CE802 - Machine Learning and Data Mining

Assignment: Design and Application of a Machine Learning System for a Practical Problem

University of Essex
School of Computer Science and Electronic Engineering

REPORT

Submitted by
Arun Ravi
Registration number: 2009918
MSc – Artificial Intelligence & its Applications

Word count: 808

1. Study of "Tosco & Spency"

The Supermarket chain wants to predict if their customers will buy few expensive products or many cheap products. This task is supported by providing the past data that comprises various information related to the customer. Hence this is a classification problem.

1.1 Data Preparation

Two data files (.csv) were given with one of them containing features and target values for training and testing and the other data file without the target values for prediction.

The dataset (CE802_P2_Data) consists of 15 features in which 14 had complete information and F15 had 50% missing values but the value distribution was normal. Hence to take all the features into account while building the model, the missing values in F15 was replaced with the mean of column F15.

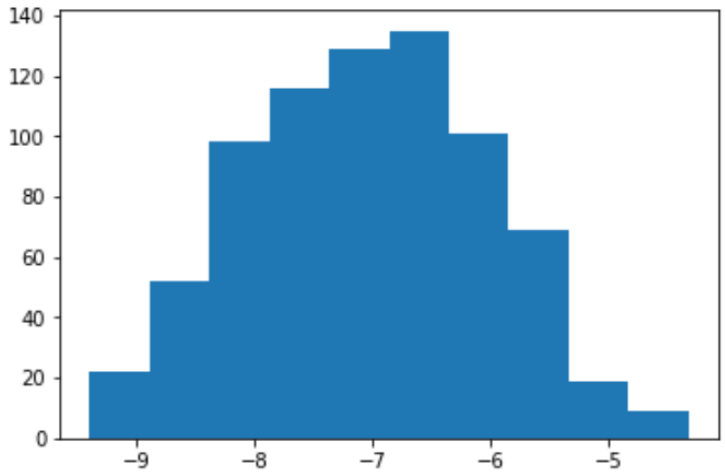


Fig 1. Value distribution of F15 (Normal distribution)

The given dataset is experimented with different Machine learning models and find the best one with highest possible accuracy. To evaluate the best model, the dataset is split into training set and testing set at a ratio of 7:3.

1.2 Various ML models:

1.2.1 Decision Tree Classifier:

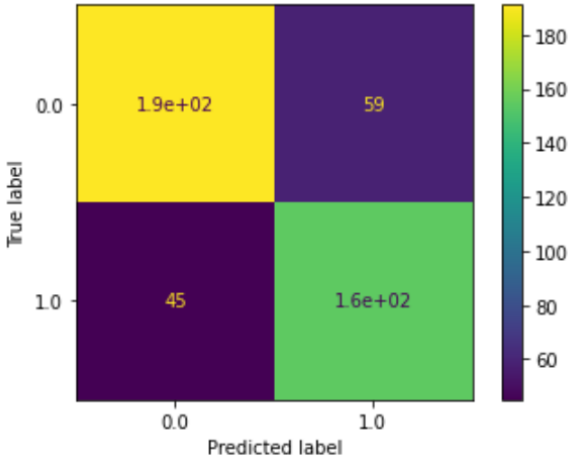
This is the simplest and most effective model to a binary classification problem. The criterion is set to “gini” which will partition the data recursively using greedy algorithm. Decision tree model can also efficiently handle missing values and feature engineering. Following figures shows the Precision and Recall table and the confusion matrix for Decision Tree classifier.

	precision	recall	f1-score	support
0.0	0.81	0.76	0.79	250
1.0	0.73	0.78	0.75	200
accuracy			0.77	450
macro avg	0.77	0.77	0.77	450
weighted avg	0.77	0.77	0.77	450

```
print(confusion_matrix(y_test, y_pred))
```

```
[[191  59]
 [ 45 155]]
```

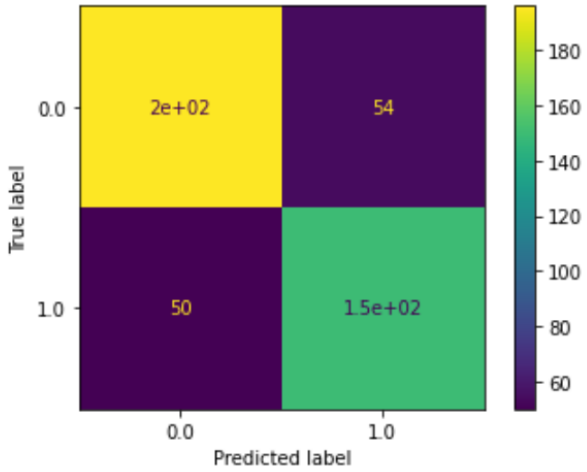
```
#plotting Confusion matrix
plot_confusion_matrix(classifier, X_test, y_test)
plt.show()
```



1.2.2 Random Forest Classifier:

It is an intuitive model and performs well because of its stochastic approach. It builds multiple trees and combines them together to examine. It gave better result than the Decision tree model. The following figures shows the Precision and Recall table and the confusion matrix for Random Forest classifier.

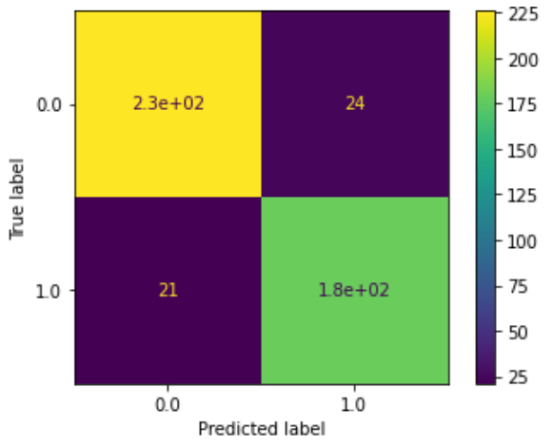
	precision	recall	f1-score	support
0.0	0.82	0.74	0.78	250
1.0	0.71	0.79	0.75	200
accuracy			0.76	450
macro avg	0.76	0.77	0.76	450
weighted avg	0.77	0.76	0.77	450



1.2.3 Support Vector Machine (SVM):

These models are majorly used for classification problem in which each data get plotted on an n-dimensional space with value of each feature and the hyper-plane classifies the data. The following figures shows the Precision and Recall table and the confusion matrix for Support Vector Machine.

	precision	recall	f1-score	support
0.0	0.91	0.90	0.91	250
1.0	0.88	0.90	0.89	200
accuracy			0.90	450
macro avg	0.90	0.90	0.90	450
weighted avg	0.90	0.90	0.90	450



1.3 Evaluation:

Three different classifier Machine learning models were built and same dataset had been used for training and testing all three models. The Support Vector Machine (SVM) clearly performed better than the Random forest and Decision tree model. The SVM is the most stable model for this classification problem and is used to predict the classification in test data (CE802_P2_Test)

2. Study of Sunsbory's:

Previously, we tried to predict whether the consumer will be buying few expensive products or many cheap products. In this study, we will try to estimate the amount of money that the customer will spend. Hence it is a regression problem.

2.1 Data Preparation:

Two data files (.csv) that has 16 features were given with one of them containing target values for training and testing and the other data file without the target values for prediction. Features like F5 and F6 were encoded using one-hot encoder and Label encoding to convert them into numerical values.

2.2 Various ML models:

2.2.1 Artificial Neural Network (ANN):

This computational Algorithm consists of large number of simple elements called perceptron or neuron. A sequential Artificial Neural Network has been used that has one input layer and 3 hidden layers. The first and third layer has 50 neurons each and the second layer has 25 neurons. The input layer will have 18 neurons. To build this ANN, a library called Keras has been used along with the Adamax optimizer and relu activation function. The data has been trained on this ANN with 700 epochs. The metrics score for ANN is:

```
Mean Absolute Error: 55.16422848410871
Mean Squared Error: 21123.938567920195
Root Mean Squared Error: 145.34076705425838
```

2.2.2 Linear Regression:

It is a statistical method to find a relationship between the dependant and independent variable. This linear regression statistic model was built using sklearn without parameters. The metrics score for Linear regression is:

```
Mean Absolute Error: 378.18599521050226
Mean Squared Error: 251087.7149777833
Root Mean Squared Error: 501.08653442073586
```

2.2.3 Random Forest Regressor:

It is an ensemble machine learning technique in decision tree. The max_depth and n_estimators parameters are set to 10 and 100 respectively. The metrics score of Random Forest Regressor is:

```
Mean Absolute Error: 440.9022450283072
Mean Squared Error: 356742.69037577574
Root Mean Squared Error: 597.27940729258
```

2.3 Evaluation:

Three different regression Machine Learning models were built and the same dataset (CE802_P3_Data) was used to train and test all the three models. Artificial Neural Network performed better with lowest Mean Absolute Error and better R2 value than the other two models. Hence this model was chosen to predict the Target value for the test data (CE802_P3_Test)

Google colab platform was used to develop the python notebook.

The github repository contains all the data files used, Python notebook and the output csv files: https://github.com/arunravi8595/CE802