Cardiff School of Computer Science and Informatics

CMT307

Applied Machine Learning

Implementation and Evaluation of a Case Study

Using

Machine Learning Techniques

Report Presented by

Sai Kalyan Kalluri (21109714)

Student ID: 21109714

Contents

Course Work Question 1

Course Work Question 2

Student ID: 21109714

Question 1:-

The below mathematic algorithm gets the following results in a classification experiment, where in the table, 'Id' is the index number, 'Target' is the ground truth that the classifier aims to achieve, 'Prediction' is the predicted results. Below we are going to calculate the "Confusion Matrix", "Precision", "Recall", "f1-measure" and "Accuracy" methods manually with respective steps and formulas.

Id	Target	Prediction
1	True	True
2	True	True
3	True	False
4	True	True
5	True	True
6	True	False
7	True	True
8	True	True
9	True	True
10	False	False
11	False	False
12	False	False
13	True	True
14	True	False
15	True	True
16	False	False
17	False	False
18	False	True
19	False	True
20	False	False

Given variables from the above classification experiment:-

- 'Id'- It is the index number
- 'Target'- It is the ground truth that the classifier aims to achieve
- 'Prediction' - It is the predicted results

Methods I am going to prove manually from the above example are:-

a) Confusion Matrix
b) Precision
c) Recall
d) F1-measure
e) Accuracy

Student ID: 21109714

a) Confusion Matrix:-

Generally, we define the Confusion matrix as a method for defining the performance of a classification algorithm where the output can be two or more variables. It is the table consists of four different combinations of predicated and actual class values.

	PREDICTED CLASS	
Actual Class	***True Positive (TP)***	***False Negative (FN)***
	False Positive (FP)	***True Negative (TN)***

True Positive (TP):- It represents the positive values that were accurately categorized by the classifier.

True Negative (TN):- It represents the negative values that were correctly labeled by the classifier.

False Positive (FP):- It represents the negative values that were incorrectly labeled as positive.

False Negative (FN):- It represents the positive values that were mislabeled as negative.

b) Precision:-

Precision is the ratio between the True Positives and all the Positives or we can also define it as a measure of being accurate.

Below is the formula for calculating Precision:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

C) Recall:-

The recall is nothing but the number of the correct positive forecast made out of all positive forecasts that could have been made.

$$Recall = \frac{True_{positive}}{True_{positive} + False_{negative}}$$

Student ID: 21109714

D) F-Measure:-

F measure is a union of Precision and Recall.

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall + Precision} = \frac{2p \times r}{p + r}$$

E) Accuracy:-

It is derived as the ratio between the no of accurately predicted values to the total no of predicted values

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}}$$

Student ID: 21109714

a) Confusion Matrix:-

	PREDICTED CLASS			TOTAL POINTS
ACTUAL CLASS		True	False	
	True	9 (TP)	3 (FN)	12
	False	2 (FP)	6 (TN)	8

(b) Precision:-

$$\text{Precision (P)} = \frac{TP}{TP+FP}$$

$$= \frac{9}{9+2} = \frac{9}{11}$$

∴ Precision (P) = 0.81 (or) 81%.

(c) Recall:-

$$\text{Recall (r)} = \frac{TP}{TP+FN} = \frac{9}{9+3} = \frac{9}{12} = 0.75$$

∴ Recall (r) = 0.75 (or) 75%

(d) F-1 Measure

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$= \frac{2 * 0.81 * 0.75}{0.81 + 0.75} = 0.77$$

∴ F Measure (F) = 0.77 (or) 77%.

Student ID: 21109714

(e) Accuracy:-

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{9+6}{9+6+3+2} = \frac{15}{20} = \frac{3}{4} = 0.75$$

Accuracy = 0.75 (or) 75%

Student ID: 21109714

Question:- 2

In this problem statement, we have a large set of data belonging to e-commerce. We are going to develop machine learning models to predict e-commerce visitors' purchasing intention. The data contains shoppers' online activity information including "clickstream" and "session information data", where the last column "Revenue" represents visitors purchasing intention. My initial tasks will include data exploration, followed by data pre-processing, machine learning method selection, Implementation, and model performance evaluation.

1. **Data exploration.**
2. **Data pre-processing.**
3. **Model implementation**
4. **Performance evaluation.**
5. **Results analysis and discussion.**

1. Data Exploration in Machine Learning:-

Data exploration is one of the crucial steps in which we will be interacting with real-time huge data. It is the beginning step in data analysis where we use data visualization and statistical methods to describe data set characters such as input and output variables, size and quantity of data, accuracy for clearly understanding the nature of the data.

As it is the initial step, it includes some of the below goals.

a) Importing Libraries.
b) Importing and Loading data.
c) Data Features (Basic information about the data, Identification of variables and data type, Numerical and Categorical features from the data).
d) Finding missing data.
e) Data Inspection (Scatter, Box, and Histogram Plots for the data).
f) Correlation analysis.
g) Class Imbalance.

A) Importing the Libraries:-

The most popular programming language for data science and Machine Learning are Python and R, both are highly flexible, open-source, data analytics languages. Python is considered as one of the foremost options for machine learning for its flexibility. I used python in this model as it has a huge number of

libraries in which some of which play a vital role while implementing algorithms in Machine Learning. Some of the common python libraries I used in the course work for machine learning are "**NumPy"**, "**Pandas"**, "**Matplotlib", and "Seaborn".**

B) Importing and loading the dataset

Before beginning any of the Machine Learning model (or) project the primary and utmost vital issue is Importing/loading information. We need to load the data for starting any ML projects. Concerning data, the familiar type of data for ML projects is CSV (comma-separated values), a simple file format used to store tabular data (number and text data in plain text).

These are three common approaches in Python to load CSV data files is

- Loading CSV file with Python Standard Library
- Loading CSV with NumPy
- Loading CSV with Pandas

There are many ways to load ML data in python. From among the three, we are going to load data CSV files using the python library "pandas" data frame. The approach to load CSV data file by Pandas is **"pandas.read_csv()function**".

This is the very user-friendly statement that returns "pandas.DataFrame" which can be used immediately for plotting. We are importing a CSV data file by using the below command.

```
dataset = pd.read_csv('Dataset.csv')
```

dataset - Variable.

pd - Calling pandas library

read.csv - Reads the loaded CSV file

Data.csv - Data file name and .csv is the format

.csv - CSV files contain plain text and are a well-known format that everyone can read, including Pandas.

C) Data Features:-

Data Features are one of the important categories which describes information about data (Describing category of each feature). In Machine learning, Data features are classified into two types

Categorical:- These features can only take on a limited, and usually fixed, several possible values. It represents a character or feature of a data object For example, if a dataset is about information related to users, then you will typically find features like Customer_id, Name, Age group, etc.

Numerical: it will represent numbers or Quantity (integer or real-valued) related to mathematical operations.

D) Finding the Missing Data:-

In real-time data, there are some cases where a particular attribute is missing because of various reasons, such as corruption of data or failure to load the available information, or incomplete extraction. There are several methods to deal with missing data, before implementing the methods we have to make sure to choose the most effective one for obtaining accurate values. In my implemented model, there were no missing values.

E) Data Inspection:-

Data inspection is nothing but the act of viewing data and verifying it carefully by examining the data and need to ensure that we are going to deal with the right information for getting a clear picture of the data at every stage of the transformation process. We can also see how the data looks like and what kind of relationship is held by the attributes of data. Also, we can understand each attribute of our dataset independently.

The following are some techniques that I used in my ML model:-

Scatter Plot:-
A scatter plot is a graph in which each data point is represented by a dot. I defined the characteristic features between them in my machine learning model. I have characterized the attribute features between "Administrative Duration vs Duration", Informational Duration vs Bounce Rates, etc.

Histogram:- Histogram Plot is represented for each value of the data set is represented by a bar. In this model, I have represented Histograms for all the feature columns separated by prediction label value. I used the Python module "Matplotlib" for the representation of the histogram for all the features. It also helps us to see possible outliers. Overall, the graph display tabulated frequencies, shown as bars. x-axis values, y-axis values represent frequencies.

Box Plot:- It is one of the useful methods to review the distribution of each attribute. It is univariate and summarizes the distribution of each attribute. The dots outside the whiskers signify the outlier values.

F. Correlation:-

Correlation is an indication of the changes between two variables. These variables can be input data features that have been used to forecast our target attribute. It gives us a clear idea about the degree of the relationship between the two variables. We can plot the correlation matrix to show which variable is having a low or high correlation concerning another variable. In the built model I represented correlation using python model seaborn with Heatmap () function which is defined as shading matrix used for graphical representation of data using colors to visualize the value of the matrix. Few features are not correlated with our prediction of target "Revenue". Hence dropped those redundant features.

G. Class Imbalance:-

In machine learning, the class imbalance is a typical issue, Particularly in classification issues. It is one of the most crucial Machine Learning principles. There is a class imbalance when one class's observation is higher than other classes' observation. Imbalanced data might influence model accuracy. I tested the class imbalance in my model and discovered that there is a class imbalance of 15.47 to 84.53 percent in favour of False.

2) Data Pre-processing:-

It is not always the case that we come across tidy and prepared data in real-world circumstances. We occasionally get data with noise, missing values, or in an unsuitable format that cannot be immediately utilized to train a machine learning model. It is necessary to clean and prepare data before doing any action with it. Pre-processing data is an important step in cleaning and preparing data for a machine learning model since it enhances the model's accuracy and efficiency.

Using the dataset to create Training and Test sets:-

We divide our dataset into a training set and a test set in machine learning data preparation. By doing so, we may increase the performance of our machine learning model. Assume we have trained our machine learning model on one dataset before setting it to the test on another. Our model will therefore struggle to comprehend the links between the models. The model's performance will deteriorate if we train it adequately and its training accuracy is good, but then give it a new dataset. As a result, we constantly strive to create a machine learning model that works well with both the training and test datasets.

```
[118] from sklearn.model_selection import train_test_split
      from sklearn.feature_extraction.text import CountVectorizer
      from sklearn.preprocessing import Normalizer
```

```
[117] x = cleaned_data.drop("Revenue", axis=1)
      y = cleaned_data['Revenue']
      X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)
```

- **x train: The initial sequence's training phase (x)**
- **x test: The first sequence's test section (x)**
- **y _train: The second sequence's training phase (y)**
- **y test: The second sequence's test section (y)**

3. Model Implementation:-

In Machine Learning, there are many algorithms available. I have chosen Decision Tree, which is one of the finest Supervised learning techniques. It can be used to solve both classification and regression issues, however, it is more commonly used to solve classification problems. It's a graphical depiction for obtaining all feasible answers to a problem/decision depending on certain parameters. Internal nodes contain dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier.

The following are the reasons why I chose the Decision Tree:

- ✓ Decision Trees are simple to grasp since they replicate human thinking abilities while making a decision. It is easy to comprehend since it follows the identical steps that a human would take while choosing real life.
- ✓ It has a tree-like structure, and it follows the same procedure that a human would take while deciding on real life.
- ✓ It may be quite helpful in addressing decision-making challenges and in considering all of the possible possibilities for a situation.
- ✓ In comparison to other methods, data cleansing is not required as much.

Linear SVC is the second classification method employed in my model. Linear SVC is the machine learni ng method that is best appropriate to our case. A Linear SVC (Support Vector Classifier)'s goal is to fit the data you input and produce a "best fit" hyperplane that divides or categorizes the data. Following that, you may input some characteristics to the classifier to see what the "predicted" class is once you've obtain ed the hyperplane. This made the algorithm particularly ideal for our needs, however, it may be used in a variety of circumstances.

The model's third algorithm is called Gradient boosting classifiers are a set of machine learning algorithm s that integrate several weak learning models to generate a powerful prediction model. When conducting gradient boosting, decision trees are commonly employed. Gradient boosting models are gaining popularit y as a result of their ability to categorize difficult information. As a result of carefully analyzing all of the above strategies, I have chosen above mentioned the algorithms.

4. Performance Evaluation:-

We can analyze performance matrices in many ways. For this model, I used AUC and ROC. The AUC (Area under the ROC Curve) is one of the most essential assessment criteria for any classificati on model's performance. It is a metric for evaluating the performance of a classification issue at various thresholds. I Utilized AUC as our metric to determine the model's performance on the test data for the methods we employed. I utilized ROC (Receiver Operating Characteristic Curve) to generate the TPR and FPR curves. ROC is a plot between TPR (True Positive Rate) and FPR (False Positive Rate) derived by t aking several threshold values from a model's reverse sorted list of probability scores.

The efficiency of AUC can increase the accuracy of the model.

Decision Tree	Linear SVC	GBDR
The maximum depth of the tree: 10 The best min no of the sample at any node: 500	The best hyperparameter value is: 0.6524336329932608	learning rate=0.2, n_estimators=50

Performance ranking for Decision tree Algorithm on Prediction Customer Purchase Intention as follows.

Modem Name	Hyper Parameter	AUC
Decision Tree	10-500	0.89

Performance ranking for Linear SVC Algorithm on Prediction Customer Purchase Intention as follows.

Model	Hyper Parameter	AUC
LinearSVC	1.587	0.87

Performance ranking for Gradient Boosting Classifier on Prediction Customer Purchase Intention as follows.

Model	Hyper Parameter	AUC
GradientBoostingClassifier	50-0.2	0.9

5. Results and Conclusion:-

In this course work, I have developed a classification model that predicts the purchase intentions of online shoppers. I have learned and implemented to do Data Exploration, Preprocessing of the data, used three classification techniques that have been investigated to resolve the addressed problem, namely Decision tree, Linear Support Vector Classification, and Gradient Boosting Classifier, and also evaluated its model performance. Moreover, I used the performance metrics (Receiver Operating Characteristic) and (Area Under Curve) to evaluate the best model, also to improve the performance and the scalability of each classifier.

Model	Hyper Parameter	AUC
Decision Tree	10-500	0.89
LinearSVC	1.587	0.87
GradientBoostingClassifier	50-0.2	0.9

Based on experimentation and comparison results, I have proven the efficiency of the Gradient Boosting Classifier and Decision tree classifier techniques are balanced and that can fit the requirements of the given problem. The Highest accuracy it has managed to achieve at 90% for Gradient Boosting Classifier and 89% accuracy for the Decision tree. Therefore, I would like to recommend that using Gradient Boosting Classifier can help to get more accuracy than other methods.