**Borrower Scoring on Agriculture Sector**

1. **Business Understanding**

Indonesia is an agricultural country where fields or plantations area are very easy to find, and about 25 percent of the population work as farmers. You have to know that most of them come from lower economic background and also low educated, which mean there are a lot of problems there. One of the crucial problem is funding, without sufficient funding their crop will not be optimal. In other hand, there are several startup (P2P) that are focused on solving agriculture problem start from provide the seed, crowdfunding (investment), lending to borrower and also trading. Let’s take one interested topic “crowdfunding”. Crowdfunding is raising many small amounts of money from a large number of people (investor). I believe all investors have mindset that the money invested will increase in the future. But in the reality, we don’t know whether the project will succeed or fail. There are several factors influence in agriculture sector as like: farming experience, education level, field location, water source, rain intensity, etc. That’s why the startup need to consider who is suitable to be funded or not to decrease the failure rate. As long as I know, several startups select the borrower manually by sending surveyor to the location. They will observe the field, environment, and try to interview the borrower to get some information from them. But there are several factors still can’t observe manually like cropping calendar, weather prediction, harvest history, agricultural conditions around that field, borrower’s loan history, and other related factor. These factors have high probability to make a failed project that can make both of investor and company get a loss. That’s why borrower scoring in agriculture sector data scientist to decide who is recommended and isn’t recommended to be funded to prevent the loss.

1. **Data Understanding**

Let’s say that our data come from the previous borrower’s company data which has been labeled and save in the database (RDBS). We also need another external data like what commodities are successful harvests in that district or city to compare with the new borrower data. We need weather forecasting and cropping calendar data that maybe noSQL data come from web scraping. We also need their loan history in another company which can be obtained by using the API provided by AFPI Company, this data will be used to ensure that the borrower has no history of blurred payments (Json format). I believe this is a large-scale data, so we need save the data on data warehouse before processing the data. Now our task is select the relevant data for our dataset.

1. **Data Preparation**

Now we take the data from Data Warehouse trough ETL proses, but this proses is not totally clean the data. So in this process we have to clean the data from missing value, outlier, redundant data, different value input, fix inconsistent data, normalize the data, handling imbalance data, etc. To make the analysis easier do exploratory data analysis, so we can see the data from another perspective. In the end of the process we will have the clean dataset which is ready to process.

1. **Modeling**

Modeling is the stage where we determine the appropriate algorithm which is suitable with our data case. Machine Learning algorithm divided into 3 main group, that is regression, classification and clustering. For our case “Borrower scoring” is a classification problem, we decide which borrower will be funded (class 1) and will not be funded (class 0), so we can use logistic regression, svm, knn, random forest, naïve buyes, XGBoost, etc. What will we do in this stage?

The first we need to split the data into train and test, the goal is to avoid overfitting in the model to be build. Train data will be used to train our model, while test data will be used to see the accuracy of the model. Then we choose our initial classifier (for ex: SVM), we need to set the hyper parameter of the classifier. Generally we don’t know which hyper parameter value are match with our data, so the best way is to do hyper parameter tuning (use gridsearchCV or randomsearchCV), with this process we will get the best hyper parameter value which give higher accuracy for our model, moreover it can also be used to avoid underfitting. After that just train the data with the model, if the accuracy of the model still low. You can back to the preparation stage and try to reanalysis the data.

1. **Evaluating**

After we build the model, let’s evaluate our model by predict the test data and compere the predict result with the true value of our test data. Then calculate the accuracy by using matrix accuracy or confusion matrix. If the accuracy is still high and our time agree with that model, we go to deployment stage. But if the accuracy is low, you can back to the preparation or build the model using other classifier. But you must remember that don't settle for just one model, you can compare with other model and take the best one.

1. **Deployment/Prototyping**

The last, we will integrate the model with our dashboard system (in the form of website or apps) to make the credit scoring department easy to use. Now the just need to input some variables, document, or something important to our system, and without wait too long the result will come out. Surely, this system will decrease the effort of determining recommended borrower and minimize the company's loss.