STATISTICAL & MACHINE LEARNING

Individual Assignment

IESEG
Arnaud Vandelaer

Table of Contents

Part1 Algorithms explained

The 5 different algorithms that are covered in this report are:

- Logistic regression
- Ada boosting
- Gradient boosting
- Random forest
- K nearest neighbor

Logistic regression:

When to use it

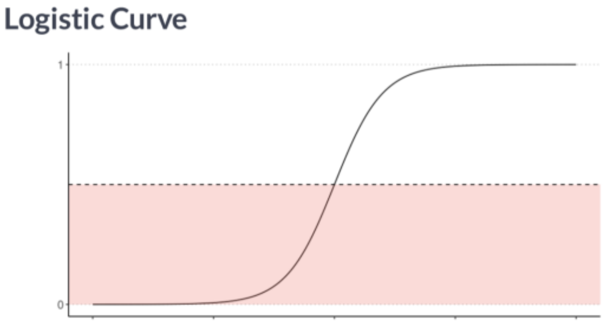
Logistic regression is a supervised learning algorithm that is widely used for classification problems. Where there is one dependent variable. The value of this variable can be visualized as one or zero that most of the times means true or false. The output we aspect for a logistic regression is that we will have a probability that something is true or false, For example a possible output could be: there is a 80% change that a person of 100kg is obese

How does it works

The equation for the logistic regression you can see in picture below. The algorithms gives a probability P(x) that the output will be 1 given with a constant e and where B0 and B1 are the coefficients. The maximum likehood estimate MLE wil help to define the parameters of the model

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Logistic regression is very similar as linear regression the only difference is that we applied a log function. By using the log, we obtain the value 1 and 0. In the graph below, you can see how it looks when you plot it.



In the following steps we will explain how a logistics algorithm works

Step 1: define the dependent and independent variable:
Step2: create a single column called features where all the independent variables are assembled.
Step3: split the data in training/test sets. This can be done by a Randomsplit() function.
Step4: train the model
Step5: run the model over the test set

Advantages

The logistic regression is that it an easy system to implement and an efficient way to train a machine learning algorithms. The result are easy to understand. It is a model that is very useful for business cases, because it gives reliable results and it is easy to understand

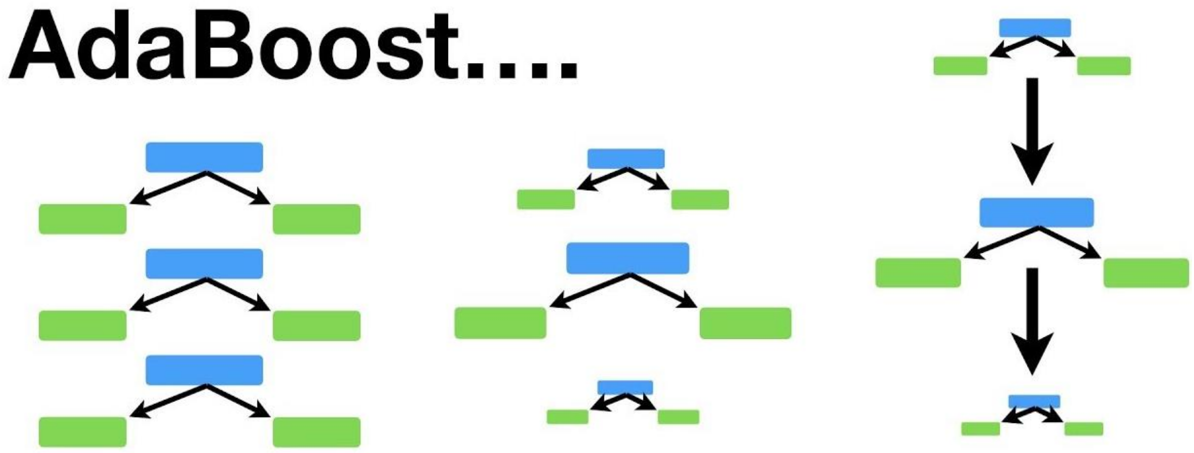
Disadvantages

Negative point of logistic regression is that you need a large sample of data; it can only predict the outcome of a categorical variable. In addition, the independent variables should not have multi-collinearity with the other variables, what means that they should not be highly correlated to each other.

Ada boosting tree

When to use it

When we used the ada Boost machine-learning algorithm it is most likely to boost the performance of a decision trees on binary classification. It often use when there a lot of weak classifier. A weak classifier is a classifier that performs poorly, but performs better then random guessing, like deciding If the person is male of female based on their heights. Ada boosting is also a supervised learning algorithm.



How does it work

In the picture above, we can visualize us how ada boost works. With ada boost, we create many different notes that are representing the variables of a data set. All variable will be considered to make a final decision. This notes always have 2 different possible outputs. Like yes/no or above average or not.

All this threes will be stacked under each other and are used to make the final predictions. Important to know it that with ada boosting, every error made in the previous three will be take into account in the next three. This also mean that some notes will have a bigger influence than other notes.

In the following steps, I will explain how ada boosting works:

Step 1: the first step that the ada boosting algorithm gone do is initialize the weights of data points. This means that if the training set has 20 observations. That each point will have an initial weight of 1/20.

Step2: we gone train a decision tree.

Step3: The algorithm gone calculate the weighted error rate of the decision tree. This represent how many wrong treated prediction are been treated. The higher the weight the more corresponding error will be. This mean if the error rate of the tree is high, it will have a less impact on the decision power of the tree.

Step4: We update the weights of wrongly classified points. This means if the tree made a correct decision with for the observation, the weight wills stays the same. However, if the model treat the data point wrong, it will give a new weight. That correspondent to the old weight multiplied by the weight of the current three. This means that the higher the weight of the tree the more accurate the tree performs is and the more important the misclassified data point by this tree will get. After the data points will be normalized so sum of all the weigh stay equal to 1.

Step5. We rerun the process for all the trees in the train set.

Step6. We make a final prediction by adding up the weights of each tree multiply by the prediction of each tree. Is clear that tree with higher weights will have more influence on the final decision.

Advantages

There are many advantages for ada boosting. First it is fast, simple and easy to program. There are also not a lot of parameters to tune only the amount of trees. Secondly, we can combine this method with any other learning algorithm. Thirdly, we do not need any prior knowledge of the weak learner. Lastly it is very versatile, because we can use it with kind of data numeric, category, …

Disadvantages

A big disadvantage of ada boosting is that it is very depending on his data and weak learners, because it can fail if the classifiers are to complex that means we go to over fit the model. Also if the classifier are too weak there is possibility to under fit the model. Lastly, the model is also very sensitive to noisy data and outliers

Gradient boosting

When to use it

Gradient boosting is supervised learning algorithm with one of the most powerful techniques for building predictive models. It can solve regression and classification problems. who produce a prediction.

How does it works

Gradient boosting is a method where we gone convert weak learner into strong learners. By boosting the new tree and let fit on a modified version of the original data.

In the following steps I will explain how gradient boosting works

Step 1: Build a decision tree and add it to ensemble

Step 2: Use the ensemble to make predictions on the training data

Step 3: Compare predictions with known labels

Step 4: Identify training instances when predictions were incorrect

Step 5: Return to the start and train another tree which focuses on improving the incorrect predictions.

Step6: As trees are added to the ensemble, its predictions improve.

Advantages

Gradient boost has many advantages, firstly it is a model that provides consistently high accuracy that not easy to be beaten by other models. Secondly, there are also different parameters that makes the model flexible. Thirdly, the model do not require many pre-processing steps. Because it treat categorical as well as numerical values. Lastly, it also handle missing data.

Disadvantages

Although the model normally has good accuracy scores, it is not always easy to interpret the results, that why we often use different models how has also good score. The many parameters brig also some disadvantages, because this means that it memory consuming and it can also take a lot of time before the models has ended with running. At the end there is also the danger of overemphasize the outliers by minimize the errors this lead to overfitting.

K-nearest neighbors

When to use it

The knn algorithm is a supervised machine algorithm and is used in classification as well as regression problems. However, it is more likely used in classification.

How does it works

To predict the values of the new data point, knn uses feature similarity. This means that a new observation will be assigned based on how closely it match to the points on the training set.

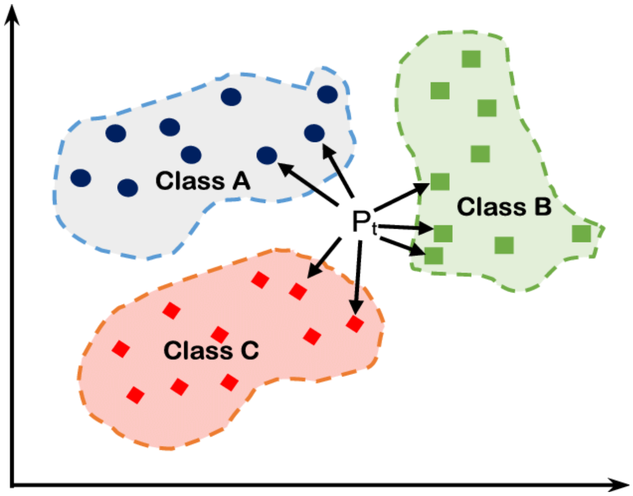
In the following steps I will explain how I works.

Step1: loading the training as well as the test data.

Setp2: choosing the parameter K

Step3: the algorithm gone then calculate a distance between each row of the training data and the test set. one method that is used to calculate this distance it Euclidean.

Step4: when we have a distance value for and we gone sort them in ascending order.



Step5. The algorithm will choose the top K(parameter) rows from the order table.

Step6: based on the most frequent class they will be assigned certain labels.

Advantages

The first advantage of Knn is that has no training period. Knn does not learn anything during the training period; it is called a lazy learner. The training dataset will be stored but it only learns something at the time of making the real time predictions. What leads to that Knn algorithm very fast is.

Secondly, because it does not have a training period it very easy to add new data.

Lastly, it is a algorithm that only have 2 parameters (k and distance function). That make it easy to implement.

Disadvantages

The first disadvantage from Knn is that it does not work well with large dataset. This is because if the data set is large there is a huge cost of calculating of the distance between a new point and an existing point. This leads in a decrease of the performance.

Secondly, if there are a lot of dimensions in the dataset it can become difficult for the algorithm to calculate the distance in ach dimension.

Thirdly, it needed to standardize and normalize the data before applying Knn algorithm. Otherwise, it can lead to wrong predictions.

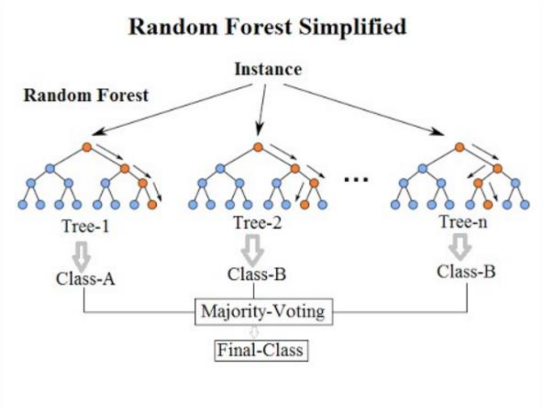
Lastly, the algorithm is also very sensitive to noisy data, outliers and missing values.

Random forest

When to use it

Random forest is one of the most used algorithm that are used for both classification as regression problems. It is a prediction based on bunch of decision trees where each three gives a vote to the prediction of a target variable. Because we use multiple random decision trees, it leads most likely to better accuracy. By random selecting the decision threes, the variance will also be reduced.

How does it works



Step1: divided you data in training and test set

Step2: define the parameters like the amount of features to use and the number of threes

Step2: train you decision tree on the training set

Set p: Repeat it for all you decision trees.

Step4: aggregate the prediction from the individual decision trees. Therefore, there are two possible ways; it takes the majority vote, but this can only if your trees produce labels. Alternatively, if they are numerical, it take the average of the trees.

<u>Advantages</u>

Random forest is able to give a high predictive accuracy, on efficient way and on a large dataset. It also handled missing data and multiple input of features. Because it has different kind of parameters, it makes the model flexible

<u>Disadvantages</u>

A disadvantage of random forest is that is not always easy to interpret the results. In addition, when you include too many decision trees in your random forest there is a high change to over fit the data

Part 2 Set of Models an results

Data summary

The data set that we use in the kaggle completion has 21 variables and 7000 Observations. The data frame that was already very clean, there were no missing values and we did not need to adapt a lot. We changed the value of the variable campaign, because it include also the last contract, so it should be reduced by one.

After splitting the data in a test validation and train data set, we continued with the feature selection.

Feature engineering:

For the feature engineering, we made a list of the importance of the variables. So if make new variables it better to focus on variables with already a high value.

The fellow variables where made:

- **Month_spring** : it represent if the observation was in spring.
- **Month_summer**: it represent if the observation was in spring.
- **Month_autumn**: it represent if the observation was in spring.
- **Month_winter**: it represent if the observation was in spring.
- **Age_ge_mean**: it represent if the age is above the mean of ages
- **Pdays_999**: it represent if the pdays is 999 or not

Processing data

We split the data in categorical, numerical and Boolean variables.

Where able to treat some categorical variables with the WOE function. The Woe function gone give the observation of a certain variable a weight. This is often used when there are more than six different labels and it loose its effect on putting them as a dummy variable.

For the numerical values, we gone discretizing the values this means we converting continuous data attribute values into a finite set of intervals with minimal loss of information. We also apply then the woe function on this data.

The Boolean variables we gone convert into dummy variables.

Variable selection

We notice that after the preprocessing part we have some Infinite values. We gone replace them by the mean of the variables. We also gone drop the categorical variables because they all have been processed and converting the Boolean variable into a dummy variable. At the end we drop also the constant variables this means where the variance = 0

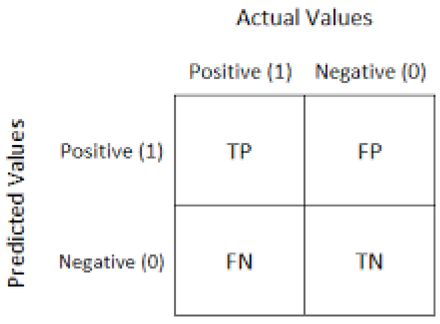
As variable selection we used Fisher Score, it will give us an score that represent the importance of every variable. By selecting only the best features. Are models will train faster and we reducing the complexity of the model. It also reduce overfitting. We were getting the best result if we selected 50 variables.

Model

To protect us from overfitting the model we gone do a k fold Cross-validation. We take a number of folds that is equal to 5. The model will run over the five folds, where every folds is once the test set and so we will be able to estimate the overall average error. With this technique, we will be able to analyze the performance of the model. For every model, I also put some parameters but I will explain them later in the report.

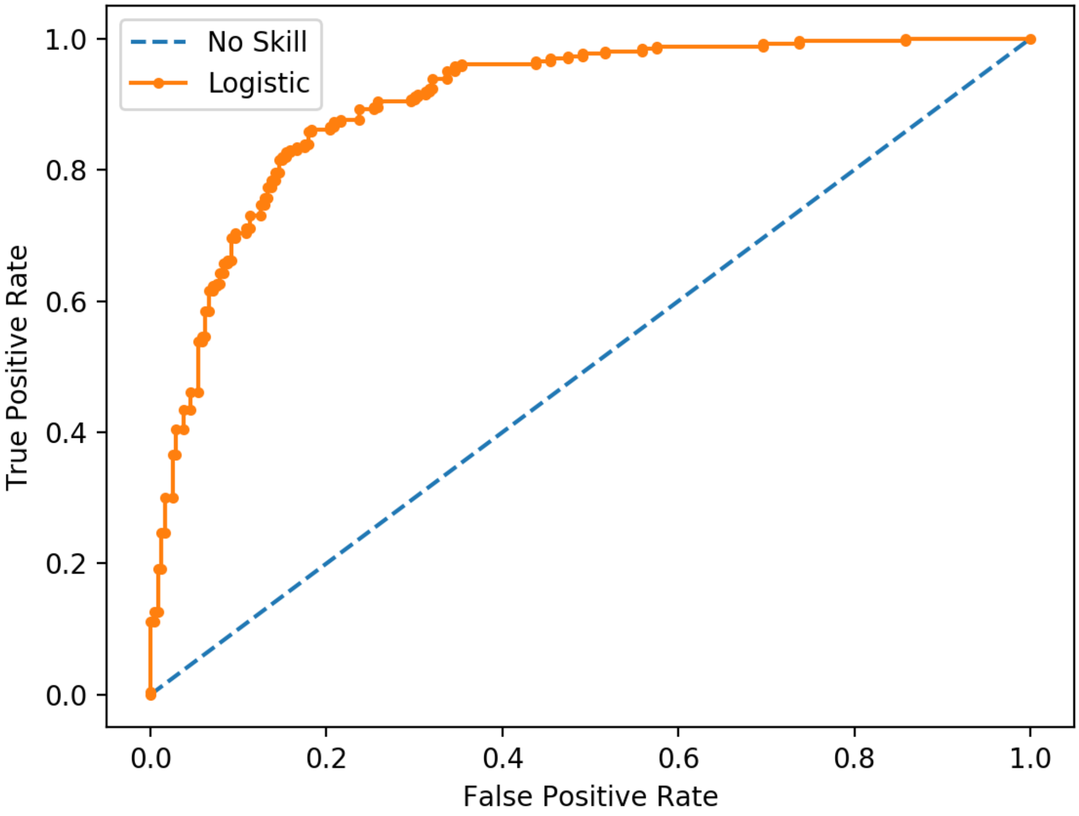
AUC

The way to analyze the results we gone use the AUC (area under curve) it gone give us an accuracy for every model, and help us to decide witch model is the best. This model is based on a confusion matrix that will give us the true positive, true negative, false positive, false negative.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

When the matrix is composed, we will be able tom make the roc curve and calculate the AUC. The AUC represent a value between 0 and 1. How closer the result is to 1 how better the prediction of the model is. Her below us an AUC curve for logistic regression (orange).



Tuning Parameters

For many models, you also have some parameters. When you try to make a model as good as possible, it is important that you take the best parameters. By setting some hyper parameter tuning. The model gone run for every-selected parameter and choose the model with the best parameters.

Below you find a list of every parameter that I selected for each model as well as the meaning and the best result.

1) Logistic regression: For logistic regression, we did not need any tuning parameter.

2) Ada boosting: the fellow parameters where used for ada boosting:
    - Maxdepth : maximum leaf nodes minus one
    - Best model was with maxdepth=20

3) RandomForest model
    - Ntree : parameter do decide the number of trees
    - Mtry : indicates the number of input variables
    - Best model was with ntree=750; mtry=14
4) Knn Model
    - K: number of nearest neighbors
    - Best model was with k=10
    - There is also a parameter that decide with method to use for calculating the distance but I didn’t include it in the model.
5) Gradient boosting

    - Important to know that gradient boosting has more parameters then I mention her below but I did not include them, because execution time was otherwise too big.
    - n.trees: parameter do decide the number of trees
    - interaction.depth: depth of tree
    - shrinkage: reduction in the effects of sampling variation
    - Best model was with n.trees=700; interaction.depth=2; shrinkage=0.01

Results

Model	AUC
Logistic Regression	0.81297520661157
Random Forest	0.736091779034363
Gradient boosting tree	0.805582862113963
KNN	0.695817746846455
Ada boosting	0.806185297955633

In the table above we have the result of the different models where we obtain a fivefold Cross validation. When we look at the result we can clearly see, that Logistic Regression outperform the other models. With an AUC of 0.81.

The models are very close with ada boosting and gradient boosting. Although this models have good scores. The models are difficult to interpret the importance of the features why is often not ideal for business related cases. In addition, the Gradient boosting took me around 8 hours before it finished running because all the parameters.

The Random Forest and KNN has performed poorly if you compare it with the other models.