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: Srikanth Reddy G

Reg.No: 2004445

Word Count: 556

As a Machine Learning Consultant, looking at the business case, it looks like Data Science can be used in the insurance use case. Now the question is, what do we want to know about the data? Do we want to know if a customer will claim the insurance or we want to know the amount which could be claimed in the future?

For building a Machine Learning algorithm, we need features. These features play a major role in making a good prediction. In the case of insurance domain, there are certain type of information which we need about the customers 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, Marital Status, Number of Kids, Home Zip Code, Previous Travel History, Previous Claim Amount, Disability Status, Existing Health Issues, Previous Accidents, Country of Travel, Country of Origin, Nationality**. These are few informative features which are good to have because they explain a lot about the customer's behaviour. Among these features, few features like Age, Purpose of Travel, Medical History plays an important role. Example, the company don't want to give any discount to someone who is going for a Sports related events, as the probability of a sports person getting an injury is high. Second example, we don't want to give discount to someone who is old and had a bad medical history.

Now, we will focus on the task where we want to know if a customer will claim the insurance or not. To handle this task, we would need to use the Classification techniques. Since we have the historical data with all the claim information, we will use the Supervised Learning techniques. Insurance is a very risky business since there is a lot of money involved. We don't have the flexibility of making many mistakes as it could cost the organization a fortune. We want to get as many attributes as we can before we use any machine learning algorithm.

Now, since we have the required data about the customers, we need to select a few algorithms which can yield us fruitful results. We can select algorithm like SVM, Decision Tree and Logistic Regression. SVMs always gives us the added advantage of dimensionality, we can have as many dimensions as we want. Generally, most of the ML algorithms don't perform well in terms of computational time since they can't deal with high dimension data, whereas SVM performs well. We can use Decision Tree as it is known for giving high accuracy results which is very much needed in the insurance field.

After using a particular algorithm, i.e SVM, we need to test its performance. We must be sure that the model works before making any live predictions. We will use the model to test on the unseen historical data that we haven't used for training. We can look at the Recall score to evaluate the performance of the ML algorithm. Why Recall

and why not accuracy?? In the insurance domain, it is okay to misclassify a good customer as a bad customer (False Negative Rate) but we cannot misclassify a bad customer as a good customer. Hence the reason we must evaluate the algorithms performance on the basis on Recall.