Logo, company name

Description automatically generated­

ANALYSIS USING SUPERVISED MACHINE LEARNING ALGORITHMS

PROJECT REPORT

**DATA SCIENCE**

MINI PROJECT

LOAN PREDICTION

2021

**ACKNOWLEDGEMENT**

Report On

**“…………………………………………………………”**

Data Science Mini Project

Submitted by:

**Mohammad Zaigam (18030141CSEO59)**

**Nivedita Biswal (18030141CSE077)**

Under the Guidance of (Faculty Guide)

**Prof. Chetan J Shelke**

in Semester VII of the Bachelor of Technology

Batch - 2018-2022

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report.  A special gratitude to our project manager, [Prof Radha R], whose contribution in stimulating suggestions and encouragement,  helped us to coordinate our project especially in writing this report.

**ABSTRACT**

It is essential for a bank to estimate the credit risk it carries and the magnitude of exposure it has in case of nonperforming customers. Estimation of this kind of risk has been done by statistical methods through decades and with respect to recent development in the field of machine learning, there has been an interest in investigating if machine learning techniques can perform better quantification of the risk. The aim of this thesis is to examine which method from a chosen set of machine learning techniques exhibits the best performance in default prediction with regards to chosen model evaluation parameters. The investigated techniques were Logistic Regression, Random Forest, Decision Tree. Artificial Neural Network and Support Vector Machine. An oversampling technique called SMOTE was implemented in order to treat the imbalance between classes for the response variable. The results showed that without implementation of SMOTE obtained the best result with respect to the chosen model evaluation metric.

CHAPTER-1

# INTRODUCTION:-

*In this chapter an overview of what aim of the thesis is provided. The topics discussed within this chapter are the thesis’ background, purpose and scope.*

* 1. **Background**

A recent development of machine learning techniques and data mining has led to an interest of implementing these techniques in various fields [33]. The banking sector is no exclusion and the increasing requirements towards financial institutions to have robust risk management has led to an interest of developing current methods of risk estimation. Potentially, the implementation of machine learning techniques could lead to better quantification of the financial risks that banks are exposed to. Within the credit risk area, there has been a continuous development of the Basel accords, which provides frameworks for supervisory standards and risk management techniques as a guideline for banks to manage and quantify their risks. From Basel II, two approaches are presented for quantifying the minimum capital requirement such as the standardized approach and the internal ratings based approach (IRB) [3]. There are different risk measures banks consider in order to estimate the potential loss they may carry in future. One of these measures is the expected loss (EL) a bank would carry in case of a defaulted customer. One of the components involved in EL-estimation is the probability if a certain customer will default or not. Customers in default means that they did not meet their contractual obligations and potentially might not be able to repay their loans [43]. Thus, there is an interest of acquiring a model that can predict defaulted customers. A technique that is widely used for estimating the probability of client default is Logistic Regression [44]. In this thesis, a set of machine learning methods will be investigated and studied in order to test if they can challenge the traditionally applied techniques.

* 1. **Purpose**

The objective of this thesis is to investigate which method from a chosen set of machine learning techniques performs the best default prediction. The research question is the following

*• For a chosen set of machine learning techniques, which technique exhibits the best performance in default prediction with regards to a specific model evaluation metric?*

* 1. ***Scope***

The scope of this paper is to implement and investigate how different supervised binary classification methods impact default prediction. The model evaluation techniques used in this project are limited to precision, sensitivity, F-score and AUC score. The reasons for choosing these metrics will be explained in more detail in section 2.10. The classifiers that will be implemented and studied are:

* pandas
* matplotlib
* seaborn
* scikit-learn
* Algorithms
* Logistic Regression
* Decision Tree
* Random Forest
* Extra Tress
* Best Model Accuracy: 81.00

***Background pattern

Description automatically generated with low confidence***

CHAPTER – 2

# PROJECT PROFILE

* 1. **OBJECTIVE**

Our aim from the project is to build predictive models to automate the process of targeting the right applicants.

We’ll make use of pandas, matplotlib, & seaborn libraries from python to extract insights from the data, & scikit-learn libraries for machine learning.

Secondly, to learn how to hyper tune the parameters using grid search cross validation for the machine learning model.

And in the end, to predict whether the loan applicant can replay the loan or not using voting ensembling techniques of combining the predictions from multiple machine learning algorithms.

* 1. **Attributes in the Dataset:**

**Dataset Information:**

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

This is a standard supervised classification task. A classification problem where we have to predict whether a loan would be approved or not. Below is the dataset attributes with description.

A screenshot of a computer

Description automatically generated with medium confidence

**Variable Description**

Loan\_ID - Unique Loan ID

Gender - Male/ Female

Married - Applicant married (Y/N)

Dependents - Number of dependents

Education - Applicant Education (Graduate/ Under Graduate)

Self\_Employed - Self-employed (Y/N)

ApplicantIncome - Applicant income

CoapplicantIncome - Coapplicant income

LoanAmount - Loan amount in thousands

Loan\_Amount\_Term - Term of loan in months

Credit\_History - credit history meets guidelines.

Property\_Area - Urban/ Semi Urban/ Rural

Loan\_Status - Loan approved (Y/N)

* 1. **Key observation from the data:**

1. Applicants who are male and married tends to have more applicant income whereas applicant who are female and married have least applicant income.
2. Applicants who are male and are graduated have more applicant income over the applicants who have not graduated.
3. Again the applicants who are married and graduated have the more applicant income.
4. Applicants who are not self-employed have more applicant income than the applicants who are self-employed.
5. Applicants who have more dependents have least applicant income whereas applicants which have no dependents have maximum applicant income.
6. Applicants who have property in urban and have credit history have maximum applicant income.
7. Applicants who are graduate and have credit history have more applicant income.
8. Loan Amount is linearly dependent on Applicant income.
9. From heatmaps, applicant income and loan amount are highly positively correlated.
10. Male applicants are more than female applicants.
11. No of applicants who are married are more than no of applicants who are not married.
12. Applicants with no dependents are maximum.
13. Applicants with graduation are more than applicants with no graduation.
14. Property area is to be find more in semi urban areas and minimum in rural areas.

A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated

CHAPTER – 3

# Theory:

*In the theory section, relevant theory behind the chosen classification methods is explained. Background needed for understanding the implemented variable selection techniques and chosen model validation methods are also provided in this chapter.*

* 1. **Formulation of a Binary Classification Problem**

Binary classification refers to the case when the input to a model is classified to belong to one of two chosen categories. In this project, customers belong either to the non-default category or to the default category. The categories can therefore be modeled as a binary random variable Y ∈ {0, 1}, where 0 is defined as non-default, while 1 corresponds to default. The random variable Yi is the target variable and will take the value of yi , where i corresponds to the ith observation in the data set. For some methods, the variable ¯yi = 2yi−1 will be used, since these methods require the response variable to take the values ¯yi ∈ {−1, 1}.

The rest of the information about the customers, such as the products the customers posses, account balances and payments in arrears can be modeled as the input variables. These variables are both real numbers and categories and are often referred to as features or predictors. Let Xi ∈ R p denote a real valued random input vector and an observed feature vector be represented by xi = [xi1, xi2, ..., xip] >, where p is the total number of features. Then the observation data set with N samples can be expressed as D = {(x1, y1),(x2, y2), ...,(xN , yN )}.

With this setup, it makes it feasible to fit a supervised machine learning model that relates the response to the features, with the objective of accurately predicting the response for future observations [14]. The main characteristic of supervised machine learning is that the target variable is known and therefore an inference between the target variable and the predictors can be made. In contrast, unsupervised machine learning deals with the challenge where the predictors are measured but the target variable is unknown. 3CHAPTER 2. THEORY The chosen classification methods in this project are Logistic Regression, Artificial Neural Network, Decision Tree, Random Forest, XGBoost, AdaBoost and Support Vector Machine. The theory for these classifiers will be explained in more detail in the sections below.

* 1. **Logistic Regression**

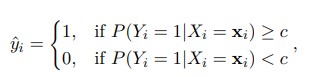
Logistic Regression aims to classify an observation based on its modelled posterior probability of the observation belonging to a specific class. The posterior probability for a customer to be in the default class with a given input xi can be obtained with the logistic function as

P(Yi = 1|Xi = xi) = e β0+β>xi /1 + e β0+β>xi ,

where the parameters β0 and β are parameters of a linear model with β0 denoting an intercept and β denoting a vector of coefficients, β = [β1, β2, ..., βp] >. The logistic function from Equation (2.1) is derived from the relation between the log-odds of P(Yi = 1|Xi = xi) and a linear transformation of xi , that is

log P(Yi = 1|Xi = xi)/ 1 − P(Yi = 1|Xi = xi) = β0 + β >xi

The class prediction can then be defined as



where c is a threshold parameter of the decision boundary which is usually set to c = 0.5 [10]. Further, in order to find the parameters β0 and β, the maximization of the log-likelihood of Yi is performed. After some manipulation of Equation (2.1), the expression can be rewritten as 

Since P(Yi = 1|Xi = xi ; β0, β) completely specifies the conditional distribution, the multinomial distribution is appropriate as the likelihood function [18]. The loglikelihood function for N observations can then be defined as

* 1. **Decision Trees**

A decision tree algorithm binary splits the feature space into subsets in order to divide the samples into more homogeneous groups. This can be implemented as a tree structure, hence the name decision trees.

The terminal nodes in the tree are called leaves and are the predictive outcomes. In this particular example, a regression tree which predicts quantitative outcomes has been used. However, classification trees that predict qualitative outcomes rather than quantitative will be used in this project.

**Chart, scatter chart

Description automatically generated**

(a) Two dimensional feature space splits into three subsets.

(b) Corresponding tree to the split of the feature space.

In a subset of the feature space, represented by the region Rm with Nm number of observations, let the indicator function be I(·) and

A picture containing text

Description automatically generated

be the fraction of class k observations in Rm [19]. Then the observations lying in Rm will be predicted to belong to class k(m) = arg maxk pˆmk. Since the Gini index, defined by

Text

Description automatically generated with low confidence

is amenable for numerical optimization [20], it will be chosen as the criterion for binary splitting.

* 1. **Artificial Neural Networks**

Artificial neural networks (ANN) is originally inspired by how a human brain works and is intended to replicate its learning process [23]. A neural network consists of an input layer, an output layer and a number of hidden layers.

The input layer is made of p predictors x1, x2, ..., xp and xi is an arbitrary ith observation such that xi = [xi1, xi2, ..., xip] >. For K-class classification problems, K is the number of target measurements Yk and k = 1, ..., K represented in as a binary variable of either 0 or 1 for the kth class [21]. The target Yk is a function derived.

**Diagram

Description automatically generated**

* 1. **Feature Selection Methods**

In the data given by Nordea, features are presented as both continuous and categorical variables. Thus, in order to understand how features correlate with each other and the response variable, implemented methods for feature selection should process both continuous and categorical variables simultaneously. That is why it is been decided to use following methods for feature selection: Feature selection with Kendall’s Tau Coefficient Analysis and Recursive Feature Elimination.

* + 1. **Recursive Feature Elimination**

Recursive Feature Elimination (RFE) is a multivariate method of variable selection [38], which performs a backward selection of predictors [30]. The algorithm works in the following way. It starts with computing an importance score for each predictor in the whole data set. Let S be a sequence of the number of variables to include in the model. For each iteration, the number of Si predictors which have been top-ranked are retained in the model [9]. Further, the importance scores are computed again and the performance is reassessed. The tuning parameter for RFE is the subset size and the subset of Si predictors with the highest importance score is then used for fitting the final model. Thus, the performance criteria is optimized by the subset size with regards to the performed importance ranking. A more detailed description of the procedure can be found in Algorithm 2.

Further, it is also relevant to highlight that improvement for some models may be seen when applying RFE, while for others no remarkable difference in performance could exhibit. For example, random forest is one of the models that might benefit when RFE is applied [30]. One of the reasons is caused by the nature of model ensembles. Random forest tends not to exclude irrelevant predictors when a split is made, which requires a prior irrelevant feature elimination.

Thus, the aim is to test how RFE will impact on the implemented models. In contrast, when applying logistic regression for RFE, it is relevant to consider that

Text

Description automatically generated

* 1. **Model Evaluation Techniques**

**Diagram, schematic

Description automatically generated**

* + 1. **Confusion Matrix**

One common way to evaluate the performance of a model with binary responses is to use a confusion matrix. The observed cases of default are defined as positives and non-default as negatives [10]. The possible outcomes are then true positives (TP) if defaulted customers have been predicted to be defaulted by the model. True negatives (TN) if non-default customers have been predicted to be non-default. 19CHAPTER 2. THEORY False positives (FP) if non-default customers have been predicted to be defaulted, and false negatives (FN) if defaulted customers have been predicted to be non-default.

**A picture containing graphical user interface

Description automatically generated**

From a confusion matrix there are certain metrics that can be taken into consideration. The most common metric is accuracy which is defined as the fraction of the total number of correct classifications and the total number of observations. It is mathematically defined as

Text

Description automatically generated with low confidence

The issue with using accuracy as a metric is when applying it for imbalanced data. If the data set contains 99% of one class it is possible to get an accuracy of 99%, if all of the predictions are made for the majority class. A metric that is more relevant in the context of this project is specificity. It is defined as



and will be used for explaining the theory behind receiver operator characteristic curve and its area under the curve

Graphical user interface, text, application

Description automatically generated

* 1. **Cross-Validation**

In order to prevent using the same information in the training phase and the evaluation phase of models, which makes the results less reliable, the data is divided into training set, validation set and test set [40]. The training set and validation set are used for finding the best model and the test set is only used for calculating the prediction performance of the best model. The test data will therefore be held out until the best model is obtained, this is called the holdout method [40]. Choosing the best model from a set of models can be done by a method called K-fold cross-validation (CV).  **A picture containing graphical user interface

Description automatically generated**

The objective is to find the α that minimizes the validation error, denoted as ˆα. This is also known as a hyperparameter search and will be implemented by looping through every combination of a chosen set of hyperparameter values for each machine learning method. When the final model f(x, αˆ) is obtained, where ˆα represents the best combination of hyperparameters, the performance of f(x, αˆ) will be calculated when predicting on the test set .

CHAPTER -4

*In this chapter, procurement of data, preprocessing of data and implementation of variable selection will be discussed.*

* 1. **Preprocessing of Data**

As mentioned in the section 3.1, the data contained different types of accounts that the customers possessed. These accounts can be related to different financial products such as mortgages, loans or credit. Thus, in order to decrease the number of variables, all different types of financial products have been aggregated into one category, see Figure 3.2.

Diagram

Description automatically generated

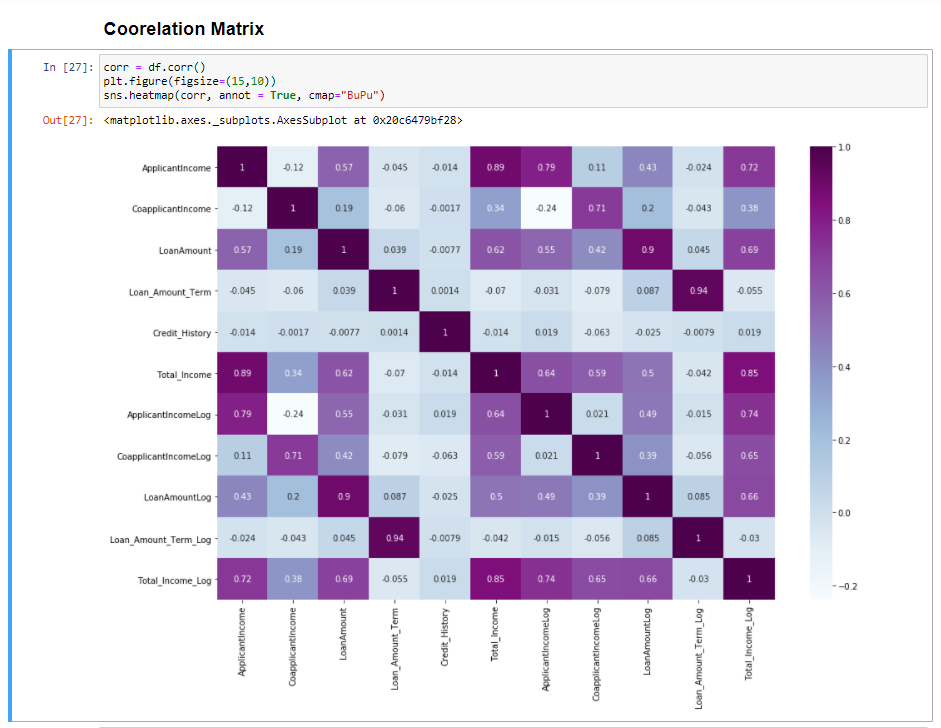
Some of the created variables describe the same feature, but for different time spans. This means that the variables, for example bf ag 1 1, bf ag 1 3, bf ag 1 6 and bf ag 1 12, describe the same feature but represent different time periods.

The input data to all of the models were standardized according to section. Missing values were treated by doing a complete case analysis.

**Graphical user interface, text, application

Description automatically generated**

* 1. **Variable Selection by correlation analysis**

****

After this selection, a significance test described in section 2.8.1 was executed in order to see if the null hypothesis can be rejected. For the final selection of variables with α-level of 0.1, all variables from Figure 3.5 were tested with regards to their independence with the response variable. The result showed that all tests proved that there is a dependence between the feature variables and the response variable, and therefore the null hypothesis can be rejected. Thus, the variables shown in Figure 3.5 are included in the final model.

**Heat Map for Selected Variable**

**Graphical user interface

Description automatically generated**

CHAPTER – 5

# RESULTS

*In this chapter, results obtained from different models will be presented and discussed.*

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceGraphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

* 1. **Graphical user interface, text, application, email

     Description automatically generated****Performance of the Methods**

CHAPTER – 6

# Discussion

* 1. . **Conclusion**

The analytical process started from data cleaning and processing, Missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.811. This brings some of the following insights about approval. Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.

* 1. **Future Work**

The potential future work for this project will be a further development of the model by deepening analysis on variables used in the models as well as creating new variables in order to make better predictions. Data available for the scope of this thesis has constraints in terms of many years are covered by the data presented as well as geographical breadth of the Nordea’s clients. The majority of customers at Nordea’s clients are from the Nordic countries, thus, it should be considered that the behavior of Nordic customers influence the results of this research. It means that the behavior of clients outside Nordics may or may not follow the same pattern and therefore one should make additional analysis and obtain a geographically-broader data set if the objective is to have a model unbiased of the geographical location. An assumption can also be made that if there is data available for longer time span as well as broader geography of clients, there is an interest to implement macro-economic variables, which in turn might open some new insights about factors impacting default of a customer as well as what machine learning methods are more suitable for this type of a problem. Further, a large part of this project was to make a grounded feature selection such that variables included in the models were valuable for prediction. Variable selection was made by RFE and correlation analysis with Kendall’s Tau, but it would be interesting to apply other variable selection methods. An alternative for dimensionality reduction could be Principal Component Analysis (PCA).

It would also be interesting to make a study concerning what metrics are the most relevant for this type of the problem. As mentioned previously, in this project the main metric all evaluations were analyzed by was F-score, because the aim was to achieve a trade-off between the sensitivity and the precision. If a deeper analysis could be performed regarding the most relevant metric for this type of problem, then potentially a weight function could be implemented if one of the metrics explored turned out to be of more importance. The example of a weight function can be to use a weighted F-score , where β is not set to 1, but the value of interest.