Analysis of Banking Marketing Campaign using Machine Learning Algorithms

Shaishav Maisuria Zalak Panchal

1 Introduction

Advancement in technology has provided various technologies that have been useful in every sector. The algorithms used to create such unique and outstanding technologies have increased the efficiency and overall outcome of various industries. Similar when such technologies are applied to advertising campaigns, it provides companies with valuable information indicating the positive and negative return on investment. The process of collecting and analyzing the data is done using some incentives by these companies.

The following paper focuses on applying machine learning algorithms to understand the customer's behavior and find the features that provide the most prominent effect on the data. In addition, the conducting the studies is based on several stages such as data pre-cleaning, data pre-processing, data preparation, model selection, model application, testing the applied model, and compare selected models. The main goal of this study is to help companies choose customers based on desired features and applied machine learning techniques to obtain the most acquired return on investment. Therefore, making it much easier for companies to select customers that provides the maximum profit.

2 Data Handling

The original dataset consists of 21 features including target variables with the missing value(labeled by 'unknown').

2.1 Data Understanding

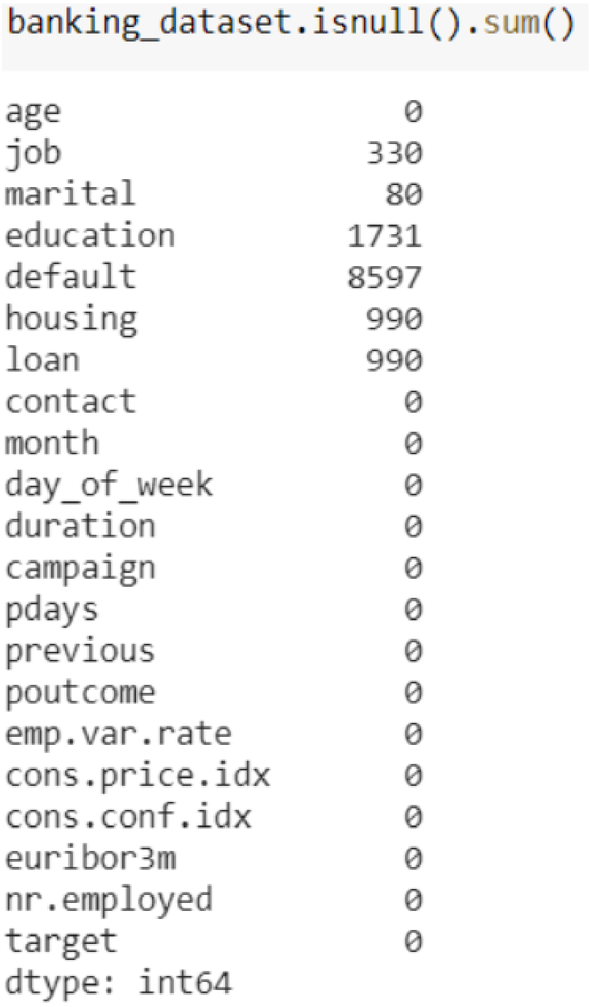
- Feature "age" has type number
- Feature "job" is of type categorical, which means it contains different classes inside it. The categorical feature present are entrepreneur, housemaid,student,admin,unemployed,self-employed,services,technician, unknown,management,retired,blue-collar
- Feature "marital" contains the marital status of type categorical. The categorical feature present are single, married, unknown, and divorced. The feature divorced can be considered as widowed or divorced.

- Feature "education" is type categorical that contains categories such as basic.6y, high.school, university.degree, basic.4y ,unknown ,illiterate ,professional.course ,basic.9y
- Feature "default", indicates the presence of credit and the feature type categorical. The categories present are yes as 1, no as 0, unknown
- Feature "housing", suggests the presence of a housing loan and the feature type categorical. The categories present are yes, no, unknown.
- Feature "loan", indicates the presence of a personal loan and the feature type categorical. The categories present are yes, no, unknown
- Feature "contact", indicates the type of contact information, and the feature is type categorical. The categories present are telephone, cellular
- Feature "month", shows when the customer's last contacted month in the past year and the feature is of categorical type. The categories present are jan,feb,mar,.....,dec
- Feature "day of the week", indicates when was the customer contacted week and the feature is type categorical. The categories present are mon,tue , wed ,thu,fri
- Feature "duration", indicates when was the customer contacted duration and the feature is type Numeric. The class it contains is seconds. One of the critical things to consider for this feature is that this feature is highly related to the target feature y; therefore, one often does not know the duration of call the customer will take, and one will know the outcome at the end of the call. Therefore, this feature must not be used in the model.
- Feature "campaign", indicates the amount of times the customer is contacted during the campaign, and the feature is type numeric. The class contains the last available contact
- Feature "pdays", indicate the number of days that have been passed without contacting the customer is contacted during the last campaign, and the feature is type numeric. Often numeric representation in the feature contains 999, which indicates the customer was not contacted before
- Feature "pervious" indicates the amount of times the customer was contacted before the current campaign, and the feature type is numeric.
- Feature "pervious" indicates the amount of times the customer was contacted before the current campaign, and the feature type is numeric.
- Feature "emp.var.rate" indicates the employment variation rate for quarterly indicator and the feature type is numeric.
- Feature "cons.price.IDX" indicates the monthly indicator for consumer price index, and the feature type is numeric.
- Feature "cons.conf.IDX" indicates the monthly indicator for consumer confidence index, and the feature type is numeric.

- Feature "euribor3m" indicates the daily indicator for Euribor 3 month rate, and the feature type is numeric.
- Feature "nr.employed" indicates the quarterly indicator for a number of employees, and the feature type is numeric.

2.2 Data preparation

First, the data cleaning is done by removing missing values from the dataset. Some of the columns like job, marital, education, default, housing, and loan have missing values. Their count as follows shown in Figure **??**



```
banking_dataset.isnull().sum()

age                  0
job                330
marital             80
education         1731
default           8597
housing            990
loan               990
contact              0
month                0
day_of_week          0
duration             0
campaign             0
pdays                0
previous             0
poutcome             0
emp.var.rate         0
cons.price.idx       0
cons.conf.idx        0
euribor3m            0
nr.employed          0
target               0
dtype: int64
```

Figure 1: Missing values inside Each Attribute

Rows with the missing values are dropped using the dropna() method. The dataset has reached the shape of from (41188, 21) to (30488, 21) by handling missing values.

As the next step, data processing is handling the categorical data. This banking dataset has almost eight features with categorical data. Dummy variables are created for categorical features to convert them to numerical. In order to avoid the interdependence of classes and for accurate classification of multi-class variables, the dummy variables are created. To avoid writing the different equations of the model, dummy variables were designed to create a model that helps obtain the utmost information. Therefore, it is required to obtain dummy variables for the dataset.

As this dataset is imbalanced, it is found that resampling is the best method to control the class distribution in the dataset. After splitting data into test and train, oversampling is used on xtrain and train to equal the distribution of the class. This study uses the scaling algorithm for all the models implemented in this study for the better performance of a model. The scaling algorithm selected for this study is Mini-Max Scaler. The reason for selecting this algorithm is to transform all the values inside the dataset between ranges (0,1). This algorithm preserves the shape of the dataset, therefore, allowing more datasets to train upon. In addition, this algorithm also handles any negative values very well by shrinking the dataset between ranges (-1,1) if there is a negative value in the dataset. The formula of how the algorithm works is provided in picture 2 (Pires, van Miguel et al 2020).

$$X_{new} = \frac{X\text{-}\ X_{min}}{X_{max}\text{-}X_{min}}$$

Figure 2: Min-Max Scaler formula

3 Methods/Case Study

For the binary classification, the following linear and non-linear machine learning approaches were attempted: Decision Tree Classification, Logistic regression, Linear Discriminative Analysis, K-nearest neighbor, Support Vector Machine(SVM and kernel-SVM).

Initially, the whole dataset was split up into two sets with a ratio of 70:30. The one set is used for the training purpose, to train the model, and the other set is used for the testing purpose, to check the model's performance on the unseen data. Through the training process, it has been made a use that there is no leakage of the testing data, therefore, ensuring the quality of testing data. The diagram shown in Figure 3 shows the model evaluation process:

After splitting the dataset, a model is created for each algorithm. Subsequently, the model is fitted on train data, and prediction is made on the test dataset. Accuracy is measured on the train and test dataset. Cross-validation is also used to validate the model's performance using

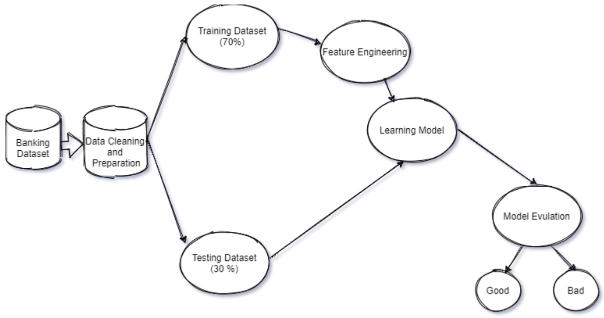
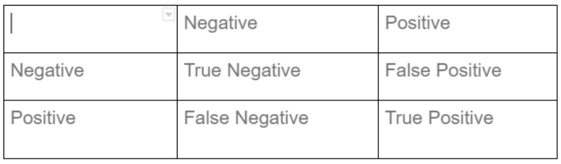


Figure 3: The architecture of model

K-fold cross-validation, followed by graphs for each model depicting the train and test accuracy.

Given Classification model can be interpreted using a confusion matrix. A confusion matrix is an n-dimensional matrix that consists of x-y labels such as the x-axis usually consists of Actual values, and the y-axis represents predicted values of the model. This study focused on obtaining a binary classification confusion matrix of size 2x2 dimension with the amount of true positive rate and false-positive rate. Model evaluation is an important process. In order to distinguish between the model performance and break ties between obtaining accuracy. This study includes focusing on factors like Precision, Recall. In order to properly understand the class distribution between negative and positive class classification effects on the dataset based on the selected model, this study has also selected an F-1 score to obtain a broader view of precision and recall on the dataset as shown in Figure 4.



	Negative	Positive
Negative	True Negative	False Positive
Positive	False Negative	True Positive

Figure 4: Table to obtain value from the Formula

Useful formulas in order to understand the calculated value in this study:

Total predicted positive Score =True positive + False positive

Precision = True positive/Total Predicted positive

Total Actual positive = True positive + False positive

NegativeRecall = True Positive /Total Actual Positive

F1-score = 2 * Precision * Recall / (Precision + Recall)

The study also uses the validation curve, one of the critical tools in order to obtain the measure of the sensitivity happening between the train and test accuracy for each model. The validation curves used in the study is using specific parameters obtained from the model and the model's actual score value. The sklearn library is used in order to perform cross-validation over different models with different cross-validation (cv) values. The aim for using validation curves is to obtain graphs results where the training and validation curves are close to each other. This study tries to tune the hyperparameters of the model in order to avoid cross-validation graphs having a training curve that has a high score relative to the validation score indicating the overfitting. The selected hyperparameters were tuned in order to minimize the overfitting of the model

The model evaluated during this study where describe the process and interpretation of obtained results. This study also provides a graph for all the model indicating the sensitivity among the changes happening inside models' accuracy over train and test datasets. This can be concluded for the models that are implemented consists of scores are not different or hugely varying while looking at classification reports and validation curves. In addition, the validation curve gives an indication that the dataset is not overfitted as there is a minor difference between the accuracies obtained from the results. The study also considers hyperparameter tuning in order to minimize the overfitting problem for each implemented model. The model evaluated during this study where describe with the process and interpretation of obtained results.

3.1 Linear discriminative analysis

On the basis of several explanatory variables, discriminant analyses are used to estimate the likelihood of a given categorical outcome (predictors).

In order to get the most detail out of the data and better match the dataset's binary classification, this method is used. The focus of this research is on the selection of features. Using the method RFE (Recursive Function Elimination) from the sklearn.feature selection library, the model is trained on the best selected feature.

To implement a linear Discriminant Analysis Model, LinearDiscriminantAnalysis() is used with solver SVD(Singular value Decomposition).In which,the cost function calculates the fit between a given matrix (the data) and an approximating matrix (the optimization variable), with the approximating matrix having a reduced rank as a constraint.

Figure 5 yielded the following result. It can easily understand the false positive and true negative rates by examining at the observed confusion matrix. The F1-score suggests that "no" is selected with stronger precision and recall with a value of 0.93 and "1" is selected with a value of 0.41. The outcome for this model illustrates the correct prediction made by the model 7624+397=8021 and incorrect predictions 744+382=1128.

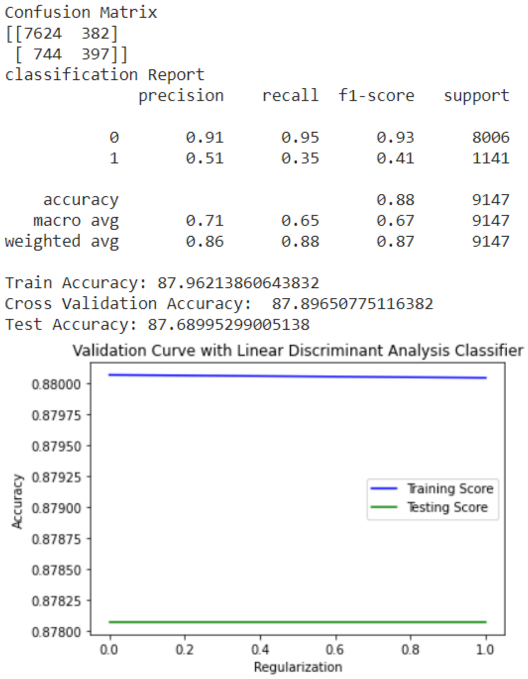
3.2 Logistics Regression

Logistic regression is a regression analysis used when the target variable is binary. It is a predictive analysis. It explains the relationship between one dependent binary variable to another independent variable.

LogisticRegression() is used with hyperparameters in this method. Values were chosen for the hyperparameters in our model, with maxIter set to 20000 and C set to 0.01 value.

The displayed confusion matrix helps us to understand the false positive rate and true negative rate.

Based on image 6, The outcome of this model aids in understanding the model's correct prediction (7911+202=8113) and incorrect predictions (939+95=1034). An F1-score of 0.94



```
Confusion Matrix
[[7624  382]
 [ 744  397]]
classification Report
              precision    recall  f1-score   support

           0       0.91      0.95      0.93      8006
           1       0.51      0.35      0.41      1141

    accuracy                           0.88      9147
   macro avg       0.71      0.65      0.67      9147
weighted avg       0.86      0.88      0.87      9147

Train Accuracy: 87.96213860643832
Cross Validation Accuracy:  87.89650775116382
Test Accuracy: 87.68995299005138
```

Figure 5: Result of Linear discriminative analysis

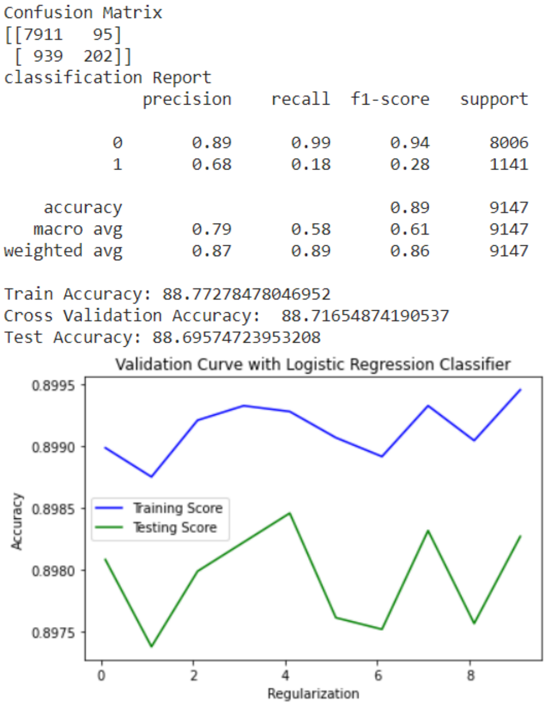
indicates that "no" is selected with greater precision and recall, while "1" is chosen with an f-1score of 0.28.

3.3 Decision Tree Classification

DecisionTreeClassifier is a supervised learning algorithm. By learning basic decision rules inferred from preliminary data, the DecisionTreeClassifier is a supervised learning algorithm. By learning basic decision rules inferred from primary data, the Decision Tree is to construct a training model that can be used to predict the class or value of the target variable (training data).

DecisionTreeClassifier() is used in this process, along with hyperparameters from the sklearn library. Values chosen for the hyperparameters in our model, with max depth set to two.

Figure 7 yielded the following result: It can better appreciate the false positive and true negative rates by looking at the shown confusion matrix. The model's correct prediction (7645+486=8131) and incorrect predictions (655+361=1016). An F1-score of 0.94 indicates that "no" is selected with greater precision and recall, while "1" is selected with an f-1score of 0.49.



```
Confusion Matrix
[[7911   95]
 [ 939  202]]
classification Report
              precision    recall  f1-score   support

           0       0.89      0.99      0.94      8006
           1       0.68      0.18      0.28      1141

    accuracy                           0.89      9147
   macro avg       0.79      0.58      0.61      9147
weighted avg       0.87      0.89      0.86      9147

Train Accuracy: 88.77278478046952
Cross Validation Accuracy:  88.71654874190537
Test Accuracy: 88.69574723953208
```

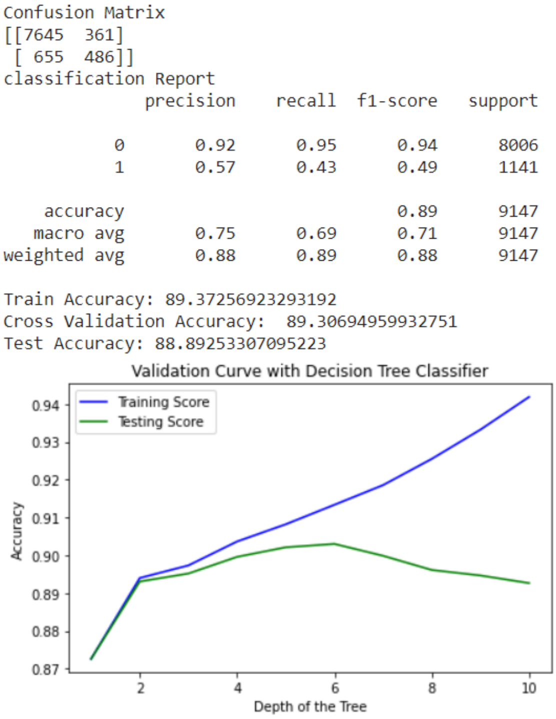
Figure 6: Result of Logistics Regression

3.4 K-nearest Neighbours Classification

K-nearest neighbor is a non-parametric classification. The k closest training examples in a dataset serve as the input. The output, in this case, is a class membership. By measuring the distance with various distance metrics, assigning weights to the contributions of the neighbors so that the neighbors who are closer contribute more to the average than those who are farther away.

Values specified for the hyper parameters in our model, with n neighbours set to a dynamic k neighbor value derived from the k-values graph figure 8 and graph figure 9. The knn accuracy rate versus k value is made up of accuracies obtained from all different k values, as shown in figure 8, while the knn error rate versus k value is made up of errors obtained from all different k values, as shown in figure 9. The model is chosen for the values with the highest accuracies and the lowest error.

It can better appreciate the false positive and true negative rates by looking at the shown confusion matrix in figure 10. The result of this model assists in understanding the model's correct prediction of (7927+179=8106) and incorrect predictions of (962+79=1041). F1-score of 0.94 implies that "no" is selected with better precision and recall, while "1" is selected with an f-1score of 0.26.



```
Confusion Matrix
[[7645  361]
 [ 655  486]]
classification Report
              precision    recall  f1-score   support

           0       0.92      0.95      0.94      8006
           1       0.57      0.43      0.49      1141

    accuracy                           0.89      9147
   macro avg       0.75      0.69      0.71      9147
weighted avg       0.88      0.89      0.88      9147

Train Accuracy: 89.37256923293192
Cross Validation Accuracy:  89.30694959932751
Test Accuracy: 88.89253307095223
```

Figure 7: Result of Decision Tree

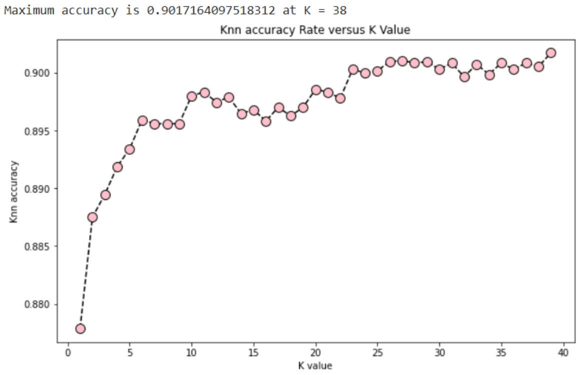
3.5 Support Vector Machine

The support vector machine algorithm's goal is to find a hyperplane in an N-dimensional space (N is the number of features) that distinguishes between data points.There are several hyperplanes from which to choose to distinguish the two types of data points. The aim is to find a plane with the greatest margin, or the greatest distance between data points from both groups. Maximizing the margin gap provides some reinforcement, making it easier to classify potential data points.Hyperplanes are decision boundaries that aid in data classification.The C parameter specifies how harshly you want your model to be punished for each misclassified point on a given curve.As c is a smaller value, Noisy points have a minor effect. And if there are any misclassifications, planes that effectively distinguish the points will be discovered.

3.5.1 SVM

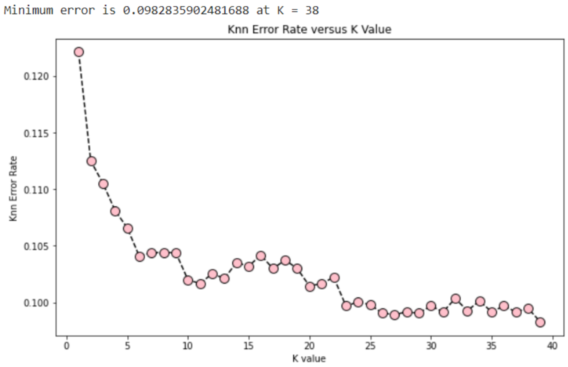
The SVM() library is used with hyperparameter C;regularization parameter set to 0.1 and probability is True.

By looking at the shown confusion matrix in figure 11, it can better understand the false positive and true negative rates. The model's correct prediction of (7887+211=8098) and incorrect prediction of (930+119=1048) are explained by the model's outcome. With an F1-score of 0.94, "0" is chosen with greater precision and recall, while "1" is chosen with an F1-score of 0.29.



```
Maximum accuracy is 0.9017164097518312 at K = 38
```

Figure 8: KNN-Accuracy Versus K values



```
Minimum error is 0.0982835902481688 at K = 38
```

Figure 9: KNN-Error Versus K values

3.5.2 Kernel-SVM

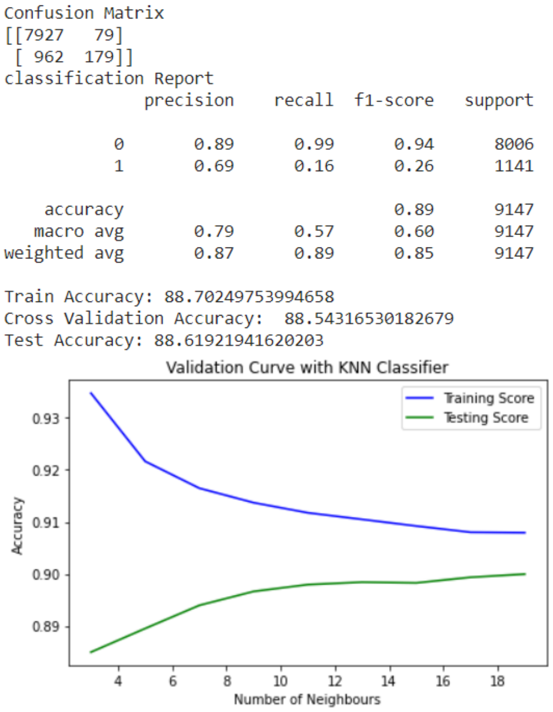
Here, SVM is used as a kernel function with the linear kerne and the regularization parameter C=0.1.

Here are the findings of the algorithm shown in figure 12

The precision and recall for the correct prediction is (7864+232)= 8096 and for incorrect prediction is (909+142=1051) ,with the f1 score for 0.94 and for 0.31 class.

This study also provides a graph for all the model indicating the sensitivity among the changes happening inside models' accuracy over train and test datasets. This can be concluded for the models that are implemented consists of scores are not different or hugely varying while looking at classification reports and validation curves. In addition, the validation curve gives an indication that the dataset is not overfitted as there is a very minor difference between the accuracies obtained from the results.

Based on the images, one can observe the accuracy obtained from the application of the model. This accuracy indicates how the model is properly fitted to the dataset. Where the Training accuracy is the highest followed by cross-validation accuracy and Test accuracy.



```
Confusion Matrix
[[7927   79]
 [ 962  179]]
classification Report
              precision    recall  f1-score   support

           0       0.89      0.99      0.94      8006
           1       0.69      0.16      0.26      1141

    accuracy                           0.89      9147
   macro avg       0.79      0.57      0.60      9147
weighted avg       0.87      0.89      0.85      9147

Train Accuracy: 88.70249753994658
Cross Validation Accuracy:  88.54316530182679
Test Accuracy: 88.61921941620203
```

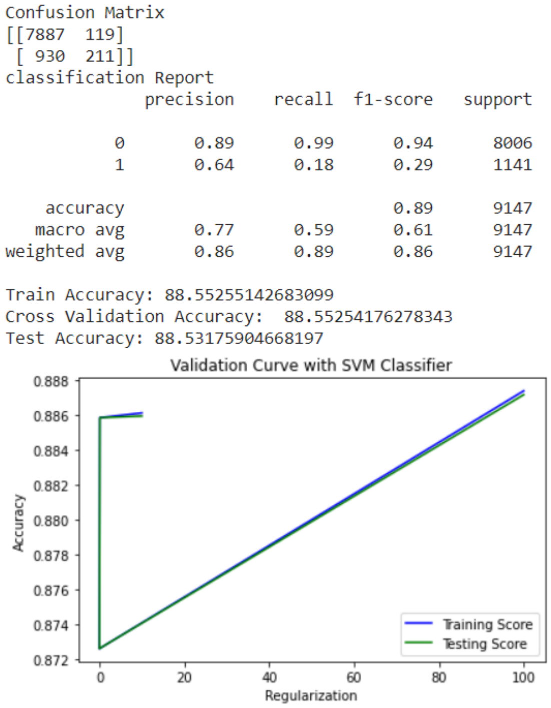
Figure 10: Result of K-nearest Neighbours Classification

4 Results and Discussion

The study's main goal is to use machine learning algorithms to choose a customer before making a call. One of the most important and time-consuming aspects of the analysis was the data cleaning and preparation. The selection of an algorithm and the construction of a model come next. The algorithms used in this analysis were chosen based on the dataset. The dataset primarily consists of categorical and numerical attributes for certain functions.

Hyper-parameter tuning, in which each model is calibrated based on the need to build a model that is not overfitted, was one of the main aspects that needed to be focused on when conducting the analysis. Decision Tree, Logistic Regression, Linear Discriminative Analysis, K-Nearest Neighbors, SVM, and SVM with Kernel are the algorithms chosen.

The findings shown in Figure 13: obtained based on the implemented model assist us in determining which model will be better for this dataset. However, this analysis takes into account the fact that no model is flawless and that the findings obtained for the dataset are correct.Based on the available proof, the decision tree has the highest percentage of correct predictions and the lowest percentage of incorrect predictions in a single test dataset. Furthermore, the F-1 score for class 1 is the highest of all other models; nevertheless, the F-1 score for class 0 is almost identical to the rest of the models. As a result, a decision tree is most likely to be used for the dataset. To find the dataset one can obtain it from the following url `https://archive.ics.uci.edu/ml/machine-learning-databases/00222/`.



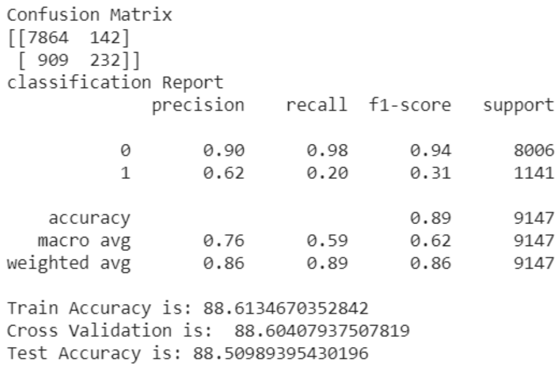
```
Confusion Matrix
[[7887  119]
 [ 930  211]]
classification Report
              precision    recall  f1-score   support

           0       0.89      0.99      0.94      8006
           1       0.64      0.18      0.29      1141

    accuracy                           0.89      9147
   macro avg       0.77      0.59      0.61      9147
weighted avg       0.86      0.89      0.86      9147

Train Accuracy: 88.55255142683099
Cross Validation Accuracy:  88.55254176278343
Test Accuracy: 88.53175904668197
```

Figure 11: Result of Support Vector Machine without Kernel

The AUC–ROC curves used in this study helps to visualize how well the machine learning classification is done by each models. ROC helps us to show the performance of of each classification model using True positive rate and False Positive Rate. Whereas, AUC stands for Area Under the ROC Curve. Is usually helps measure the ROC cure area underneath from range (0,0) to (1,1). A classifier with a high AUC may often perform worse in a particular region than one with a lower AUC. However, Figure 14 shows knn has low area and is a better. Which also makes the statement is true that no algorithm is perfect based on one certeria.From the standpoint of precision, the decision tree is considered to be the winning algorithm with the maximum accuracy of 0.8937256923293192 and F1Score.However, The roc curve, on the other hand, indicates that KNN is also a better algorithm for this dataset since its auc is lower than the other algorithms.

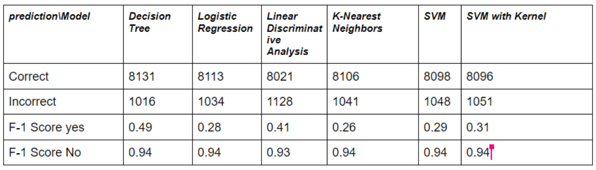
5 Conclusion

Despite the fact that neither model produced perfect results, some of the algorithms were tried and found to be the best fit for the data after various methodology for the data preparation. One algorithm in particular, a decision tree, performs significantly better on this banking dataset. Based on the above findings, it is reasonable to conclude that models with this large amount of data will produce far more sophisticated results. This project is noteworthy because it goes deeper into the basics of machine learning algorithms. The project's main goal is to find the



```
Confusion Matrix
[[7864  142]
 [ 909  232]]
classification Report
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      8006
           1       0.62      0.20      0.31      1141

    accuracy                           0.89      9147
   macro avg       0.76      0.59      0.62      9147
weighted avg       0.86      0.89      0.86      9147

Train Accuracy is: 88.6134670352842
Cross Validation is:  88.60407937507819
Test Accuracy is: 88.50989395430196
```

Figure 12: Result of Support Vector Machine with Kernel



prediction\Model	***Decision Tree***	***Logistic Regression***	***Linear Discriminative Analysis***	***K-Nearest Neighbors***	***SVM***	***SVM with Kernel***
Correct	8131	8113	8021	8106	8098	8096
Incorrect	1016	1034	1128	1041	1048	1051
F-1 Score yes	0.49	0.28	0.41	0.26	0.29	0.31
F-1 Score No	0.94	0.94	0.93	0.94	0.94	0.94

Figure 13: Overall Model Results

best-fit model well with highest accuracy while avoiding overfitting.

Reference

Pires, van Miguel et al. “Homogeneous Data Normalization and Deep Learning: A Case Study in Human Activity Classification.” Future internet 12.11 (2020): 1–. Web.

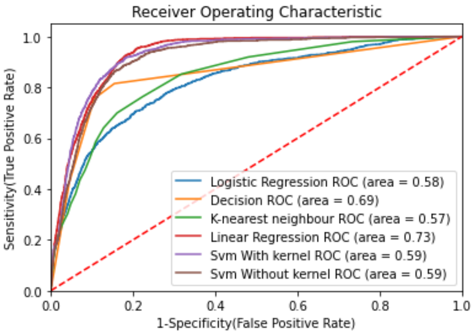


Figure 14: Overall Roc-Auc Curve for All model