Kinara Capital-Lending Decision Modelling Exercise

Learnings from Analysis and Modelling Exercise:

1. I was able to get the **flavour of the data** that I might come across for devising strategy and lending decision for MSME.
2. Data consisted of lot of categorical attributes. I did some EDA to understand the **distribution of categories** among different attributes.
3. Based on distribution we could see **categorical variables** were skewed towards specific categories. I was expecting them to act as powerful predictors.
4. I tried both **label encoding and one-hot encoding**. I found that one hot encoding was giving better result (same is implemented in code).
5. One hot encoding led to creations of lot of variables. In order to utilise maximum information gain, I decided to use nonlinear algorithm capable of handling large attributes such as **XGBOOST**.
6. In order to, benchmark the model, I also wrote **logistic model from scratch** using just Numpy in problem 2.
7. I could clearly see that XGBoost was acting as far **better predictor than Logistic Regression**.

Things I have done if I had more time:

1. Spend more time in cleaning the data and creating features using PCA if possible.
2. More EDA on data in terms of **PDP plots, bi-variate plots, and correlation plots**.
3. Optimizing **the XGBoost hyper parameters** for better model fit and results.
4. Looking at more detailed performance metrics apart from KS and AUC such as **Precision, recall, FPR, accuracy etc** and optimising them accordingly.
5. Since it consists of lot of binary variables with high correlation, we could have tried fitting **Bayesian Models** and compare for performance.