2021 -22

JOEL JOHN
UNIVERSITY OF ESSEX

Abstract:

In this project, for the first part (part 2 in project description), we were asked to consider a real-world scenario of opening a new establishment and asked to perform different Machine Learning techniques to predict whether the establishment will be profitable. For the second part (part 3 in project description), we were asked to investigate and implement an ML algorithm to predict the annual profit of an establishment being profitable. The data was provided, so we followed the supervised training model for this project. In this report, we contain all the details with regards to the project and steps followed to achieve the final output.

PART 2

Investigation:

The data provided was thoroughly studied. Upon investigation in the first part of the task, two data sets were provided, first for training (CE802_P2_Data) and the second data set was for conducting the prediction with the trained model.

Upon analyzing the data, it was determined that the ML classification algorithms are most suitable for achieving the optimal output. After deciding that classification is the way to go upon further analysis, it was found that the provided data for both the test and train there were 21 different features to be considered and had missing values in one of the features (F21). Few options were considered to handle the missing data but, in the end, it was decided to try two different imputation methods they were:

1. Dropping the whole column altogether so that out of the 21, only 20 features will be considered to train the model
2. Replacing the missing values with the mean of that column/feature

Also, upon evaluating the heat map of the 21 features, it was determined the output (Class) had almost a similar dependency on all the features, so that was one other reason for considering the second imputation method.

Choosing Classifiers and Training:

Four different classifiers (all from the sklearn library) were selected for testing they are:

1. Random Forest Classifier
2. Decision Tree Classifier
3. K-Nearest Neighbour
4. Support Vector Machines

For training, the data received was split in two, one for training and the other for determining the accuracy of the trained model. All the four classifiers were trained and tested for the imputation method considered. So that we can select the one combination with which we observe the best values as the optimal result.

Testing and Obtained Values:

The pattern followed while testing the accuracy is that the feature F21 was dropped as it had missing values. All other features had no issues, so the data was first split into input and output (input: F1 – F21, output: Class), and these were divided into training and testing. After training the models with training data. Accuracy was determined of prediction was determined by using the testing data, and from this, the obtained values are as follows:

1. Random Forest Classifier : 82%
2. Decision Tree Classifier : 85%
3. K-Nearest Neighbour : 65%
4. Support Vector Machines : 68%

Followed by this, the second imputation method was used. So the missing values were replaced with the mean value, and then the data was split similar to the previous form. Followed by this, the model was trained, and the accuracy was determined as follows:

1. Random Forest Classifier : 88%
2. Decision Tree Classifier : 84%
3. K-Nearest Neighbour : 66%
4. Support Vector Machines : 72%

Conclusion:

From determining and observing the values, it was determined that the best way for handling the missing values was to impute them with the mean value of feature F21. This seems to increase the accuracy of all the classifiers.
Upon further analyzing we can see that Random Forest Classifier obtains the best result.
The final result or the output for the test data set provided by the client was made using the trained Random Forest Classifier with missing values in feature F21 replaced with mean values.

PART 3

Investigation:

The data provided was thoroughly studied. Upon investigation in the first part of the task, two data sets were provided, first for training (CE802_P3_Data) and the second data set (CE802_P3_Test) was for conducting the prediction with the trained model.

Evaluating the data, it was determined that there are two columns of categorical data being the feature F5 and F34, so for training the ML algorithms, it's compulsory that all the data should be represented with numbers, so a data cleanup should be done for this all the unique values in the F5 and F34 were replaced by a corresponding integer making all the data in numerical format.

As a result, to predict a continuous value, we will be using regression algorithms.

Choosing Algorithms and Training:

Four different regression methods (all from the sklearn library) were selected for testing they are:

1. Linear Regression
2. Ridge Regression
3. Random Forest Regression
4. Gradient Boosting Regression

For training, the data received was split in two, one for training and the other for determining the accuracy of the trained model. All four regression algorithms will be trained with the same data, and tests will be done using the exact data for all the regression algorithms

Testing and Obtained Values:

For testing, the same data was used to determine the RMSE (Root Mean Squared Error, Mean Square Error, and coefficient of determination (r2 Score) of all the four regression methods. The acquired values are shown in the below table:

Table 1

Training	r2 Score	RMSE	MSE
Linear Regression	0.7140	646.760	418298.796
Ridge Regression	0.7140	646.760	418298.912
Random Forest Regression	0.9506	268.792	72249.368
Gradient Boosting Regression	0.99150	111.495	12431.242

Table 2

Testing	r2 Score	RMSE	MSE
Linear Regression	0.6095	779.862	608185.310
Ridge Regression	0.6095	779.896	608238.350
Random Forest Regression	0.6575	730.393	533473.993
Gradient Boosting Regression	0.8520	480.000	230400.684

Conclusion:

From Table-1 and Table-2, we can see that the best values are observed when determining the RMSE, MSE, and r2 score is given by Gradient Boosting Regression, so we can conclude that this is the best method. So, in the end, the Gradient Boosting Regression trained model is used to predict the output for the test data set provided.

Reference: https://scikit-learn.org