Report

Exploratory Analysis of data

Link:https://github.com/yashchoubey/Enlightiks-Business-Solutions/blob/master/ExploratoryAnalysis.ipynb

- Clearly data contains many null field.
- Feature_3 population only contains 1 (total count ~110)
- Feature_9 contains only null.
- Feature_14 population only contains 1 (total count ~800)
- Feature_4/Feature_5/Feature_6 has nearly same data distribution.

Strategies:

There can be 2 ways to solve this classification problem.

- Strategically impute missing value.
- Use algorithm which is resistant to missing value.

LDA(Linear Discriminant Analysis):

Link: https://github.com/yashchoubey/Enlightiks-Business-Solutions/blob/master/PCA_LDA.ipynb

LDA is a method used in machine learning to find a linear combination of features that characterizes or separates two or more classes. It is not resistant to missing value.

I've used **StandardScaler** to standardize the data and imputed the columns with values:

Feature no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Replace strategy	M	M	1	0	0	1	0	0	0	1	M	M	0	0	M	M	M	0	0

- M: Mean of column

Also I have used PCA as a dimensionality-reduction technique apart from dropping Feature_3, Feature_9 and Feature_10.

The most appropriate values n_components =12.

Results:

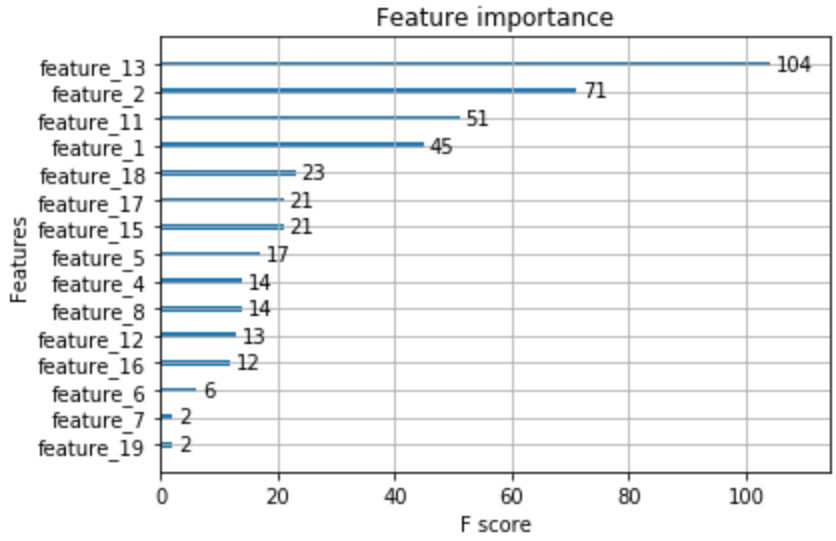
accuracy_score: 0.9207033842070338

f1_score: 0.6306027820710974

XGBoost:

Link : https://github.com/yashchoubey/Enlightiks-Business-Solutions/blob/master/xgboost.ipynb

I've used XGBoost as it is resistant to missing values and through 'plot_importance' we can easily know important features in ascending order.



Results:

accuracy_score: 0.9409422694094227

f1_score: 0.7420289855072464