Hello Mr and Ms Good Morning, Thank you for this opportunity, and let me start the presentation. Lending money has become trend in digital era right now, because of the ease of lending money and human desires that must be fulfilled such as building a bussines, fulfill needs, and even fulfill their desire. So many clients will borrow money from the bank, the bank will find it difficult if there is no system to know the costumer’s interest in lending money from the bank, because of what? Because if the bank accept all the clients to lend their money it will be risky and inefficient. So now in this presentation I will explain the filter to the costumer to find out their Intereset Rate for lending money in Loan Prediction Analysis.

Let’s see how I describe sequence start from introduction until conclusion

Here I am Enggar Kristian with background Aeronautical Engineering with passion in Data Science because love of Artificial Technology, and the first step I learmed in bootcamp at hacktiv8, so many project that I have made and done you may see my own project from bootcamp in my repositories github.

Second chapter about data loading, in this chapter we will to know deeply about the data well, there are template to be some Data Science with pricessing data use programming language python is import support libraries, load the data set we have data train and data test, we should know information and stuff that we have to understand that data, after we know of course we are as a Data Science should have clean data like handling missing values and handling outliers, and after we got structure data and all know the information of data well we do Model definition with the definition each algorithm, do model training and then we got result we should evaluation final.

Many insight it should be exist in this dataset have many size with 280.000 more rows and 14 columns inside , but I am make some insight just 4 visualization there are

First graph diagram pie is Count of Gender in Data Loan in this data set describe how many gender of client that loan to our bank, in the reality may male than female with comparison 72% appeal 28%, Yes indeed, we know that men take a lot of risks for their plans so they need capital to realize it and the goal is our bank.

Second graph is bar plot of Home Owner each Gender client for loan, for this graph we want to know how much client have their own home to be guarantee client to lending for the bank. The graph show most client is mortagage or we call KPR that is mean the client guarantees the property they bought to the bank as an asset for the mortgage loan for the purchases of the house. After credit period, the property will fully become the property of the credit borrower.

Third graph is scatter plot of Dependency Annual Income with Loan Amount by Gender client, that is summary of Annual Income client to know how much their income each gender, it shows all equal but there’s one person that’s a Male have annual income more than 7 million, this is should be outliers of the data, but this is normal, after being traced in the graph, the four clients are indeed someone who will open a new business and need capital and borrow from this bank.

Fourth graph is bar plot of Average Annual Income with purpose of Loan by Income Verified, that’s mean client lending for their need but at the most they used for make some business and home credit.

The target label from raw data set show 1, 2, and 3 but I replace number to be 0, 1, and 2 because model training can read fit training start from 0 or like an index, the target label is column Interest Rate is a levels for clients. We want to know how machine predict our clients to be efficient to make decision to their interest rate level. After replace first level of interest rate is 0, second 1, and third 2, bar plot show is imbalance target, my expectation is for make model is best result we should have to improve model with rebalancing target with oversampling or undersampling, but the reality is the target just exist in data train, not in the data test so we cannot compare betweet both data and rebalancing in data train because rebalancing for manipulation of data not for analysis data just for do better performance prediction of machine.

Next chapter is data preprocessing these are my template to doing modeling for machine learning,

the first we doing is splitting data train and test to be X\_train\_loan, y\_train\_loan, and X\_test\_loan, there is no y\_test\_loan because dataset there’s no target label.

Second we should have to handling missing values, ya because training data is not complete if contains missing values inside. There are total 4 column contain missing values these are feature Months\_Since\_Deliquency (integer) 53%, Home\_Owner (object) 15,4%, Length\_Employed (object) 15,2%, and Annual\_Income (Integer) 4.4%. In this case I should have to drop 1 column that is Months\_Since\_Deliquency because if feature have missing values have more 50% is bad feature to did manipulation data for this column contain 53% missing values. And then feature type numerik I fillna with median, and if feature type is object I filled use most frequent of object.

Third is handling outlier many outliers for numerik feature to this data, then almost all have skewed distribution so I treat outliers use trim and cap outliers.

Fourth do feature selection like split feature to be 2 types numerical and nominal categoric for easy treat to make a pipeline

Fifth after we did feature selection we know the data that have type all numerik, then from feature numerik we do scaling with minmax scaler

Sixt my expectation and as my template is doing oversampling for rebalancing data set to have better result from our data prediction. But in this case we don’t have comparison of data between data train and data test.

Seventh after we did scaling all numerik features, also we do encoding all nominal categorical features with onehotencoder, that’s mean all unique from feature nominal categorical will be dummy, of course we will do so much feature on that.

Eighth we have so many feature from onehotencoder that’s no efficient for training model, other side make running time is over. So we do dimensionality reduction with PCA and I did cut columns from the beginning is 41 feature and reduction 93% efficient column features to be last feature is just have 15 column features.

Yash at last but not least we have data that ready to modelling, that will be start from model definition of all algorithms in this case I just used Random Forest and XGBoost, but initially I did with 5 algorithm and do Cross Validation but running time until 5 hours and error, so I decide to used best algorithms for my own speculation are Random Forest and XGBoost. After I did cros validation to looking for best one algorithm base on result, I do gridsearchcv to finding best parameter alias hyperparameter tuning, and I did hyperparameter tuning for my algorithm speculation is Random Forest, wow Running time is so many waste time, I took running time until 20 hours 51 minutes, after that the result is so bad out of my expectation. So I fit the algorithms with default parameters.

Model training with default parameters is surprisingly because beside running time is fast just 6 minutes and the result have the best is 100% as we can see from classification of Random Forest beside is all 1.00 which mean machine correctly predict with all learning for training data, and then from bar plot feature importance from algorithms random forest all features are equally but we can see from feature 3 alias feature or column Annual\_Income is mostly impactfull for this prediction. Model evaluation use XGBoost is more bad prediction using this algorithm as we can see from classification reports we just have 57% accuracy that correct for prediction, and then feature importance shows tend to be dominant for feature 3 or Annual\_Income. In the other words to this prediction use machine to looking for best interest rate for candidate client please be more focus on their annual income.

Well these step are done then I’ll do model saving with the pipeline, and final data set writing after we do cleaning. So these are the conclusion