Loan Status

Objective:- To Build a model to predict if a customer would be approved/rejected for a loan application based on the details provided by him,These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others.

Steps used are:- Exploratory data analysis - Data Preprocessing steps - Model Selection - Model Evaluation - Model Interpretation

---

Exploratory Data Analysis

- In this step we check for the dependance of the features like Gender, Education, Credit History etc. on the label, if loan should be approved or not.
- In this step, we check the distribution of the data.

Data Preprocessing Steps:

- During data preprocessing, I found some values in feature columns missing. So I replaced the missing values with mode of that particular column. I did it as the mode represent the most frequently occuring value in a data. And it is dominant in this case, as compared to mean and median.

Model Selection:

- Actually here I used a combination of models. As the dataset is small for machine learning, So I chose to experiment with various model.
- The models are:
    - Random Forest Classifier
    - Decision Tree Classifier
    - Adaboost Classifier
    - Gaussian NB
    - KNeighboursClassifier
    - XGBClassifier
    - Logistic Regression

Model Evaluation:

- Next step as model evaluation, I set evaluation parameters as accuracy score, precision, recall and F1 score. These parameters give good insight of how well the model is performing.
- The following are the scores of different classification models:

model_name	accuracy_score	precision_score	recall_score	f1_score
Gaussian Naive Bayes	0.789189	0.816462	0.717628	0.73444
XGBoost	0.789189	0.825282	0.714103	0.731075
Random Forest	0.778378	0.792248	0.709295	0.724192
AdaBoost	0.772973	0.780998	0.705128	0.71913
Logistic Regression	0.783784	0.843722	0.699359	0.71477
Decsision Tree	0.72973	0.702448	0.692949	0.696761
K Nearest Neighbor	0.589189	0.436121	0.471795	0.424337

Model Interpretation:

As the task is to classify the Loan Status into YES/NO category, The machine learning libraries for classification task are used. And as we can see the table above, which give us insights into the model performance in decreasing order. Amond these, Gaussian Naive Bayes and XGBoost have performed almost equally. And the accuracy is quite good. When data amount is not issue, the performance will increase as the machine learning algorithms will have more data to train on. But when data exceeds certain limit, we should shift to Neural Networks/Deep Learning. They work very well with big data.