University of Essex

Department of Mathematical Sciences

Part 1: Pilot Study

Subject:

CE802 Machine Learning and Data Mining

Registration number: 2003303

Supervisor: Dr Luca Citi

Date of submission (January 20 2021)

Word count:655

Executive Summary

Nowadays number of people traveling through many mediums i.e., by bus, by air.travel insurance company help travelers to secure their travel or trip from unforeseen losses. Since A travel insurance agency wants to offer a discounted premium to clients that are less inclined to make a claim in future. We also have historical data of policies compony were using and data like destination of traveler, insurance claiming details. We need to predict whether customer will claim in the future.

1. Solutions

Four main points discussed below are predictive task, selective informative features, learning procedures and evaluation.

1.1 Type of Predictive Task

In this problem, we need to categorize on machine learning model. As in supervised learning, we have a feature variable that comprises of labeled preparing information and an ideal target variable. At the point when data are being utilized to predict categorical variable, managed learning is additionally called classification. When there are just two labels then it's called binary classification. When there are multiple classes, the issues are called multi-class classification. Whereas regression predictive model uses for predicting continuous values. Therefore, for predicting whether customer will claim an insurance in time to come or not falls under classification predictive task (True or False).

1.2 Selective Informative Features

The research illustrates that data collection and feedback are main characteristics for predicting that travel insurance will get claimed by customer or not. Travel company should get basic information age of the customer, source of customer travel, mode of travel(by air, by road and by water),sex of the customer, customer claimed details, Destination, customer

baggage details and health condition. Using basic data along with mandatory information we can analyse the travel frequency per customer and frequency of customer traveling destination. At last feedback of customer experience would help predict accurate.

1.3 Learning Procedures

For selecting the best Machine Learning algorithm there are various factors need to be consider for instance 1. size of the training data 2.Training time 3. Linearity 4. Number of features. Learning procedures for Logistic regression, Naïve bayes and decision tree as follows:

- **Logistic regression algorithm**: We can find many different ways on how to regularize our model, without having to worry about whether or not features are correlated which can be an issue with Naive Bayes. Moreover, it has good probabilistic interpretation and enables us to easily update our model to acquire new sets of data, which we may not experience with Support vector machines or decision trees.[1]

- **Support vector machine** (SVMs): The advantage of SVMs is its high-accuracy and high-performance characteristic. It provides excellent theoretical guarantees regarding overfitting and has flexible selection of kernels for data that isn't linearly separable. It is particularly popular in text classification problems, especially when very high-dimensional spaces exist.[1]

- **Decision tree**: Decision trees are basically quite simple to interpret and explain. It is also non-parametric and does not require any distribution. Having said this, when we use decision trees, we don't have to worry about outliers or whether or not the data set can be separated linearly. Considered a heuristic algorithm, decision trees do not suffer multicollinearity and are good for few categories variables.[1]

For this particular problem we use SVMs model to predict will customer claim in future or not because of SVMs ability to divide classes into two groups.

1.4 Evaluation approach

The actual output of many binary classification algorithms is a prediction score . The score indicates the model's confidence that a given observation belongs to a certain class. To make the decision about whether the observation should be classified as positive or negative, as a consumer of this score, we will interpret the score by picking a classification threshold (cut-off) and compare the probability score against it. Any observations with scores higher than the threshold are then predicted as the positive class and scores lower than the threshold are predicted as the negative class.[2]

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Negative
Predicted Negative	False Positive	True Negative

1.5 References

1. https://pythonprogramminglanguage.com/what-are-the-advantages-of-different-classification-algorithms/
2. https://medium.com/analytics-vidhya/10-essential-ways-to-evaluate-machine-learning-model-performance-6bf6e11f9502