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**.**

Pilot - Study Proposal

Created By:

Name: Jayani Dipalkumar Bhatwadiya

Reg. No: 2004351

Word Count: 697

University of Essex
School of Computer Science and Electronic Engineering

CE802 - Machine Learning and Data Mining

- There are many problems that will be solved using machine learning as an example if we would like to predict something on the premise of the past, we are able to use machine learning techniques to create future predictions. The most advantage of machine learning is to predict the long run so actions may be taken so as to pander to consequences. one amongst the applications of machine learning is to predict carcinoma among females on the premise of some variables **i.e.,** age, red blood cells level, white blood cells deficiency, etc

- As a Machine Learning Consultant, watching the business case, it's like Data Science will be employed in the insurance use case. Now the question is, what will we want to grasp about the data? will we want to grasp if a customer will claim the insurance or we wish to understand the number which can be claimed within the future?

- There are two different kinds of machine learning, supervised and unsupervised. This dataset has an output within the variety of false/true which states it's a controversy of supervised. within the supervised form of machine learning, there are further two sorts of regression and classification. Regression problems have output within the sort of numeric or continuous value whereas if the output is given within the style of category so it's a controversy of classification and model during this case classify the unseen point during the prediction phase. This problem has two categories true and false so it's a problem of classification with two categories so it is a binary classification problem.

- For building a Machine Learning algorithm, we want features. These features play a serious role in making an honest prediction. within the case of the travel insurance domain, there's a particular form of information which we would like about the shoppers to perform better than the rivals by providing better quotes. The important information could be **Age, Gender, Travel Purpose, Mode of Travel (Flight, Bus, Ship), Travel Class (Business Economy), Duration of Trip, Income, legal status, Number of youngsters, Home postal code, Previous Travel History, Previous Claim Amount, Disability Status, Existing Health Issues, Previous Accidents, Country of Travel, Country of Origin, Nationality**. These are some informative features that are good to own because they explain a lot about the customer's behaviour. Among these features, few features like Age, Purpose of Travel, medical records play a vital role.

- Now, since we have got the desired data about the shoppers, we want to pick some algorithms which may yield fruitful results. one amongst the three used models is compulsory to implement which may be a pruned decision tree, the other two chosen models are the support vector machine and Random Forest. SVMs always gives us the added advantage of dimensionality, we will have as many dimensions as we would like. Generally, most of the ML algorithms do not perform well in terms of computational time since they can not pander to high

dimension data, whereas SVM performs well. Random Forest is additionally famous for its good accuracy and performance within the case of supervised machine learning problems.

- Dataset is divided into test/train, the ratio of 80/20 is chosen, 80% for training, and 20% for testing. After employing a particular algorithm, i.e. SVM, we would like to check its performance. We must take care that the model works before making any live predictions. We will use the model to check on the unseen historical data that we have not used for training. We can look at the Recall score to judge the performance of the ML algorithm. Why Recall and why not accuracy?? within the insurance domain, it is okay to misclassify an honest customer as a foul customer (False Negative Rate) but we cannot misclassify a nasty customer as an honest customer.

- Hence the explanation we must evaluate the performance of the algorithms on the idea of Recall.

- Some steps are done before applying models. like missing va1ues within which are replaced with average, target class has been converted into numeric values 0 for false and 1 for true, conversion of textual data to numerical data. Other steps include splitting data into xtrain, ytrain, xtest and ytest.