**FINAL PROJECT - GUVI DATA SCIENCE**

# **Problem Statement 1:**

1.Below is the table containing the information of Customer price index, discounts, offers where the organization wants to predict the sales based on the cpi, discounts, offers.

|  |  |  |  |
| --- | --- | --- | --- |
| CPI | discounts | offers | Sales |
| 2600 | 3 | 20 | 550000 |
| 3000 | 4 | 15 | 565000 |
| 3200 | 5 | 18 | 610000 |
| 3600 | 3 | 30 | 595000 |
| 4000 | 5 | 8 | 760000 |
| 4100 | 6 | 8 | 810000 |

Given below information find out the Sales that has

* 5000 cpi, 3 percentage discounts, 20 rewards offer.
* 4000 cpi, 8 percentage discounts, 19 rewards offer.

**Solution:**

The Model Used for this problem statement are **Multiple Linear Regression (MLR),** **K Nearest Neighbour Regressor** & **Random Forest Regressor** from the **Scikit learn library**.

1.**MLR:**

* + **R2 Score:** 0.9517053277318955
  + **5000 cpi, 3 percentage discounts, 20 rewards offer - Sales is** 826645.34
  + **4000 cpi, 8 percentage discounts, 19 rewards offer - Sales is** 732680.36

2. **K-Neighbour Regressor:**

* + **R2 Score:** 0.3121500559910415
  + **5000 cpi, 3 percentage discounts, 20 rewards offer - Sales is** 668000
  + **4000 cpi, 8 percentage discounts, 19 rewards offer - Sales is** 668000

3. **Random Forest Regressor:**

* + **R2 Score:** 0.9553874300111982
  + **5000 cpi, 3 percentage discounts, 20 rewards offer - Sales is** 649400
  + **4000 cpi, 8 percentage discounts, 19 rewards offer - Sales is** 680450

Out of these 3, **K Nearest Neighbour Regressor** performed **very poorly** so we can't choose that model, but we can go with either **Linear Regression** model or **Random Forest Regressor** both are **good**.

# **Problem Statement 2:**

2.Below table contains the bank details of customer – need to predict the how we can offer loan to customer based on the Cibil score, Age, insurance, debit card, cards.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Customer id | Cards | Debit card | Insurance | Age | Cibil Score | Loan offer |
| 5 | 0 | 1 | 0 | 50 | 34.94 | 0 |
| 3 | 1 | 0 | 0 | 18 | 0.891 | 1 |
| 66 | 0 | 1 | 0 | 5 | 0.33 | 1 |
| 70 | 0 | 1 | 1 | 31 | 0.037 | 0 |
| 96 | 0 | 1 | 0 | 30 | 0.038 | 1 |

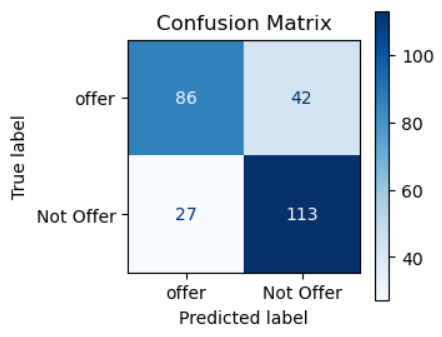
**Solution:**

The Model Used for this problem statement are **Logistic Regression**, **Support Vector Machine (SVC)** & **K Nearest Neighbour classifier** from the Scikit learn library.

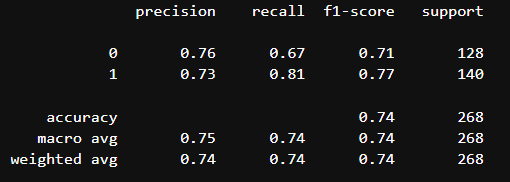
1.**Logistic Regression:**

**Accuracy Score:** 0.7425373134328358

**Confusion Matrix:**

****

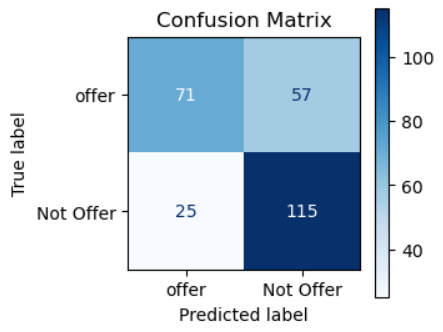
**Classification Report:**

****

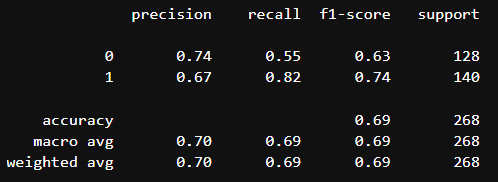
2.**Support Vector Machine:**

**Accuracy Score:** 0.6940298507462687

**Confusion Matrix:**

****

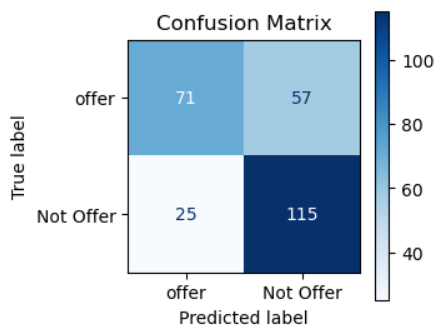
**Classification Report:**

****

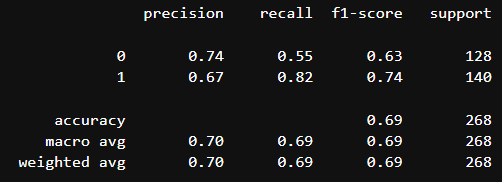
3.**K Nearest Neighbour Classifier:**

**Accuracy Score:** 0.6865671641791045

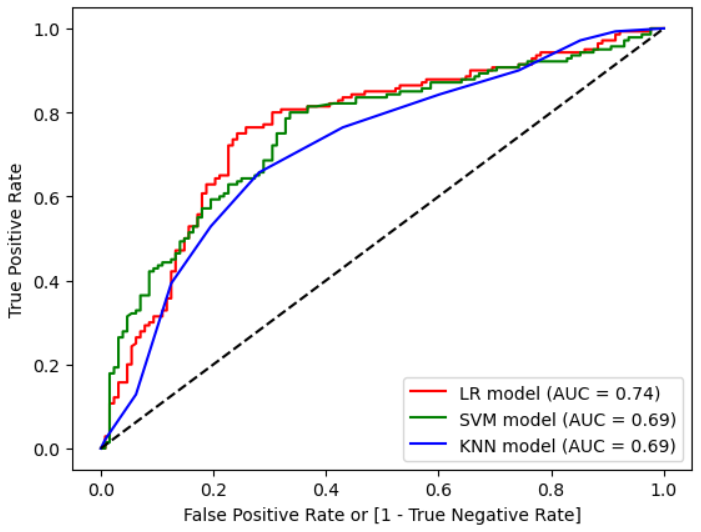
**Confusion Matrix:**

****

**Classification Report:**

****

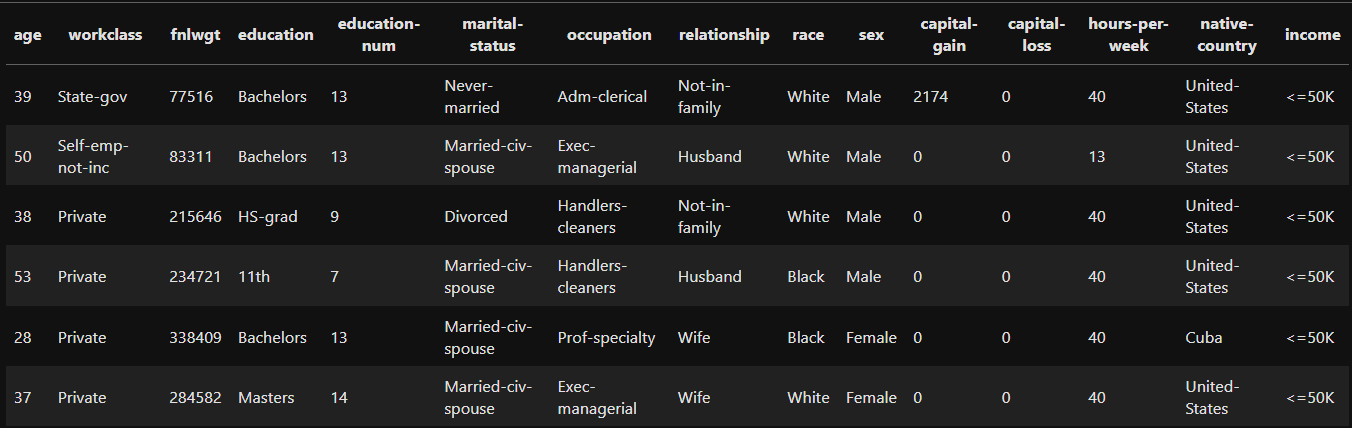
**ROC Curves:**

****

Out of these 3, only **Logistic Regression** Performed **good** so we are choosing Logistic Regression**.**

# **Problem Statement 3:**

3.Below table contains the information of customer and their geographical and personal details, Build the classification models like Decision Tree, Random Forest, KNN, K Means and SVM to classify the data in to proper groups.



**Solution:**

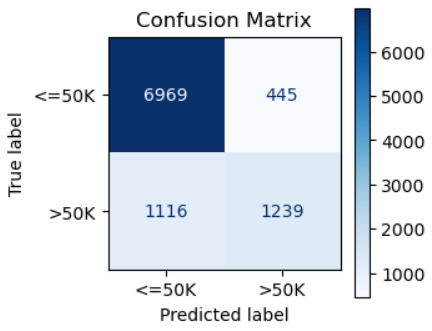
* + Every Models that is mentioned in the question is completed.
  + We have used Frequency Encoding for this Dataset.

**Supervised / Classification Model:**

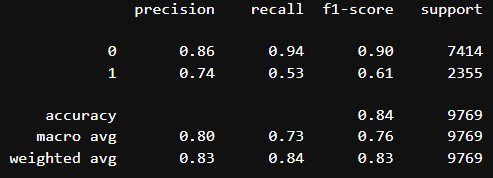
1**.Decision Tree:**

**Accuracy Score:** 0.8402088238304842

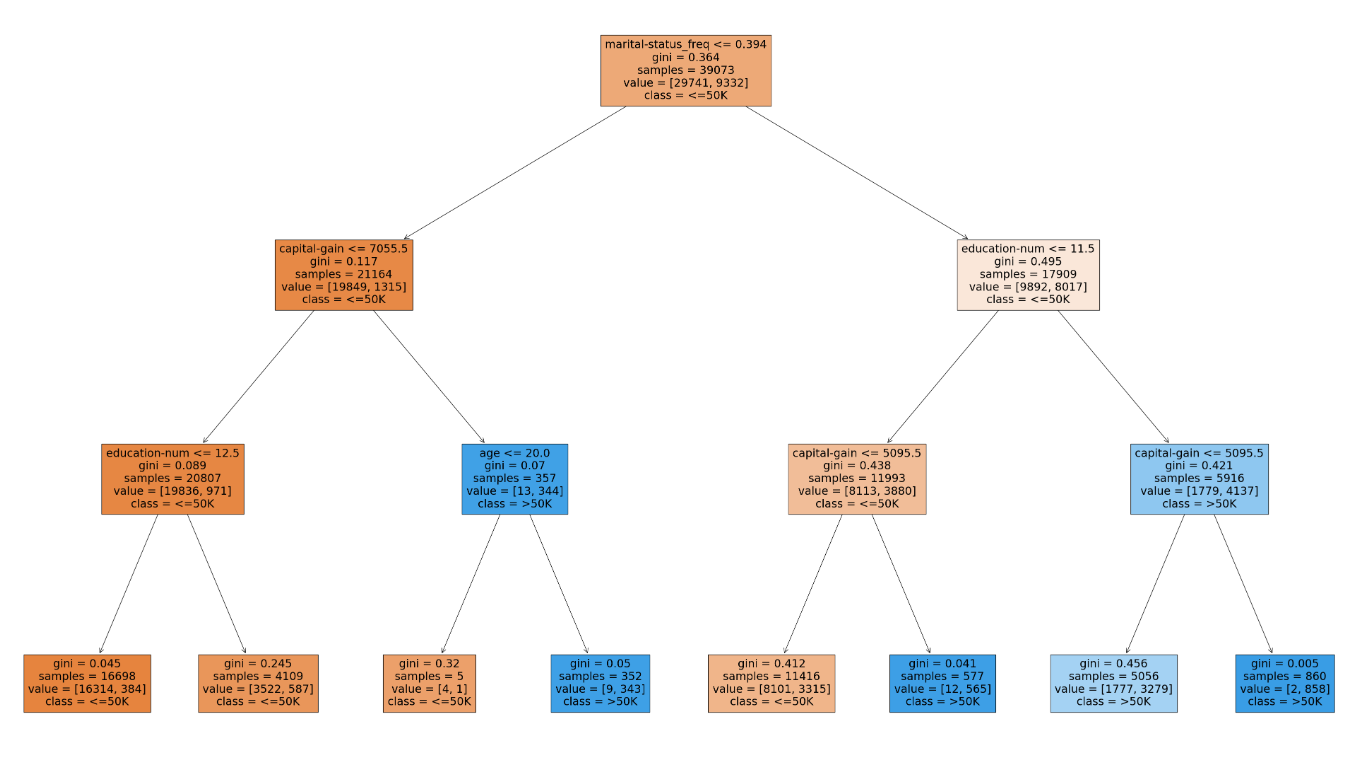
**Confusion Matrix:**

****

**Classification Report:**

****

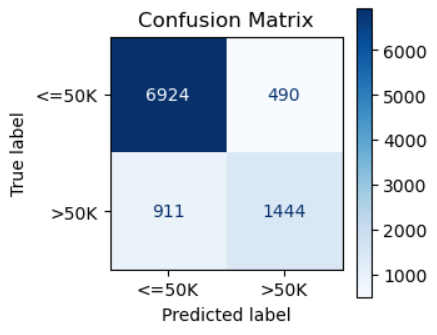
**Decision Tree Diagram:**

****

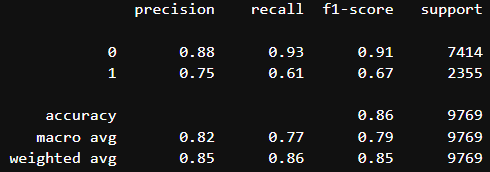
2. **Random Forest Classification:**

**Accuracy Score:** 0.8585320913092436

**Confusion Matrix:**

****

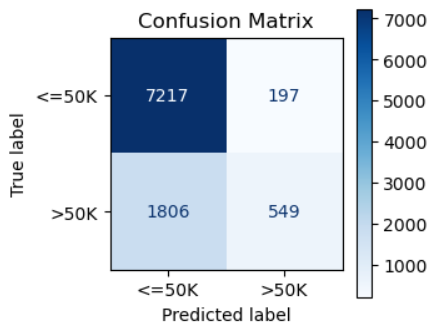
**Classification Report:**

****

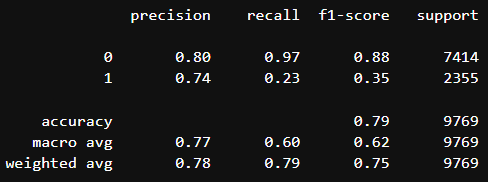
**3. KNN:**

**Accuracy Score:** 0.7949636605589109

**Confusion Matrix:**

****

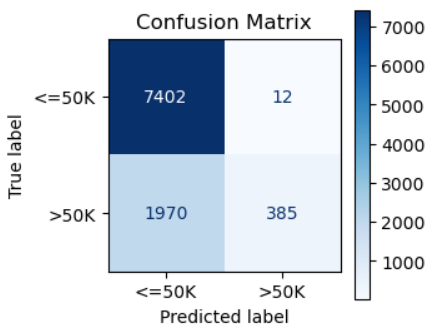
**Classification Report:**

****

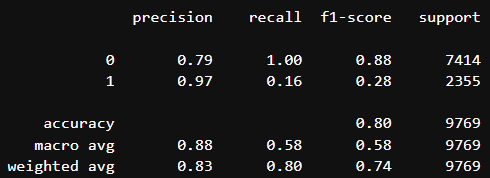
**4. Support Vector Machine:**

**Accuracy:** 0.7971133176374245

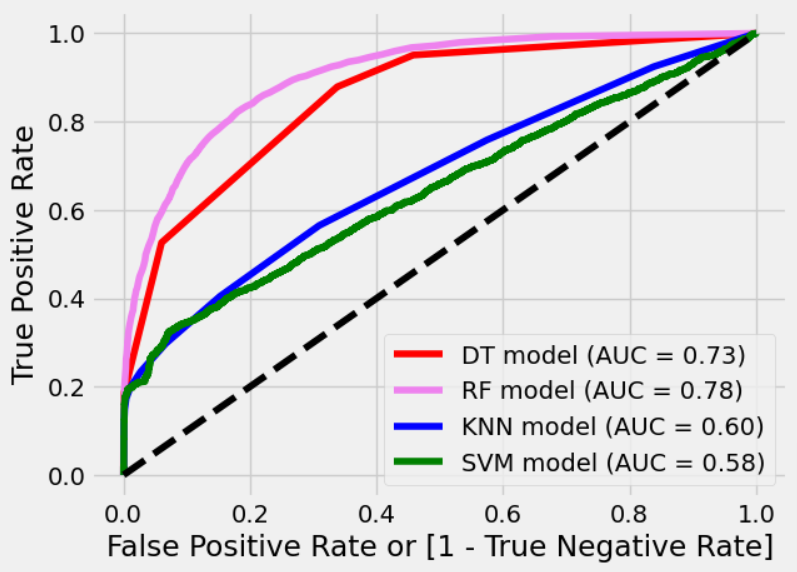
**Confusion Matrix:**

****

**Classification Report:**

****

**ROC Curve for these models:**

****

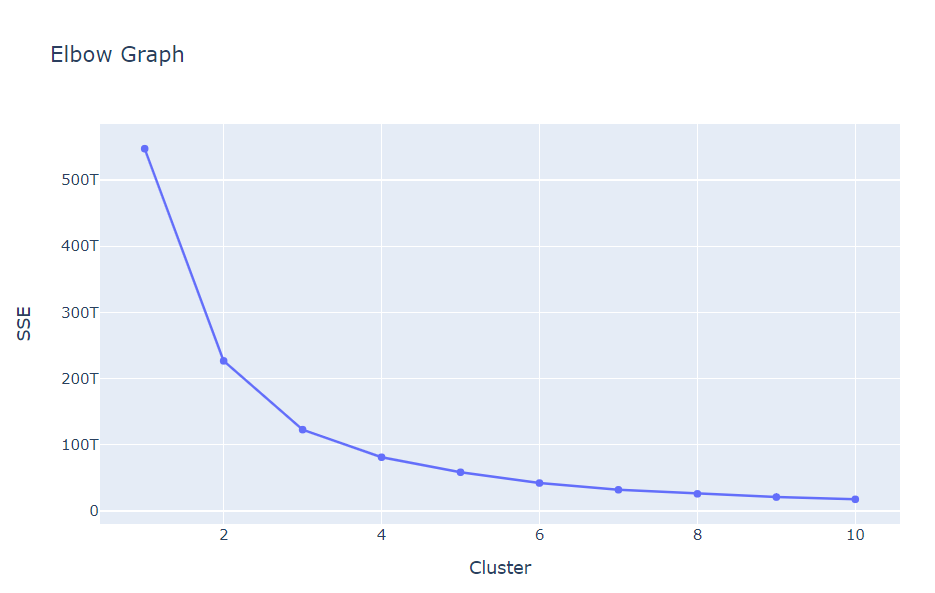
We have performed 4 classification models here out of these **Random Forest classifier** performed slightly better than other Models. But all the models have good Accuracy Score.

**Un Supervised / Clustering Model:**

**K-Means Clustering:**

**Cluster Selection:**

We have selected **2** clusters using **elbow plot**.

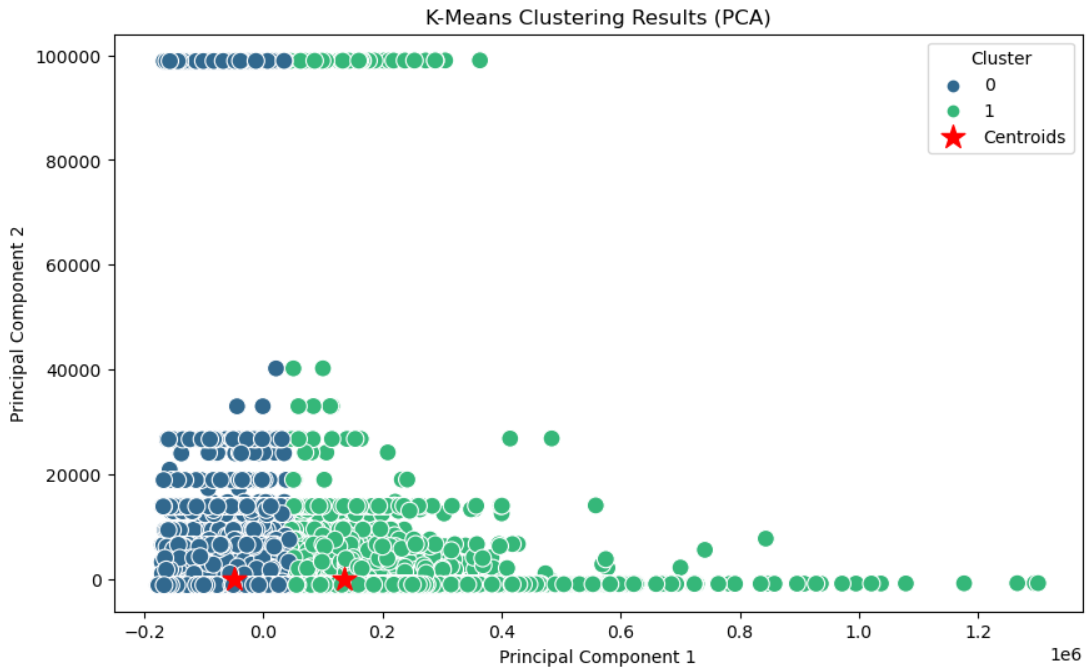


Like Classification Model we can’t evaluate clustering model with accuracy score we have to evaluate with **Silhouette score**.

**Silhouette Score:** 0.5838667124538208,

A reasonable structure has been found.

As we have many variables in Dataset which we are using we have done **Principal Component Analysis (PCA).**



* We have used 5 Models for this problem 4 Classification & 1 Clustering Model.
* Out of 4 Classification Model **(Decision Tree, Random Forest, KNN, SVM)** almost every model performed well though **Random Forest Classifier** has higher accuracy score than others.
* K-Means Clustering have the decent silhouette score for segmentation.

# **THANK YOU**

**Regards,**

**Shanmugagouthaman KR**

**Batch:** DTM – 10