**Bank Loan Default Project**

Here is the Basic Approach to Solve this Project using Crisp-dm process as:

1. Business understanding
2. Data understanding
3. Data preprocessing
4. Model development
5. Model evaluation
6. Model deployment

1. **Businessing understanding**- here in this project as Bank loan default we have to find wheather a customer is default or not just by train model using past data. Past data has 8 variables in which 7 are independent and 1 is target variable. So from this business problem this can be stated as supervised classification problem.

2. **Data Preprocessing**- data preprocessing is a part of exploratory data analysis. Here we use data find all the patterns from it and relation among those. For data preprocessing has missing value analysis, outlier analysis, feature selection, feature scaling is done.

Here from dataset there are 17 % data is missing from the default variable. For filling this data many techniques such as mean, median, mode and KNN imputation can be used. Since by using mode and KNN technique data becomes biased that’s why I use here mean imputation method. Mean of all the default variable is imputated in place of missing values.

Feature scaling requires because there is a big difference in scaling of feature because of these model doesn’t predict the correct values. For feature scaling techniques such as Normalization and standredization are used. Here I used normalization technique for scaling.

Feature selection is process of selecting only important features from whole features . Relations between features can be derived using correlation. Correlation is a range between -1 to +1 where -1 shows highly negative correlation and +1 shows highly positive correlation b/w variables. So If two variables shows highly positive correlation then there is no need to put both features ito models we can use only one feature there because both are providing the same meaning.

Here in this project none variable shows high positive correlation that’s why I didn’t remove any feature.

Until now data preprocessing is done now modeling is the next step. Follow are the some models that are used for supervised classification problem.

**Logistic Regession** - Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘**Sigmoid function**’ or also known as the ‘logistic function’ instead of a linear function.

**Random Forest**-Random Forest is a supervised learning algorithm. Like we can already see from it’s name, it creates a forest and makes it somehow random. The „forest“ it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. The random-forest algorithm brings extra randomness into the model, when it is growing the trees. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model. Therefore when you are growing a tree in random forest, only a random subset of the features is considered for splitting a node. You can even make trees more random, by using random thresholds on top of it, for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

**Model Evaluation**- For model evaluation following parameters are used-

**Confusion matrix**-A confusion matrix gives you a lot of information about how well model does, but there a way to get even more, like computing the classifiers precision.

**F 1 score**- we can combine precision and recall into one score, which is called the F-score. The F-score is computed with the harmonic mean of precision and recall. Note that it assigns much more weight to low values. As a result of that, the classifier will only get a high F-score, if both recall and precision are high.

**Precision recall curve** - For each person the Random Forest algorithm has to classify, it computes a probability based on a function and it classifies the person as survived (when the score is bigger the than threshold) or as not survived (when the score is smaller than the threshold). That’s why the threshold plays an important part.

**SUMMARY**

I started with the data exploration where I got a feeling for the dataset, checked about missing data and learned which features are important. During this process I used seaborn and matplotlib to do the visualizations. During the data preprocessing part, I computed missing values, Afterwards I started training different machine learning models, picked one of them (random forest) and applied cross validation on it. Then I discussed how random forest works. Lastly, I looked at it’s confusion matrix and computed the models precision, recall and f-score.