Abdon Ahuile

Statistical and Machine learning Approaches for Marketing - Individual Project

Summary

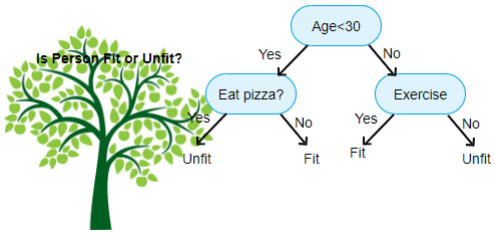
For this project, we are given a dataset from a bank in which it contains:

1. Variables related to the client itself, like its education, housing information or if he has a loan or not (Group 1 – Bank client data)
2. Contact information of the current campaign campaign (Group 2 –Related with the last contact of the current campaign)
3. Other variables such as the number of days that passed after he was contacted (Group 3 – Other Attributes).
4. Sociodemographic variables such as employment rate or consumer price index (Group 4 – Social and economic context attributes)
5. Target variable which determines if the client has subscribed or not

To determine if the client will subscribe or not beforehand, several Machine Learning classification algorithm have to be put into place, the chosen ones for this project are Decision Tree, Random Forest, Logistic Regression, Support Vector Machine and Naïve Bayes Classifier. All of them with their respective setup and evaluation metrics.

Decision Tree

Decision trees is a very simple process in which an observation is classified into one class depending on the path that it follows during the whole tree. Below it can be observed an illustrative image of a decision tree in order to determine if a person is fit or unfit.



Process:

Different decision trees based on all the variables and also based on the different groups of variables described in the summary

Auc Scores :

All variables **0.771564534924723**
Group 1 **0.552698568353321**
Group 2 **0.730490749540751**
Group3 **0.58381832897313**
Group 4 **0.755748583633622**

It can be seen, that the bet accuracy obtained by a decision tree is using all the variables, although Group 4 and Group 3 variables are relevant predictors for classifying the observations.

Random Forest

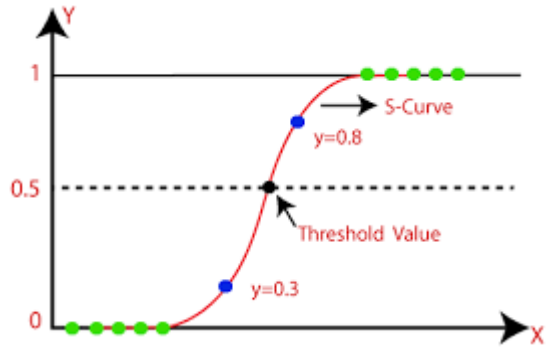
Random forest could be said that it is an upgrade of a decision tree. The main purpose of this algorithm is to obtain a big collection of uncorrelated small decision trees (this process is called bagging) and then obtain an average in order to classify an observation into a certain class.

Auc Score: **0.785**

Logistic Regression

Logistic regression is a classification algorithm which finds the probability of an observation (given the characteristics of its predictors) to fall into a certain category depending on the desired threshold.

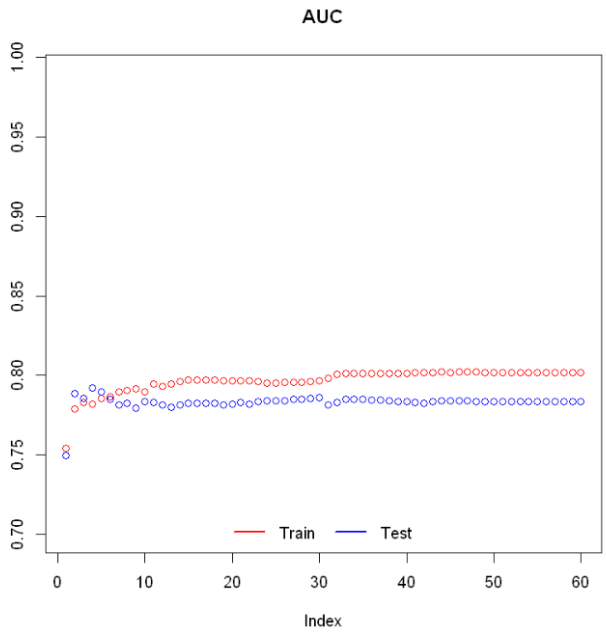
Below it can be observed an image that clearly illustrates the logistic regression concept, any point that falls above the 0.5 threshold when getting the probability using the sigmoid function, will be classified in class 1 and 0 in the other case.



Process:

- To start the categorical variables where dummy encoded
- The approach given to select the features in this case is stepwise feature selection using the forward method.
- The best model out of all of this would be the one chosen in order to classify the variables.

Evaluation:



We can see that the logistic performs bordering 80% with the training data and is smaller for the case of the test set. In addition, the maximum AUC found is using 30 variables, which gives a score of **0.786** in the test set.

Support Vector Machine

Support vector machine can deal very well with the bias-variance trade off in order to classify observation better. It creates a margin between the threshold and the closest point.

Since Support vector classifiers allows misclassification, the margin between the data points is called a “soft margin” and the best of these can be determined using cross validation with many dimensions and determine in which one there are less misclassified observations (the maximum amount of dimensions are the amount of predictors). Unfortunately, support vector classifiers do not deal very well with complex data.

That is when support vector machines comes in handy because it takes the data points to another dimension for better classification and finds a support vector classifier that separates the data the best way.

SVM uses kernel functions to find support vector classifiers in higher dimensions, which calculates the relationship between every point.

Process:

- Dummy encode the categorical variables for both the train and test datasets
- Input the correlation matrix within each predictor
- Divide the “bank_mkt_train” dataset into train and test
- Train the support vector machine algorithm with a radial kernel and using the the variables obtained from the correlation matrix using a cutoff point of 0.5 (13 predictors left) using a 5 fold cross validation due to the fact that the dataset is not large and to reduce bias.

Auc Score:**0.743**

Linear Discriminant Analysis

In linear discriminant analysis, we model every single predictor (independent variables except identifier) in the target variable to then use the Bayes theorem to obtain the probability given that independent variable $Pr(Y=k|X=x)$.

One predictor can be decisive for classifying the observation into one class, but we can also use more predictors to determine better the category. Therefore, LDA focuses in maximizing the separability among the classes.

LDA does this by projecting the data into an axis that maximizes the separability between the means and at the same time minimize the variance within each category. If we optimize this both criteria at the same time, we can get a good separation to make a classification.

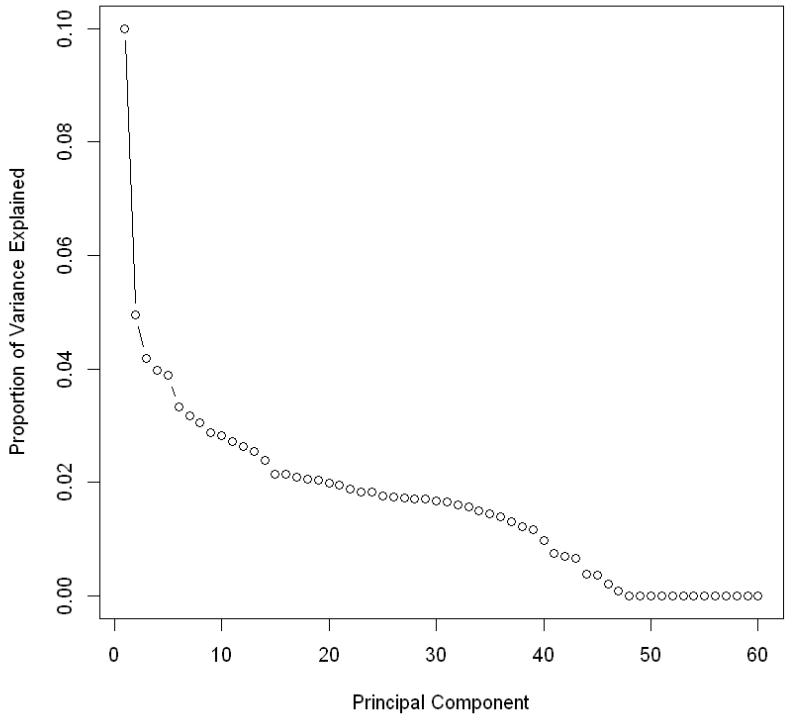
Advantages:

- Very stable compared to logistic regression
- Good for when there is more than two classes (not the case of this project)

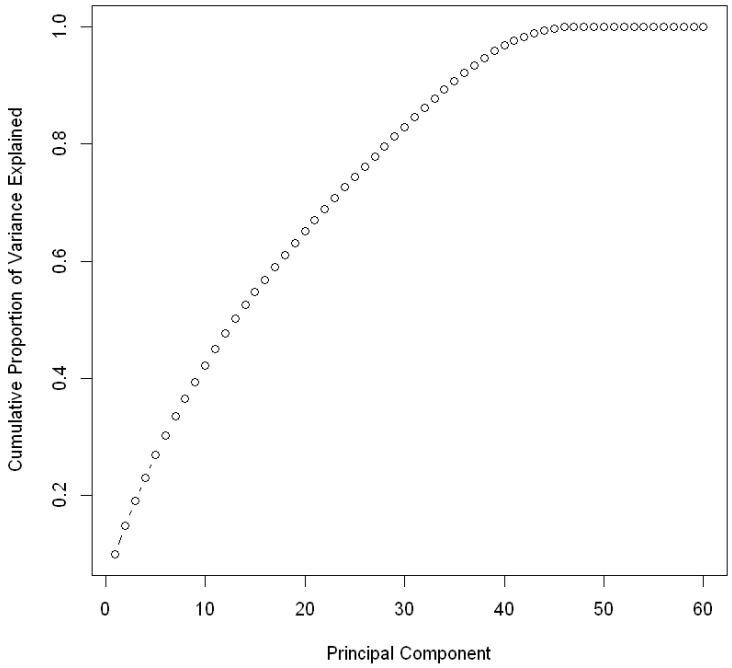
Process:

- Dummy encode the variables since its not convenient to have categorical variables for this algorithm.

- Divide the train dataset (bank_mkt_train) into train and test to see be able to see an accuracy measurement
- Perform a principal component analysis



We can observe from the graph above that the variance is explained until component number 36, in the graph below of the cumulative variance, component number 46 explains about 99% of the total variance. Therefore for this model we will reduced the number of components from 60 to 46.



- Later we train the model in the train dataset with the chosen components and later we predict it in the test set.
- For the Kaggle submission, another prediction is made using the "bank_mkt_test" dataset.

Auc score**: 0.763**

Bibliography

Al, G. J. (n.d.). An introduction to statistical learning with applications in R. In 127-372. New York: Springer.

Brownlee, J. (2014, September 22). *Feature Selection with the Caret R package*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/

Lopes, M. (2017, 01 09). *Dimensionality Reduction — Does PCA really improve classification outcome?* Retrieved from Towards Data Science: https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-classification-outcome-6e9ba21f0a32

Starmer, J. (n.d.). *Stat Quest*. Retrieved from https://www.youtube.com/user/joshstarmer/videos