University of Essex
School of Computer Science and Electronic Engineering
CE802 - Machine Learning and Data Mining

CE-802
Machine Learning and Data Mining

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

Report on the Investigation

Created By:

Name: Jayani Dipalkumar Bhatwadiya

Reg. No: 2004351

Word Count: 784

University of Essex
School of Computer Science and Electronic Engineering
CE802 - Machine Learning and Data Mining

This report comprises two main sections, one, where I will be able to discuss the techniques used for building the **Classification algorithms** and their results, and also the second, where I will explain the **Regression task**. The objective of the primary task is to predict whether a customer would claim the insurance in order that discounted insurance quotes might be given for those that have little or no chance of claiming. Since we have already got their past records that tell about their claim history, we will be using the Supervised Learning technique to spot the 2 distinct customer segments.

Classification Task:

• Pre-Processing Phase

The data consists of 15 different characteristics and a dependent class. The last function,' F15', contains almost half of the missing values of all the features available. It is good to discard it when we have a function where almost half of the values are missing. Since the function's distribution varies with a big margin, it becomes very difficult to impute the values. I performed the mean imputation in the experiment and I trained the algorithm.

• Model Building

80 percent of the data was taken for training the model and the rest for research.

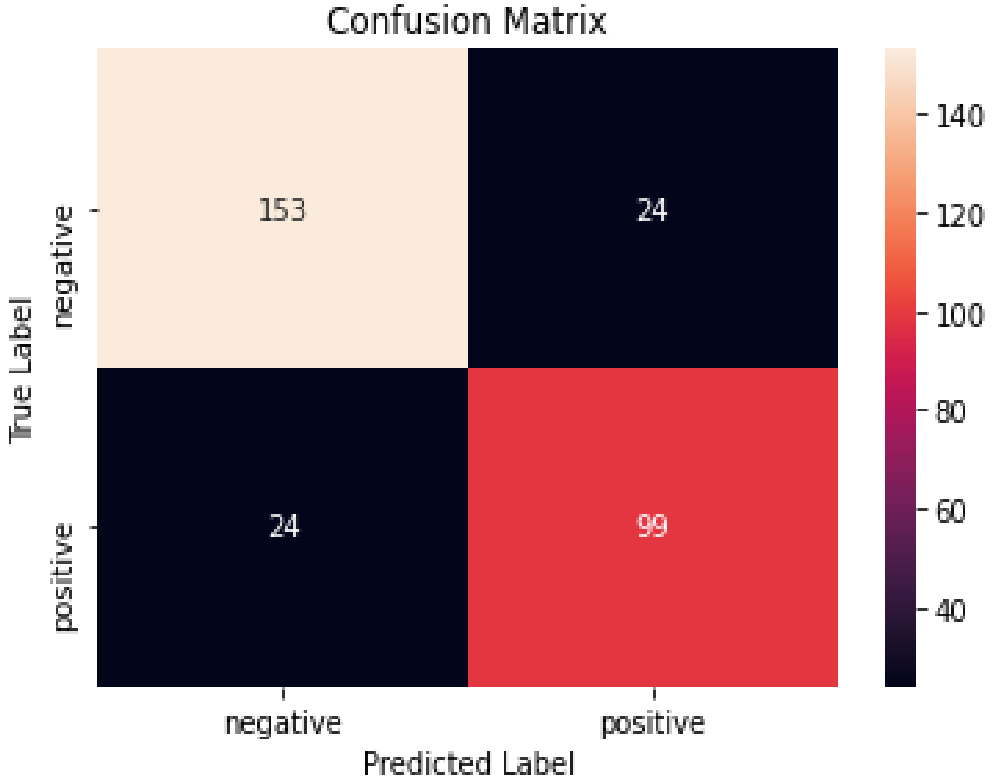
The **Decision Tree** algorithm is the first model the data has been trained with. A decision tree is a technique of supervised machine learning that uses a tree to display steps, and output class categories are shown in the last bottom one. There are three essential sections in a decision tree, nodes for attribute checking, edges for the purpose of branching by selected variable values, and then the leaves marking the output class where each leaf has a specific class attached to it. Decision tree faces overfitting issue when the tree goes down more so pruning is a must take action to resolve this issue.

In this case study, the second model employed is the **Support Vector Machine**. As we want to know a lot about the clients, insurance is a domain where we are supposed to have a lot of features. SVM, however, triumphs over those scenarios where the given characteristics are converted into larger dimensions and the correct hyperplane is found. For the given data, the trained SVM model has been proven to make the best predictions.

Random Forest is the third and final algorithm that I have been using for preparation. It is one of the best algorithms that give you a high degree of precision.

- ## Model Evaluation

Now that the information is learned, we need to test it on the new dataset to make live predictions. Accuracy might not be the best way to show a success. For the assessment process, measures such as Uncertainty Matrix, Recall Score and Precision were used.



- ## Result Interpretation

Random Forest provided the best results from the three Supervised Learning algorithms, followed in both cases by Support Vector and Decision Tree. In order to measure the performance with more bias against the recall score, we use all the metrics. The explanation why we concentrate more on the recall score is that we want a model to have a much lower false-positive rate (FPR). We do not want a poor customer to be listed as a good client in the area of insurance.

	Algorithm	Accuracy	Precision	Recall	F1-Score
0	Decision Tree Classifier	75.667%	0.775	0.669	0.718
1	Support Vector Classifier	80%	0.788	0.777	0.783
2	Random Forest Classifier	83%	0.819	0.813	0.816

Regression Task:

- **Problem Statement:**

We want to classify the claim sum of the customer now that we have divided the clients into 2 segments. To find out the argument, we will be applying regression algorithms.

- **Pre-Processing Phase:**

We also identified categorical variables in the data and, sadly, categorical variables cannot be solved by regression algorithms. We have to encode the variables into numeric characteristics. We will be using the function **'get_dummies'** to translate the values into new functions.

F4_Rest	F4_UK	F4_USA	F12_Low	F12_Medium	F12_Very high	F12_Very low
1	0	0	0	0	0	0
1	0	0	0	0	0	1
0	1	0	1	0	0	0
0	1	0	1	0	0	0
0	0	1	0	0	0	1

We can see in the diagram above that the categorical variable values have been encoded as new functions.

- ## Model Training:

Three algorithms, **Linear Regression, Decision Tree Regressor**, and **Random Forest Regressor**, have been selected for data training. Linear Regression has proved to be the best one out of the three algorithms.

- ## Result Analysis:

It can be clearly seen that the least of all the algorithms is the MSE of Linear Regression. This is why we chose Linear Regression to train the invisible dataset. The remaining algorithms also performed well, but if the dataset has too many features, it is a proven fact that Tree-based algorithms are poor learners. The transformation of Categorial Features has increased the dimensionality of the data in the Regression task, so the Tree Learners did not function properly.

	Algorithm	R^2	MSE
0	Linear Regression Model	0.79409	270126
1	Decision Tree Regressor	0.670026	642479
2	Random Forest Regressor	0.735976	491179