

Machine Learning Models for a Predictive Classification Problem – Benchmarking

Tatiane Dutra Bruno

**Individual Project for the Module Statistical and Machine Learning Approaches for Marketing**

Professor Minh Phan

**Master of Science in Big Data and Analytics for Business**

IESEG Business School

# Introduction

Statistical learning refers to a set o tools used to model and understand complex datasets. These tools can be classified as supervised or unsupervised.

The goal of supervised learning is to predict or estimate an output (or a target variable) based on one or more inputs. Here we use a labelled dataset to train the model, where there is the information regarding the inputs and the outputs. After that, we can apply this model on a new dataset, that only has the inputs in order to predict the value of the outputs (target). For example, we want to know if a customer will pay their bills based on their age, salary, education etc. In this example, the inputs are age, salary and education and the output is if a customer will pay his bill (Yes or Not). Examples of supervised learning include:

* Classification: The target variable is categorical, for example, customer will pay the bill (Yes or Not), Color of the Hair (Blonde, Black or Ginger) etc.
* Regression: The target variable is continuous, for example the salary of a professional, the price of a product etc.

The goal of unsupervised learning is to understand the relationship and the structure of the dataset. Here we can also use one or more inputs to train the model, but we don’t expect that our model predicts any value for a target variable, however we expected that our model groups the datapoints that have similar behavior. For example, we can analyze a clickstream data of a website that records the action of all users and separate them as users who visits the website daily and users who visits the website monthly or separate them as their interests, like users who reads articles related to culture, trips or news etc. Examples of unsupervised learning include:

* Clustering, grouping together data points with similar data.
* Association, understanding how certain data features connect with other features.

In this paper, we will discuss about five algorithms used on **supervised learning** approach for **classification** problem.

# 5 machine learning predictive algorithms

Variable selection, dimensional reduction method (Stepwise, F-score, Boruta, etc.)

F-score – Geneneral

Lasso – Logistic Regression

Boruta – Classification Tree and Random Forest

Logistic Regression

Linear Discriminant Analysis

Decision Tree

Random Forest

Support vector machines

# Evaluations Methods

Accuracy

Error Rate

Confusion Matrix

ROC Curve and AUC (Area Under the Curve)

# Benchmarking

## The Data Set

o Cross-validation method (holdout, k-fold CV, etc.)

o Evaluation metric (AUC, Accuracy, etc.)

o Any other methods

# Conclusion

# Sources

Machine Learning Package: <https://mlr.mlr-org.com/articles/tutorial/integrated_learners.html>

https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/