern to relate changes in input variables to changes in the output. The framework has input features a target coefficients and an intercept to analyze data. It predicts outcomes by fitting a line to the data to minimize error and provide accurate results. This method explains how the target variable depends on independent variables and how adjustments in inputs influence the output. It provides a simple approach to understand and model variable relationships [1].

Linear regression use one or more input variables to predict a dependent variable. Input variables are often represented by X while the output is denoted by Y. Linear Regression itself has various types but the two types which revolve heavily around number of features are: single/simple linear regression (single layer) and multiple linear regression (multi layer) [1,2].

2. Dependent & Independent Types in Linear Regression:

In **simple linear regression** where only one input feature is used the input is represented by x. This method models the relationship between a single predictor and the target variable. The goal is to fit a line that best represents how changes in the input feature affect the output. In simple or single linear regression we can say that x (a feature) denotes the customers spending (can be current or previous) amount and in light of this the output variable can be y which denotes the customer spending prediction we can use single regression to predict spending results [3].

Where-as **multiple linear regression** uses multiple input features to predict the output. These features are represented as x1 x2 till xn where each feature contributes to the prediction of the dependent variable Y. Multiple linear regression allows for more complex models that consider the influence of several variables at once providing a more accurate prediction when multiple factors affect the target variable. This approach helps capture interactions between variables and provides a better understanding of how each input affects the outcome. We may also use group techniques or piece wise linear techniques for further better analytical results [4].

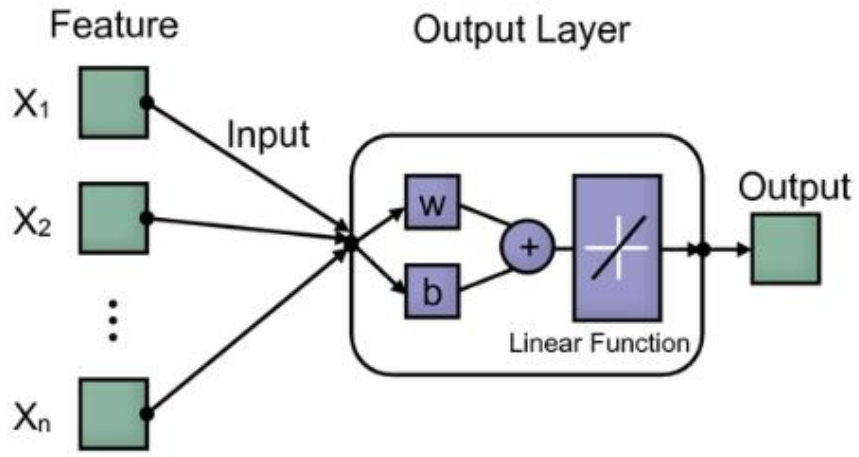


Fig. 1.0. Diagram for Multiple Linear Regression [2].

Both methods rely on the same principle but the difference is calculate as an error term which the algorithm seeks to minimize to create the most accurate prediction model.

3. Problem Statement

As we know to predict customer spending behavior is a challenge for business which aims to optimize their market strategy and improve customer engagement. This tutorial aims to address the challenge by develop a model using linear regression that predicts to estimate the Spending Score of customers based on their demographic and financial attributes like age, gender and annual income. Our analysis will help see the factors influence spend patterns which enables more effective results.

4. Implementation

4.1 Tutorial for Single/Simple Linear Regression:

Below is the single regression tutorial which use customer/s data to predict the spending score with respect to given ages in the dataset. Prediction of spend is the concern w.r.t age.

By use jupyterNotebook linear regression algorithm (single linear) is apply on the customer dataset which is available on Kaggle [4].

Step#1:

Use the pandas library and import relevant libraries in Python and load the dataset efficiently within a Jupyter Notebook environment. This allows you to work with structured data enable easy manipulation exploration directly from the notebook which is ideal for both exploratory preprocess tasks in data science workflow.

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import pandas as pd

#Load dataset
mall_data = pd.read_csv('F:/Mall_Customers.csv')
```

Step#2:

Analysis looks at age as the independent variable and spending score as the target. The dataset is split into two parts with 80% use to train the model and 20% for test its performance. Linear regression model learns the relationship between age and spending score by find the slope and intercept. Slope is -0.5917, showing a slight negative relationship where spending scores drop as age increases. The intercept is 74.80 which means the predict spending score for someone at age zero starts at this value.

After training the model is tested on new data to see how well it works. Its performance is measured using mean squared error (MSE) and R-squared (R^2). MSE which shows the average error in predictions 468.05 indicates a significant gap between the predictive and actual scores. The R^2 value is 0.0518 shows age explains only 5.2% of the variation in spending scores suggests the relationship between age and spending scores is very weak.

```
#Select features and target
X = mall_data[['Age']]  #Predictor:Age
y = mall_data['Spending_Score']  #Target:Spending_Score

#Split the data into training and testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#Apply Linear Reg.
model = LinearRegression()
model.fit(X_train, y_train)

#Predictions
y_pred = model.predict(X_test)

#Evaluate Model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error (MSE):", mse)
print("R-squared (R^2):", r2)
print("Coefficient (Slope):", model.coef_[0])
print("Intercept:", model.intercept_)
```

```
Mean Squared Error (MSE): 468.04993405528967
R-squared (R^2): 0.05107060648915396
Coefficient (Slope): -0.5917744816161146
Intercept: 74.80242451588705
```

Step#3:

A plot that visualizes the relationship between age and spending score use a linear regression model. Plot is initialize with a size of 10 by 6 inches to ensure it is large enough for clear visualization. A scatter plot is create use scatterplot() function where the x-axis represents the ages from the test data (X_test[Age]) and the y-axis represents the actual spend scores (y_test). These data points are plotted in black and the label for this data in the plot legend is set as Data. A regression line is then added to the plot using plt.plot where the x-axis still represents the ages and the y-axis represents the predicted spending scores (y_pred).

Overall, the code provides a visual representation of the relationship between age and spending score shows both the actual data and the regression line predicted by the model.

```
#Graph
plt.figure(figsize=(10, 6))

#Scatter plot
sns.scatterplot(x=X_test['Age'], y=y_test, color='black', label='Data')

#Regression line
plt.plot(X_test, y_pred, color='red', linewidth=2, label='Regression Line')

plt.title('Linear Regression: Age vs Spending Score', fontsize=16)
plt.xlabel('Age', fontsize=14)
plt.ylabel('Spending Score', fontsize=14)
plt.legend(fontsize=12)

plt.show()
```

Result gained from above analysis:

These results suggest that age is not a strong predictor of spending behavior in this dataset. Weak performance metrics indicate that other factors like annual income or gender may better explain the variability in spend scores. Include additional features or explore more complex relationships might improve the performance of the model.

4.2 Tutorial for Multiple Linear Regression:

For multiple linear regression let us use Age and Annual_Income_(k$) as predictors to predict Spending_Score.

Step#1:

Use the pandas library and import relevant libraries in Python and load the dataset efficiently within a Jupyter Notebook environment. This allows you to work with structured data enable easy manipulation exploration directly from the notebook which is ideal for both exploratory preprocess tasks in data science workflow.

```python
#Multiple Regression
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

#Load dataset
mall_data = pd.read_csv('F:/Mall_Customers.csv')

#Preprocessing: Convert Gender to numeric values
mall_data['Gender'] = mall_data['Gender'].map({'Male': 0, 'Female': 1})

#predictors (Age, Annual Income) and target (Spending_Score)
X = mall_data[['Age', 'Annual_Income_(k$)']]  #Predictors
y = mall_data['Spending_Score']  #Target
```

Here we use multiple predictors to predict accurate values for spending score. We use 0 for male and 1 for female to associate gender data with the predictors if needed.

Step#2:

Process of split data, train a model and make predictions by use multiple linear regression. The dataset X (features) and Y (target variable) are divide into train and test sets use the train_test_split() function. The test_size=0.2 argument indicate that 20% of the data will be use for test while 80% will be use for train. The random_state=42 ensures that the split by fixing the random seed. A multiple linear regression model is create use LinearRegression(. The model is then train by fit it to the train data (X_train and y_train) use the fit() method.

```python
#Splitting the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#multiple linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

#predictions
y_pred = model.predict(X_test)
```

Step#3:

Code evaluates the performance of the multiple linear regression model and visualizes the results. It calculates two key evaluation metrics Mean Squared Error (MSE) and R-squared (R^2). MSE measures the average square difference

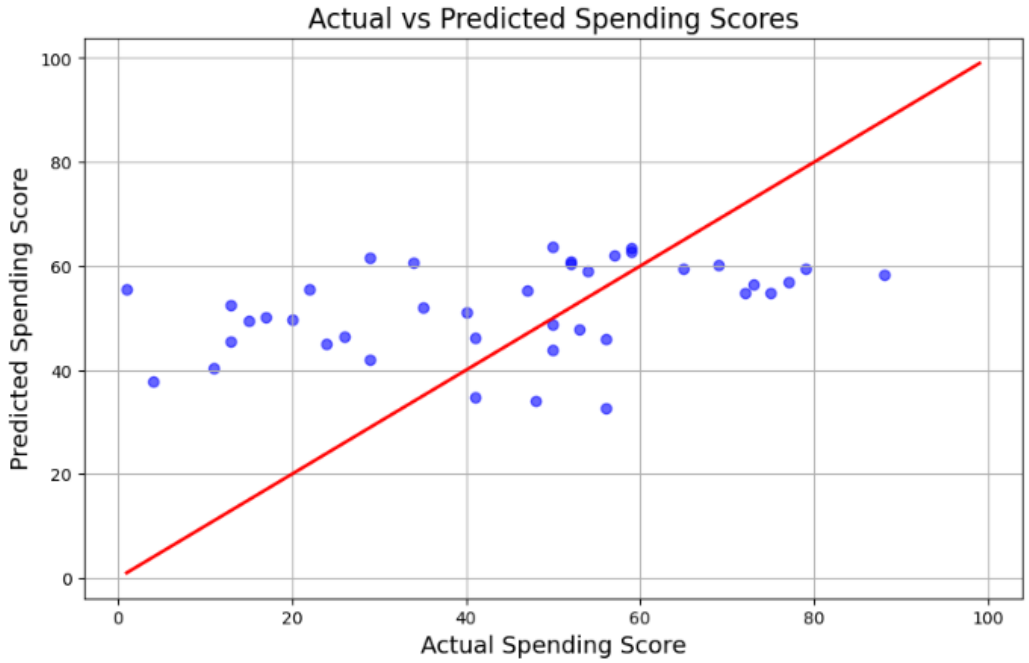
between actual and predict values while R^2 indicates how well the model explains the variance in the target variable. Models coefficients (slopes for each feature) and intercept are also print to understand the relationship between the features and the target. After evaluation the model the code visualizes the results by plotting a scatter plot compare the actual vs predicted spending scores with a red line representing a perfect prediction.

```
#evaluate
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
coefficients = model.coef_
intercept = model.intercept_

print("Mean Squared Error (MSE):", mse)
print("R-squared (R^2):", r2)
print("Coefficients (Slopes):", coefficients)
print("Intercept:", intercept)

#visualize
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, color='blue', alpha=0.6)
plt.plot([y.min(), y.max()], [y.min(), y.max()], color='red', linewidth=2)
plt.title('Actual vs Predicted Spending Scores', fontsize=16)
plt.xlabel('Actual Spending Score', fontsize=14)
plt.ylabel('Predicted Spending Score', fontsize=14)
plt.grid()
plt.show()
```

```
Mean Squared Error (MSE): 483.55682175408344
R-squared (R^2): 0.01963177813218009
Coefficients (Slopes): [-0.58929193  0.05235827]
Intercept: 71.5325839533133
```



Result gained from above analysis:

Analysis suggests that Age and Annual Income predict better Spending Score than single linear regression since more features are in use but additional features might be need to improve the model predictive power if we want to achieve further better predictions or different types of predictions.

5. Conclusion

Comparison between single and multiple linear regression analysis highlights the differences in model performance and predictive power when we use one vs multiple predictors. In single linear regression analysis age was in use as the only predictor of spend score. The model revealed a weak relationship with a low R-squared value of 0.0518 suggests that age alone explains only 5.18% of the variability in spend scores. The high mean squared error indicates prediction errors demonstrate the limitations to rely on a single variable to predict spending behavior. Multiple linear regression model incorporates both age and annual income as predictors. The model R-squared value 0.0196 remains low indicates that these two variables together explain only 1.96% of the variance in spend scores the additional predictor enhances the models complexity. The MSE was high reflects significant discrepancies between the actual and predict values. The multiple regression model provides a more comprehensive analysis compare to the simple linear regression model. It incorporates multiple features offers a broader perspective on the relationship between the predictors and spending scores. The improvement in model performance reinforces the idea that these features alone may not be sufficient to accurately predict spend scores. Multiple regression enhances the models explanatory power it suggests that additional factors or more relevant features might be need to capture the full complexity of the spend score analysis as the current set of predictors seems not enough.

6. Conclusion

[1] Kumari, Khushbu, and Suniti Yadav. "Linear regression analysis study." *Journal of the practice of Cardiovascular Sciences* 4.1 (2018): 33-36.

[2] Min, Dae-Hong, and Hyung-Koo Yoon. "Suggestion for a new deterministic model coupled with machine learning techniques for landslide susceptibility mapping." *Scientific reports* 11.1 (2021): 6594.

[3] Dang, Tran Tri, et al. "Constructing and understanding customer spending prediction models." *SN Computer Science* 4.6 (2023): 852.

[4] Sharma, Prashant, Aratrika Chakraborty, and Judhajit Sanyal. "Machine learning based prediction of customer spending score." *2019 Global Conference for Advancement in Technology (GCAT)*. IEEE, 2019.

[5] MechLearn, Shruti. *Customer Data*. 2018. Kaggle. https://www.kaggle.com/datasets/shrutimechlearn/customer-data?resource=download.

[6] https://www.geeksforgeeks.org/simple-linear-regression-in-python/ (For single linear regression)

[7] https://www.kaggle.com/code/emineyetm/multiple-linear-regression-in-python (For multiple linear regression)