Data Science Project

CRISP-DM (Cross Industry Standard Process – Data Mining)

1. Business Problem Understanding
2. Data understanding
3. Data pre-processing
4. Modelling
5. Evaluation
6. Presentation

Business Problem Understanding

- Understand the extract problem and the requirements of the client

Ex: Identify the loan defaulters in a bank

Data Understanding

- Understand the each and every variable clearly (if required take the help of domain expert)
- Understand the dataset (shape & info)
- list the discrete variables and continuous variables

Data Pre-processing

- **EDA (Exploratory Data Analysis)**
  - Understanding given raw data clearly by applying pandas, Matplotlib & Seaborn
  - Both Discrete and continuous variables separately
    - Univariate Analysis
    - Bivariate Analysis
    - Multivariate Analysis
  - Check & list how many variables are having
    - missing values
    - outliers
    - skewed

        - discrete nominal
        - discrete ordinal

- **Feature Selection**
    - Drop unimportant variables (constant valve, uniques)

- **Data Cleaning**
    - Treat the missing values (mandatory)
    - Treat the outliers

- **Data Wrangling**
    - Convert skewed to normal
    - Convert discrete categorical to discrete count (mandatory)

- **Train test Split**

- **Feature Scaling**

- **Dimension Reduction (PCA)**

Modelling & Evaluation

- Select various Algorithms which are applicable for given problem.

For Each and Every Algorithm

- **with default parameters**

  Import Algorithm
  +
  Fit the train data
  +
  Predict on train, Predict on test
  +
  Calculate the train, test accuracy & cross validation score
  +
  Evaluation

    - Check for Overfit or Underfit or good model
      (Good Model will have train accuracy = test accuracy)
    - Check test accuracy = cross validation score

Apply Hyper parameter tuning for each algorithm separately

+

Identify the best parameters

Rebuilt the model with same algorithm by considering the best hyperparameters

+

Calculate the train, test accuracy & cross validation score

+

Evaluation

- Check for Overfit or Underfit or good model (Good Model will have train accuracy = test accuracy)

  and

- Check test accuracy = cross validation score

- Save the train accuracy, test accuracy and CV values for ML Algorithm

Model Selection

- **Once each and every algorithm is completed, identify which is the best model with hyperparameters which satisfies the business requirements**

Save Model

- Save the final model (which has given maximum accuracy) as a pickle file or joblib file as per requirements of deployment team