

PREDICT CREDIT RISK USING SOUTH GERMAN BANK DATA

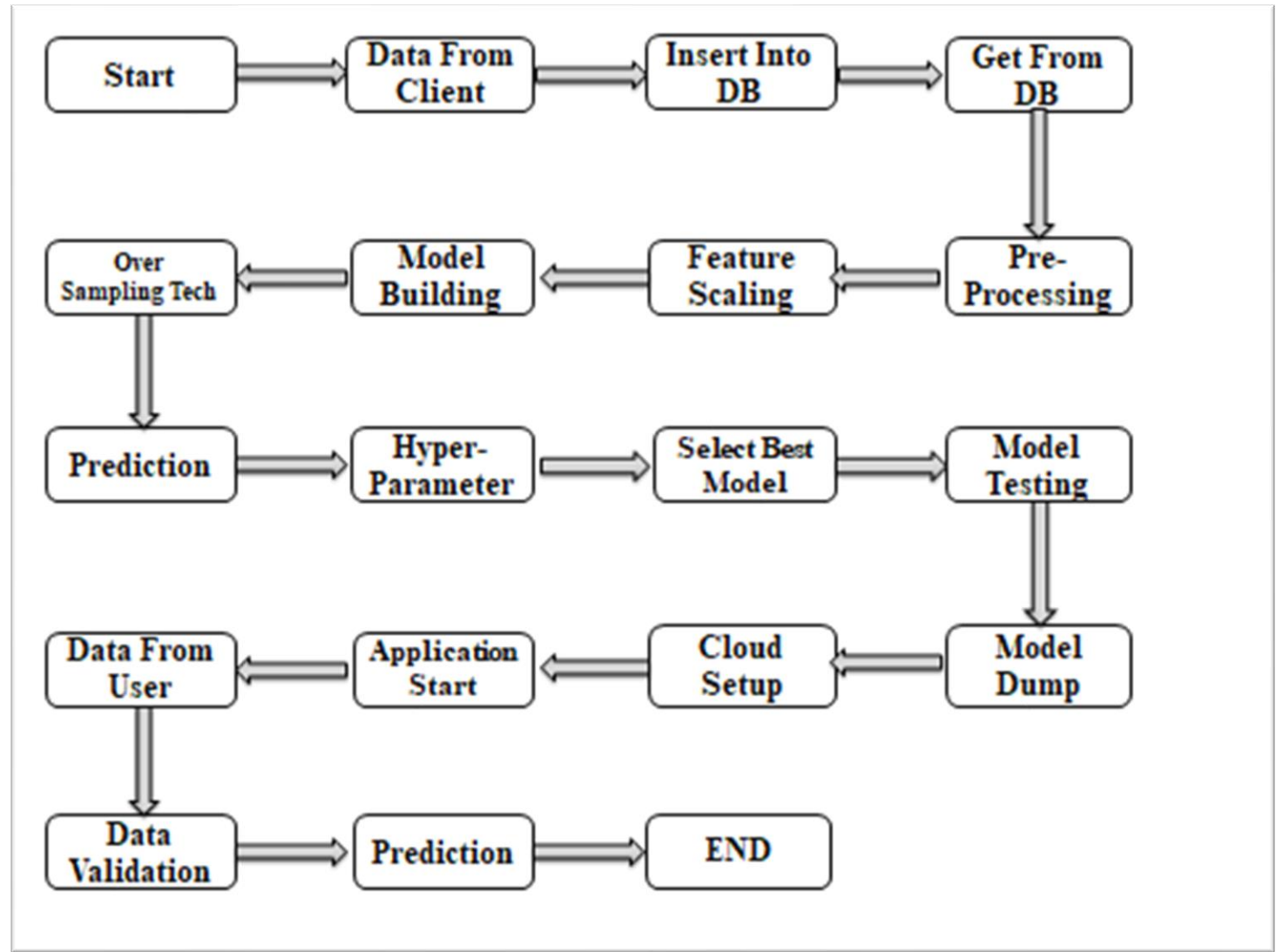


OBJECTIVE

Credit analysis draws conclusions by evaluating the available quantitative and qualitative data regarding the creditworthiness of a client and making recommendations on whether or not to approve the loan application. The objective of credit analysis is to determine the risk of default that a client presents and assign a risk rating to each client. The risk rating will determine if the company will approve (or reject) the loan application, and if approved, the amount of credit to be granted.



- ARCHITECTURE



- Data From Client:-

We collect the data from client in .asc format in folder. In that we have get different file format, out of that that the .txt format file contain the feature name in German and English language.

- Insert Into Database & Get From Database:-

After gathering the data from client we have to put that data in database for future purpose. The database we use is Cassandra database for that we use DataStax Astra website. Our database name is “South German Bank Data” then the keyspace name is “credit” and last table is “credit_data”. Then we extract the data from database with help of query “SELECT * FROM credit.credit_data” and save in the file.

```
token@cqlsh> use credit;
token@cqlsh:credit> select * from credit.credit_data;
```

id	age	amount	credit_history	credit_risk	duration	employment_duration	foreign_worker	housing	installment_rate	job	number_credits	other_debtors	other_installment_plans	people_lia
ble	personal_status_sex	present_residence	property	purpose	savings	status	telephone							
769	24	1275	2	0	15	3	2	1	4	3	1	1		3
2			2	2	3	4	5	1	1					
23	26	1424	4	1	12	4	2	2	4	3	1	1		3
2			3	3	2	4	1	2	1					
114	35	3976	2	1	21	4	2	2	2	3	1	1		3
2			3	3	3	2	5	2	2					
660	33	1414	2	1	8	3	1	2	4	3	1	3		3
2			3	2	1	3	1	2	1					
893	33	1131	2	0	18	1	2	2	4	3	1	1		3
2			2	2	3	2	1	1	1					
53	49	2331	4	1	12	5	2	2	1	3	1	2		3
2			3	4	1	3	5	4	2					
987	20	674	2	0	12	4	2	2	4	3	1	1		3
2			4	1	2	3	2	1	1					
878	22	1366	2	0	9	2	2	1	2	3	1	1		3
2			2	4	2	3	1	1	1					
110	29	3959	2	0	15	3	2	2	3	3	1	1		3
2			2	2	2	0	1	1	2					
91	25	2991	2	1	30	5	2	2	2	3	1	1		3
2			2	4	3	3	5	2	1					
128	30	1820	2	1	18	3	2	2	2	4	1	1		3
2			4	2	2	0	1	4	2					

- **Pre-Processing:-**

In pre-processing there is different method are involved such as, filling up nan value, encoding categorical column etc. But in our dataset there is no nan value and categorical value. In our dataset the feature name is in German language so we have to convert it into English language that are given in dataset file.

- **Feature Scaling:-**

In feature scaling we had use Standard Scalar method for normalization purpose. In that we convert all the data in between 0 & 1 range. So that the data can show the proper distribution.

- **Model Building:-**

In the given step we divide the data into train test split. So that we can apply train and test data to the model. We are use various ML model for our project, such as Logistic Regression, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, Bagging Classifier, Gradient Boosting Classifier, XG Boost Classifier & Light GBM. But problem is that our target feature is not equally distributed, so it will affect on the model accuracy.

- Over Sampling Technique :-

In these method we use the ADASYN technique for sampling the target feature. After sampling the good risk value count is 562 & bad risk value count is 575. But before use of sampling technique the value count of good risk value count is 700 & bad risk value count is 300.

- Prediction :-

After the over sampling technique we again build the model with new train & test value. Then we fit the data to the model and see the result and prediction of the data.

- Hyper-parameter :-

In these section we tune some of the model that we use at the time off model building. In model tuning we use best 5 model such as, Random Forest Classifier, XG Boost Classifier, Bagging Classifier, Gradient Boosting Classifier and Light GBM etc.

- Select Best Model :-

After hyper parameter we select the best model from the accuracy for the further purpose.

- **Model Dump:-**

The higher accuracy model is dump in pickle file (Random Forest Regressor).

- **Cloud Setup:-**

In that we upload code on GitHub for deployment purpose. The HTML content to create the front page & backend we will be using the flask python framework.

- **User Interface:-**


The user interface is divided into two pages, first one is homepage where user can enter the value for calculating the credit risk & second one is result page where you can the result that whether the credit risk is good or bad.

Homepage

Please fill the following details in order to Predict Bank Credit Risk

<div>Status</div> <div>STATUS</div>	<div>Duration</div> <div>Enter Integer Value</div>
<div>Credit History</div> <div>CREDIT HISTORY</div>	<div>Purpose</div> <div>PURPOSE</div>
<div>Amount</div> <div>Enter Integer Value</div>	<div>Savings</div> <div>SAVINGS</div>
<div>Employment Duration</div> <div>EMPLOYMENT DURATION</div>	<div>Personal Status Sex</div> <div>PERSONAL STATUS SEX</div>
<div>Installment Rate</div> <div>INSTALLMENT RATE</div>	<div>Present Residence</div> <div>PRESENT RESIDENCE</div>
<div>Property</div> <div>PROPERTY</div>	<div>age</div> <div>Enter Integer Value</div>
<div>Number Credits</div> <div>NUMBER CREDITS</div>	<div>Telephone</div> <div>TELEPHONE</div>


Click to Submit

 RASHMI DUBEY

Result Page

Predict Bank Credit Risk Using Credit Data

The Credit Risk Is Good

in 

Thank you*

Q & A

Q1) What's the source of data?

➡ The data for training is provided by the client .

Q 2) What was the type of data?

➡ The data type is numerical in nature.

Q 3) What's the complete flow you followed in this Project?

➡ Refer slide 3rd for better Understanding

Q 4) How logs are managed?

➡ We are using different logs as per the steps that we follow in, Data Insertion, Model Training log , prediction log etc.

Q 5) What techniques were you using for data pre-processing?

➡ We don't use many technique because there is no need of that except the Normalization

Q 6) How training was done or what models were used?

- ➡ 1] Before diving the data in training and testing set we performed scaling operation on the training and testing set.
- 2] Algorithms like Logistic Regression, RF, XG Boost, Bagging, Gradient Boosting, Decision Tree, SVM, Light GBM were used to find out the accuracy.

Q 7) What are stage for deployment?

- ➡ 1] The 1st stage is to create your app.py file so that we can see our model.
- 2] The 2nd stage is to update your project on GitHub.
- 3] The 3rd stage in which you can deploy your model on any server such as GCP, AWS, Azure

The background features a gradient from light blue on the left to light orange on the right. In the top-left corner, there are several overlapping, wavy, semi-transparent shapes in shades of blue and white. In the bottom-right corner, there are similar overlapping, wavy, semi-transparent shapes in shades of orange and white.

THANK YOU