CAE-II

Name: Mohnish Devaraj

Reg No: 39110636

Roll No: 195115398

Date: 08 Apr 2022

Subject Name: Data Science

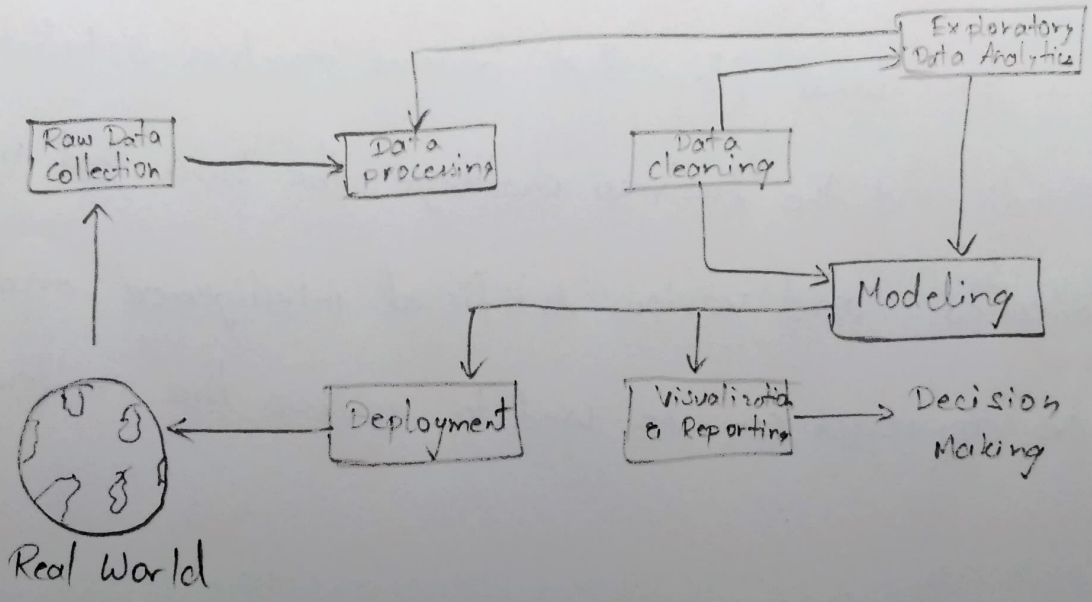
subject code: SCSA3016

Year : III year

Total pages: 6

PART-B

(6)



The key steps involved in Data Science Modeling are:

step 1: Understanding the Problem

step 2: Data Extraction

step 3: Data cleaning

step 4: Exploratory Data Analysis

step 5: Feature selection

step 6: Incorporating Machine Learning Algorithm

step 7: Testing the Models

step 8: Deploying the Model

Step 1 : Understanding the problem

The first step involved in Data Science Modeling is understanding the problem. A Data Scientist listen for keywords and phrases when interviewing a line-of-business expert about a business challenge. The Data Scientist breaks down the problem into a procedural flow that always involves a holistic understanding of the business challenge, the Data that must be collected, and various Artificial intelligence and Data Science approach can be used to address the problem.

step 2: Data Extraction

The next step in Data Science Modeling in Data Extraction Not just any Data, but the unstructured Data pieces you collect, relevant to the business problem you're trying to address. The Data Extraction is done from various sources online, surveys and existing Databases.

step 3: Data Cleaning

Data cleaning is useful as you need to sanitize Data while gathering it. The following are some of the most typical causes of Data Inconsistencies and Errors:

- Duplicate items are reduced from a variety of Databases
- The error with the input Data in terms of Precision.
- changes, Updates and Deletions are made to the Data entries
- Variables with missing values across multiple Databases.

---

(8) There are three types of machine learning algorithm

a) Supervised Learning

It is based on the results of a previous operations that is related to the existing business operations based on previous ~~algorithm~~ pattern, supervised learning aids in the prediction of an outcome. Some of the Supervised Algorithm are:

→ Linear Regression: Regression models a target prediction value based on independent variables

→ Random Forest: Random Forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision.

→ Support Vector machine: SVM is a supervised machine learning algorithm that can be used for both classification or regression challenges.

b) Unsupervised Learning

This form of learning has no pre-existing consequence or pattern. Instead, it concentrates on examining the interaction and connection between the presently available Data points. Some of the unsupervised learning algorithms are:

→ KNN (K-Nearest Neighbor): k-nearest neighbors algorithm is a non-parametric supervised learning method

→ Hierrarchical Clustering: It is an algorithm that groups similar objects into groups.

→ Anomaly Detection: It is the identification of unexpected events.

c) Reinforcement Learning

It is a fascinating Machine Learning technique that uses a dynamic Dataset that interacts with the real world. In simple terms, it is a mechanism by which a system learns from its mistakes and improves over time. Some of the Reinforcement Learning algorithm are:

→ Q-Learning: It is an off policy reinforcement learning algorithm that seeks to find the best action to take given the current state.

-> SARSA (state-action-reward-state-action): It is an algorithm for learning a Marker decision process policy, used in the reinforcement learning area of machine learning.

-> Deep Q Network: Replaces the regular Q-table with neural network. Rather than mapping a state-action pair to a q-values.

PART - A

(1) Exploratory Data Analytics (EDA) is a crucial process of using summary statistics and graphical representations to perform preliminary investigations on data in order to uncover patterns, detect anomalies, test hypotheses and verify assumption.

(2) a) Misleading Color Contrast

b) Improper Use of 3D Graphics

c) Too much Data

d) Omitting Baselines and Truncating Scale

e) Biased Text Descriptions

f) choosing the wrong visualization method

g) Confusing Correlation

h) Zooming in on Favourable Data

i) Eschewing Common Visual Association

j) Using Data Visualization in the First place.

(3) Feature selection is about selecting the subset of the original set, whereas feature extraction creates new features.

→ chi-square test

→ Fisher's score

→ Correlation coefficient

→ Dispersion ratio

(5) A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

n=165	Predicted: NO	Predicted: YES	
Actual: No	TN=50	FP=10	60
Actual: YES	FN=5	TP=100	105
	55	110	

(4)

Regression Algorithm	classification Algorithm
In Regression, the output variable must be continuous nature or real value	In classification, the output variable must be a discrete value
Regression Algorithms are used with Continous data	Classification Algorithms are used with discrete data.
The task of the regression algorithm is to map the input value (x) with the continous output variable (y).	The task of the classification algorithm is to map the input value (x) with the discrete output variable (y).