CE802 Machine Learning and Data Mining

YASWANTH MARAM

Registration Number: 2003287

University of Essex

Mathematical Sciences Department

December 31, 2020

This report contains 744 words

1. **Pilot-Study Proposal**

# Problem Statement:

The task of this project is to explore the feasibility of using Machine Learning to forecast whether a customer will file a claim on their travel.

# Predictive Task:

In solving any type of problem, selecting a specific type of predictive task plays a vital role. Predictive tasks such as classification, regression, clustering, rules mining have their own significance. For example, classification techniques are used for predicting a discrete class label output. Similarly, Regression techniques are used for predicting a continuous output and finally Clustering and rules mining are for unsupervised and market-based analysis, respectively. As problem statement refers to predicting categorical output with two categories, ***‘Binary Classification’*** predictive task suits better.

# Possible Informative Features:

Informative Features are the variables required for any given problem that can sufficiently help us build an accurate predictive model. A few examples of possible features are listed below:

* Name of the travel insurance products
* Duration of travel
* Destination of travel
* Mode of transport
* Number of previous insurance policies
* Number of previous insurance claims
* Previous insurance destinations and durations
* Sales of travel insurance policies
* Commission received for travel insurance agency
* Gender of insured
* Age of insured
* Marital Status of insured
* Children (Yes or No)
* Target: Claim Status
* Target: Claim Value

# Learning Procedures:

We can get solution of the given problem by using different supervised classifier models. Here we use ***‘Decision tree Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, Support Vector Machine classifier and XG-boost Classifier’***. Each model contains several hyperparameters, tuning them to best parameters contribute to higher accuracy of the model. Finally, the evaluation metrics will be compared for each individual model and the model which possess high accuracy will be selected.

Among the ML models we used for this data set, I think SVM’s and XG-boost works way better because:

* SVM’s have the powerful "kernel trick", which allows them to map the data to a very high dimension space in which the data can be separable by a hyperplane.
* XG-boost is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It uses a gradient descent algorithm to minimize the loss.

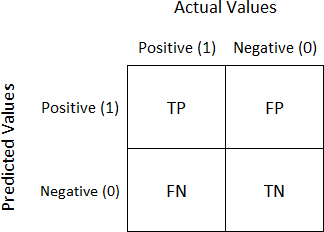
# Performance Evaluation:

The next step is to find out how effective is the model based on performance evaluation metrics such as ***‘Confusion Matrix, Accuracy,* *Precision*, *Recall*, *F1 score & Area Under Curve’*** for Classification problem and ***‘RMSE, MSE, R-square’***for regression problem. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. We can use classification performance derived from simple confusion matrix.

**Accuracy:**

It is the percentage of correctly classifies instances out of all instances. It is more useful on a binary classification than multi-class classification problems because it can be less clear exactly how the accuracy breaks down across those classes.

# Confusion matrix:

The Confusion matrix contains metrics used for finding the correctness and accuracy of the model. It is a table with 4 different combinations of predicted and actual values.

It is extremely useful for measuring Accuracy, Recall, Precision, F1 score.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

It is the ratio of the number of correct predictions to the total number of input samples

Recall/Sensitivity = TP / (TP + FN)

Out of all the positive classes, how much we predicted accurately is referred as recall

Precision= TP / (TP + FP)

Out of all the classes, how much we predicted correctly is referred as precision.

F1-score: 2 / ((1 / Precision) + (1 / Recall))

It is difficult to compare two models with low precision and high recall or vice versa. So, to make them comparable, we use F1-Score. It helps to measure Recall and Precision at the same time.

# AUROC:

AUROC is Area Under the Receiver Operating Characteristics. It is one of the most important evaluation metrics for checking any classification model’s performance. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0’s as 0’s and 1’s as 1’s.

# MSE and RMSE:

MSE is the average squared difference between the estimated values and the actual value. It tells you how close a regression line or curve is to a set of points. RMSE is just the square root of the mean square error.

**References:**

1. [www.kaggle.com](http://www.kaggle.com)
2. <https://scikit-learn.org/stable/>
3. <https://github.com/>
4. <https://machinelearningmastery.com/>
5. <https://towardsdatascience.com/>
6. <https://en.wikipedia.org/>
7. <https://medium.com/>
8. <https://stackoverflow.com/>