**ASSIGNMENT-3**

**COURSE ID: CS564**

**Foundations of Machine Learning**

**Submitted by:**

**Ankit Anand**

**Roll No: 2311MC04**

**Design a predictive regression model that forecasts sales based on the "Advertising.csv" dataset. Afterwards, employ logistic regression and Support Vector Machines (SVM) to predict defaulters using the "Credit.csv" and "Credit-Modified.csv" datasets. Perform a 70-30 train-test split for model evaluation and measurement of performance. Create a scatter plot with a clear separation line to visualize the data distribution. Generate a table that assesses the significance of the dataset features using the Anova test and test the significance of the derived model parameters**.

**Introduction:**

This assignment aims to construct predictive regression models and perform classification using machine learning techniques on specific datasets. The tasks involve creating a sales forecasting model based on the "Advertising.csv" dataset using regression methods and employing logistic regression and Support Vector Machines (SVM) for default prediction on the "Credit.csv" and "Credit-Modified.csv" datasets.

**Dataset Overview:**

The datasets— "Advertising.csv," "Credit.csv," and "Credit-Modified.csv"— encompass various features and target variables. Preliminary exploration and potential preprocessing steps, such as handling missing values or categorical data encoding, may be necessary.

**Predictive Regression Model for Sales Forecasting:**

For the "Advertising.csv" dataset:

1. Data exploration and Correlation Analysis: Understanding the relationships between features and sales using theoretical concepts such as linear regression assumptions (e.g., linearity, homoscedasticity).

2. Train-Test Split: Implementing a 70-30 split for model evaluation while emphasizing the significance of unbiased evaluation and avoiding overfitting.

3. Regression Techniques: Employing methods like linear regression or random forests, emphasizing the concept of ensemble learning and feature importance analysis.

4. Model Performance Evaluation: Assessing model performance using standard regression metrics (e.g., RMSE, MAE), ensuring interpretations beyond metrics.

**Logistic Regression and SVM for Default Prediction:**

Utilizing "Credit.csv" and "Credit-Modified.csv":

1. Data Preprocessing: Incorporating theoretical concepts such as scaling and addressing class imbalance to enhance model performance.

2. Train-Test Split: Emphasizing the significance of data partitioning for unbiased evaluation.

3. Classification Algorithms: Implementing logistic regression and SVM, focusing on understanding the sigmoid function in logistic regression and the margin optimization in SVM.

4. Visualization and Model Evaluation: Creating scatter plots with separation lines to visualize data distribution and model predictions. Evaluation using classification metrics (e.g., accuracy, precision, recall, F1-score).

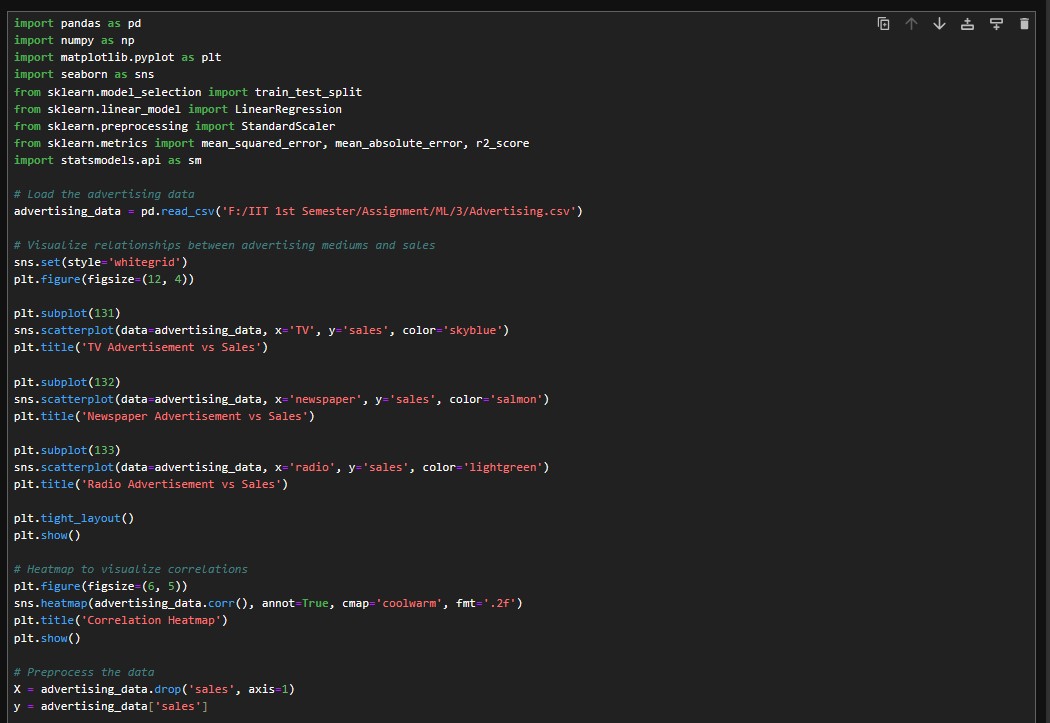
**Assessment of Feature Significance:**

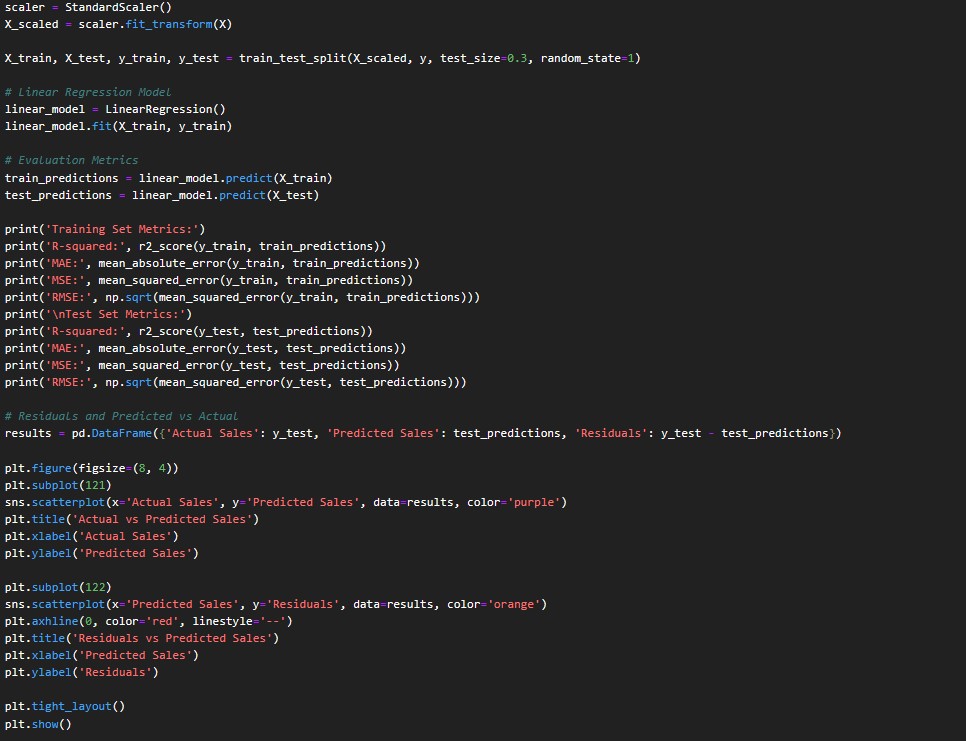
For the predictive regression model:

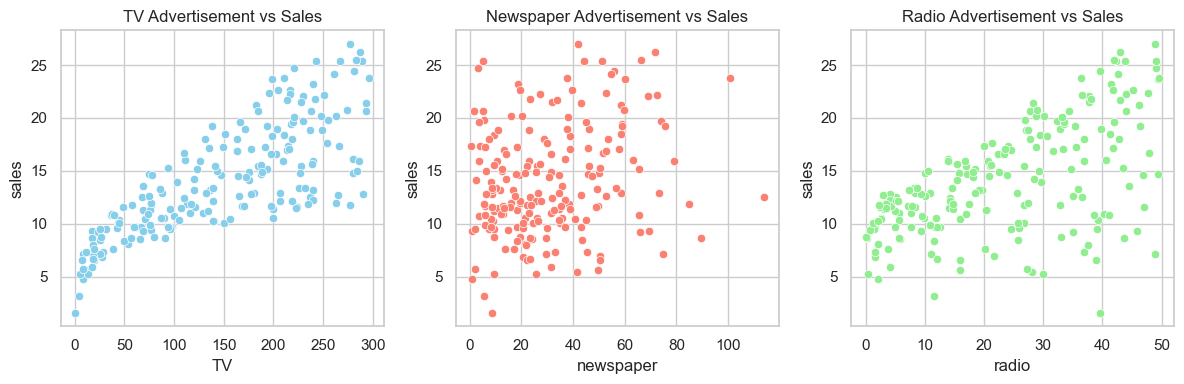
1. Anova Tests: Theoretical foundation of Anova tests to evaluate the significance of dataset features concerning sales forecasting.

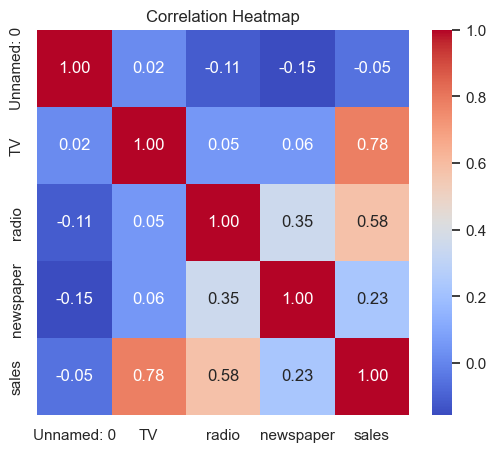
2. Model Parameter Significance: Discussing the significance of model parameters (coefficients) through hypothesis testing or confidence intervals, ensuring the understanding of parameter interpretation.

**Sample Code with Output:**









Training Set Metrics:

R-squared: 0.8850071142546371

MAE: 1.3749370490107342

MSE: 3.2030691013148522

RMSE: 1.789712016307331

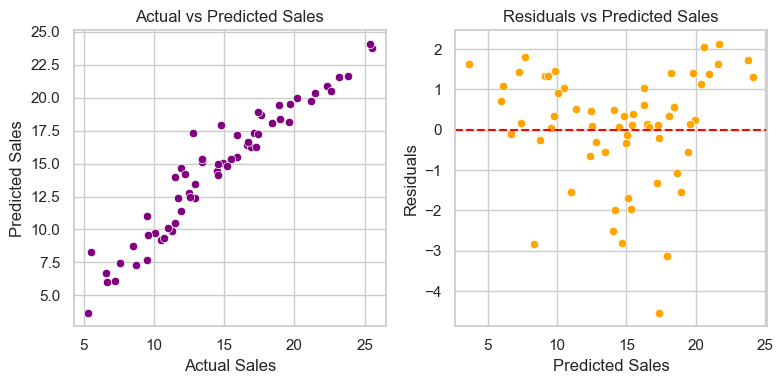
Test Set Metrics:

R-squared: 0.9225191550357023

MAE: 1.0543666890342978

MSE: 1.9274675206987768

RMSE: 1.38833264050759



OLS Regression Results

==============================================================================

Dep. Variable: sales R-squared: 0.885

Model: OLS Adj. R-squared: 0.882

Method: Least Squares F-statistic: 259.7

Date: Mon, 20 Nov 2023 Prob (F-statistic): 2.39e-62

Time: 20:09:13 Log-Likelihood: -280.14

No. Observations: 140 AIC: 570.3

Df Residuals: 135 BIC: 585.0

Df Model: 4

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 14.0057 0.154 90.784 0.000 13.701 14.311

x1 -0.0073 0.162 -0.045 0.964 -0.329 0.314

x2 4.0210 0.153 26.334 0.000 3.719 4.323

x3 2.6148 0.165 15.885 0.000 2.289 2.940

x4 0.0389 0.173 0.225 0.822 -0.303 0.381

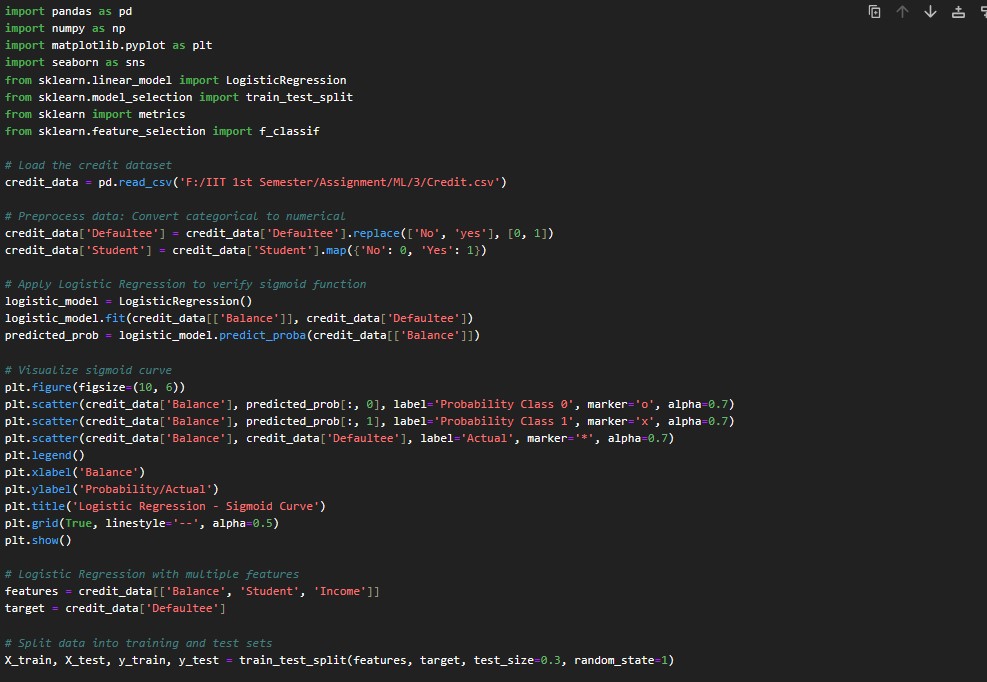
==============================================================================

Omnibus: 38.658 Durbin-Watson: 2.094

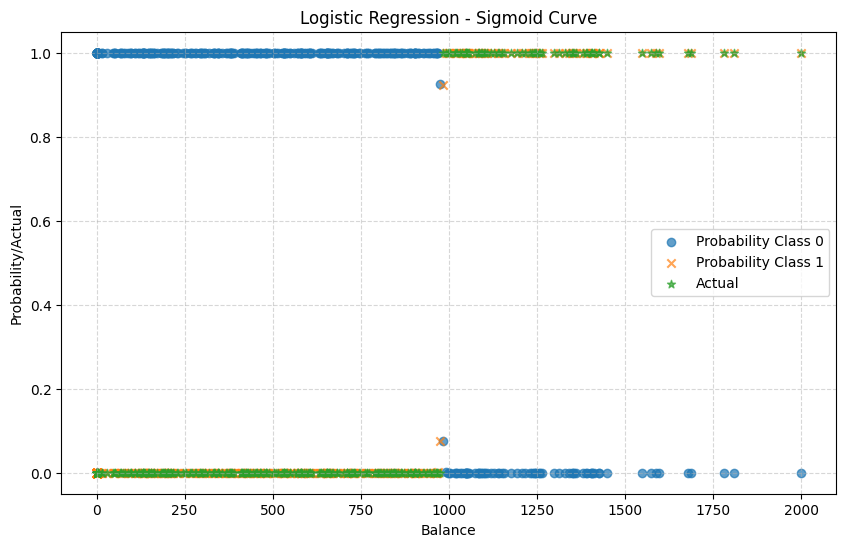
Prob(Omnibus): 0.000 Jarque-Bera (JB): 73.200

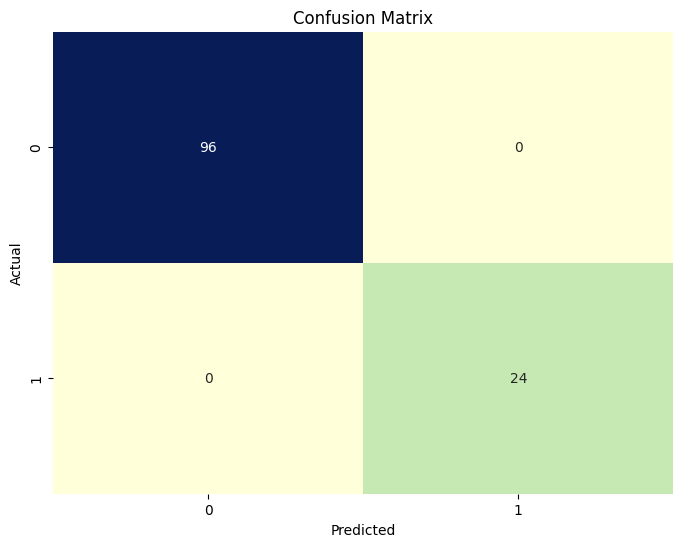
Skew: -1.242 Prob(JB): 1.27e-16

Kurtosis: 5.526 Cond. No. 1.62







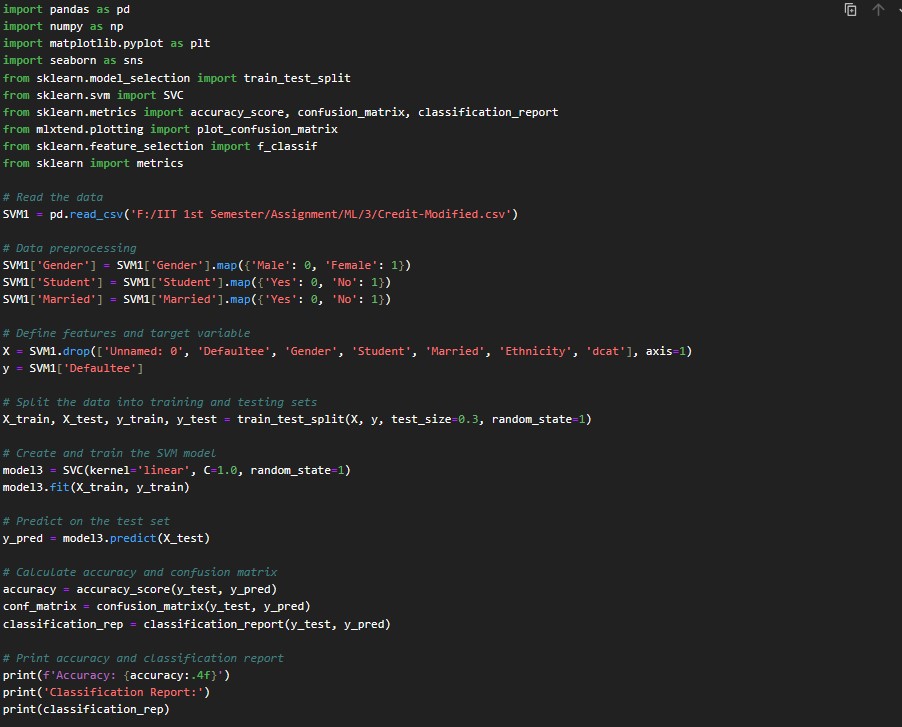


Feature F-Value P-Value

0 Balance 169.903057 1.326267e-24

1 Student 6.464834 1.229753e-02

2 Income 30.587183 1.947321e-07





Accuracy: 1.0000

Classification Report:

precision recall f1-score support

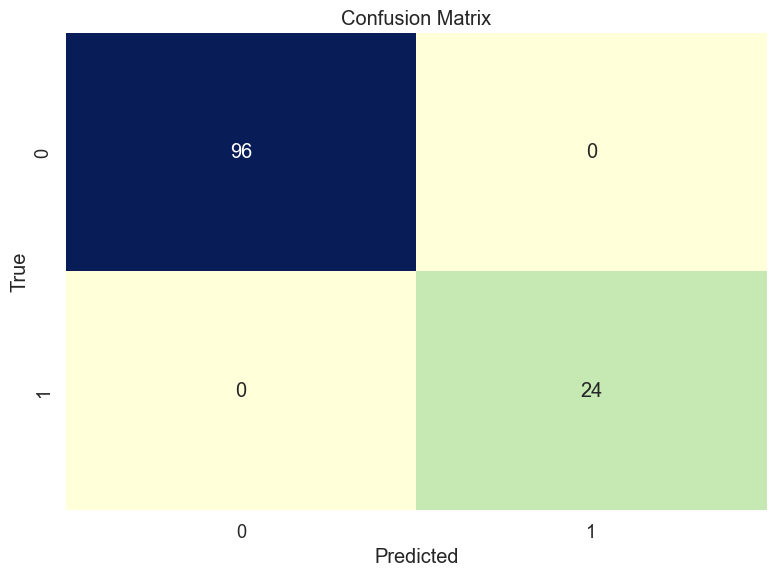
0 1.00 1.00 1.00 96

1 1.00 1.00 1.00 24

accuracy 1.00 120

macro avg 1.00 1.00 1.00 120

weighted avg 1.00 1.00 1.00 120



Feature F-Value P-Value

0 Income 30.587183 1.947321e-07

1 Limit 90.378059 2.955557e-16

2 Rating 91.075672 2.421534e-16

3 Cards 0.133908 7.150689e-01

4 Age 0.002142 9.631614e-01

5 Education 1.482109 2.258739e-01

6 Gender-num 2.754467 9.963856e-02

7 Student-num 6.464834 1.229753e-02

8 Balance 169.903057 1.326267e-24