**Capstone Project**

**Credit card Default Prediction**

**Technical Documentation**

**By**

# 

**Team**

**Vikrmaditya sah & Sivaramaguhan**

**AlmaBetter**

****

**Date: 12th May 2022.**

**Table of Contents**

[Abstract: 2](#_heading=h.1t3h5sf)

[Problem Statement: 3](#_heading=h.2s8eyo1)

[Data Summary: 3](#_heading=h.17dp8vu)

[Steps involved: 3](#_heading=h.26in1rg)

[1. Exploratory Data Analysis 3](#_heading=h.2xcytpi)

[2. Null values Treatment 3](#_heading=h.1pxezwc)

[3. Outliers treatment 3](#_heading=h.2p2csry)

[4. Encoding of categorical columns 3](#_heading=h.32hioqz)

[5. Standardization of features 3](#_heading=h.1hmsyys)

[6. Fitting different models 3](#_heading=h.2grqrue)

[7. Tuning the hyperparameters for better recall 4](#_heading=h.vx1227)

[Algorithms: 4](#_heading=h.1ksv4uv)

[1. Logistic Regression: 4](#_heading=h.3fwokq0)

[2. SVM: 4](#_heading=h.1v1yuxt)

[3. Random Forest Classification: 4](#_heading=h.4f1mdlm)

[Model performance: 4](#_heading=h.wru8aieqzpal)

[1. Accuracy 4](#_heading=h.2u6wntf)

[2. Precision 5](#_heading=h.19c6y18)

[3. Recall 5](#_heading=h.3tbugp1)

[4. F1-Score 5](#_heading=h.28h4qwu)

[1. Grid Search CV-Grid: 5](#_heading=h.tr9uqhwkmc0q)

# Abstract:

# 

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. In classification problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lied in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. Data level resampling techniques are employed to overcome the problem of data imbalance. Various undersampling and oversampling methods are used to resolve the issue of class imbalance. Different machine learning models are also employed to obtain efficient results. The split method is utilized to validate the results in which data has been split into training and test sets.

# Problem Statement:

This project is aimed at predicting the case of customers' default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the [K-S chart](https://www.listendata.com/2019/07/KS-Statistics-Python.html) to evaluate which customers will default on their credit card payments

# Data Summary:

The data set contains the account holder information (Age, Marital Status, Gender, history of payments, etc), and whether customer will default his next month payment or not

Attribute Information:

* X1 - Amount of credit(includes individual as well as family credit)
* X2 - Gender (1 = male; 2 = female).
* X3 - Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
* X4 - Marital Status (1 = married; 2 = single; 3 = others)
* X5 - Age(year).
* X6 to X11 - History of past payments from April to September
* X12 to X17 - Amount of bill statement from April to September
* X18 to X23 - Amount of previous payment from April to September
* Y - Default payment next month

# Steps involved:

## Exploratory Data Analysis

After loading the dataset we compared our target variable that is “Default payment next month” with other independent variables. This process helped us figure out various aspects and relationships among the dependent and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the dependent variable.

## Null values Treatment

Our data set didn’t have any null values to be treated.

## Outliers treatment

Isolation Forest Algorithm: Isolation forest is a tree-based algorithm that is very effective for both outlier and novelty detection in high-dimensional data.

## Encoding of categorical columns

We used One Hot Encoding(converting to dummy variables) to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to the numerical format.

## Standardization of features

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

## Fitting different models

For modeling, we tried various classification algorithms like:

* Logistic Regression
* SVM
* Decision Tree
* Random Forest Classification

## Tuning the hyperparameters for better recall

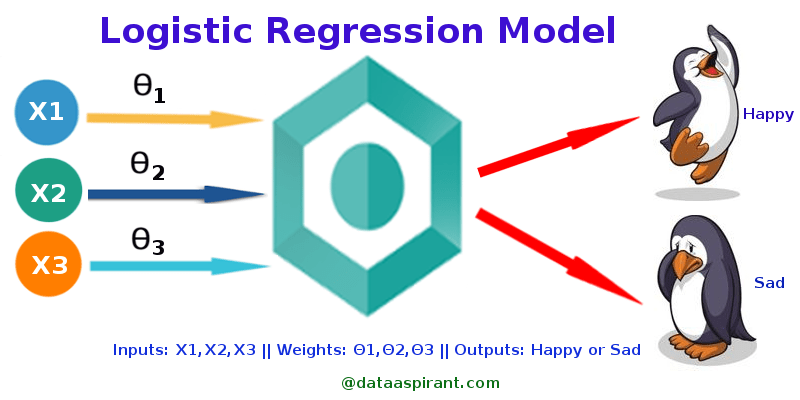
Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in the case of tree-based models like Random Forest Classifier

# Algorithms:

## Logistic Regression:

Logistic regression was used in the biological sciences in the early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

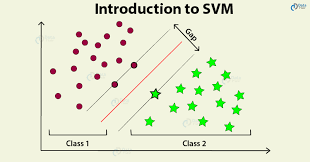
For example,

* To predict whether an email is spam (1) or (0)
* Whether the tumor is malignant (1) or not (0)

## 

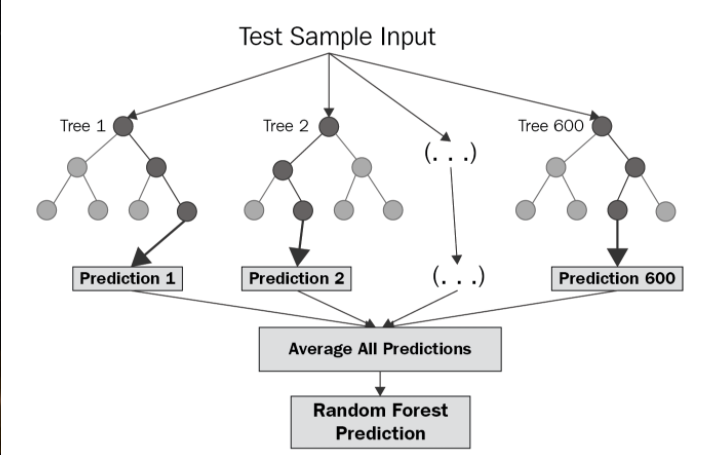
## SVM:

SVM is a supervised machine learning algorithm that can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to separate your data based on the labels or outputs you've defined.



## Random Forest Classification:

Random Forest is a bagging type of Decision Tree Algorithm that creates several decision trees from a randomly selected subset of the training set and n features, collects the values from these subsets, and then averages the final prediction out of all n number of decision trees



# Model performance:

The model can be evaluated by various metrics such as:

## Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost the same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

.

## Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/TP+FP

## Recall

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to all observations in the actual class.

Recall = TP/TP+FN

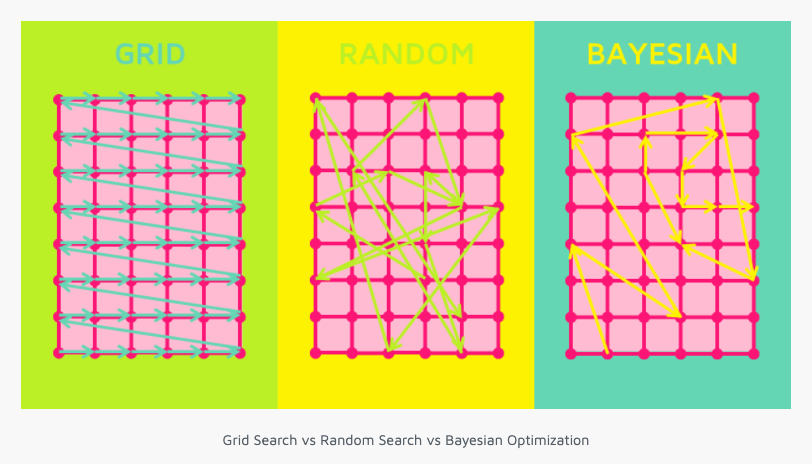
## F1-Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar costs. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

## Grid Search CV-Grid:

Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.



# 

# Conclusion

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection, and then model building.

The default rate is higher for males, increases as the education increases, also increase as the age of a person increases. i.e clients whose age over 60 was higher than mid-age and young people.

In all of these models, our recall revolves in the range of 75 to 83%.with the best fit model as random forest