Low Level Design (LLD)

South German Bank Credit Risk Prediction

Abhishek Kumar, Aravindh C, Subramanyam Sista

**Contents**

Abstract 3

[1](#_heading=h.gjdgxs) Introduction 3

[1.1](#_heading=h.30j0zll) Why this Low-Level Design Document? 3

[1.2](#_heading=h.1fob9te) Scope 3

[1.3](#_heading=h.3znysh7) Constraints 3

[2](#_heading=h.2et92p0) Technical specifications 3

[2.1 Dataset](#_heading=h.tyjcwt) 3

[2.1.1 Dataset overview](#_heading=h.3dy6vkm) 4

[2.1.2](#_heading=h.1t3h5sf) [Input schema](#_heading=h.4d34og8)4

[2.1.3](#_heading=h.4d34og8) [Attribute Information](#_heading=h.1t3h5sf)5

[2.2 Data Preparation](#_heading=h.2s8eyo1) 6

[2.3 Logging 8](#_heading=h.17dp8vu)

[2.4 Database 8](#_heading=h.3rdcrjn)

[2.5 Data format](#_heading=h.26in1rg) [8](#_heading=h.lnxbz9)

[3](#_heading=h.35nkun2) Technology stack 8

[4](#_heading=h.1ksv4uv) Proposed Solution 8

[4.1 Models used](#_heading=h.44sinio) 8

[4.1.1 Robust Scalar](#_heading=h.2jxsxqh) 8

[4.1.2 SMOTETomek](#_heading=h.z337ya) 8

[4.1.3 Random Forest Classifier](#_heading=h.3j2qqm3) 8

[4.1.4 Xg Boosting Classifier](#_heading=h.1y810tw) 9

4.2 Machine learning pipeline [10](#_heading=h.4i7ojhp)

4.3 Evaluation 10

4.4 Conclusion [11](#_heading=h.2xcytpi)

**Abstract**

Normally, most of the bank's wealth is obtained from providing credit loans so that a marketing bank must be able to reduce the risk of non-performing credit loans. The risk of providing loans can be minimized by studying patterns from existing lending data. One technique that can be used to solve this problem is to use data mining techniques. Data mining makes it possible to find hidden information from large data sets by way of classification. The Random Forest (RF) algorithm is a classification algorithm that can be used to deal with data imbalancing problems. The purpose of this study is to discuss the use of the RF algorithm for classification of South German Credit data. This research is needed because currently there is no previous research that applies the RF algorithm to classify South German Credit data specifically. Based on the tests that have been done, the optimal performance of the classification algorithm RF on South German Credit data is the comparison of training data of 84%.

Keywords – data mining, classification, random forest, bank credit receipts.

# Introduction

## Why this Low-Level Design Document?

This study uses classification modeling to be used ondata South German Credit, where the resulting label is whether credit applications are accepted or rejected. The classification model is built using a process split validation to divide South German Credit into data training and data testing data

## Scope

This software system will be a scheduled batch processing system, which trigger’s on demand to classify whether a transactionis at potential credit risk or not?

## Constraint

This dataset is imbalanced. This refers to the fact that the training set contains more instances of Good Credit than Bad Credit. This is prevalent in many real-world problems.

## Technical specifications

## 2.1 Dataset

| **Dataset** | **Finalized** | **Source** |
| --- | --- | --- |
| South German Credit Prediction | yes | https://www.kaggle.com/competitions/south-german-credit-prediction/data |

## 2.1.1 Dataset overview

This dataset classifies people described by a set of attributes as good or bad credit risks. There are 1000 entries in the dataset, which have been sampled as train and test data. This dataset is Multivariate, Numeric/Categorical. Nature of the problem is Binary Classification

'**Kredit**' is our target variable, the one whose value must be predicted.

Please note that the feature names are in German to preserve the authenticity of the data.

## 2.1.2 Input schema

| **Column Name** | **Variable Name** | **Type** | **Content** |
| --- | --- | --- | --- |
| laufkont | status | categorical | status of the debtor's checking account with the bank |
| laufzeit | duration | quantitative | credit duration in months |
| moral | credit\_history | categorical | history of compliance with previous or concurrent credit contracts |
| verw | purpose | categorical | purpose for which the credit is needed |
| hoehe | amount | quantitative | credit amount in DM |
| sparkont | savings | categorical | debtor's savings |
| beszeit | employment\_duration | ordinal, discretized quantitative | duration of debtor's employment with current employer |
| rate | installment\_rate | ordinal; discretized quantitative | credit installments as a percentage of debtor's disposable income |
| famges | personal\_status\_sex | categorical | combined information on sex and marital status |
| buerge | other\_debtors | categorical | Is there another debtor or a guarantor for the credit? |
| wohnzeit | present\_residence | ordinal; discretized quantitative | length of time (in years) the debtor lives in the present residence |
| verm | property | ordinal | the debtor's most valuable property, i.e. the highest possible code is used. Code 2 is used, if codes 3 or 4 are not applicable and there is a car or any other relevant property that does not fall under variable sparkont |
| alter | age | quantitative | age in years |
| weitkred | other\_installment\_plans | categorical | installment plans from providers other than the credit-giving bank |
| wohn | housing | categorical | type of housing the debtor lives in |
| bishkred | number\_credits | ordinal | number of credits including the current one the debtor has (or had) at this bank statement |
| beruf | job | ordinal | quality of debtor's job |
| pers | people\_liable | binary,discretized quantitative | number of persons who financially depend on the debtor (i.e., are entitled to maintenance) |
| telef | telephone | binary | Is there a telephone landline registered on the debtor's name? |
| gastarb | foreign\_worker | binary | Is the debtor a foreign worker? |
| kredit | credit\_risk | binary | Has the credit contract been complied with (good) or not (bad) ? |

## 2.1.3 Attribute Information

**Id** = Id of individual entries, for evaluation

**laufkont** = status

1 : no checking account

2 : … < 0 DM 3 : 0<= … < 200 DM 4 : … >= 200 DM / salary for at least 1 year

**laufzeit** = duration

**moral** = credit\_history

0 : delay in paying off in the past

1 : critical account/other credits elsewhere

2 : no credits taken/all credits paid back duly

3 : existing credits paid back duly till now

4 : all credits at this bank paid back duly

**verw** = purpose

0 : others

1 : car (new)

2 : car (used)

3 : furniture/equipment

4 : radio/television

5 : domestic appliances

6 : repairs

7 : education

8 : vacation

9 : retraining

10 : business

**hoehe** = amount

**sparkont** = savings

1 : unknown/no savings account

2 : … < 100 DM 3 : 100 <= … < 500 DM 4 : 500 <= … < 1000 DM 5 : … >= 1000 DM

**beszeit** = employment\_duration

1 : unemployed

2 : < 1 yr 3 : 1 <= … < 4 yrs 4 : 4 <= … < 7 yrs 5 : >= 7 yrs

**rate** = installment\_rate

1 : >= 35

2 : 25 <= … < 35

3 : 20 <= … < 25

4 : < 20

**famges** = personal\_status\_sex

1 : male : divorced/separated

2 : female : non-single or male : single

3 : male : married/widowed

4 : female : single

**buerge** = other\_debtors

1 : none

2 : co-applicant

3 : guarantor

**wohnzeit** = present\_residence

1 : < 1 yr 2 : 1 <= … < 4 yrs 3 : 4 <= … < 7 yrs 4 : >= 7 yrs

**verm** = property

1 : unknown / no property

2 : car or other

3 : building soc. savings agr./life insurance

4 : real estate

**alter** = age

**weitkred** = other\_installment\_plans

1 : bank

2 : stores

3 : none

**wohn** = housing

1 : for free

2 : rent

3 : own

**bishkred** = number\_credits

1 : 1

2 : 2-3

3 : 4-5

4 : >= 6

**beruf** = job

1 : unemployed/unskilled - non-resident

2 : unskilled - resident

3 : skilled employee/official

4 : manager/self-empl./highly qualif. employee

**pers** = people\_liable

1 : 3 or more

2 : 0 to 2

**telef** = telephone

1 : no

2 : yes (under customer name)

**gastarb** = foreign\_worker

1 : yes, 2 : no

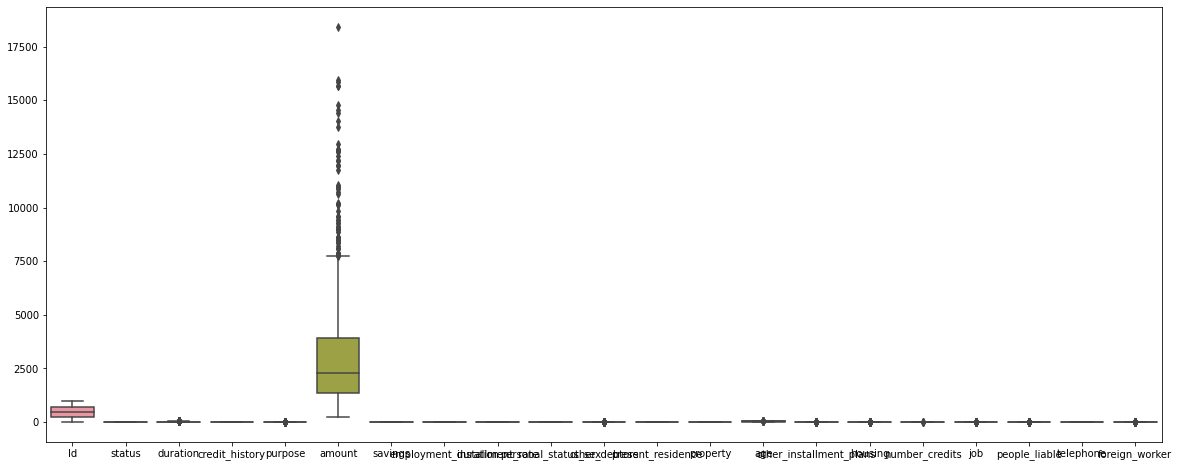
**kredit (target column)** = credit\_risk

0 : bad, 1 : good

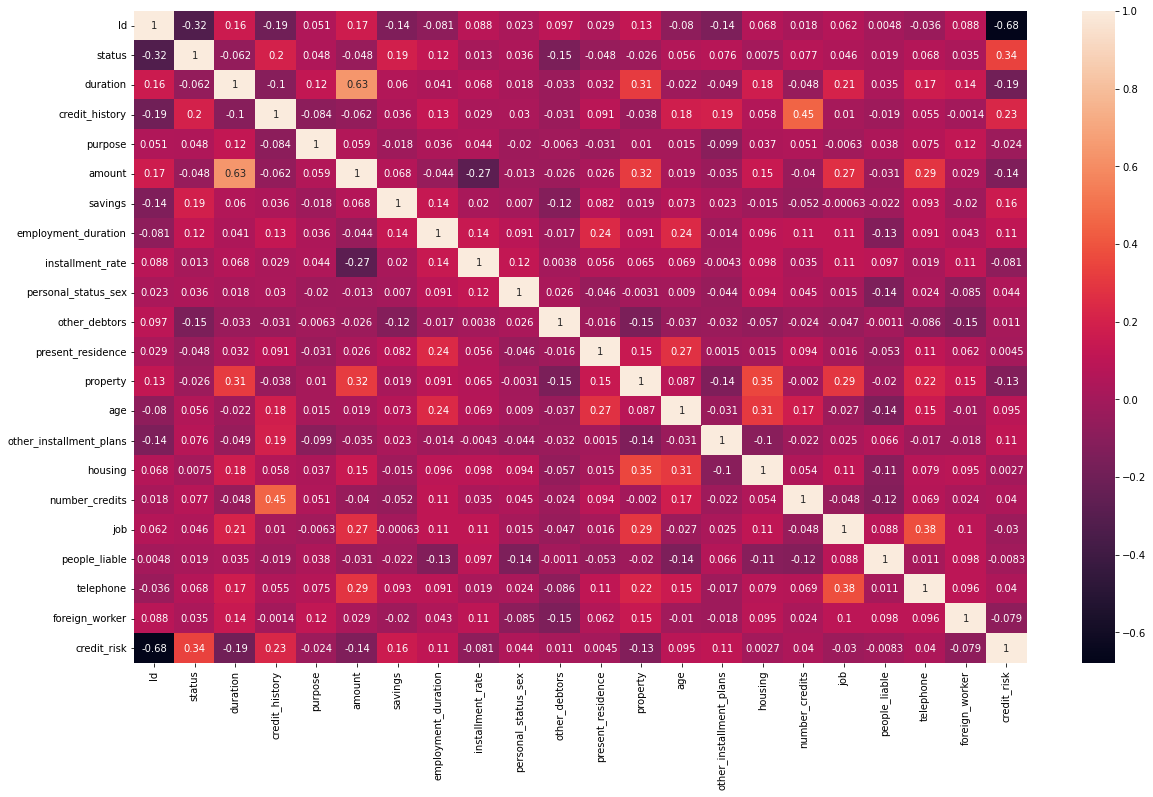
## 2.2 Data Preparation

After performing exploratory data analysis following information obtained on the dataset

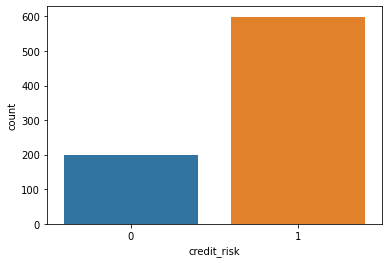
1. No null values were found
2. No duplicate records were identified
3. Outliers are present at amount field



1. Certains Independent features can be removed.
   1. Id is irrelevant which is just a sequence number
   2. "**present\_residence**", "**housing**", "**people\_liable**" are found to be having low correlation with the data



1. Dataset seems to be imbalanced



## 2.3 Logging

The system will be able to log every activity wherever required. Also log each and every system flow.

## 2.4 Database

Dataset information is loaded and pulled from MongoDB

## 2.5 Data format

This dataset contains 22 variables (columns) including target column class, feature names are in German to preserve the authenticity of the data which are in csv (comma separated value) format

**2.6 Deployment**

The system has been deployed on the AWS EC2 instance using github CI/CD pipelines.

## 3 Technology stack

1. **Database** **MongoDB**
2. **Deployment pipeline** GitHub Actions
3. **Deployment Platform : AWS EC2**
4. **Cloud Components : AWS S3,ECR**
5. **Version Control System** **GitHub**
6. **Design Workflow** **Airflow**

## 4 Proposed Solution

## 4.1 Models used

## 4.1.1 Robust Scalar

This is a preprocessing technique which scales and removes outliers in the features by applying interquartile range

Standardization of a dataset is a common requirement for many machine learning estimators. Typically this is done by removing the mean and scaling to unit variance. However, outliers can often influence the sample mean / variance in a negative way. In such cases, the median and the interquartile range often give better results

## 

## 4.1.2 SMOTETomek

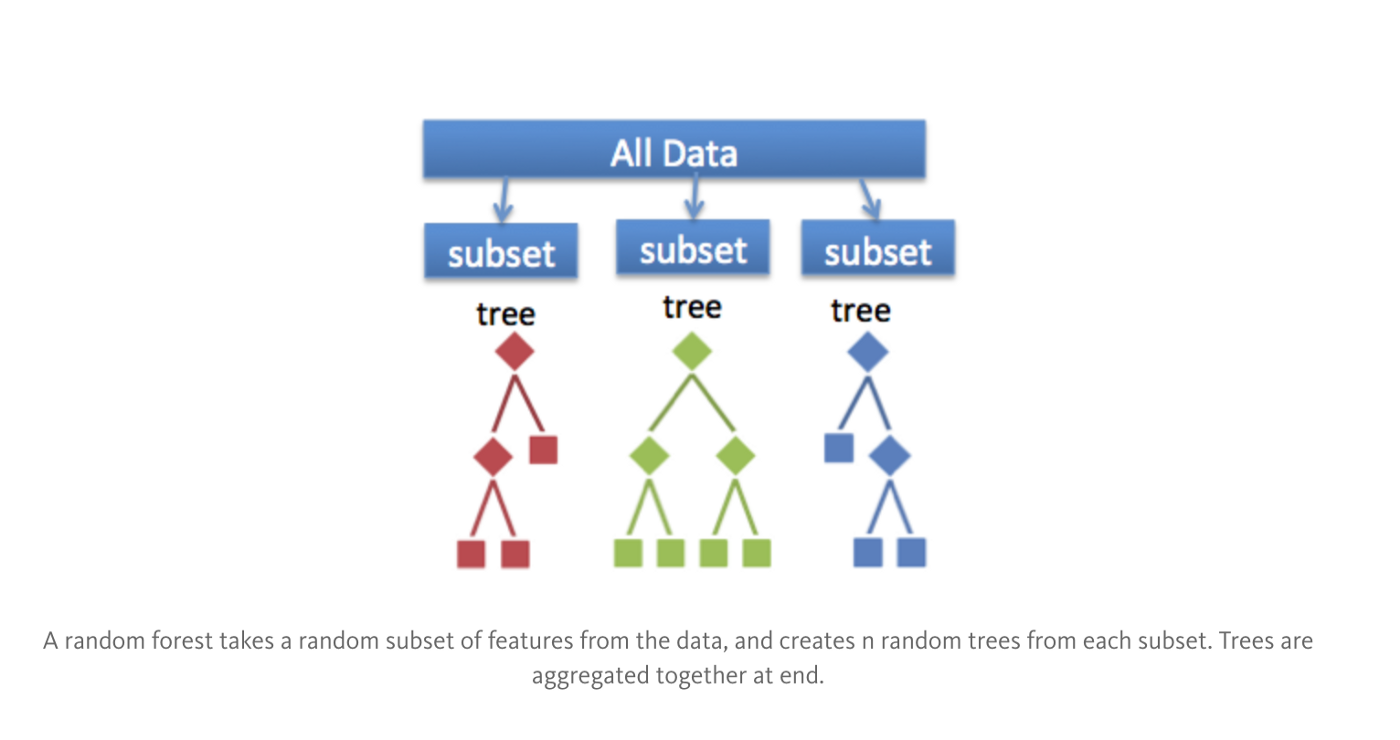
This is the technique to handle the imbalance dataset with the combination of over-sampling and under-sampling i.e., inshort either to increase/decrease the nearest datapoint around the target variable. As our dataset contains imbalanced dataset where most of the loans are risk free and very less or moderate risk prone loans are to be expected in the dataset which is more realistic and real time scenario of risk prone loans, so by applying this technique which increase the noise of the imbalance data in our case risk prone data points will be added which are nearby to actual datapoint with a decimal difference from the actual point

For information: <https://imbalanced-learn.org/dev/combine.html#combine>

## 4.1.3 Random Forest Classifier

A random forest classifier is a tree based ensemble technique. An ensemble technique is a machine learning technique that combines several base models in order to produce one optimal predictive model.

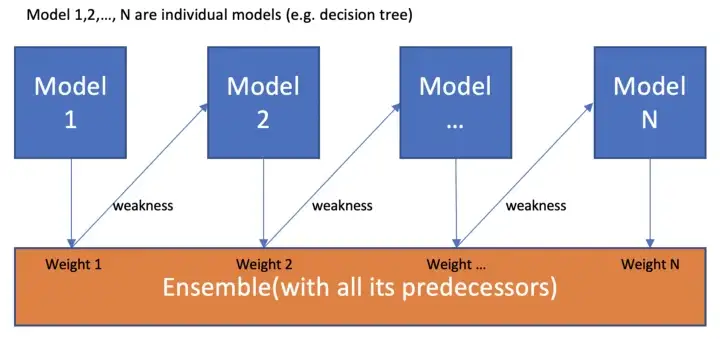
Random Forest models decide where to split based on a random selection of features. Rather than splitting at similar features at each node throughout, Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, ergo producing a more accurate predictor.



## 4.1.4 Xg Boosting Classifier

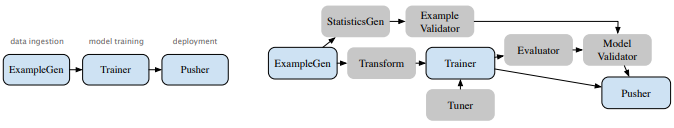
Xg Boosting is a boosting technique which is widely used in most of the machine learning techniques. Ensembling trees randomly was time-consuming and computationally inefficient. This technique follows the approach of build trees sequentially and improve the performance where previous trees failed, later these boosting algorithms started utilizing the gradient descent algorithm to form sequentially and minimize the error in predictions hence this algorithms are called as Gradient Boosting. Later researchers proposes model, algorithmic and hardware optimization to further improve the Gradient boosting algorithms performance thus this optimizations over the Gradient Boosting is known as XG-Boosting

XG Boosting also known as Extreme Gradient Boosting which supervised learning technique.



## 4.2 Machine learning pipeline

* **Data Ingestion:** This is the beginning of every machine learning pipeline. In this step, we process the data into a format that the following components can digest
* **Data Validation:** Before training a new model, we need to validate the new input. Data validation focuses on checking that the statistics of the new data are as expected. It also alerts the data scientist if any anomalies are detected
* **Data Transformation:** Data transformation techniques refer to all the actions that help you transform your raw data into a clean and ready-to-use dataset.
* **Model Trainer:** Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved.
* **Model Evaluation:** Model evaluation is the process of using different evaluation metrics to understand a machine learning model’s performance, as well as its strengths and weaknesses.
* **Model Pusher:** Now that all the things are working fine, it’s time to deploy the machine learning model to the real world. You can monitor how your model is performing, and see how to improve the workflow.



## 4.3 Evaluation

Credit comes from the Italian language, namely Credere, which means trust. Trust in question is the trust of the creditor that the debtor will repay the loan and the interest in accordance with the agreement of both parties. The implementation of credit extension usually goes through several stages, namely credit application, checking credit applications, credit analysis, credit approval, credit realization, and finally credit monitoring. Normally, most of the bank's wealth is in the form of credit which is the source of bank income, therefore it is often referred to as productive assets. In channeling credit, management must use the principle of prudence so that loans are granted in the current category. However, there are often some and non-target samples correctly predicted and reflects the

classifier ability to define the entire sample.

The accuracy can be measured by the following equation:

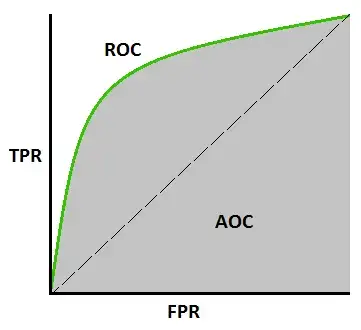
**Accuracy** = ((**TP** + **TN**) / (**TP** + **FP** + **TN** + **FN**)) × **100**

**Note:** True Positive (TP) is the number of positive samples that are predicted to be correct; False Positive (FP) is the number of positive samples whose predictions are wrong; True Negative (TN) is the number of negative samples correctly predicted; False Negative (FN) is the number of negative samples that are predicted to be wrong.

**AUC - ROC Curve:**

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.



**Over all model performance evaluation under various machine learning algorithms**

| **Model** | **Accuracy** | **Balanced Accuracy** | **ROC AUC** | **F1 Score** | **Time Taken** |
| --- | --- | --- | --- | --- | --- |
| RandomForestClassifier | 0.84 | 0.84 | 0.84 | 0.84 | 0.20 |
| LogisticRegression | 0.77 | 0.77 | 0.77 | 0.77 | 0.02 |
| BaggingClassifier | 0.76 | 0.77 | 0.77 | 0.76 | 0.06 |
| XGBClassifier | 0.75 | 0.75 | 0.75 | 0.75 | 0.08 |

## 4.4 Conclusion

This study has not been able to provide a good enough accuracy value for the classification of data South German Credit, so further research is needed to obtain a better classification model. Based on the findings of this study, it is suggested that further research can apply optimization methods to further optimize the performance of the RF algorithm