Date: 19th January 2021
Word count: 678

Course Teacher:
Dr Luca Citi

Pilot-Study Proposal

Syed Omar Faruk (2003385)

Contents

1 Pilot Study Proposal 1
1.1 Abstract . . . 1
1.2 Predictive Types Identification . . . 1
1.3 Examples of Possibly Informative Features . . . 1
1.4 The Learning Procedures . . . 2
1.5 Evaluating the Performance of the System . . . 3

1 Pilot Study Proposal

1.1 Abstract

A pilot-study proposal for investigating the feasibility of using machine learning to predict whether a customer will file a claim on their travel using historical data of past policies, information about the insured, the purpose and destination of the travel, and whether a claim was filed. The main reason for this part of the assignment is to study and investigate whether machine learning procedures could be used to successfully solve this problem. This report is under no circumstance a real report but it was done to fulfill the criteria of the CE802 course by using Latex I have.

1.2 Predictive Types Identification

After looking at the data files and from the description, I decide to use classification which goes very well with the problem and strategies are best in this regard as we need to find to which category an object belongs. Also, we can consider binary classification as we need to figure out that if the customer will file a claim or not.

1.3 Examples of Possibly Informative Features

From the manager of the travel company I will expect to gather the possible informative features as follow:

- The customer's details like age, sex, geographical location etc.

- Historical data of past policies
- The insured data
- Destination of the travel
- How many times the claims were filed
- The satisfaction level of the customer after the tour
- The total cost of the travel
- The details about the discounted premium offer

1.4 The Learning Procedures

There are a few learning procedures we can use to find out if the customer will claim a fail or not. From my experience and the data I got for this project, I would use supervised classifiers and see how much accuracy I can get. Depending on the accuracy scores, we can figure out which model will be best for this scenario. So, for this project I am going to work with top five learning procedures:

- Decision Trees
- Naive Bayes
- K-Nearest Neighbours
- Support Vector Machines and
- Multilayer Perceptron

The reason behind choosing these classifiers are because of the following reasons.
Decision Trees: Being an example of a white-box model, Decision Trees is very close to the human decision-making process. It can work with categorical and numerical features which suit our problem too. It requires little data processing and non-parametric model whereas the feature selection happens automatically which will save time. (Alpaydin, 2020)
Naive Bayes: For independent assumptions, Naive Bayes gives really good accuracy in the outcome. It is easy to implement. For text classification, it works best but I think for our case it will work fine too. (Mitchell, 2017)
K-Nearest Neighbours: Being KNN a lazy algorithm it gives a very good accuracy on the prediction for the type of data we are going to work with. The algorithm is really simple. It is very useful for classification types of problems. One con I can consider is the high memory requirement. But I since I have a good CPU and GPU I am not considering that for now.(Witten et al., 2017)

Support Vector Machines: For clear marginal separation, SVM works amazingly. For high dimensional paces, SVN is very effective. Since it uses the subsets of training points it is very much suitable for our project. One con I should consider for SVN is for a very large data set it is not good but when I looked at the data sets I got for the project, it will go well with SVN.
Multilayer Perceptron: For the project multilayer perceptron will work nicely as our dataset is small and we need to find the yes/no type predictions.

1.5 Evaluating the Performance of the System

Using the accuracy scoring and kappa statistics score I can evaluate the performance of the system before I deploy it to the system. The higher score it is the better in the sense of the accuracy of the predictions. For the model evaluation, we can also check the k-fold cross-validation score. Since it is not possible to predict which model will fit the best just by looking at the data I am going to try the above procedures and see how much score I can achieve for the individual procedure and then I will deploy it to the system once I see it satisfies all the criteria and can actually predict if a customer will file a claim or not.

References

Alpaydin, E. (2020). *Introduction to machine learning.* The MIT Press.

Mitchell, T. M. (2017). *Machine learning.* McGraw Hill.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal (2017). *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann.