Dataset : <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

This is the dataset collected in marketing campaigns done by the bank. The dataset has both numerical and categorical columns. Goal is to build a model which will help the marketing team to identify potential consumers who will be more likely to invest in new financial products ex: term deposit offered by bank.

Summary of things completed:

1. Doing the extract discovery analysis work.
2. Features datatypes and distribution of data analysis.
3. Features selection.
4. Data Cleanup and preparation.
5. Use of Pandas profiling module.
6. Data visualization using matplotlib.pyplot and seaborn packages.
7. Building the prediction models.
8. Evaluating the Accuracy, precision, recall for each class across models.
9. Selecting the best model with highest accuracy, precision, and recall.
10. Saving the model into pickle file and exposing the best model functionality through flask/streamlet Apis.

* EDA work
  + - During the Exploratory Data Analysis work, performed the single and multiple variables analysis. Gained some insights into how unique features are correlated with response variables and data distribution between various features.
    - EDA work conducted by custom queries ex: between education and balance :

for i in bank\_df["education"].value\_counts().index:

for j in bank\_df["y"].value\_counts().index:

education\_balance\_amt = bank\_df[(bank\_df["education"] == i) & (bank\_df["y"] == j)]["balance"].sum()

percentage = round(education\_balance\_amt\*100 / bank\_df["balance"].sum(),3)

print(f"Education level '{i}' and Deposit status '{j}' amount: {education\_balance\_amt}, percentage: {percentage}")

* Understanding the categorical variables : job, marital , education, default, housing, loan, contact, poutcome, month by using countplot of seaborn package.
* Panda’s profiling
* Data Cleaning and Feature Construction
* Around 37K records having poutcome=’unknown’ and contact=’unknown’ which are not of significant have been discovered and deleted.
* Custom encoder for Boolean value type have been implemented.
* Custom Dummifier for Categorical nominal variables have been completed.
* Feature selection:
* CustomCorrelationChooser, SelectKbest, SelectFromModel have been utilized to understand the features importance.
* Select the best parameters for estimators using GridsearchCV.
* Model Building:
* From the dataset , we are considering 20% as test data and 80% as train data.
* All below ML algorithms have been applied for finding the accuracy,TP,TN,FP,FN,TPR, TNR values.
* Best value of K for knn being determined by using plot of accuracy vs k param function.
* For Randomforest classifier as well, same methodology has been applied.

Text, table

Description automatically generated

* Pickle file and deploy the model:
* Deploying the model using the pickle file.
* Streamlit library is being used to deploy the model.

CODE FILES:

* 1. CS677\_Rajkurki\_project.ipynb : **Run through Jupyter notebook**.
  2. C:\Users\User\PycharmProjects\StreamLit\_ModelAPI :

Streamlit webapp using streamlit package.

COMMAND TO RUN **=> C:\Users\User\PycharmProjects\StreamLit\_ModelAPI> python -m streamlit run main.py**

A picture containing application

Description automatically generated

Background pattern

Description automatically generated

Background pattern

Description automatically generated