Statistical & Machine Learning Approaches for Marketing

Individual Assignment

Clement Bannem Jr.

1. Logistic Regression

Logistic regression is a supervised classification algorithm. During a classification situation, the target variable often take discrete values for a given set of features. The logistic regression becomes a classification technique when a decision has to be made. The classification problem has a high influence on the setting of the threshold.

Logistic regression can have multiple naming convention depending on the type of study you want to make. Binomial, Multinomial and ordinal. Through regression, it implicitly estimate the probabilities of the classes of each features in order to come up with a classified data points according to a decision scale.

Advantages:

- It is easier to implement, interpret and very efficient to train
- It performs well when the dataset is linearly separable
- It is less prone to over-fit except in high dimensional datasets.

Disadvantages:

- It is not advised to be used when the number of observations is lesser than the number of features. It can lead to overfitting.
- Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

2. Random Forest

The general idea behind the random forest is that it combines numerous decision trees that functioning together. Inside the random forest, each individual decision tree has a class prediction and the class with the highest amount of votes becomes the prediction model.

To fit the data in a random forest classifier, we use **Scikit-Learn.** During training, we give the model the features and have it make predictions about the labels.

Advantages:

- Random forest can solve both type of problems that is classification and regression and does a decent estimation at both fronts.
- It has the power to handle large datasets with higher dimensionality.
- It is effective and efficient to estimate missing data and maintains a decent accuracy when large proportion of the data are missing.

Disadvantages:

- It can be seen as a black box approach.
- It is not efficient for regression problems.

3. eXtreme Gradient Boosting

XGBoost is a machine-learning algorithm for which the main goal is to improve the efficiency of compute time and memory resource. The engineering behind was also to make the best use of available resources to train the model.

It is an improve implementation of gradient boosting decision trees.

Advantages:

- It allows user to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run.

- It has an effective tree pruning. Indeed, a XGBoost make splits upto the max_depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain.
- When XGBoost encounters a missing value at a node, it tries both the left and right hand split and learns the way leading to higher loss for each node.

Disadvantages:

- If it's not tuned properly, it will learn the noise in the training set

4. Logit Leaf Model

The LLM has been created to enhance the data classification. The main idea behind LLM is that analysis on portion of data are usually more accurate and performants in term of prediction rather than analysis on the whole dataset.

The implementation of the LLM is done around two phases: Segmentation phase and prediction phase. The AUC and the TDL are often used as evaluation metrics in the LLM.

Advantages:

- It performs well in term of both accuracy and interpretability
- Since it is a Hybrid algorithm, we can say that it gather the best of multiple algorithms.

5. Naïve Bayes Classifier

Naïve Bayes is one of the simplest supervised learning algorithm. The general idea behind the model is that the incidence of the feature in a class does not depend of the other features. This is why this model is called as "naïve".

Advantages:

- Naive Bayes has very low computation cost.
- It can efficiently work on a large dataset.
- It can also be used in a text analytics problem and perform well on top of that.

Disadvantages:

- Its way of thinking about independent features. It is impossible to have in a dataset, features that are entirely independent.
- If there is no training tuple of a particular class, this causes zero posterior probability. In this case, the model is unable to make predictions