CE802 Machine Learning Assignment

Report

Word Count: **1223 words**

Reg no:2101139

I)Comparative Study (Classification)

Introduction

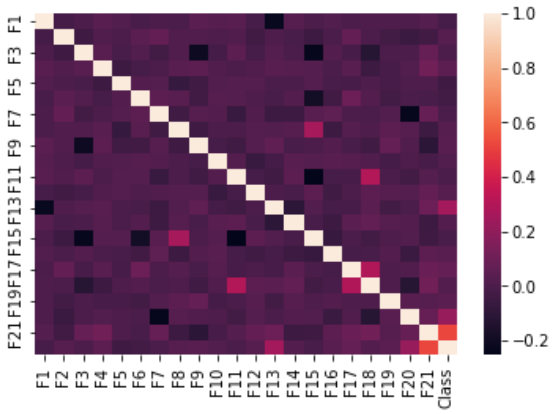
The aim of this study was to prove that machine learning algorithms can be used efficiently to predict whether opening a hotel at a particular location would be profitable, provided with the geographical and socio-economic data and the data regarding other hotels including whether they were profitable or not.

Since the data that is available is labelled, we are able to use supervised machine learning techniques. Also, since there are two values of output possible: profitable and not profitable this is a classification problem. So, we can use different classification algorithms for this.

The classification algorithms that were tried were Decision Tree Classifier, Random Forest Classifier, Support Vector Machine Classifier and K Nearest Neighbour Classifier.

The Data

The data from the file"CE802_P2_Data.csv" that was given contains 21 features and one target variable: Class.



The figure above is the seaborn correlation heatmap for the data. As you can see there is not much correlation between the features.

There are no categorical features in the data. All values are numeric. So, there is no need for encoding. There are no null values in the data, except for the 21st feature('F21'). Almost half of the values for this feature are null values. Since this is a problem, I had to replace them with something. I tried different approaches for this. I tried to replace the null values with the mean of the values in this column. Also, I tried replacing with median, mode, minimum and maximum of the values in that column. I also tried replacing the null values with 0. Finally, I tried dropping that column altogether. Replacing with mode gave the maximum accuracy. So, I moved forward with this approach. The process is summarised below:

Handling Null values in “F21”	Accuracy for Random Forest Classifier
Replacing with mean	0.89
Replacing with median	0.895
Replacing with mode	0.905
Replacing with minimum value	0.89
Replacing with maximum value	0.89
Replacing with 0	0.89
Dropping the column	0.885

The train-test split

In order to train with enough data, but making sure there is no overfitting and the accuracy is maximum, the data was split into two independent parts: 80% for training data and 20% for validation. At the end of training, the performance is evaluated by testing on the validation data.

The Classification Algorithms

Four algorithms were tried for the sake of this problem: Decision Tree Classifier, Random Forest Classifier, Support Vector Machine Classifier and K Nearest Neighbour Classifier. The performance of all these models were compared by changing different hyperparameters. For Decision Tree Classifier, when the “criterion” parameter was changed to “entropy” and the “max_depth” increased to 10, it gave the best performance. In the case of Random Forest Classifier, when the “max_depth” was increased to 10 it gave maximum performance. For Support Vector Machine Classifier, there was not much change in performance when the hyperparameters were changed. For K Nearest Neighbour Classifier, it achieved best performance when the “n_neighbors” was increased to 20.

Evaluating the models

In order to evaluate the models, different criteria were used:

- Accuracy: ratio of number of correct predictions to total number of predictions
- Precision: TP/(TP+FP), a measure of quality
- Recall: TP/(TP+FN), a measure of quantity
- F1 Score: 2*(precision*recall)/(precision+recall)

Where TP=True Positives

FP=False Positives

FN=False Negatives

Precision denotes the ratio of true positives predicted to the total number of samples that were predicted positive.

Recall denotes the ratio of true positives predicted to the total number of samples that are actually positive.

F1 Score is a score based on precision and recall.

The best performance was achieved by Random Forest Classifier, but the Decision Tree Classifier also produced a performance very similar to Random Forest. Support Vector Machine and K Nearest Neighbour classifiers exhibited much lesser accuracy. The results are summarised below.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Decision Tree	0.875	0.863	0.872	0.867
Random Forest	0.905	0.894	0.904	0.899
SVM	0.715	0.793	0.531	0.636
K Nearest Neighbour	0.665	0.648	0.627	0.637

From this, I concluded that Random Forest Classifier is the best among them, and it was used to predict the test data.

The test data was taken from the file "CE802_P2_Test.csv", in the column 'F21' null values were replaced with mode of the values, and prediction was done using Random Forest Classifier and the results were copied to the file "CE802_P2_Test_Predictions.csv".

II)Additional Comparative Study (Regression)

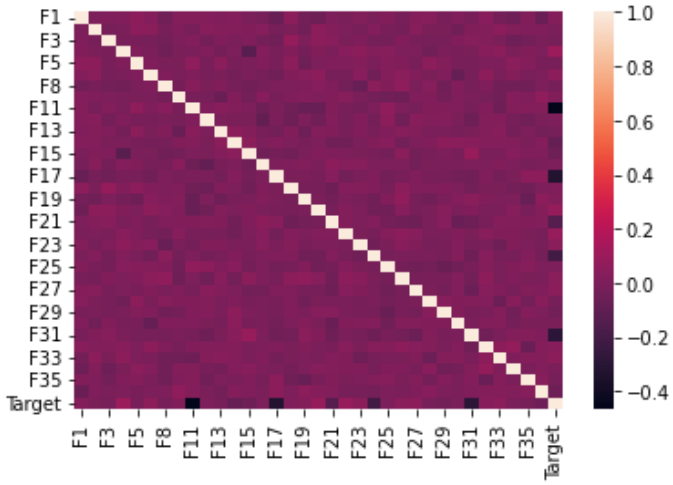
Introduction

The aim of this study was to predict the annual profit (or loss) of a company when data regarding business is given and historical data regarding the business and the numerical value for profit is provided as training data. Since the profit is given in the training data, it is labelled data. So, this is supervised learning. Also, since we have to predict a continuous value, we use regression.

We can use different regression techniques for this like: Linear Regression, Ridge Regression and Gradient Boosting Regression.

The Data

The data that was given in the file "CE802_P3_Data.csv" contains 36 features and one column for the target variable which denotes annual profit and it can be positive or negative.



The above figure is the seaborn correlation heatmap for the data. As you can see there is not much correlation between the features.

From the first glance itself we can understand that the data contains categorical data for the features: “F6” and “F10”. These needed to be encoded so label encoding was used. Each label was assigned a unique integer. For “F10”, the labels were “Very low”, “Low”, “Medium”, “High”, “Very high”. So these were assigned integer values, 1,2,3,4,5 respectively. For “F6”, the labels were “USA”, “Europe”, “UK”, “Rest”. These were assigned integer values 1,2,3,4 respectively. There were no null values in the data, so that problem was not there.

The train-test split

The data had to be split into 2: one for training and the other for validating the performance of the model. Here the split was 90% for training and 10% for validation. This was so that there was enough data for training.

The Regression Algorithms

Three algorithms were used for this part: Linear Regression, Ridge Regression and Gradient Boosting Regression. Linear Regression didn’t have much change in performance when the parameters were changed. Ridge Regression gave maximum performance when alpha was set to 1.0. Gradient Boosting Regression gave maximum performance when the parameter “loss” was changed to “huber” and “learning_rate” was changed to 0.3.

Evaluating the models

In order to evaluate the models different metrics were used. These were:

R^2 Score: It is the coefficient of determination. The best score is 1.0. It can be negative also.

Mean Squared Error: The mean of the squared values of error for each sample.

Root Mean Squared Error: It is the square root of the Mean Squared Error.

The best performance was achieved by Gradient Boosting Regression with both Linear and Ridge Regression producing lower performance. This is summarised below:

Regressor	R^2 Score	Mean Squared Error	Root Mean Squared Error
Linear Regressor	0.691	490919.195	700.656
Ridge Regressor	0.691	490947.437	700.676
Gradient Boosting Regressor	0.863	217286.118	466.139

From this we can conclude that Gradient Boosting Regressor gives the best performance. So, it was used on the testing data.

The data was used from CE802_P3_Test.csv, the same label encoding was done for features “F6” and “F10” and the output was predicted using Gradient Boosting Regressor. The results were copied to the file, CE802_P3_Test_Predictions.csv.

References

https://scikit-learn.org/

https://www.geeksforgeeks.org/how-to-create-a-seaborn-correlation-heatmap-in-python