Name: Ghodrat Rezaei
Personal Code: 952128
Date: 14/11/2020

Machine Learning - Classification Challenge Report

Politecnico di Milano

Prof. Carlo Vercellis

Introduction

This report is aimed at explaining the methodology that was chosen to analyze and do the classification for a data set concerning an insurance company interested in offering its existing customers vehicle insurance. The approach that was followed in order to reach the training and test results are briefly presented in the following pages, along with how the data was processed.

Preprocessing and Data Preparation

As the first step in the data processing and preparation, the column of the attribute license type was removed as a first attempt, as it has an already balanced distribution. Then, all rows containing missing data were removed from the set to avoid any noise. They weren't replaced by for example the mean or median value to contain the accuracy of the data.

To visualize the categorical data (Gender, license type, previously insured, vehicle age and vehicle damage), they were represented in histograms showing the different records vs the frequency. Then all categorical attributes were represented as dummies.

Then the visualization of numerical data (Age, annual premium, driving licence, policy sales channel, region code, seniority, target and the log of the annual premium) was also represented in histograms and the log of the annual premium was taken to better understand it at a lower scale. Then all numerical attributes were standardized and presented in a boxplot.

From the numerical data we decided to keep the following attributes for our analysis: age, regional code, the log of the annual premium, the policy sales challenge and the seniority. The driving license attribute was treated as a dummy as it's true for the majority, because it

varies only between 0 and 1 and the Standard Scaling of such attribute as Numerical will be not very efficient during the calibration process . Hence, not indicating anything.

Training-Test Sets

The data was split then into a training and test set to apply the different classification methods and test their accuracy. By default the test set was set to be 30% of the total data. For each model, we defined the Model Structure itself and Parameterization space by which the greedy search(gs) approach can iteratively examine through the defined space of parameters and pick up the best parameter which can give us the minimum amount of error(Objective_Function(J)). By defining the cross-validation (cv) equal to 3 in greedy search, probability of over-fitting will decrease significantly.

Classification Models

Model	**The obtained f1 scores**
KNN	0.6556
Decision Tree Model	0.6896
Naive Bayes	0.6873
Logistic Regression	0.6474
Support Vector Machines (SVM)	0.676
Multi-layer Perceptron Classifier	0.6618

At first the roc curve was set in order to test the usefulness of each model applied. From applying the six different models, we can see that all of them have very similar accuracies (f1 scores). They're all around 67.8% , with the decision tree model being the highest with an accuracy of 69.25%. This can be justified with the fact that the different attributes can be easily split in different clusters. The similar f1 scores of the data and the fact that it's always around 67.8% can tell us that the behavior of the data is not easily predictable, resulting in a higher uncertainty. This could be a result of inaccurate or biased data. Applying the SVM model took a lot of time to run as it's based on a recursive optimization model. On the contrary the Naive Bayes model that relies on probabilistic methods was able to run in a shorter time. Finally, the logistic regression model is a mix between the two methods, combining probabilistic and recursive optimization approaches.

Improvement

To suggest an improvement in the results obtained a for loop can be defined to test every parameter and hypermarater on its own. Currently, this solution could not be applied due to the fact that it will take a lot of time to run.