Smart Lender
Applicant Credibility Prediction for Loan Approval

PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY AND ENTREPRENEURSHIP

Submitted By

Anirudh T E (195001016)
Pranav Raj SB (195001078)
Tushar Shah (195001117)
Yashwanth M (195001130)

ACADEMIC YEAR 2022-2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SSN COLLEGE OF ENGINEERING, KALAVAKKAM 603110

November 2022

PROJECT REPORT

Smart Lender - Applicant Credibility Prediction for Loan Approval

Team ID:	PNT2022TMID53029	**Project ID:**	13820-1659532356
Team Members:	Anirudh T E Pranav Raj SB Tushar Shah Yashwanth M	**Roll Numbers:**	SSNCE195001016 SSNCE195001078 SSNCE195001117 SSNCE195001130
Industry Mentor(s):	Nidhi	**Faculty Mentor**:	Lakshmi Priya S

1. INTRODUCTION

1.1 Project Overview

A loan default occurs when a borrower takes money from a bank and does not repay the loan. People often default on loans due to various reasons. Borrowers who default on loans not only damage their credit but also risk being sued and having their wages garnished. Many individuals utilize debt to pay for things they wouldn't be able to buy otherwise, such as a home or a vehicle. While loans may be beneficial financial instruments when utilized correctly, they can also be formidable foes. Lending is a vital tool that propels all enterprises and individuals worldwide to greater financial success. The need for capital has risen dramatically as the world's economies become increasingly integrated and interdependent. In the last decade, the number of retail borrowers, SMEs, and commercial borrowers has increased dramatically. Though most financial institutions have seen an increase in revenue and profit due to this rising trend, not everything is green. In recent years, there has been an increase in loan defaults, which has already begun to affect the bottom lines of several financial institutions.

1.2 Purpose

We aim to make use of machine learning to make better financial predictions and understand the banking sector's lending applications and the creditworthiness of individuals and organizations. Machine Learning techniques are very crucial and useful in the prediction of these types of data.

2. LITERATURE SURVEY

2.1 Existing problem

One of the most important factors which affect our country's economy and financial condition is the credit system governed by the banks. The process of bank credit risk evaluation is recognized at banks across the globe. As we know credit risk evaluation is very crucial, there are a variety of techniques used for risk level calculation. In addition, credit risk is one of the main functions of the banking community. The prediction of credit defaulters is one of the difficult tasks for any bank.

2.2 References

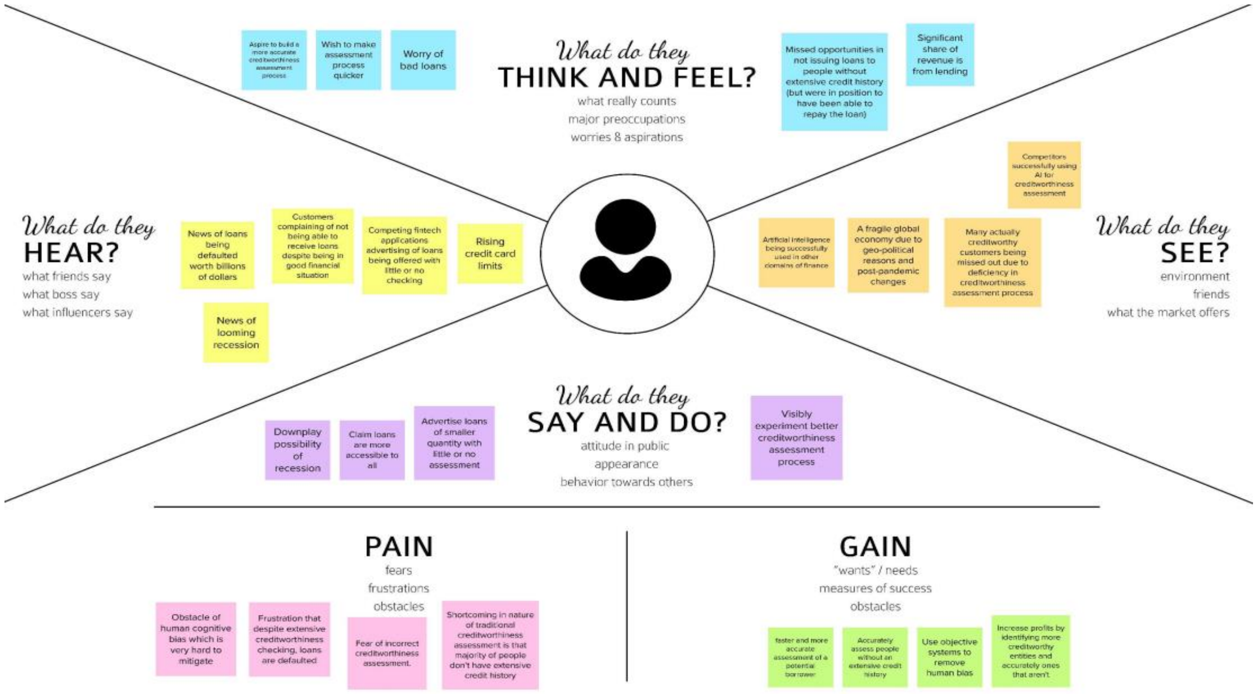
- Soni P M, Varghese Paul,International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
- VivekBhambri "Application of Data Mining in Banking Sector", IJCSt Vol. 2, ISSue 2, June 2011, ISSN : 2229- 4333(Print) | ISSN : 0976-8491(Online)
- Challita N., Khalil M., Beauseroy P. 2016 IEEE IntMultidiscipConfEng Technol. 2016. New feature selection method based on neural network and machine learning

2.3 Problem Statement Definition

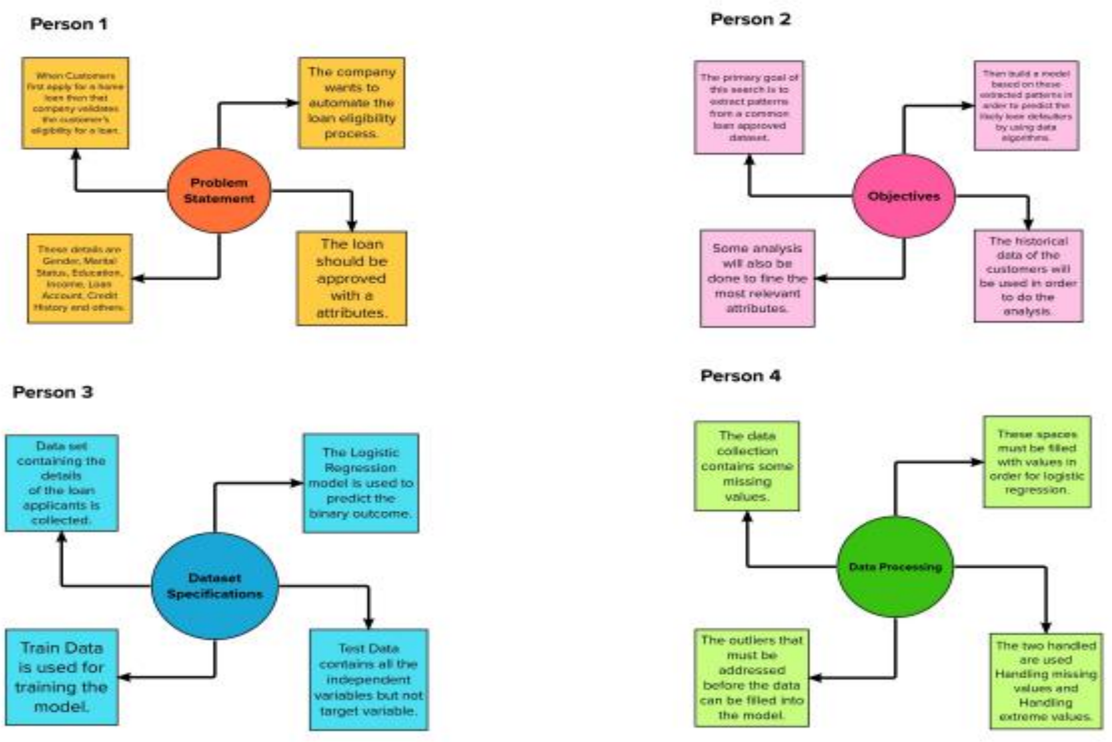
The prediction of credit defaulters is one of the difficult tasks for any bank. But by forecasting the loan defaulters, the banks definitely may reduce their loss by reducing their non-profit assets, so that recovery of approved loans can take place without any loss and it can play as the contributing parameter of the bank statement.

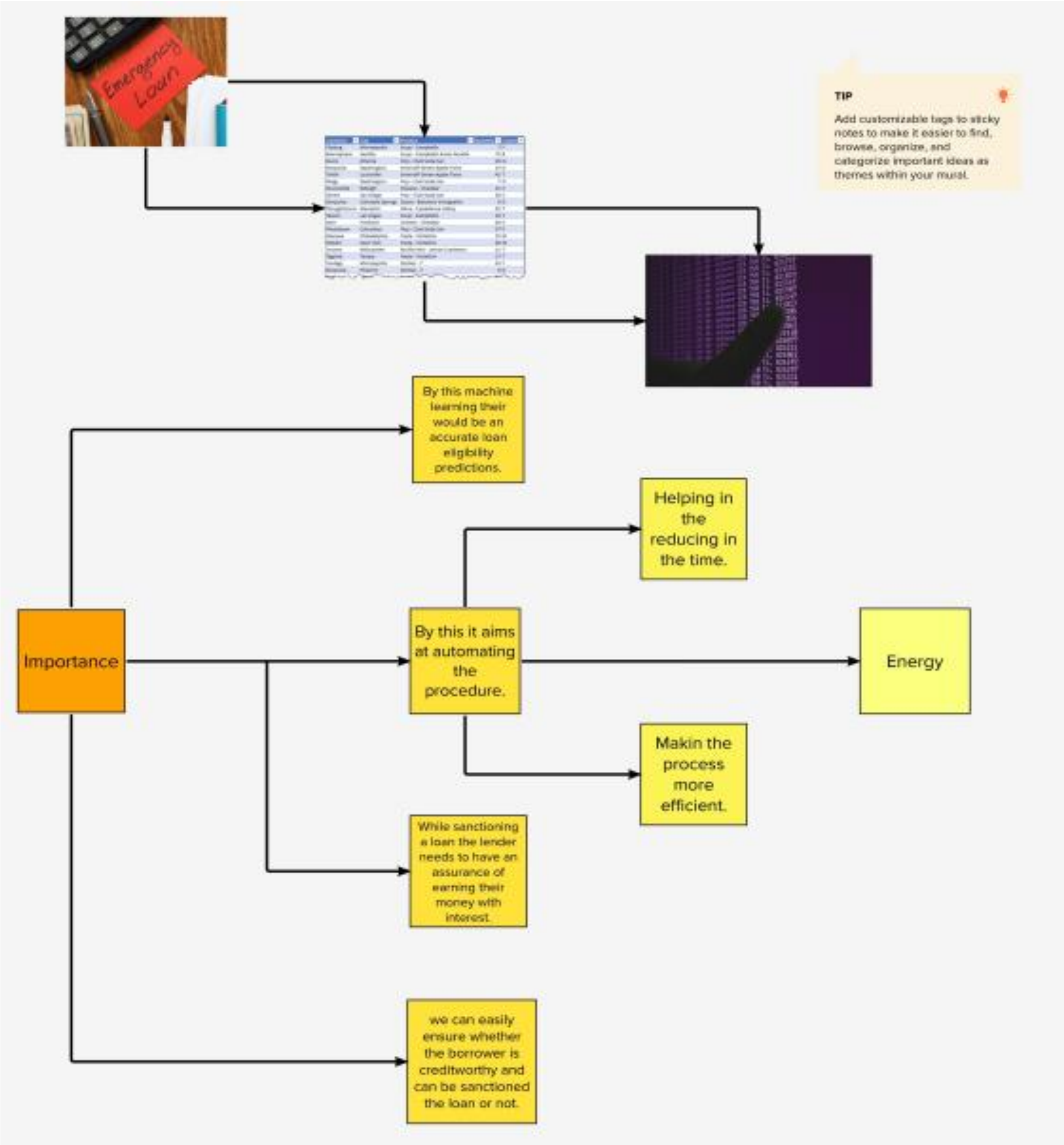
3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

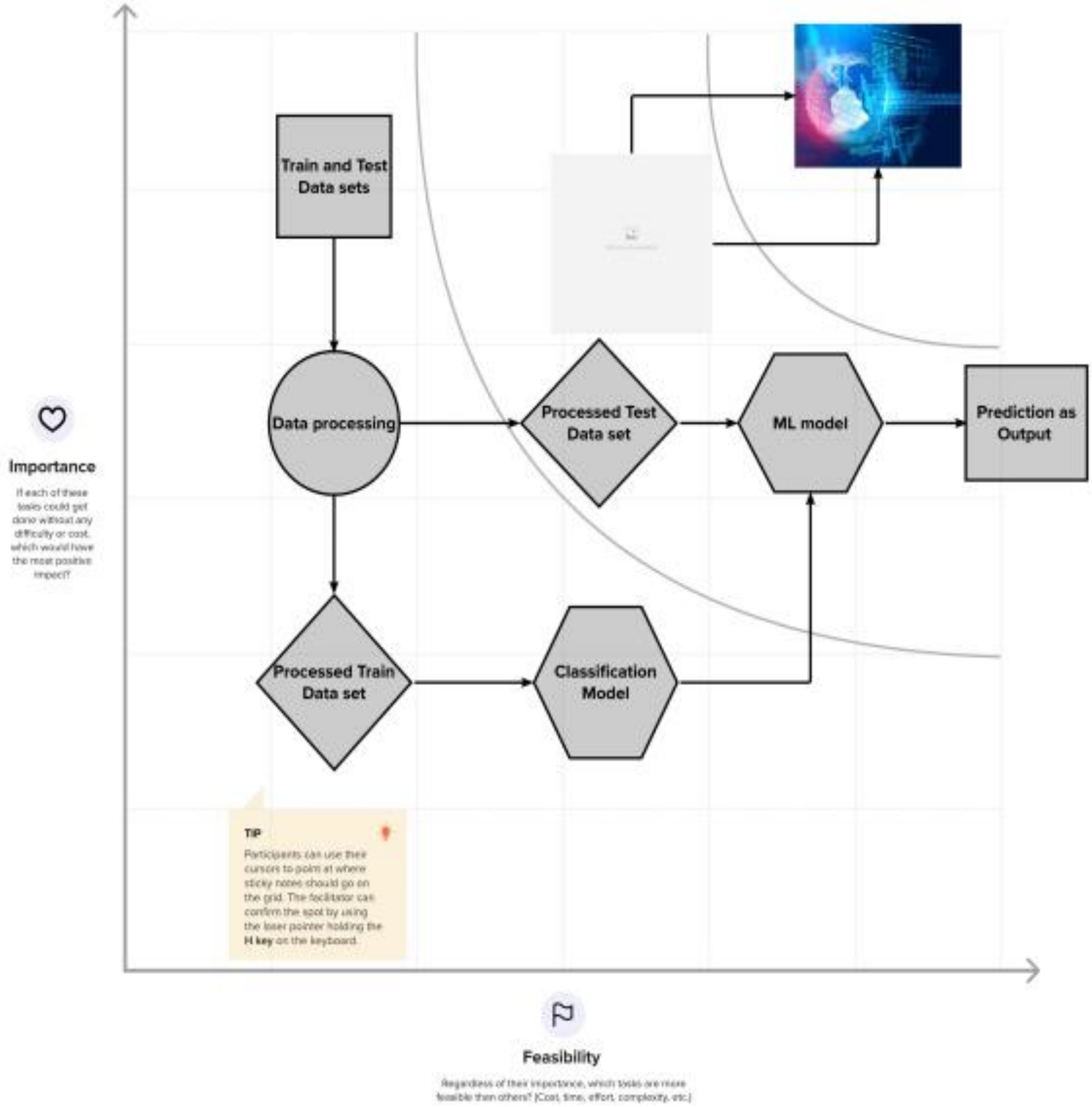


3.2 Ideation & Brainstorming





Emergency Loan
TIP
Add customizable tags to sticky notes to make it easier to find, browse, organize, and categorize important ideas as themes within your mural.
By this machine learning their would be an accurate loan eligibility predictions.
Helping in the reducing in the time.
Importance
By this it aims at automating the procedure.
Energy
Makin the process more efficient.
While sanctioning a loan the lender needs to have an assurance of earning their money with interest.
we can easily ensure whether the borrower is creditworthy and can be sanctioned the loan or not.



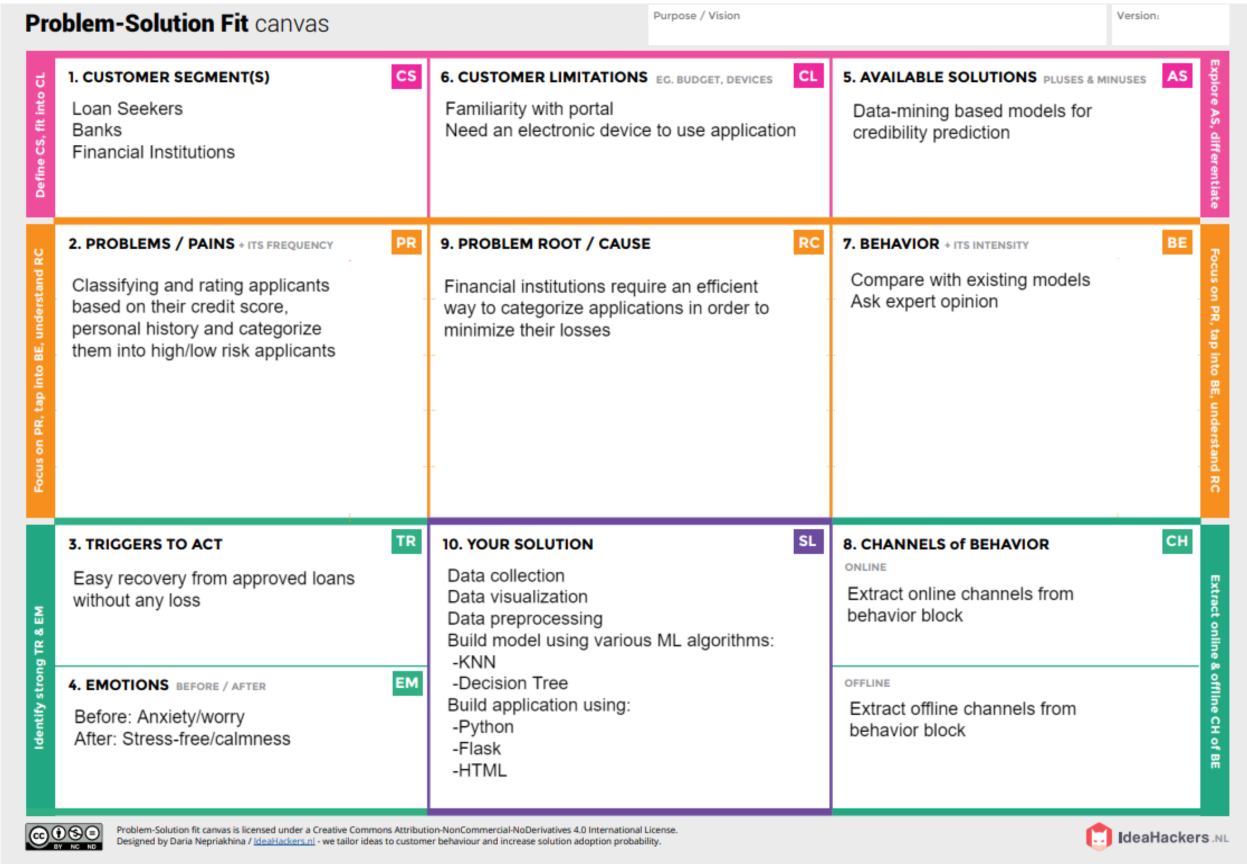
Train and Test Data sets
Data processing
Processed Test Data set
ML model
Prediction as Output
Processed Train Data set
Classification Model
Importance
If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?
TIP
Participants can use their cursors to point at where sticky notes should go on the grid. The facilitator can confirm the spot by using the laser pointer holding the H key on the keyboard.
Feasibility
Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)

3.3 Proposed Solution

S. No.	Heading	Details
1.	Problem Statement	One of the most important factors which affect our country’s economy and financial condition is the credit system governed by the banks. The process of bank credit risk evaluation is recognized at banks across the globe. The prediction of credit defaulters is one of the difficult tasks for any bank. But by forecasting the loan defaulters, the banks definitely may reduce their loss by reducing their non-profit assets, so that recovery of approved loans can take place without any loss and it can play as the contributing parameter of the bank statement.
2.	Proposed Solution/Idea	We aim to make use of machine learning to make better financial predictions and understand the banking sector's lending applications and the creditworthiness of individuals and organizations. Machine Learning techniques are very crucial and useful in the prediction of these types of data.
3.	Novelty/Uniqueness	We will be using classification algorithms such as Decision tree, Random forest, KNN, and XGBoost. We will train and test the data with these algorithms. From this, the best model is selected and saved using pickle format. We will be doing flask integration and IBM deployment.
4.	Social Impact	Using this system would significantly improve the banking ecosystem and reduce the loan defaulting rates that banks currently see. This would also allow the clients to borrow based on their past records and better understand their borrowing capacity.

5.	Business Model	This system would be used by both banks as well as the clients of the bank. It would reduce time and effort from both ends thus making it monetizable at both ends. The banks would spend to ensure that this process helps them screen loan applicants better and the customer could pay a nominal processing fee if his loan is approved after the screening done by our system.
6.	Scalability	The system, being dynamically and modularly developed, allows for much modification and large scalable operations. More data when made available can be processed and produce efficient results. This system is easily and efficiently scalable.

3.4 Problem Solution fit



4. REQUIREMENT ANALYSIS

4.1 Functional requirements

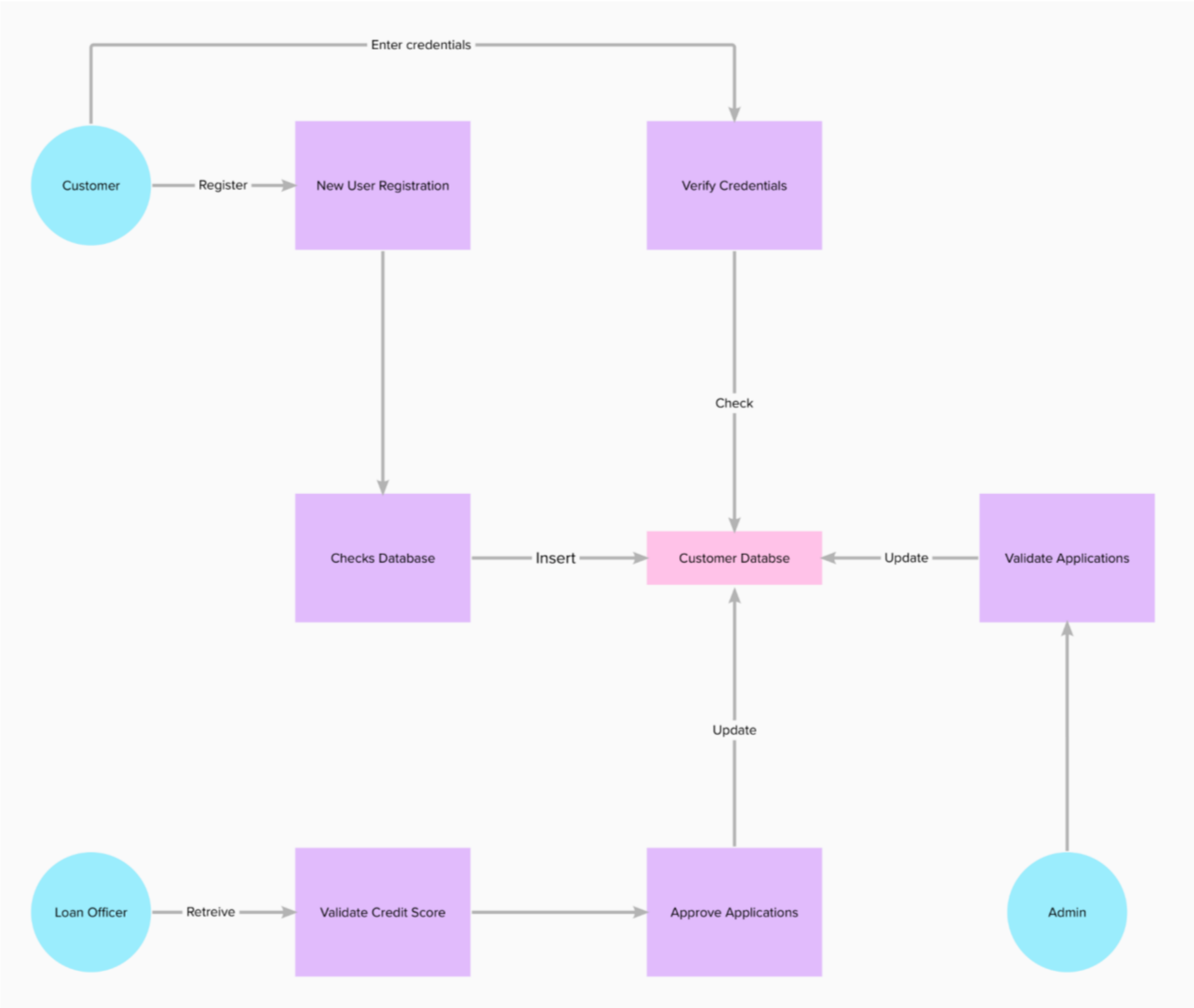
S. No.	Functional Requirement	Sub-Requirement/Sub-task
1.	User Registration	Can be performed in either of the following ways: a. Through Registration Form b. Through Email - Gmail c. Through LinkedIn d. Through Apple ID
2.	User Confirmation	Confirmation can be performed by: a. Through Email OTP b. Through Telephonic OTP c. Through SMS OTP d. Biometric Verification
3.	User Loan Application	a. Filling the Application form. b. Modify Application c. Verifying the application form d. Submitting application form
4.	Loan Issuance	a. Checking the current Loan status b. Loan Approval c. Loan Rejection
5.	Credit History Check	a. Auditing previous loan history. b. Income statement analysis
6.	User Management	a. Recommend appropriate Loan Schemes for users b. Categorize users based on income and credit history slabs for easier analysis

4.2 Non-Functional requirements

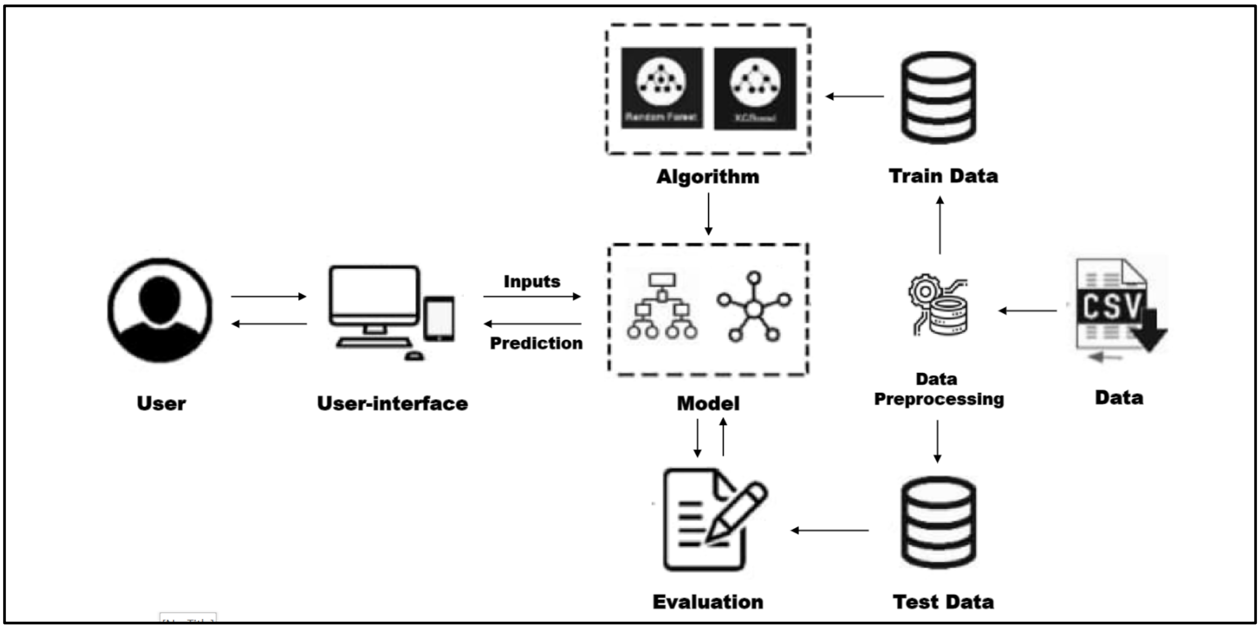
S. No.	Functional Requirement	Sub-Requirement/Sub-task
1.	Usability	a. Simple and eye-pleasing UI b. Seamless navigation and clutter-free tabs c. Displaying relevant info only based on user’s choice d. Light web pages for fast access to data
2.	Security	a. OTP verification on login b. The privacy of users is to be taken care of c. Limited and necessary access to data only d. Verification of users
3.	Reliability	a. All data must be backed up for easy retrieval. b. Light pages to ensure fast data-access, minimal lag.
4.	Performance	a. Web-based application. b. Requires minimum Intel Pentium 4 processor, 4 GB RAM, 1280x1024 screen with application window size 1024x680
5.	Availability	a. Cross-platform support independent of the device used
6.	Scalability	a. Can be used on larger sets of data and by growing user base with varying system specifications to ensure scalability

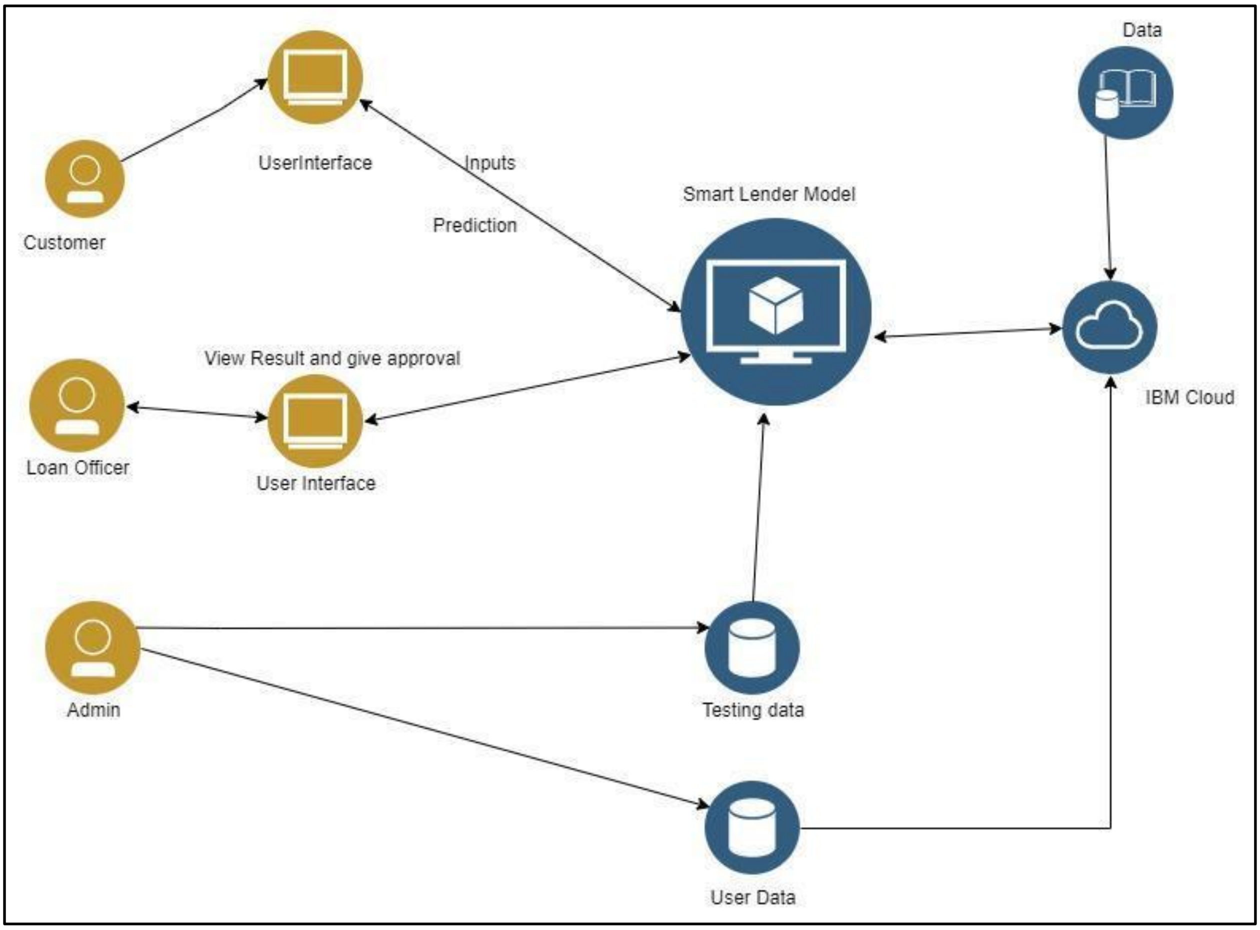
5. PROJECT DESIGN

5.1 Data Flow Diagram



5.2 Solution & Technical Architecture





5.3 User Stories

User Type	Functional Requirement	User Story Number	User Story/Task	Acceptance Criteria	Priority	Release
Customer	Registration	USN-1	Enter credentials to register	Access account	High	Sprint-1
		USN-2	Receive confirmation mail for registration	Use account	High	Sprint-1
	Login	USN-3	Log on using registered credentials	Access dashboard	High	Sprint-1
	Dashboard	USN-4	Create and submit application	Request Loan	High	Sprint-2
		USN-5	Re-apply for loan	Re-request Loan	Low	Sprint-3
		USN-6	View application status	View status	Medium	Sprint-2
Admin	Login	USN-7	Log on using given credentials	Access dashboard	High	Sprint-2
	Dashboard	USN-8	View applications received	Validate applications	High	Sprint-3
Loan Officer	Model	USN-9	Automated generation of credit score	Separate applications into groups	Medium	Sprint-4
	Dashboard	USN-10	Analysis of credit score	Approve loans	High	Sprint-4

6. PROJECT PLANNING & SCHEDULING

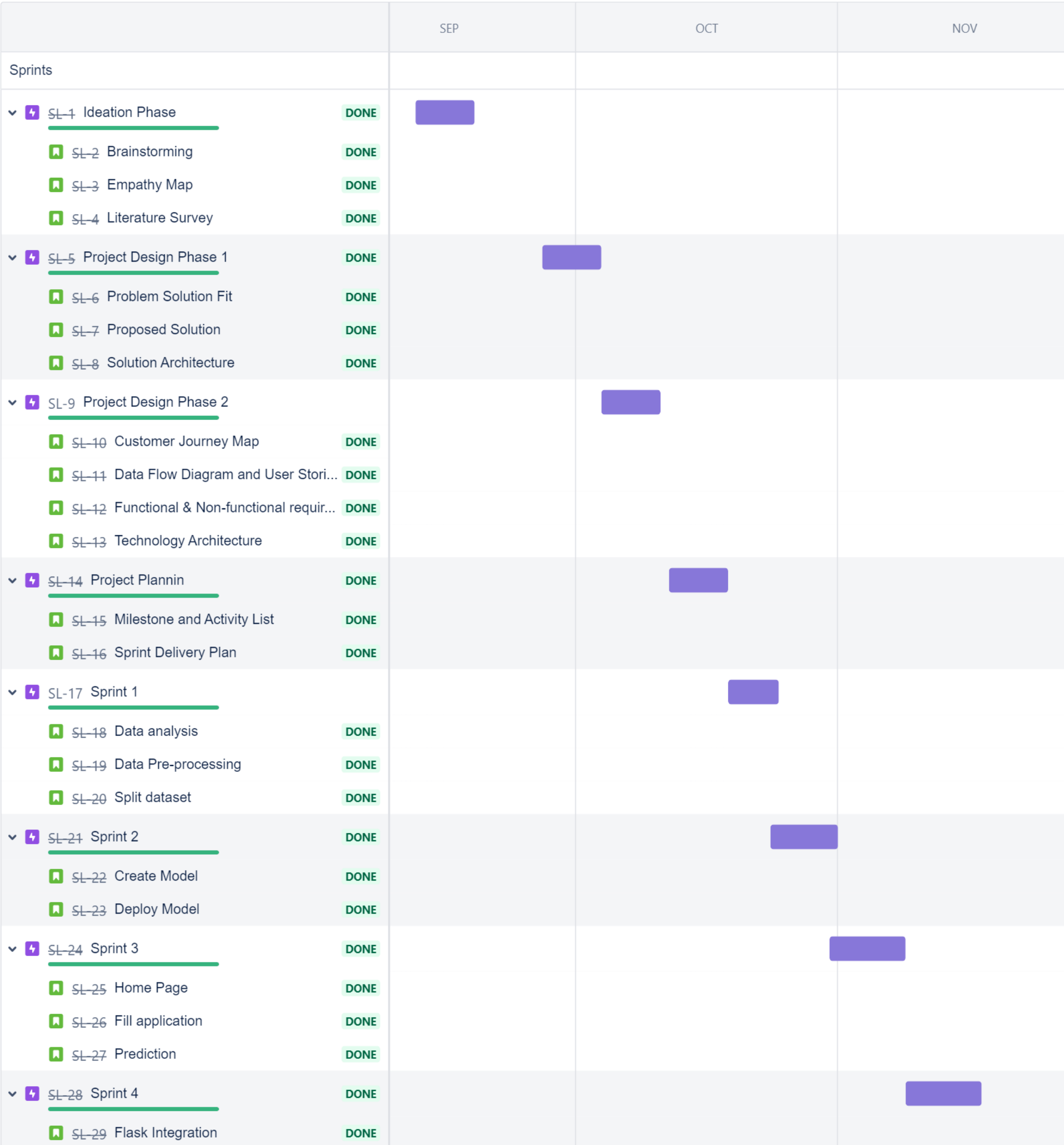
6.1 Sprint Planning & Estimation

Sprint	Functional Requirement	User Story Number	User Story/Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	Enter credentials to register	2	High	Anirudh T E Tushar Shah
Sprint-1	Registration	USN-2	Receive confirmation mail for registration	2	High	Pranav Raj SB Yashwanth M
Sprint-1	Login	USN-3	Log on using registered credentials	2	High	Anirudh T E
Sprint-2	Dashboard	USN-4	Create and submit application	2	High	Tushar Shah Yashwanth M
Sprint-3	Dashboard	USN-5	Re-apply for loan	1	Low	Tushar Shah
Sprint-2	Dashboard	USN-6	View application status	1	Medium	Yashwanth M
Sprint-2	Login	USN-7	Log on using given credentials	2	High	Pranav Raj SB Tushar Shah
Sprint-3	Dashboard	USN-8	View applications received	3	High	Anirudh T E Pranav Raj SB
Sprint-4	Model	USN-9	Automated generation of credit score	2	Medium	Anirudh T E
Sprint-4	Dashboard	USN-10	Analysis of credit score	3	High	Pranav Raj SB Yashwanth M

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed	Sprint Release Date (Actual)
Sprint-1	6	7 Days	24th Oct 2022	30th Oct 2022	6	30th Oct 2022
Sprint-2	5	7 Days	31st Oct 2022	6th Nov 2022	5	6th Nov 2022
Sprint-3	4	7 Days	7th Nov 2022	13th Nov 2022	4	13th Nov 2022
Sprint-4	5	7 Days	14th Nov 2022	20th Nov 2022	3	20th Nov 2022

6.3 Reports from JIRA



7. CODING & SOLUTIONING

The Loan Credibility of the Applicant is calculated based on multiple factors which help us generate an overall score acting as a good indicator of the users credit worthiness. Taking into consideration multiple factors which are assigned separate weights leads to a better overall score not heavily dependent on a single feature alone. Some of the features taken into consideration and their impact are shown below.

Feature 1: Income of the user

The income of the user gives a very general overview of the users capability to repay a loan, if sanctioned. The higher the income is, the more easily one can pay back the amount borrowed from the bank. This feature is given an important weightage and is a good indicator of the worthiness of the individual.

Feature 2: Income of the co-applicant of the user

The loan might be taken by more than one person jointly. In such a scenario, the co-Applicants income plays an important role too. Our model combines this income along with the income of the primary applicant to produce a joint income score which is then added to the overall score used to calculate the credit worthiness of the borrower.

Feature 3: Number of dependents of the user

An applicant might have a family consisting of multiple members and be the sole breadwinner of the family. In this case, his expenses might be higher than someone with lesser dependents. This helps us get an idea about how the user spends which gives us a clearer picture of the amount the user will be left with to repay the monthly installments. A user with 2-3 dependents might have more expenditure and likely to have lower money in hand to repay EMIs than a user with 0-1 dependents under his umbrella. A score equal to the number of the dependents is allocated to the user to understand their credit worthiness.

Feature 4: Educational history of the user

An applicant might be a graduate which opens doors to higher employment opportunities and possibly a higher income too. This would make them a better candidate to disburse the loan to and thereby is a good feature to take into consideration. A non-graduate applicant, although employed might have a lower future growth prospect, at least on paper which would place him under the educated applicants when this feature is looked at in an isolated manner. A score 1 or 0 is allocated to the applicant based on whether he is a graduate or not.

Feature 5: Source of income of the user

The applicant might be either self-employed or a salaried employee. Self-Employed applicants also carry the risk of varying incomes and the possibility of lower than expected incomes during some months. This risk is relatively lower in the case of employees with fixed salaries. We assign a score of either 1 or 0 based on this feature, where a preference is given to one with a fixed income over someone who is self-employed.

Feature 6: Loan amount and Tenure

The loan amount desired by the user and the tenure of the loan can be used to calculate the monthly EMI for the loan amount. This is then compared with the income of the user to determine if the user will be able to repay the loan. This feature, although simply goes a long way in determining the creditworthiness of the applicant when they apply for a loan.

Feature 7: Guidelines pertaining Credit Score

The credit score of an individual is based on a score assigned to them based on their repayment history and current credit credit card usage amongst other factors. A credit score guideline helps us understand if the user's past credit history is clean and if they have defaulted on any loans earlier. When an individual who has met the guidelines applies, they are more likely to repay the loan when compared to someone who has flouted them earlier. We allocate a score of 1 for those who meet these guidelines and 0 for those who do not.

Feature 8: Property owned by the Applicant

An applicant might have property which can act as collateral in case of a loan default. The property is analyzed and we use the property location as a factor to evaluate the property value. A property may be in the Urban, Semi-Urban, Rural areas for which scores of 3,2,1 are given respectively. An applicant with no property is given a score of 0 in this category.

Components & Technologies:

S. No.	Component	Description	Technology
1.	User Interface	Customer, Admin and Loan Approval Officers will be able to register login and use the functionalities available to them.	HTML
2.	Data Collection	User gives input to the model to check the prediction.	Python
3.	Approval	The loan officer can check the results produced from the model and can either accept or reject the loan.	Python, Watson Studio
4.	User data	Data that is provided to the user.	Test data
5.	Data	Data used to train the model.	Train data
6.	Smart Lender Model	Performs automated loan approval decision and gives details supporting decision.	Javascript, Python
7.	Machine Learning model	Automated decision making for loan approval and giving details for decision.	Decision tree, Random forest, KNN, XGBoost.
8.	Infrastructure (Server/Cloud)	Default	Flask

Application Characteristics:

S. No.	Component	Description	Technology
1.	Model Analysis	Predictive modeling is commonly used as a statistical technique to predict future behavior.	Regression
2.	Availability	Decentralized storage and distribution along-with web application approach make the service highly available.	IBM Cloud
3.	Train & Test split of data	The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based algorithms. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results.	Dataset
4.	Scalable Architecture	Since the application servers can be deployed on many machines. Also, the database does not make longer connections with every client – it only requires connections from a smaller number of application servers. It improves data integrity.	3-tier architecture

8. TESTING

8.1 Test Cases

ID	Test Case	Test Steps	Test Data	Expected Outcome	Outcome	Pass/Fail
TC1	Checking input criteria	1. Click on the Apply button 2. Input various fields present 3. Leave out an empty field	On the Form page: Name field is left empty	Please fill in the name	As expected	Pass
TC2	Pass Loan Scenario - Ideal	1. Input all the parameters on the form page. 2. Input the criteria in a manner where the loan amount and tenure lead to an EMI lesser than monthly EMI with 2 dependents, self-employed, married and non-graduate employee and with no property.	Name: XYZ Gender: Male Married: Yes Number of Dependents: 2 Education: Non-Graduate Self-Employed: Yes Applicant income: 2500 Co-Applicant income: 4600 Loan Amount: 176 Loan term: 360 Credit Guidelines Met: Yes Property Area: No Property	Eligible for loan	As expected	Pass
TC3	Loan Rejected	1. Input all the parameters on the form page. 2. Input the criteria in a manner where applicant is male, married with 3 dependents and a salaried employee graduate with no credit history and property in rural area.	Name: XYZ Gender: Male Married: Yes Number of Dependents: 3 Education: Graduate Self-Employed: No Applicant income: 4890 Co-Applicant income: 0 Loan Amount: 121 Loan term: 360 Credit Guidelines Met: No Property Area: Rural	Not eligible for loan	As expected	Pass

8.2 User Acceptance Testing

Multiple User profiles and loan applications are tested by the bank employees who enter the data post verification. The loans of many applicants are accepted in case their overall score is healthy and above a threshold and rejected otherwise. The application user tests various applications and here are some profiles and the outcome produced by the system. The end user, the bank loan manager, uses this system to simplify the process of screening loan applicants and has tested this system. Based on the feedback we come to a conclusion about the acceptability of our system.

Defect Analysis: This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

Resolution	Severity-1	Severity-2	Severity-3	Severity-4	Subtotal
By Design	10	4	2	3	20
Duplicate	0	2	0	0	2
External	1	3	1	0	5
Fixed	7	4	4	10	25
Not Reproduced	0	0	0	0	0
Skipped	0	0	0	0	0
Won't Fix	0	0	0	0	0
Totals	18	13	7	13	52

Test Case Analysis: This report shows the number of test cases that have passed, failed, and untested.

Section	Total Cases	Not Tested	Fail	Pass
Form submission	5	0	0	5
Wrong information in form	5	0	0	5
Security	2	0	0	2
Loan approval	10	0	0	10
Loan rejection	10	0	0	10
Final report output	4	0	0	4

9. RESULTS

9.1 Performance Metrics

```
randomForest(x_train, x_test, y_train, y_test)
```

```
Random Forest Classifier:
Confusion matrix
[[ 83  24]
 [ 20 106]]
Classification report
              precision    recall  f1-score   support

           0       0.81      0.78      0.79       107
           1       0.82      0.84      0.83       126

    accuracy                           0.81       233
   macro avg       0.81      0.81      0.81       233
weighted avg       0.81      0.81      0.81       233

score
0.8111587982832618
```

10. ADVANTAGES & DISADVANTAGES

10.1 Advantages

- **Scalable System**

Our system is built on python and uses a csv file to train and test the model. A CSV file can hold a very large number of rows and this model is easily scalable and usable for larger banks with multiple databases and users available.

- **Usability**

Our system works on a web application with a very simple and usable interface making it easy to use for anyone, even without a tech background. The usage is very easy to navigate and data is presented in a clean and usable manner to the end user.

- **Minimizes Risk for the Bank**

The model replaces a system where all decisions are made manually. In such a system the risk of processing a loan with lower repayment possibility is high thus putting the banks money at risk. Our system automates and uses the power of python and machine learning to eliminate the risk and predict if an applicant is worthy by considering multiple factors in detail.

- **Automates an otherwise tedious process**

The traditional method of loan prediction involves multiple human employees and takes a long time to process each loan application. Our system cuts down the time to a few seconds and helps automate and fastens the process manifold.

- **Avoids Bias in the Loan Disbursal System**

Traditional loan approval systems involved bias when known members with poor credit histories were approved for a loan only because of societal status. Our system eliminates that as the prediction is made solely on the basis of data without any bias.

- **Does not require high level management involvement**

Our system does not require involvement of people from higher ranks in the bank. Anyone can input data into the system and record the status which can then be reverified. This saves time for the people in the banking industry.

- **Records data which can be used to increase the accuracy of prediction later**

All the data predicted and past history of approval can be saved and later used as a training set to better train the model and make better and more accurate predictions. This also helps the bank maintain a record and also gradually progress in the accuracy department thereby reducing the risk and loss over time.

- **Automated choosing of weights of factors for enhanced accuracy**

The weights for the individual factors are chosen by the model and not by the person inputting data. This reduces human error and also increases the accuracy of the model overall.

10.2 Disadvantages

- **Emphasis on multiple features weights**

The disadvantage of this model is that it emphasizes different weights to each factor but in real life sometimes loans can be approved on the basis of a single strong factor only, which is not possible through this system.

- **Complexity**

Another disadvantage is that the system, although easy to use, the code is complex for an average user to comprehend in case of any errors.

- **Maintenance**

Updating the code to make use of new technologies and updating it in case of any errors can be considered as a disadvantage due to the complexity of the code.

11. CONCLUSION

A system that is used to predict the credibility of an applicant for loan approval was designed to eliminate the manual process which is predominant in the current scenario. The system was built using multiple machine learning algorithms and processes, web app was developed using HTML, CSS and JavaScript and the primary coding and integration was done using Python. This application can be used by banks to enter relevant information about a loan applicant and our models would make a prediction instantly about the creditworthiness of the applicant by taking into consideration a wide range of features. This system can be used by anyone owing to the simplicity and ease of access. It allows banks to automate the loan approval process and speed it up considerably to enhance the efficiency and effectiveness of loan applications in the current scenario. The application was developed after a tremendous amount of research and understanding of the current model and how there was room for massive improvement. The final system is deployed on IBM Watson and is available for use by the banks and other possible users to understand their credit worthiness.

12. FUTURE SCOPE

The system built is a very scalable and deployable model. This broadens the scope of the project considerably. The system can be easily utilized by existing banks and other loan providers to speed up their loan approval processes and also reduce the error due to human negligence and bias. The dataset used can be constantly updated to make the predictions more accurate for a stronger hold over the banking loan industry and the systems scope is endless in this respect.

13. APPENDIX

GitHub Repository:
https://github.com/IBM-EPBL/IBM-Project-13820-1659532356

Project Demo:
https://drive.google.com/file/d/1anT2f5Si_mXbudCT4VeQdnvOsSZYPPWR/view?usp=drivesdk