Email Spam detection:

Problem Description:

Here in this particular dataset we are required to predict whether the message present in the dataset is spam or not. Since the target here is to predict only yes or no(1 or 0) this here is a classification problem to predict the target(label).

First we are required to import all the necessary libraries from python to import , analyze and solve the problem at hand which in this case is classification(whether the person is defaulter or not) . After importing all the libraries, we are required to load the dataset into the notebook using pandas.

Exploratory data analysis (EDA):

After importing the dataset comes the analysis part. Detailed Exploratory data analysis must be performed so that we can understand the problem and choose the necessary steps required for model training. Here analysis was done about dataset

- Some null values were observed
- Message column was sting format which had many characters
- ‘1’ represents yes whereas ‘0’ represents no
- Subject column had negligible importance

Data preprocessing(Feature Engineering):

Feature engineering plays a vital role in proper functioning of a model. So it is absolutely necessary that the features must be treated precisely. From EDA we concluded many things and how we must act on it. Subject feature had no importance and had to be dropped. Message column had many characters in string format so we had to split them and convert them into computer readable format. That was done using countvectorizer.

Model Training:

Now as we have defined our dependent and independent variables or features as y and x respectively. ‘x’ is all the features excluding ‘label’ which is out dependent variable ‘y’. Now using train_test_split method we split the data into 67% for training and 33% with random state as 42. Hyperparameter tuning was done on KNNeighborsClassifier(n_neighbors=1) and RandomForestClassifier (n_estimators=13) with GridSearchCV and best_params_ was applied to give the best parameter for approach. The algorithm we applied gave pretty descent scores.

Report :

We trained the model and found the results for various algorithm . But LogisticRegression gave the best result. It is wise to save the model with the same parameters and conditions.

Model	Accuracy Score	Cross_val_score	Roc_auc_curve
KNeighborsClassifier	94.725511	93.602383	90.343745
SVC	95.048439	94.420383	84.563758
DecisionTreeClassifier	95.263724	95.699881	90.935725
LogisticRegression	99.461787	98.685672	98.593616
RandomForestClassifier	97.847147	97.298907	93.560059