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

Reg# 2101142 |Word count: 984

Table of Contents

ABSTRACT:

The purpose of this project is to design machine learning model to predict whether the hotel will be profitable or not, for this we do preprocessing of data and applied different classification algorithms like decision tree, random forest, Support vector machine, k nearest neighbor, and regression algorithms like decision tree, linear regression, polynomial regression, and Support vector machine and analyze the results, out of which results of decision tree are more accurate for classification.

INTRODUCTION:

Following important steps of this project.

Data preprocessing:

After collecting dataset for model, basic EDA operations are performed on it which includes shape description, description of data, and outlier detection by plotting seaborn boxplot and then detected outliers are deleted. Then data is ready for implementation of algorithms.

Implemented algorithms:

The different algorithms which are implemented on the data are classification algorithms like decision tree, random forest, Support vector machine, k nearest neighbor, and regression algorithms like decision tree, linear regression, and Support vector machine.

Linear Regression:

It is the supervised Machine learning model, here we find the best fit line between the response and explanatory variables. In the second part of the exercise, for getting the profits prediction I got the best results using linear model than other regressors I used. I got a root mean squared error of around 600 which was better than other algorithms I used.

Logistic regression:

Logistic regression is a classification machine learning technique. A logistic function is used to describe the probability of the probable outcomes of a single trial in this technique. I tested the algorithm's accuracy with several C values and found that C=0.01 provided the best results. I got almost 79% accuracy using this algorithm.

Random forest:

As the name indicates, a random forest is made up of a huge number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes become the prediction of my model. In random forest algorithm, I tried different hyper parameter settings and got best with following setting:

RandomForestClassifier(n_estimators = 100). I got 85% accuracy with this.

Decision tree:

A decision tree is a decision-making aid that employs a tree-like representation of options and their outcomes. I got best accuracy with this parameter settings:

DecisionTreeClassifier (criterion = 'entropy', min_samples_leaf=5, min_samples_split = 6).

I got high efficiency using this algorithm which is 88%.

K-Nearest Neighbours:

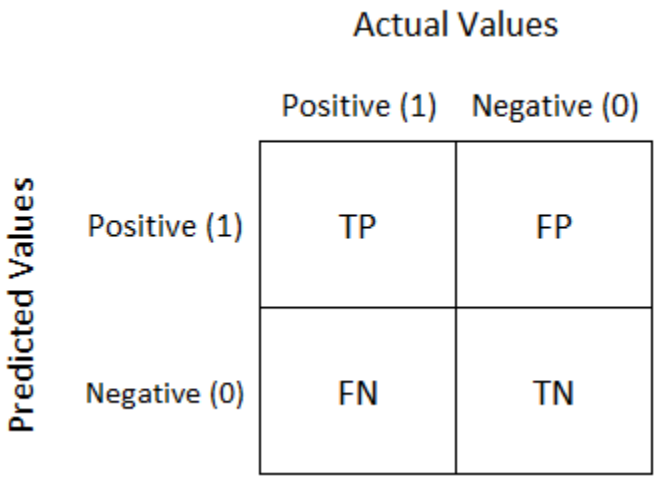
Neighbor-based categorization is a sort of lazy learning since it doesn't try to build a general internal model and instead only saves instances of the training data. The classification is determined by a simple majority vote of each point's k closest neighbors. I have got best accuracy with Euclidean distance with n=5.

Results Analysis:

The result from these implemented algorithms is analyzed based on following parameters:

Confusion matrix:

To detect the performance of classification model, a confusion matrix of NxN order is formed which gives the information about the actual positive or negative and predicted positive or negative classes.



Confusion Matrix for Decision Tree is below from classification technique:

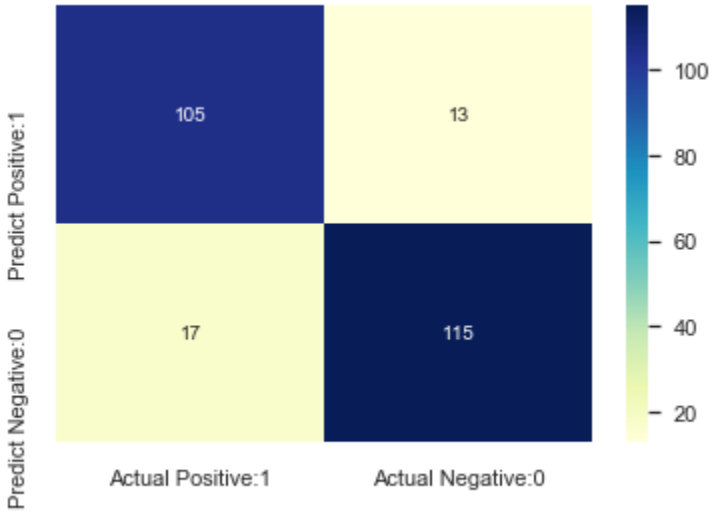
Confusion matrix

True Positives(TP) = 105

True Negatives(TN) = 115

False Positives(FP) = 13

False Negatives(FN) = 17



Precision:

Precision is the ratio of actual positive by the sum of actual positive and the one who are not positive but predicted as positive which is false positive

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

Recall:

Recall is the ratio of actual positive by the sum of actual positive and which are not actually negative but predicted as negative.

$$\text{Recall} = \frac{True\ positive}{True\ positive + False\ negative}$$

Accuracy:

Accuracy is the ratio of sum of actual positive and actual negative to the total target class.

$$\text{Accuracy} = \frac{True\ positive + True\ negative}{Total}$$

Classification Report

```
In [705]: from sklearn.metrics import classification_report
          print(classification_report(y_test, y_pred_DT_entropy))
```

```
              precision    recall  f1-score   support

       False       0.86      0.89      0.88       118
        True       0.90      0.87      0.88       132

    accuracy                           0.88       250
   macro avg       0.88      0.88      0.88       250
weighted avg       0.88      0.88      0.88       250
```

Mean Absolute Error

The average of the difference between the Actual value and the predicted value is the Mean Absolute Error. It shows us how far the projections differed from the actual outcome.

Mean Squared Error

MSE is calculated by averaging the square of the difference between the original value and the predicted value. MSE has the advantage of being easy to calculate the gradient, but mean absolute error requires the slope to be calculated using a complex linear programming tool. When you square the error, the effect of the larger error is more pronounced than the smaller error, and the model can focus on the larger error.

RESULTS:

The results determined for all algorithms are:

Algorithm	Accuracy (%)	Mean squared error
Decision tree	88	1050
SVM	78	1200
Linear regression	79	600
Random forest	85	

CONCLUSION:

In this project we designed machine learning model to predict whether the hotel will be profitable or not, by doing first preprocessing of data and then applied different classification algorithms like decision tree, random forest, Support vector machine, k nearest neighbor, and regression algorithms like decision tree, linear regression, polynomial regression and Support vector machine and analyze the results on the basis of precision, recall, accuracy, mean squared error and confusion matrix out of which results of decision tree are more accurate.

REFERENCE:

[1] Kotsiantis, Sotiris & Zaharakis, I. & Pintelas, P.. (2006). Machine learning: A review of classification and combining techniques. Artificial Intelligence Review. 26. 159-190. 10.1007/s10462-007-9052-3.

[2] Akritidis, Leonidas & Bozanis, Panayiotis. (2013). A supervised machine learning classification algorithm for research articles. Proceedings of the ACM Symposium on Applied Computing. 115-120. 10.1145/2480362.2480388.

[3] Maulud, Dastan & Mohsin Abdulazeez, Adnan. (2020). A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends. 1. 140-147. 10.38094/jastt1457.