SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

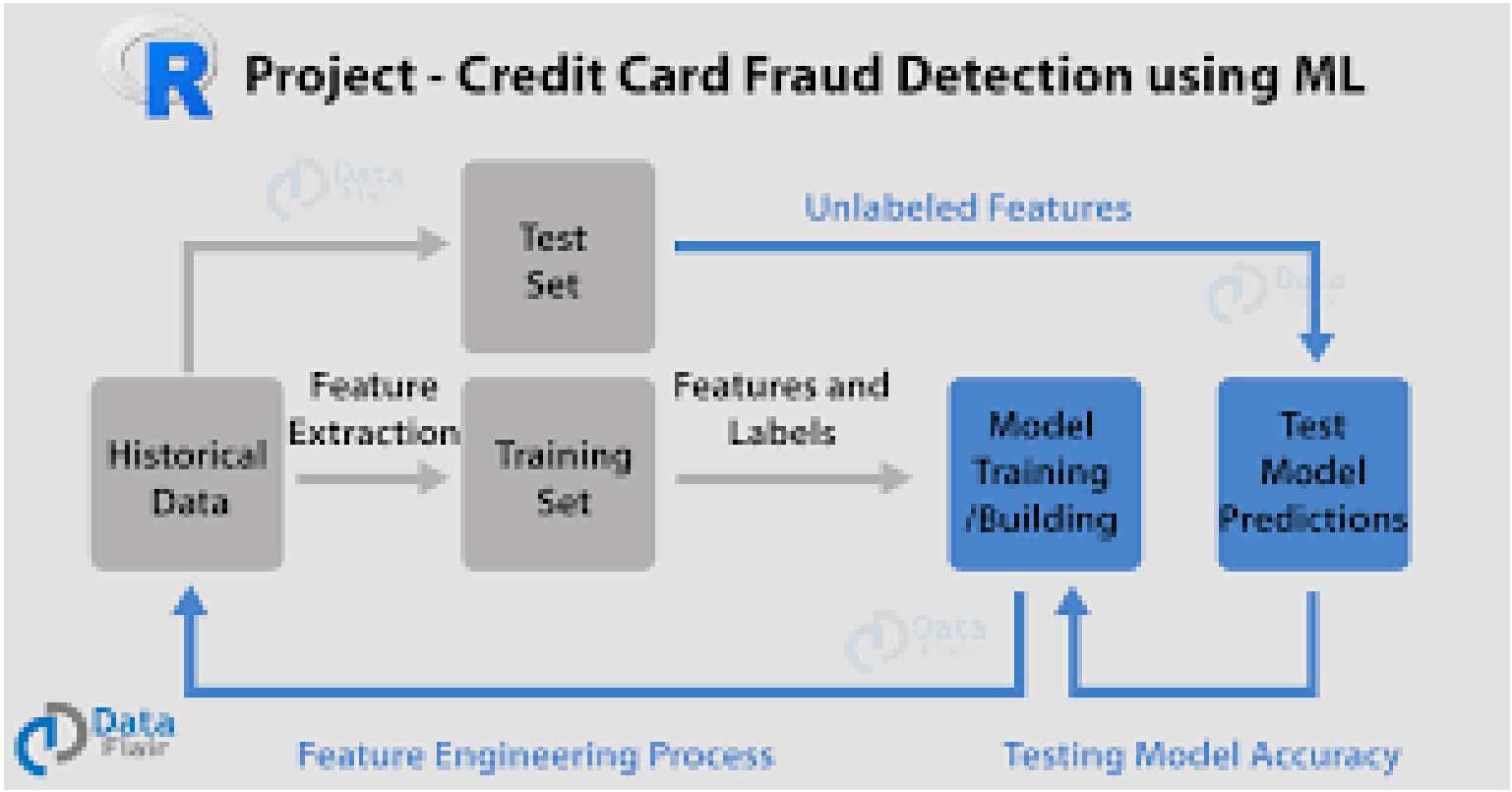
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME :19CS407 DATA ANALYTICS WITH R

II YEAR / IV SEMESTER

Unit 1- Introduction

Topic : A Project on Data Analytics - A Little History on Methodologies for Data Analytics


Project - Credit Card Fraud Detection using ML
Test Set
Unlabeled Features
Historical Data
Feature Extraction
Training Set
Features and Labels
Model Training /Building
Test Model Predictions
Feature Engineering Process
Testing Model Accuracy

A Project on Data Analytics

- ✓ Every project needs a plan. Or, to be precise, a methodology to prepare the plan.
- ✓ A project on data analytics does not imply only the use of one or more specific methods

A Project on Data Analytics

- ✓ understanding the problem to be solved
- ✓ defining the objectives of the project
- ✓ looking for the necessary data
- ✓ preparing these data so that they can be used
- ✓ identifying suitable methods and choosing between them
- ✓ tuning the hyper-parameters of each method
- ✓ analyzing and evaluating the results
- ✓ redoing the pre-processing tasks and repeating the experiments

hyper-parameters

- The values of the hyper-parameters are set by the user, or some external optimization method.
- The parameter values, on the other hand, are model parameters whose values are set by a modeling or learning algorithm in its internal procedure.
- When the distinction is not clear, we use the term parameter. Thus, hyper-parameters might be, for example, the number of layers and the activation function in a multi-layer perceptron neural network and the number of clusters for the k-means algorithm.

Examples of parameters

- ✓ weights found by the backpropagation algorithm when training a multi-layer perceptron neural network and the distribution of objects carried out by k-means

- ✓ a methodology from Academia, KDD
- ✓ a methodology from industry, CRISP-DM

Breast Cancer in Wisconsin

- Breast cancer is a well-known problem that affects mainly women. The detection of breast tumors can be performed through a biopsy technique known as fine-needle aspiration.

- This uses a fine needle to sample cells from the mass under study. Samples of breast mass obtained using fine-needle aspiration were recorded in a set of images .

- Then, a dataset was collected by extracting features from these images. The objective of the first problem is to detect different patterns of breast tumors in this dataset, to enable it to be used for diagnostic purposes

Polish Company Insolvency Data

- The second problem concerns the prediction of the economic wealth of Polish companies. Can we predict which companies will become insolvent in the next five years?

- The answer to this question is obviously relevant to institutions and shareholders.

A Little History on Methodologies for Data Analytics

- ✓ Machine learning, knowledge discovery from data and related areas experienced strong development in the 1990s. Both in academia and industry, the research on these topics was advancing quickly.
- ✓ Naturally, methodologies for projects in these areas, now referred to as data analytics, become a necessity.
- ✓ In the mid-1990s, both in academia and industry, different methodologies were presented

A Little History on Methodologies for Data Analytics

- ✓ The most successful methodology from academia came from the USA. This was the KDD process of Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth . Despite being from academia, the authors had considerable work experience in industry.
- ✓ The most successful tool from industry, was and still is the CRoss-Industry Standard Process for Data Mining (CRISP-DM) [8]. Conceived in 1996, it later got underway as an European Union project under the ESPRIT funding

A Little History on Methodologies for Data Analytics

- In 1999 the first version was presented. An attempt to create a new version began between 2006 and 2008 but no new discoveries are known from these efforts.
- CRISP-DM is nowadays used by many different practitioners and by several corporations, in particular IBM. However, despite its popularity, CRISP-DM needs new developments in order to meet the new challenges from the age of big data.
- Other methodologies exist. Some of them are domain-specific: they assume the use of a given tool for data analytics. This is not the case for SEMMA, which, despite has been created by SAS, is tool independent. Each letter of its name, SEMMA, refers to one of its five steps: Sample, Explore, Modify, Model and Assess

A Little History on Methodologies for Data Analytics

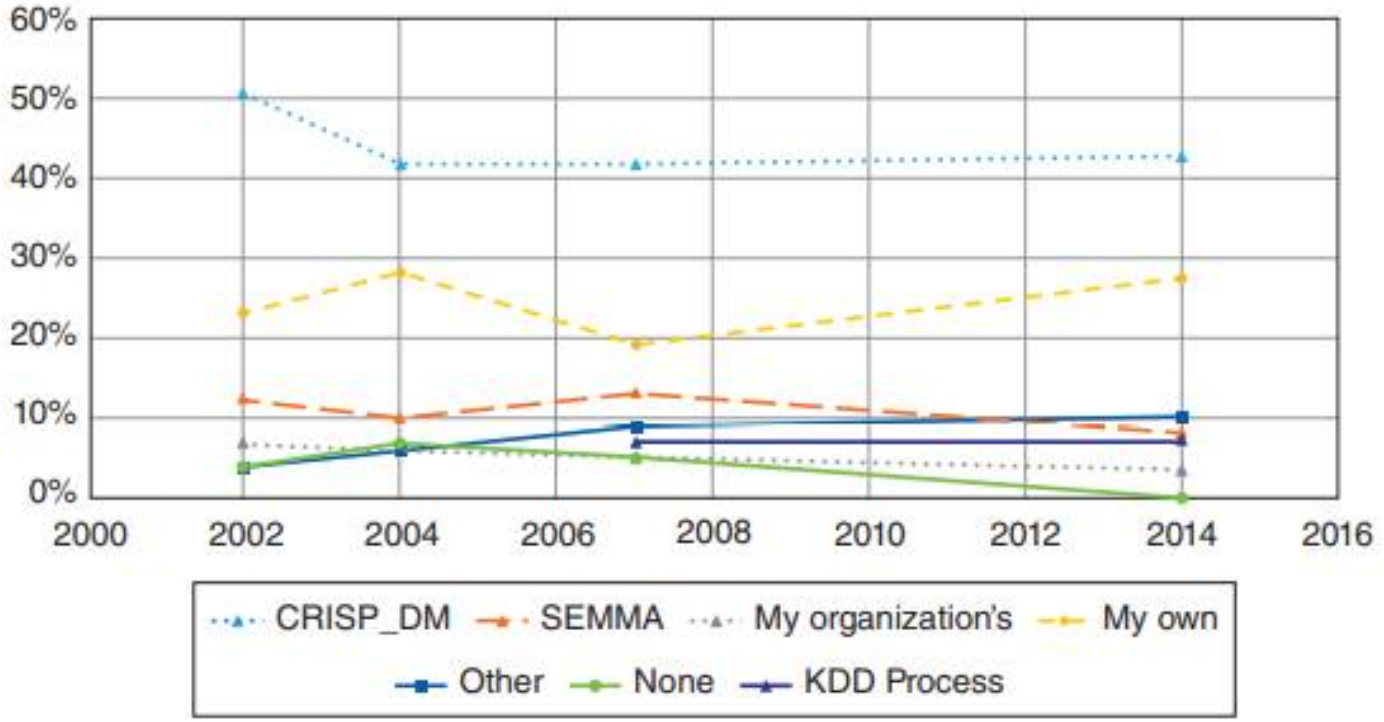


Figure 1.2 The use of different methodologies on data analytics through time.

Assessment 1

To create your own Data Model for above 2 problem

References

1. João Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

Thank You