Satges of Analytics
----

1. Descriptive Analytics
----

what happened in the past and present

Example: Number of Covid cases to across various countries

2. Diagnostic Analytics

Why It has happened

Example: why Covid-19cases are increasing

3. Predictive Analytics

What will happen

Example:What will be the no of covi-19 cases for the next month

4. Prescriptive Analytics

Try toprovide remedies and solution for what will happen future?

Example: what should we do toavoid the spread of Covid-19

CRISP DM
----

cross industry standard process for data mining

1. Business Understanding/Defining the problem

Problem Statement: significant proportion of customers who take loan are unable to repay
Business Objective: Minimize defaulters
Business Constraint: Maximize profits
Project Charter

Problem Statement:Significant proportion of customers are complaining that they did not do the credit card transaction
Business objective : Minimize fraud/minimize fault transaction
Business constraints: Maximize customer convenience

Smart data platforms can bring together customer transactions data and data from realtime communication streams to disclose the insights concerning customers feelings about the services which allows addressing the satisfaction-related issues and churn prevention.

Business Objective:Min customer churn
Maximize customer services
minimize down time
Business constraint:Max customer satisfaction

Collection of positive & negative reaction to the service or product from social media sources, recent trends via Customer sentiment analysis. This may provide an opportunity of utilizing mechanisms for direct responding.

objective:

Maximize customer openion

Maximize customer's sentiments

Maximize direct responding

Constraint:

Maximum customer satisfaction

Data Collection
----
primary:
pros: features are as per our requirement
cons: time taking process
methods: survey,experiment,interview

Secondary:
pros: data is easily available
cons: you have to do a lot of data cleansing
methods: collect data from net or any database

Business Objective and constraints should be SMART

Key Deliverable:Project Charter

Specific
Measurable
Attainable
Reliable
Time-bound

2. Data Collection
----
understand various data types

Data Types

Continuous Vs Discrete
----
any number which can be represented in decimal it is continuous and can't be represented it is discrete

Categorical

Example:
1.Country ofResidence
2.Gender
3.Pay on time/Late

Count data

1.Total number of Laon defaulter
2.No.of peoplewho claimed insurance
3.Number of cancer patients
4.Number of attendance in a concert
5.No of Covid cases

Qualitative Vs Quantitaive
===========================
Qualitative means non numerical data while Quantitaive means numerical data

Example for Qualitative

-pretty- ugly/ long-short
-This kitten is small
-This weighs heavy

Example of quantitaive

- weight is 85 kg
- Height is 180cm

Qualitative/Categorical

-Binary
-Multiple
----
nominal--->where order does not matter

ordinal----> in some order

Structured Vs Unstructured

Structured means in tabular form while other than tabular is unstructured data
for ec: text speech video etc

cross sectional Vs Timeseries data

-cross sectional data is that where date time sequence does not matter while it matters in case of time series.
- cross sectionaldata actually contains more than one table

Balance data vs Imbalance data
----
output-->2class--->Default/Non default----> equal proportion--->Balance data other wise Imbalnced data

for example: 53 %deafult and 47% non default(Balnced data)
83 % default and 17 % non default (Imbalanced data)

minor data couldnot be less than 30%

to overcome this issue

1.Randome oversampling or undersampling(Resampling)---- undersample(when data is sufficient)
2. K-fold cross validation
3. SMOTE

4. Cluster based sampling
5.Ensemble technique
6. Use right evaluation metrics

Data collection Sources
=======================
1.Primary : as per the requirement, no much data cleansing is required here
forex:data collected through surveys, list of experiments, from IoTsensors
2.Secondary:not as per the requirement,data cleansing is required

3A. Data Cleansing/Data Preperation
===================================
Data organization,Data Munging and Data wrangling

1. Outlier treatment

- the show abnormal distance from the other data points
- when we have outlier in our dataset then we use to prefer median instead of mean because
 mean are affected by outlier.
- if you have outlier then they will affect the correlation values sowhenever you have to use this correlation
  use data without outlier because correlation values are very sensitive to outliers
Rectify
Retain
Remove

A.Winsorization

-Minimizing the influence of outliers
-Winsorization is the technique which modifies the sample distribution of random variables
 by removing outliers.
- For example 90% outliers means all the data below 5th percentile is set at 5th percentile
  and all the data above the 95th percentile is set at 95th percentile.
90%   5% 90 95%

B.Alpha Trimming
----
lets you set an aplha value.
-for example if alpha=5% then all the lower and upper 5% values are trimmed or removed.

2.Missing Values

=================

Types of missing values can generally be classified as:

Missing Completely at Random (MCAR)

This happens when the missing values have no hidden dependency on any other variable or any characteristic of observations. If a doctor forgets to record the age of every tenth patient entering an ICU, the presence of missing value would not depend on the characteristic of the patients.

Missing at Random (MAR)

In this case, the probability of missing value depends on the characteristics of observable data. In survey data, high-income respondents are less likely to inform the researcher about the number of properties owned. The missing value for the variable number of properties owned will depend on the income variable.

Missing Not at Random (MNAR)

This happens when the missing values depend on both characteristics of the data and also on missing values. In this case, determining the mechanism of the generation of missing value is difficult. For example, missing values for a variable like blood pressure may partially depend on the values of blood pressure as patients who have low blood pressure are less likely to get their blood pressure checked at frequently.

-we delete the row in which the missing value is present
-we delete the column in which the missing value is present
- we shall impute the cell with mean/median/mode/random/Hot deck/Regression/KNN imputation
Hot deck Imputation
----
we replace the missing values with an observed responses from a similar unit.
KNN Imputation

K sample in the dataset will be taken which are similar to the data point which is missing value.we will replace the missing values with these k samples.
3.Normalization/Standardization
----
to make your data unit free and scale free

4. Dummy variable

----

converting categorical into numerical data

1.One hot encoding

2.Label encoding

3B. Exploratory Data Analysis

4.Model Building

5. Model Evaluation

6. Deployment