Machine Learning Project – UCI Spambase Data

Learning Objectives

This program enables the participants to review the learnings of the Machine Learning using R / Python program.

The primary objective of the project is to enhance the participant's knowledge of machine learning skills.

Background

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography

Our collection of spam e-mails came from our postmaster and individuals who had filed spam. Our collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

Refer https://archive.ics.uci.edu/ml/datasets/spambase

Project Objectives

The goal of this project is to predict spam or good mail. This is the "class" variable in the training set. You will create a report showing how you built your model, how you used cross validation, confusion matrix and why you made the choices of the final model. Finally, we will also use our prediction model to predict 20 different test cases.

Procedure

- The training data for this project are available here:
  Provided herewith as spambase.csv. Original available at
  https://archive.ics.uci.edu/ml/datasets/spambase
- Perform necessary EDA, VDA.
- Perform cleaning & imputation activities as required. Outliers, if any, should be converted to value of Lower Range & Upper Range.
- Collinearity is to be eliminated.
- Use the training dataset and create machine learning model using the best / optimum algorithm.
- Considering you would be using some ensemble algorithm, what would be the tree count & max feature count. Why?
- Clearly show method of benchmarking performance (accuracy, cross validation and confusion matrix, AUC) for each method evaluated.
- Use the testing dataset, evaluate the performance of your model.
- Show method of evaluation of performance during testing stage.

Submission

Please submit the project in a Jupiter Notebook. This will be evaluated on your own laptop.

Acknowledgement

The data for this project comes from this source: https://archive.ics.uci.edu/ml/datasets/spambase. Since you are using the dataset for your project, please cite the source as UCI Machine Learning Datasets have been very generous in allowing their data to be used for this kind of assignment.

Summary

- Describe your experience of using R / Python for Data Analytics.