Smart Lender - Applicant Credibility Prediction for Loan Approval

Under the supervision of

Industry Mentor: Nidhi

Faculty Mentor:T.Sampath

Team ID – PNT2022TMID24793

Team Members:

1. Lurdes Infant Joe J - 210419106058
2. Akilesh V A - 210419106005
3. Ajay Balan B - 210419106003
4. Ajay M - 210419106002

NOV/DEC 2022

1. INTRODUCTION

1.1 Project overview

Smart Lender - Applicant Credibility Prediction for Loan Approval

All the banking sectors across the country are providing the loan for various sectors like Agriculture,Finance,Real Estate,Home Loans and also education Loan . Every individual of an organisation has a different Financial Statement. The Important parameters to be analysed like capacity,capital,conditions and the collateral which plays a major role in determining the credibility for the loan approval prediction. Smart Lender is a novel approach that helps the user to know about their status of Bank Accounts based on their previous Transaction of bank data. It predicts the loan approval credibility by using the attributes like bank name,credit score,rate of interest,penalty etc The Smart Lender Project is being implemented by using various Machine learning models since the loan approval model comes under the category of the binary classification. We are able to make the prediction model by using the supervised learning Classification models like Logistic regression , decision tree,random forest,support vector machine, KNN algorithms. As building the machine learning model requires data preprocessing steps to know the clarity of the data and also extracting the important features from the dataset is very important for the model prediction.

1.2 Purpose

A novel hybrid feature selection approach is proposed to predict the loan repayment capability behavior of a customer in a cost effective way. Complex set of decision making needs to be taken by bank officers to determine whether to approve loan applicants or not. Normally classification techniques solved the problem up to an extent. Now the experiment proved that a model that use feature selection before classification can help the bank officers to take proper decision more accurately. This proposed methodology will protect the bank from further misuse, fraud applications etc by identifying the customers whose repayment capability status is risky especially in the co-operative banking sector. The experiment proved that the classification accuracy have considerably increased after feature selection. The proposed algorithm had produced better accuracy than existing methods. Experiments on standard data sets proved that the proposed algorithm for loan credibility prediction system outperforms many other feature selection methods.

2. LITERATURE SURVEY

2.1 Existing problem

Nowadays people approach or select bank loans to fulfill their needs, which are very common. This practice has been increasing day by day especially for business, education, marriage, agriculture as well. But several people take advantage and misuse the facilities given by the bank. With technology developing at such a peak stage these days, data mining plays a key role in computer science to solve such issues. Classification is the most suitable predictive modeling technique in data mining to predict the loan repayment capability of a customer in a banking industry.

2.2 References

1.In this paper, author did a comprehensive study on predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default.

M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

2.In this paper, author did a comprehensive study on the Extending credit to individuals is necessary for markets and society to function smoothly. Estimating the probability that an individual would default on their loan, is useful for banks to decide whether to sanction a loan to the individual or not.

A survey on Ensemble Model for Loan Prediction Anchal Goyal [1], Ranpreet Kaur [2] Research Scolar [1], Assistant Proffesor [2] Department of Computer Science and Engineering RIMT –IET (PTU), Mandi Gob

3.In this paper Author the lending industry, investors offer loans to lenders for the purpose of repaying interest. If the

borrower pays the loan, then the lender will make a profit on the interest
A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION Prateek Dutta*1 *1Student, B.tech Artificial Intelligence, G.H. Raisoni College of Engineering, India

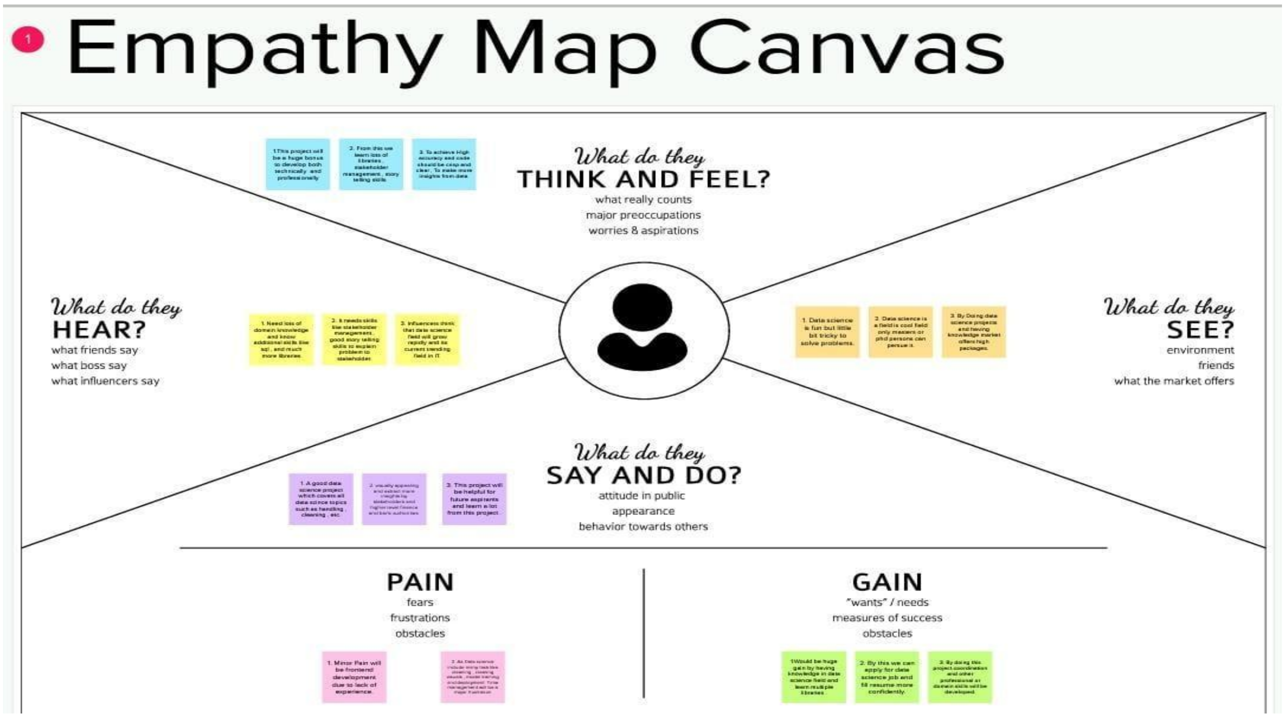
4.Credit risk management is essential to financial institutions as it directly affects business results. Although artificial intelligence (AI) and machine learning are not new, microcredit organizations are shy in accepting these methods in their credit risk assessment.
Z. Ereiz, "Predicting Default Loans Using Machine Learning (OptiML)," *2019 27th Telecommunications Forum (TELFOR)*, 2019, pp. 1-4, doi: 10.1109/TELFOR48224.2019.8971110.

2.3 Problem Statement Definition

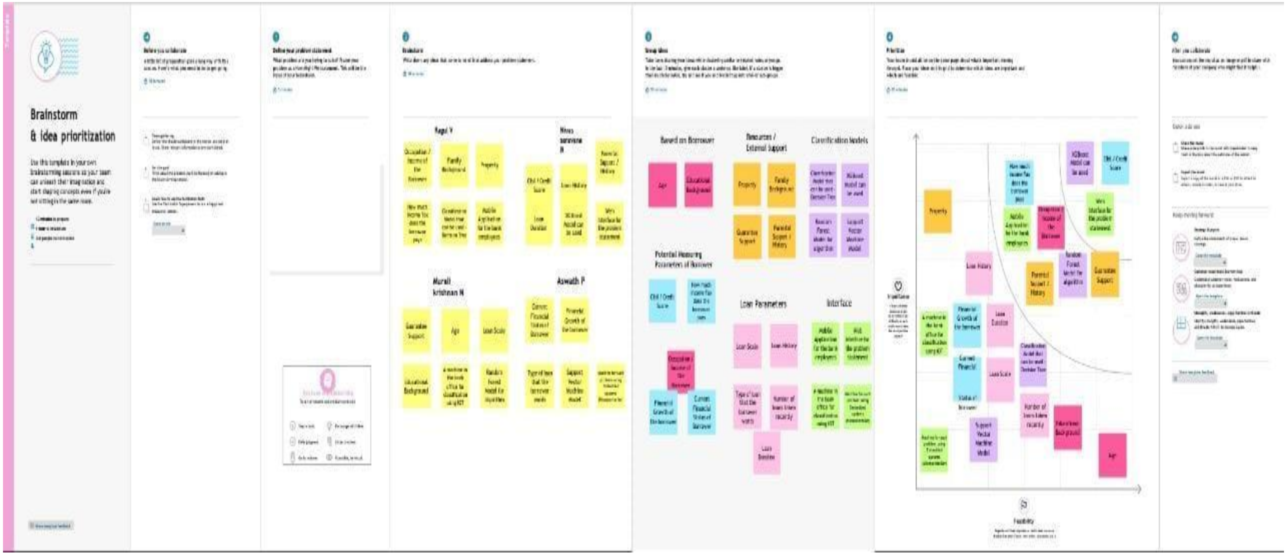
Nowadays, people approach banks to fulfill their needs via bank loans, this practice has been increasing day by day all over the globe especially for business , education, marriage,agriculture etc. But several people take advantage and misuse the facilities given by the banks, so banks realize that retaining customers and preventing fraud should be a strategic tool for healthy competition. One of the important factors affecting the economy and financial condition of our country is the credit system operated by banks. Bank credit risk evaluation is recognized in banks all over the world. There are various methods used for risk level calculation.as we know that the credit risk assessment is very crucial. Everyday a large number of people make application for loans, but all these applicants are not reliable and everyone cannot be approved for loan. We read and heard no of cases that the people can't repay the loan amount and this causes the bank huge loss. So, the main source of income of any bank from the customers is their credit line. By using applied data science techniques and machine learning algorithms, we will check the credit score of the person and predict whether the loan is approved or not.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



3.2 Ideation and Brainstorming



3.3 Proposed Solution

Novelty / Uniqueness:

The uniqueness of the current study is that the model eliminates the two most pressing problems in the banking industry, which are determining whether a borrower is unsafe and lending money to a borrower who isn't a risk. The client information submitted when filling out the online application form is used to automate the loan eligibility procedure. Gender, marital status, education, the number of dependents, income, loan size, credit history, and other information are included. Three key criteria—customer identification, credit underwriting, and fraud underwriting—are used to screen clients. For better filtering, we use the applicant's prior records, and we push clients toward low-interest loans based on their income.

Feasibility:

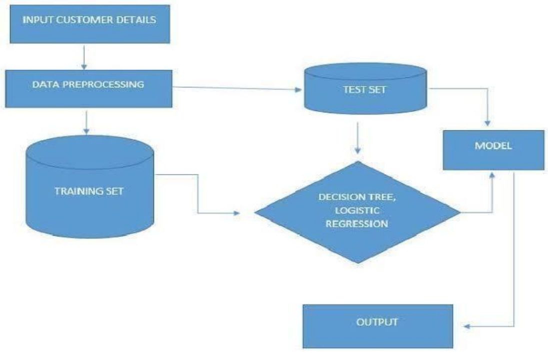
Before making a loan, the ideal candidate must be screened. It is impossible to manually select or anticipate the ideal applicant when there are several loan requests. Data mining techniques like logistic regression and random forest are employed for this selection. When used to predict a binary outcome based on a set of independent variables, logistic regression aids in classification. A supervised learning algorithm is random forest. An ensemble of decision trees, typically trained using the "bagging" approach, make up the "forest" that it constructs. The bagging method's general premise is that combining learning models improves the end outcome.

Social Impact / Customer Satisfaction:

Due to the fact that low interest loans are accepted for the applicants based on their income and there would be no societal consequences. It will be simple for the consumers to pay their interest, and no loan defaulters will be found. This model also contributes to the conclusion that a bank should not just target wealthy consumers for loan approval but should also consider other customer characteristics that are crucial in determining credit approval and identifying loan defaulters.

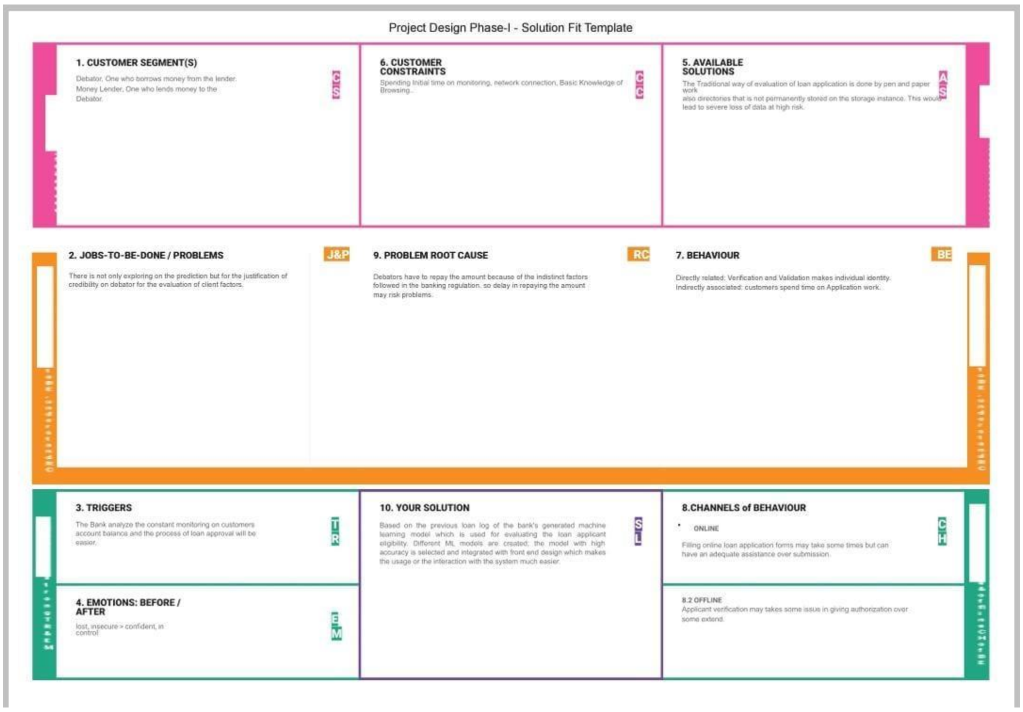
Scalability Of The Solution:

This model supports a huge data set and is subjected to numerous data mining techniques. Applicants who have high incomes and request smaller loan amounts are more likely to be authorised and to repay their debts. The corporation doesn't appear to take into account other factors like gender or marital status. But in this case, all factors are taken into account when issuing a loan, thus candidates with low incomes can also be granted a loan with favourable terms in addition to the wealthy. They are able to



repay their obligation in full thanks to this.

3.4 Problem Solution fit



4. REQUIREMENT ANALYSIS

4.1 Functional requirement

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Home Page	• Smart Lender Applicant Credibility description • Information about Credibility details required for loan approval • if new user , REGISTER • if already exist , SIGN IN
FR-2	User Registration	Enter Mail Id and other personal details required for Registering
FR-3	User login	User Mail Id and Password for Login
FR-4	Loan Approval form	Credibility details should be entered for prediction
FR-5	Result	if Approved - It displays the information about what is done to be next. if Not Approved - It displays the information about what rejection criteria you are not eligible for the loan.

4.2 Non-Functional requirements

Non-functional Requirements:

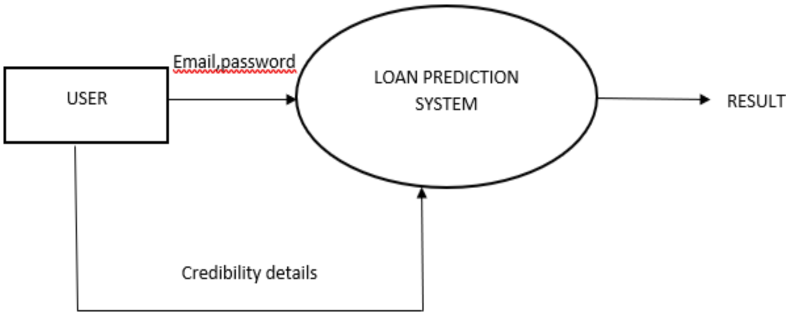
Following are the non-functional requirements of the proposed solution.

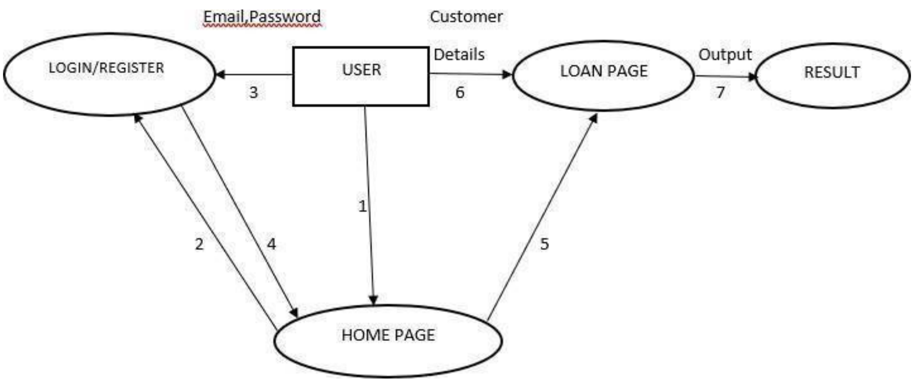
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	It describes the context of Who,What,When,Where and Why. The specific activities the requirements describe should reflect the both range of goal that the system must support business goals for creating a new system.
NFR-2	**Security**	Security functionality that ensures one of many different security properties of software is being satisfied.Security requirements are derived from industry standards,applicable laws,and a history of past vulnerabilities.
NFR-3	**Reliability**	It is the measure of the stability or consistency of the test score
NFR-4	**Performance**	It defines how well the software system accomplishes certain functions under specific condition.
NFR-5	**Availability**	It defines how long the IT system can be unavailable without impacting operations.
NFR-6	**Scalability**	It is the measure of a system ability to increase or decrease in performance and cost in response

5. PROJECT DESIGN

5.1 Data Flow Diagrams:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.





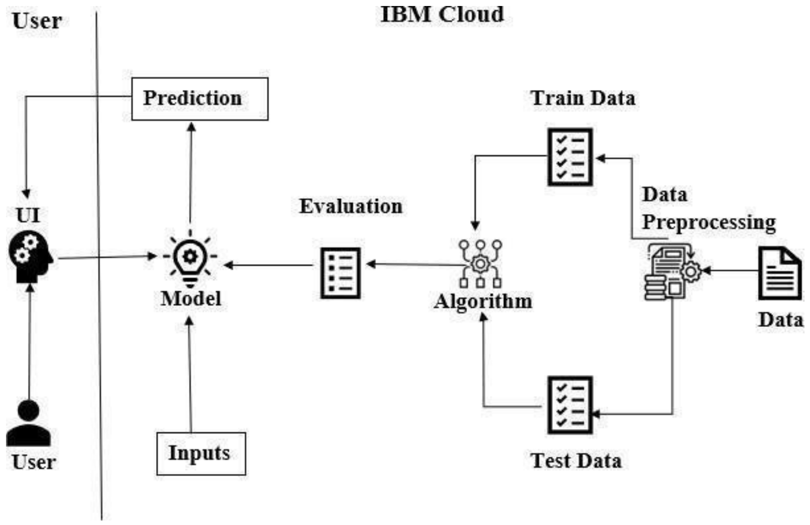
1st Level DFD Data flow:

1. User open the application then the homepage is appear.
2. User open the login/Register page from Homepage.
3. User can register through email id and password.
4. User is redirected to the Homepage once they login.
5. User open the loan page.
6. User enters the required details for loan approval prediction.
7. Result will be displayed in the result page.

5.2 Solution & Technical Architecture

Technical Architecture:

The Deliverable shall include the architectural diagram as below and the information as per the table1 & table 2



Guidelines:

Include all the processes (As an application logic / Technology Block)
Provide infrastructural demarcation (Local / Cloud)
Indicate external interfaces (third party API's etc.)
Indicate Data Storage components / services
Indicate interface to machine learning models (if applicable)

Table-1 : Components & Technologies:

S. No	Component	Description	Technology
1.	User Interface	User interact with our application through web User Interface.	HTML, CSS and Python flask.
2	Application	When the user click on the login button , he/she	HTML ,CSS, Python

.	Logic-1-Login.	is directed to login page, if they are registered already.	flask.
3.	Application Logic-Registration	When the user click on the Register button , he/she is directed to Register page for further process.	HTML,CSS, Python flask.
4.	Application Logic-Credibility details	After Logged in , when the user click on the credibility details form button,he/she directed to the form page to enter the details of applicant for prediction.	Front end- HTML ,CSS , MySQL, Pythonflask Back end-Python
5.	Database	Data type - String ,Numeric.	MySQL.
6.	Cloud Database	Database Service on Cloud	IBM.
7.	File Storage	File storage requirements	NIL
8.	External API-1	Purpose of External API used in the application	NIL
9.	External API-2	Purpose of External API used in the application	Aadhar API
10.	Machine Learning Model	Get the data from the user and predict the data with tested and trained dataset models	Data Recognition Model, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	NIL

5.3 User Stories

User Stories

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Home Page	USN - 1	Loan approval prediction description	I can view /access my homepage.	Low	Sprint - 3
		USN - 2	Information about the credibility details required for the prediction		Low	Sprint - 3
	User Register	USN - 3	Enter Email ID and other personal details required for Register.	I can successfully register by receiving mail.	Medium	Sprint - 2
	User Login	USN - 4	Uses Email ID and Password for login	I have successfully logged in.	Medium	Sprint - 2
	Loan approval Form	USN - 5	Credibility details required for loan should be entered for prediction.	I can access the customer details form	High	Sprint - 1

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
	Result	USN - 6	Results will be displayed.	I got my result successfully.	High	Sprint - 1
		USN - 7	• If Approved, The information about what is done to be next is displayed. • If Not approved, The information about which rejection criteria you are not eligible for the loan is displayed.	I got useful information	Low	Sprint - 4

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement	User Story	User Story / Task	Story Points	Priority	Team Members

	(Epic)	Number				
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Lurdes Infant Joe J Akilesh V A
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	1	High	Ajay M Ajay Balan B
Sprint-1		USN-3	As a user, I can register for the application through Gmail	2	Medium	Akilesh V A Ajay Balan B
Sprint-1	Login	USN-4	As a user, I can log into the application by entering email & password	1	High	Lurdes Infant Joe J Akilesh V A
Sprint-1	Dashboard	USN-5	As a user, I can access the dashboard to check my loan available status.		High	Lurdes Infant Joe J Ajay M

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Form	USN-6	As a user, I can enter the data which I have and also the data which the website asks to me to predict	6	Very High	Ajay M Ajay Balan B
Sprint 3	Prediction	USN-7	As I have given the data into the webpage now the data can be predicted for the loan avail	4	Medium	Akilesh V A Ajay Balan B
Sprint-4	Deployment of the webpage in cloud	USN-8	As a user, I require global access to the web page as a user	3	Low	Lurdes Infant Joe J Ajay Balan B
Sprint-5	Deployment of AI model in the cloud	USN-9	Model could be running on the cloud	3	Low	Ajay M Akilesh V A
Sprint-6	Model building	USN-10	I REQUIRE AN ML model that can credit	5	High	Lurdes Infant Joe J Akilesh V A

			defaulters			

6.2 Sprint Delivery Schedule

Project Tracker, Velocity & Burn down Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	23 Oct 2022	28 Oct 2022	11	29 Oct 2022
Sprint-2	20	6 Days	30 Oct 2022	04 Nov 2022	11	06 Nov 2022
Sprint-3	20	6 Days	08 Nov 2022	13 Nov 2022	11	15 Nov 2022
Sprint-4	20	6 Days	15 Nov 2022	20 Nov 2022	11	21 Nov 2022

Velocity:

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let’s calculate the team’s average velocity (AV) per iteration unit (story points per day)

7. CODING & SOLUTIONING

7.1 Feature Selection

Feature Selection Many feature selection techniques are available. Before doing classification feature selection can be applied so as to increase the accuracy level. During the last two decades, feature selection techniques have become an active and fruitful research field in machine learning . Feature selection is an important data mining task which can be effectively utilized to develop a knowledge based model for the Loan Credibility Prediction System. Feature selection plays a major role in data preprocessing. Generally the dataset consists of relevant, irrelevant and redundant features. Irrelevant and redundant features do not contribute anything to determine the target class and at the same time reduces the accuracy of the model created. The process of eliminating such features from a dataset is termed as feature selection.

The search strategy usually employs feature ranking or subset search [9] techniques. In feature ranking, a weight or score is assigned to each feature according to its individual merit. Ranking methods are not able to remove redundant features within the dataset. Subset search evaluates the quality of subsets of features.

The best performance of the selected features can be achieved when both the feature selection and classification stages are optimized together using the same criterion function . Criteria function can be either classifier independent (i.e., filter approach) or classifier specific (i.e., wrapper approach or embedded method).

8. TESTING

8.1 Test Cases

The user will be provided with the following data attributes such as Loan_ID,Gender,Married,Dependents,Education,Self_Employed,ApplicantIncome,CoapplicantIncom4e,LoanAmount,Loan_Amount_Term,Credit_History,Property_Area, Loan_Status so first proceeding on to the prediction the test cases should be ensured that the following data is provided for the prediction .

8.2 User Acceptance Testing

As the user enters the data into the html page the model builded with the various algorithms will take all the inputs in the frontend and it will be processed in the backend with the algorithms make sure it undergoes all the rules of data providing so that user will get the output to be accepted.

9. RESULTS

9.1 Performance Metrics

True Positive, True Negative, False Positive and False Negative are usually presented in a tabular format in the so-called Confusion Matrix, which is simply a table organizing the four values.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Accuracy

Accuracy is the fraction of predictions our model got right out of all the predictions. This means that we sum the number of predictions correctly predicted as Positive (TP) or correctly predicted as Negative (TN) and divide it by all types of predictions, both correct (TP, TN) and incorrect (FP, FN).

Accuracy ranges between 0 and 1. These extreme cases correspond to completely missing the predictions or having always correct predictions. For instance, if our model is able to perfectly predict, the model will have no False Positives or False

Negatives, making the numerator be equal to the denominator, bringing the Accuracy to 1.

Conversely, if our system is always off, incorrectly predicting each time, the number of True Positives and True Negatives will be zero, making the equation be zero divided by something positive, leading to an Accuracy equal to 0.

In real life, Accuracy technically ranges between 0.5 and 1, because if the Accuracy falls below 0.5, we can simply flip the predictions labels to obtain a better prediction.

Accuracy, however, is not a great metric, especially when the data is imbalanced. When there is a significant disparity between the number of positive and negative labels, Accuracy does not tell the full story. For instance, let's consider an example where we have 100 samples, 95 of which labelled as belonging to Class 0, and 5 labelled as Class 1. In this case, a poorly built "dummy" model which always predicts Class 0, achieves a 95% Accuracy, which indicates a very strong model. However, this model is not really predictive and Accuracy is not the right performance metric to evaluate the power of this model. If we used only Accuracy to evaluate this model, we would end up providing stakeholders, and clients eventually, with a model that is not performant or predictive.

Precision

To overcome the limitations of Accuracy, Data Scientists usually use Precision, Recall and Specificity. Precision tells what proportion of positive predictions was actually correct. It achieves this by counting the samples correctly predicted as positive (TP) and dividing it by the total positive predictions, correct or incorrect (TP, FP).

Recall = Sensitivity = True Positive Rate = Hit Rate

Similarly to Precision, Recall aims at measuring what proportion of actual positives was identified correctly. It does so by dividing the correctly predicted positive samples (TP) by the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (TP, FN).

Specificity = True Negative Rate = Selectivity

Symmetrically to Recall (also known as Sensitivity), Specificity aims at measuring what proportion of actual negatives was identified correctly. It does so by dividing the correctly predicted negative samples by the total number of negatives, either correctly predicted as negative or incorrectly predicted as positive (TN, FP).

Considering the example to show the shortcomings of Accuracy, if we use Precision, Recall and Specificity, we get:

- Accuracy: 0.95
- Recall: 0

By using additional performance metrics instead of Accuracy, we can better understand that a model predicting the majority class all the time is actually a low-performance model (*Recall = 0*) even though Accuracy is high (*Accuracy = 0.95*).

Area Under the ROC Curve (AUC)

As we've seen, one of the issues of Accuracy is that it can lead to overly inflated performance if the distribution of the classes is not very well balanced. AUC, which stands for "Area under the ROC Curve" (more about this later), measures the entire two-dimensional area underneath the entire ROC curve.

It is an aggregate measure of performance across all possible classification thresholds. Another way of interpreting AUC is as the probability that the model ranks a random positive sample higher than a random negative sample.

AUC is a great metric, especially when dealing with imbalanced classes, and is one of the most frequently used performance measures in classification, even though it can be used only in binary classification settings (i.e. not with more than 2 classes as target).

Some of the properties that make it a preferred metric are:

- **Scale-Invariance**. AUC measures how well predictions are ranked, instead of their absolute values.
- **Classification-Threshold-Invariance**. AUC measures the quality of the model's predictions regardless of what classification threshold is chosen.

F1 Score

The F1 score is a less known performance metric, indicating the harmonic mean of Precision and Recall. The highest value of an F1 Score is 1, indicating perfect Precision and Recall, and the lowest possible value is 0 if either the Precision or the Recall is zero.

Performance Charts

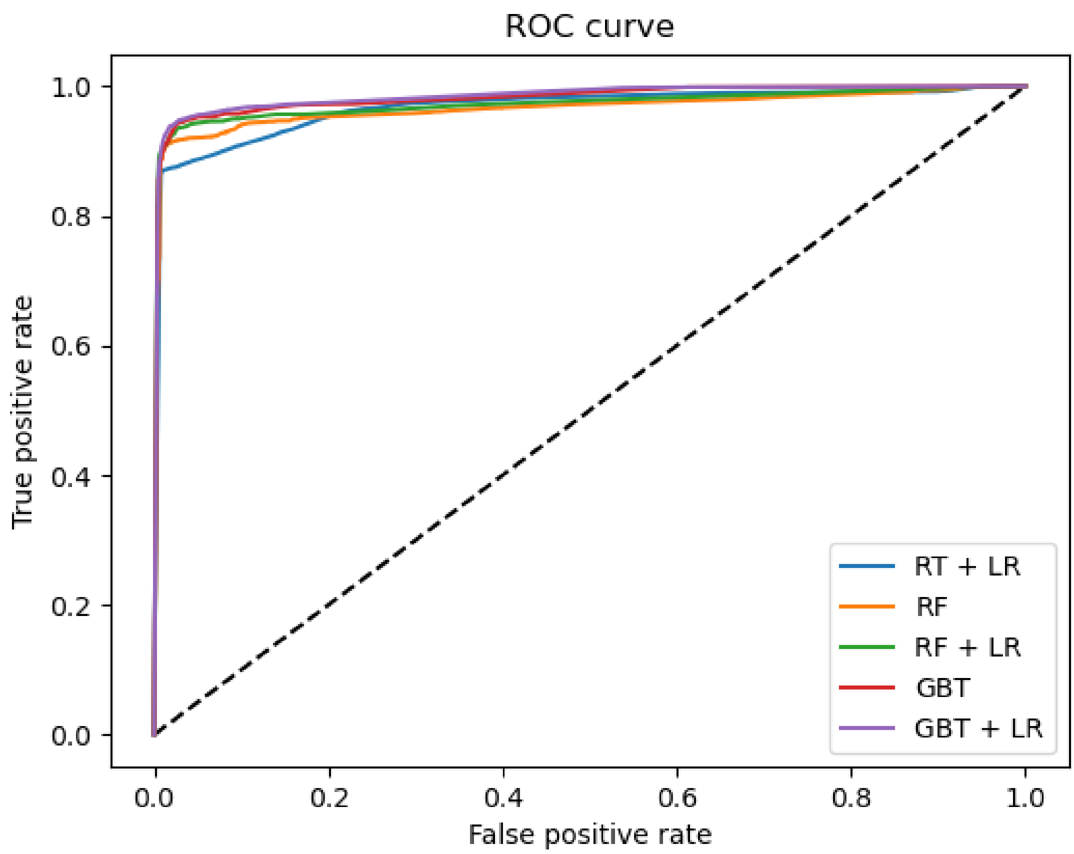
Additionally, performance measures can be not only communicated as single numbers but also as charts. Some common charts showing a Machine Learning Model's performance are the ROC Curve and the Precision/Recall Curve.

ROC Curve (Receiver Operating Characteristic Curve)

A **ROC curve** is a graph showing the performance of a classification model at all classification thresholds. The chart's y-axis is the True Positive Rate, while the x-axis is the False Positive Rate and the plot consists of the TPR and FPR values varying the threshold.

The worst-case scenario (random chance) consists of a 45 degrees diagonal line. The best-case scenario consists of an angled line, going vertically first and horizontally after.

Lowering the classification threshold, the model classifies more items as positive, increasing both False Positives and True Positives.

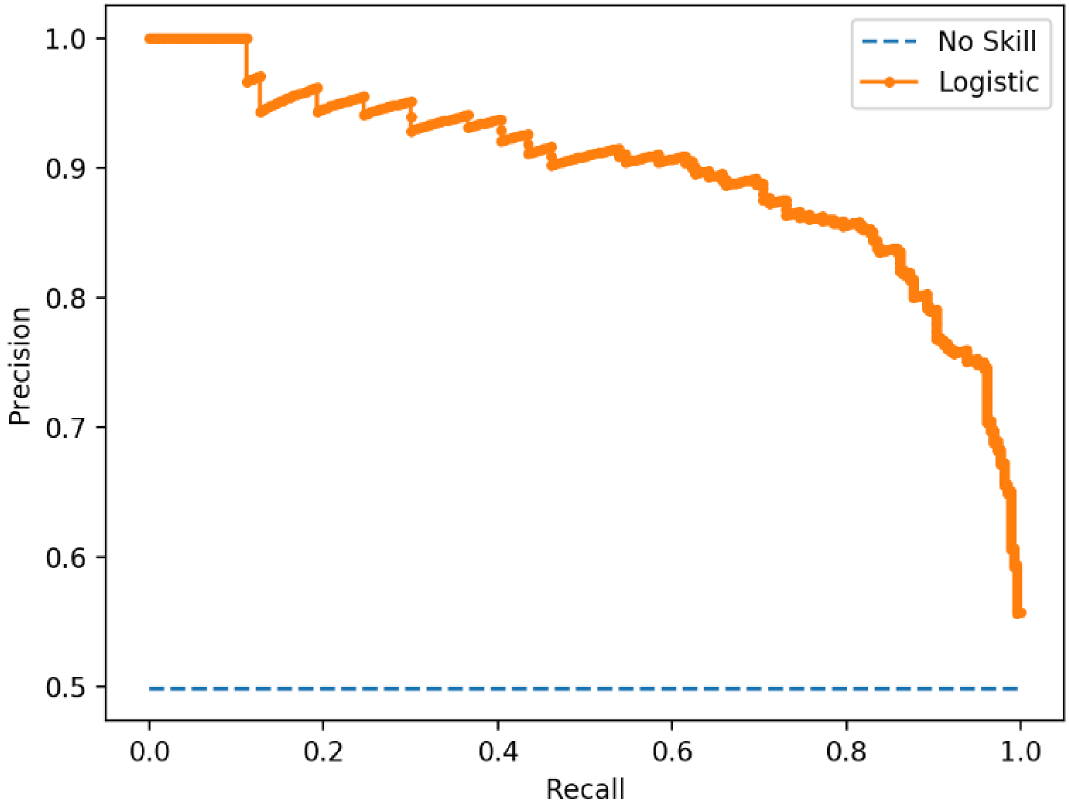


Source: [Eugenio Zuccarelli]

Precision/Recall Curve

Similarly to the ROC Curve, a Precision/Recall Curve plots performance over a y-axis showing Precision and an x-axis which is Recall. Each point is evaluated at different threshold values.

The best-case scenario is a flipped version of the ROC Curve's best-case scenario, basically consisting of a horizontal line then becoming vertical. Differently, the worst-case scenario, random chance, is seen as a horizontal line at *Precision = 0.5*.



Source: [Eugenio Zuccarelli]

Impact of Choosing the Right Performance Metric

Choosing the right metric is key, especially in cases where False Positives and False Negatives do not have the same impact. Ideally, we would want to have a perfect prediction both in terms of False Positive and False Negative (both zero), but with Machine Learning models there is usually a tradeoff between detecting False Positives or False Negatives well.

For instance, if our model predicts whether a person has got a deadly disease, like cancer, it could be said that False Positives are more important. We want to make sure that if that person has the disease, we correctly flag them. We are less concerned if we accidentally misclassify a person as having the disease even though they didn't have it.

10. CONCLUSION

The proposed algorithm had produced better accuracy than existing methods. Experiments on standard data sets proved that the proposed algorithm for loan credibility prediction system outperforms many other feature selection method.

11. FUTURE SCOPE

In future I wish to develop a Data mining application using various feature selection algorithms and surely it helps the bank officers to make proper decisions when a new customer approaches the bank for taking the loan. The proposed hybrid feature selection algorithm for classifying the loan credibility behavior of a customer in a banking industry can also be used for several other applications in the future especially binary classification problems such as prediction of various diseases, prediction of various examination results etc

12. APPENDIX

Source Code GitHub:
IBM-EPBL/IBM-Project-43216-1660714264: Smart Lender - Applicant Credibility Prediction for Loan Approval (github.com)

Project Demo Link:
https://drive.google.com/file/d/1kMNCOQSLEzf25tuwbGopRqus0miF911R/view?usp=share_link