Faculty of Graduate Studies and Research

Introduction to Data Science

CS 842

University of Regina

Project Proposal

Title: Vehicle loan approval system

Submitted by:

Dharmik Rampariya

Course Instructor:

Alireza Manashty

Date:

August 23, 2020

Team members with Role

- **Name :** Dharmik Rampariya
- **Email:** ddr402@uregina.ca

Role :

- Discover Data
- Web-scrapping
- Data cleaning
- Creation of model
- Training the model
- Building front-end
- Integration
- Testing
- Deploying

Table of content

Index **Page No.**

1 INTRODUCTION……………………….………….……. 1

2 PROBLEM STATEMENT…………………………. 1

3 SOLUTION OVERVIEW………………………….. 2

4 DATA……………………………………….……….…. 3

5 TOOLS……………….……………….……….……… 4

6 TIMELINE…………………………………………….. 5

7 EXPECTED OUTCOMES……………………….….... 6

8 DATA ANALYTICS LIFE CYCLE……………….. 8

9 IMPLEMENTATION ………………………….……… 11

1. INTRODUCTION

In this digital era, Data Science is the field that outstands the most and is admired by all the tech-giants in the world. Data science includes the techniques and methods that can be used for getting the data, extract meaningful information, and predict the future. So data science can help build such kind of applications which allows predicting certain patterns and analysis which can be used for various purposes like business, commercial, healthcare, financial, etc.

The approval software systems work as an automated workflow that considers various factors as input and predicts approval.

2. PROBLEM STATEMENT

There are lots of loan applications received by the bank sector daily. Approval of the loan depends on various factors such as income, credit score, employment status, previous credit history, asset cost, number of active loans, etc. Hence, it could take a lot of time for a bank employee to review the application and make the decision. On the other hand, there is a possibility of human error while making a decision.

To solve this problem data science concepts and techniques can be used. It allows us to predict the approval of the loan.

3. SOLUTION OVERVIEW

Nowadays, in the banking sector, there are a bunch of applications received daily. However, going through each application manually can be a tedious task and might not be accurate enough. In this case, to simplify this process, this project uses data science and machine learning concepts and techniques to build an application that can predict whether an applicant is eligible for the loan and how much EMI he will get to pay. Hence this application can be useful for the banking sector to determine whether to approve the loan. Besides, it can be used in general by normal people to check if they stand to qualify for the loan or not.

At first, data will be web scrapped from the data source (rpubs.com) and store in .csv formate. Afterward, some data cleaning techniques will be applied for cleaning the dataset. Secondly, a machine learning model will be created using classification and regression techniques. Also, there are different classifications models like Naïve Bayes, Random forest, KNN, and Support vector machine. After that, the model will be trained over a set of data that can recognize the different types of patterns, analyze and predict the result. For training measures, cross-validation will be used. Ultimately it will provide accuracy to the model which can help to predict appropriate results.

Components of the project :

- Fully functional website
- Web scrapped dataset
- Data science problem
- Model

Website components :

- Html
- CSS
- Django

Website publisher :

- Pythonanywhere

4. DATA

- **Source:** rpubs.com
- **Data format:** CSV file
- **Access:** The data will be web-scrapped form the data source (rpubs.com) and will be stored in a CSV file. Furthermore, data cleaning will be done to achieve accuracy.
- **Data storage:** Data will be stored on a web-server.

Data will get split into two parts. Training Data and Testing data. In this case, cross-validation will be used. So a certain portion of the data will get trained and at the same time other portions of data will get tested and vice versa. By using this technique every portion of data will get trained and tested as well. This will ultimately enhance the accuracy of the model.

Here are some of the important attributes of the dataset.

Unique id	An identifier for the customers
Disbursed amount	Amount of loan disbursed
Asset cost	The total cost of an asset
Branch id	Branch where the loan was disbursed.
Supplier id	Vehicle dealer where the loan was disbursed
Date of birth	Date of birth
Employment type	Employment type of customer
Disbursed date	The date of loan disbursed

[Table 1.1: Data dictionary]

Mobile num	Mobile number of the customer
Perform CSN score	Bureau score
Pri no of accts	Total number of loans taken by the customer
Pri active accts	Total number of active loans by customer
Pri disbursed amount	Count of default accounts
Pri current balance	The total principal outstanding amount
Credit history length	Time since the first loan

[Table 1.2: Data dictionary]

5. TOOLS

Below mentioned are some of the tools that will be used through the entire building of this project.

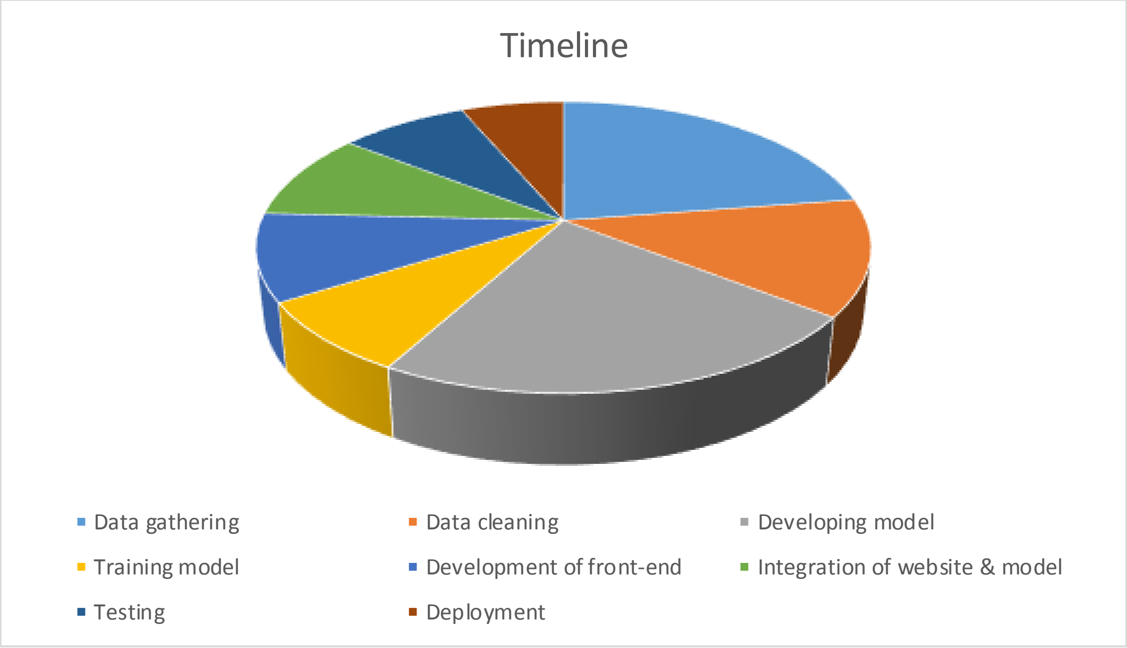
- **Python:** Python will be used for developing the web application as well as for writing script. It allows using various packages and libraries.
- **Pycharm:** For writing the code default IDE (Integrated Development Environment) by the python will be used.
- **Numpy:** For data cleaning purposes this software will be used.
- **Django:** For the development of the web application Django will be used.
- **Spyder:** I will use the Spyder IDE by Anaconda for the data analysis.
- **Matplotlib:** For Data visualization purposes Matplotlib library will be used. It allows us to represent different statistics.
- **Scikit-learn:** To build the machine learning model this library will be used.
- **PythonAnywhere:** To publish the website I will use PythonAnywhere web server.

6. TIMELINE

Here is the timeline for the project.

Task	Start Date	End Date	Duration
Data Gathering	8/6/2020	20/6/2020	12 days
Data Cleaning	23/6/2020	30/6/2020	8 days
Creating a model	3/7/2020	15/7/2020	12 days
Training model	16/7/2020	20/7/2020	5 days
Developing front-end	21/7/2020	26/7/2020	6 days
Integration of website & model	27/7/2020	1/8/2020	6 days
Testing an application	2/8/2020	6/8/2020	5 days
Deployment	7/8/2020	10/8/2020	4 days

Pie-chart :



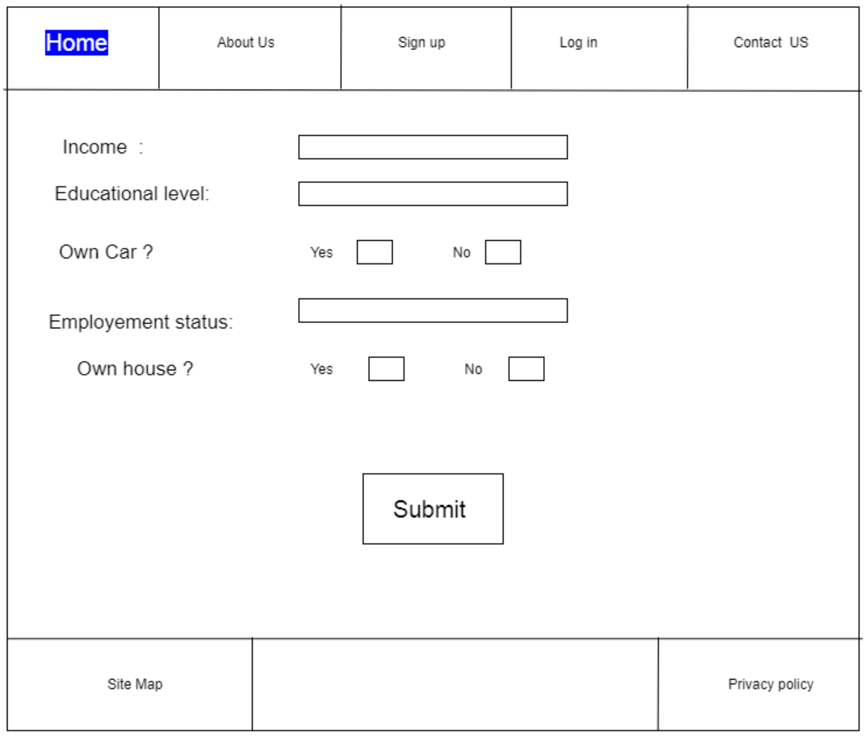
7. EXPECTED OUTCOME

After the deployment of the web applications, it can be used by the bank to determine to check the eligibility for getting a loan. The system will also predict the EMI of the loan.

Normal people can also use this application to get the basic idea of whether they are eligible for a loan or not.

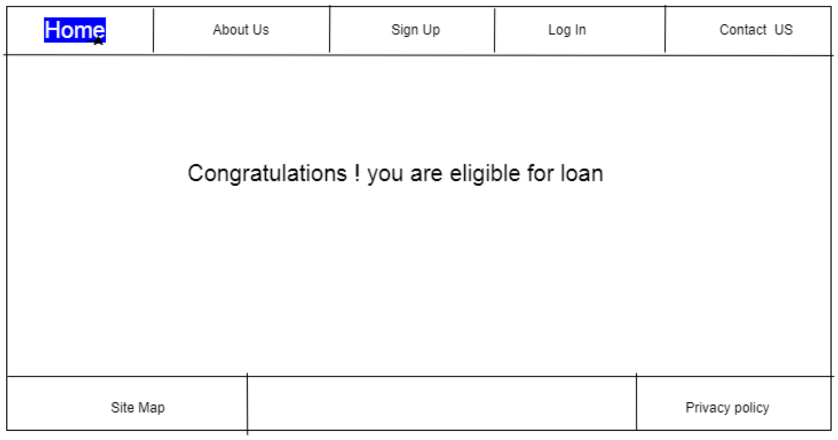
High-fidelity prototype :

1. Check eligibility



[Figure 1.1 – check eligibility]

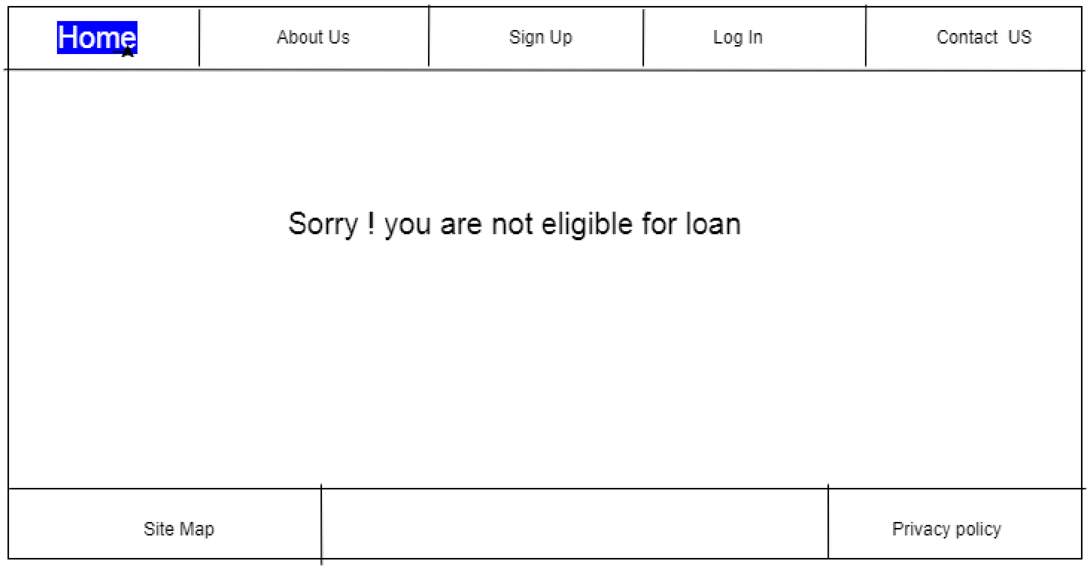
2. Approval page



Congratulations ! you are eligible for loan

[Figure 1.2 - approval page]

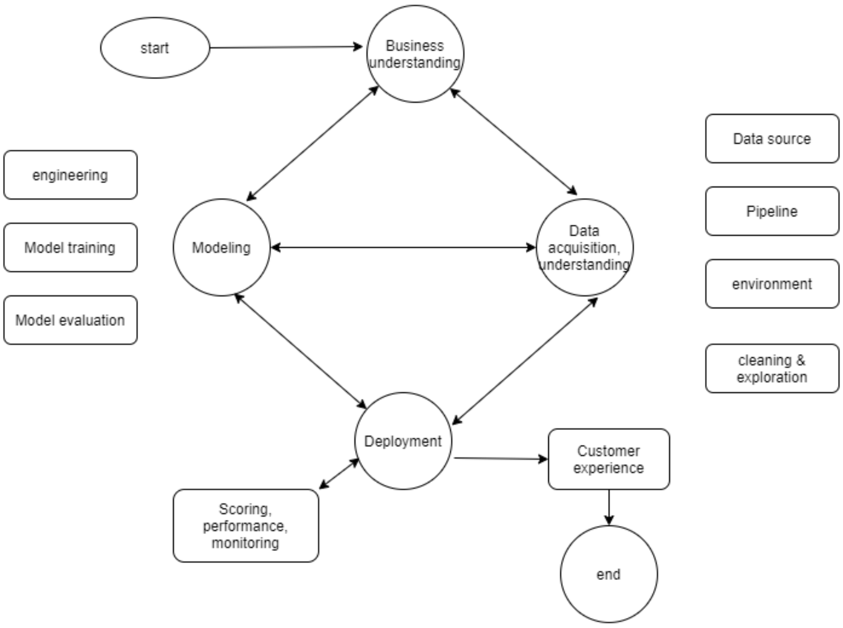
3. Decline page



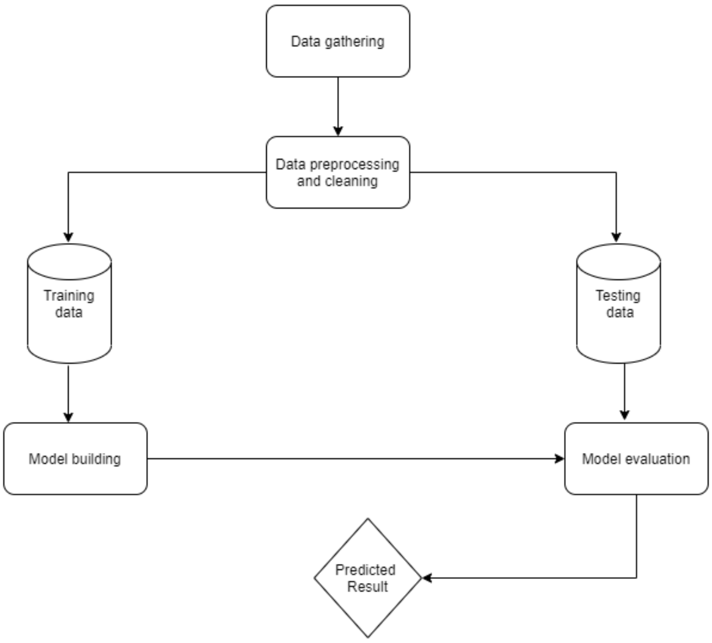
Sorry ! you are not eligible for loan

[Figure 1.3 – Decline page]

8. DATA ANALYTICS LIFE CYCLE



Data flow diagram :



Discovery :

Data discovery refers to the collection of datasets and analysis of the dataset to determine the patterns. This project requires data like education level, income, house owned, car owned, etc. So a similar type of dataset is available on Rpubs.com that is used in the project. In this dataset, there are 233156 rows and 42 features are available.

Data preparation :

Data preparation is considered a crucial step in the data analytic life cycle. It is also known as data preprocessing. Data preprocessing means transforming the raw data into a structured form that is accessible for model training. It ultimately enhances the quality of the data. It has several steps, for example, data cleaning, data integration, data transformation, etc. It removes the duplicate entries form the dataset, fills the missing values that increase the accuracy of the model.

The primary aspect of data preparation is data cleaning. It refers to fill the missing values and deleting duplicate entries. There are some options for data cleaning like using average value, the previous value, the next value, zero, or any specific number. Interpolation and regression algorithms can also be used. The ideal approach to fill the missing value is using the most frequent value.

The second phase in this process would be label encoding. Label encoding refers to transforming non-numerical value into its equivalent numerical data. It is an efficient approach for the training model. So we used label encoder from the Scikit-learn library to deal with empty values.

Model planning:

The primary purpose of the solution is to predict whether a user is eligible for a loan or not. So the classification approach will be used to solve this problem. In the given dataset there is a feature that contains labels so a supervised learning model can be used. There are several models available to use like KNN, decision tree, etc.

We will use the K nearest neighbor (KNN) model to overcome this issue. The reason for using is that it is widely used in the finance sector, banking sector, and in loan disbursement. it is simple and easy to understand.

Model building :

Every data science problem is different and requires a different solution based on a different business perspective. The goal is to design a website that can determine a person will get a loan or not. In this case, I am using a classification model K nearest neighbor model. It is a classification supervised learning algorithm. This model works on similarity as it stores all the available cases and classifies the new data using a similarity measure. 'k" in KNN refers to the number of nearest neighbors available. If the new data is added it will be chosen based on the value of k and if a majority of neighbors have similar value, that value will be assigned to the new data.

Communicate results :

It refers to analyzing the end product and considering the key finding of the project. Both stakeholders and analysts scale the output of the product whether it is successful or not.

In this phase, we have done the integration of the machine learning model and website developed in Django using python. By accessing the website user can register and check the eligibility for getting a loan by entering all the necessary details in the field. For checking the accuracy, we have tested the website by entering details of all kinds.

Operationalization :

It refers to deploying the end product in the production environment and analyzing the outcomes and changing the model if necessary.

Below is the link for GitHub where the project is deployed and along with that link for youtube to check how the system works.

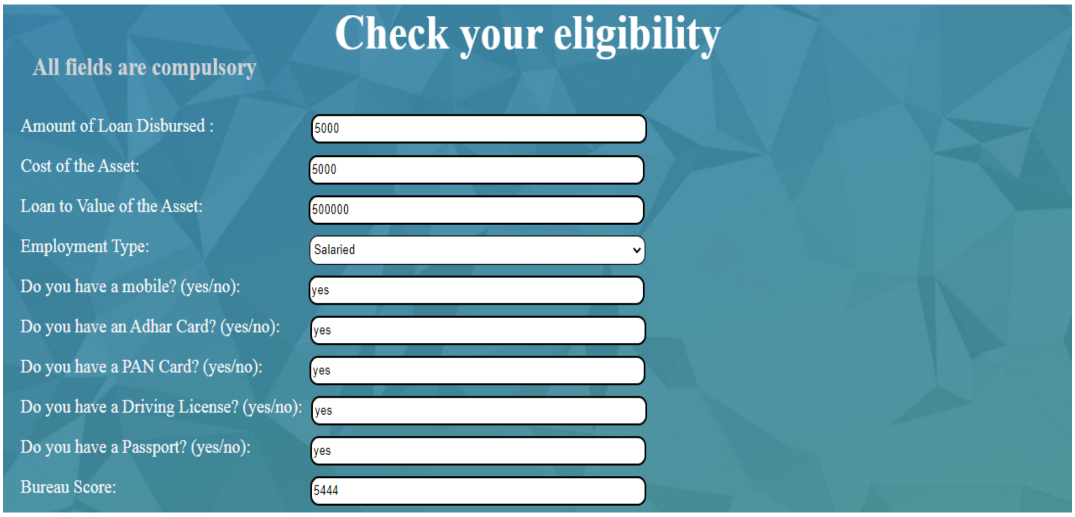
Github link : https://github.com/Dharmik003/DS

Youtube link : https://youtu.be/bgrJtT_e74I

9. IMPLEMENTATION :

Below given are the screenshots of the implementation of the project.

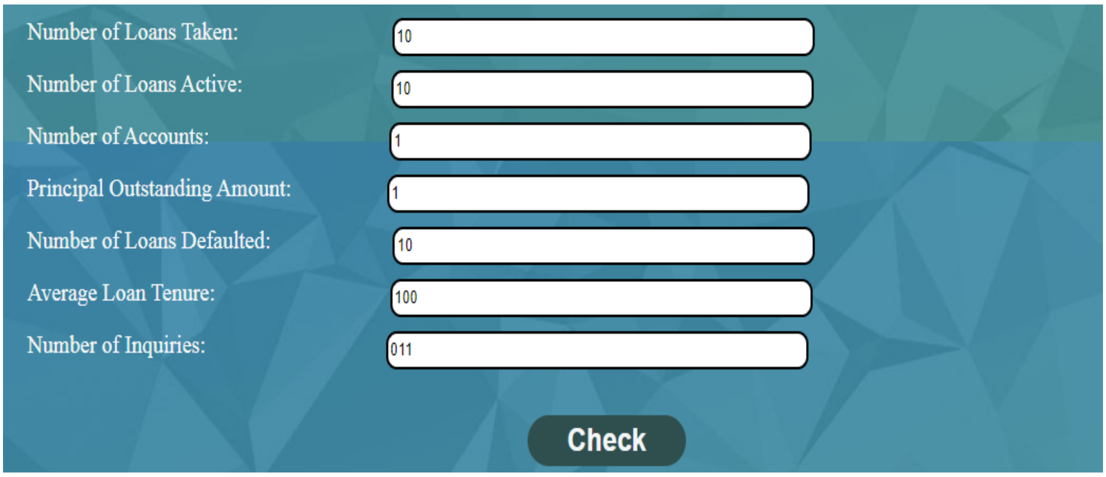
9.1 Check eligibility



Check your eligibility

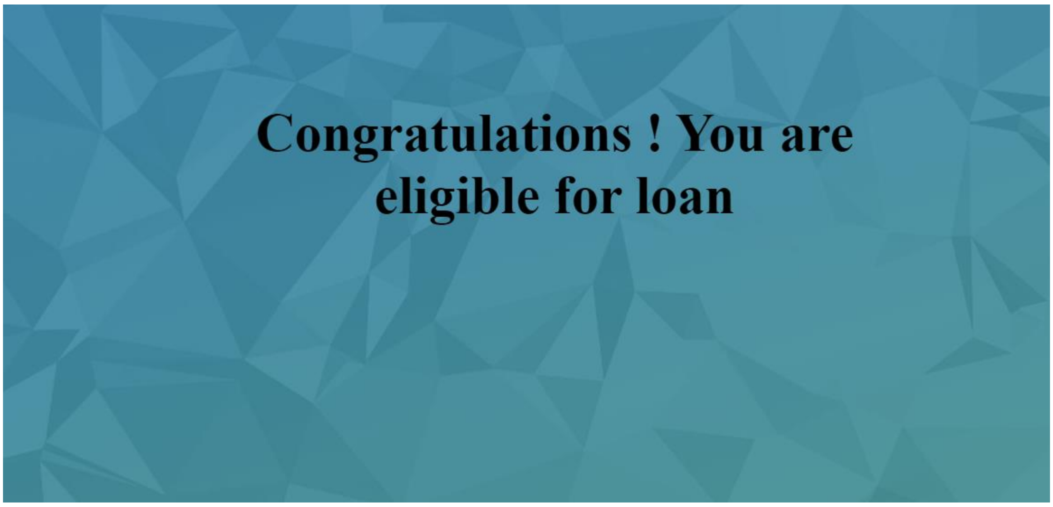
All fields are compulsory

Amount of Loan Disbursed :	5000
Cost of the Asset:	5000
Loan to Value of the Asset:	500000
Employment Type:	Salaried
Do you have a mobile? (yes/no):	yes
Do you have an Adhar Card? (yes/no):	yes
Do you have a PAN Card? (yes/no):	yes
Do you have a Driving License? (yes/no):	yes
Do you have a Passport? (yes/no):	yes
Bureau Score:	5444



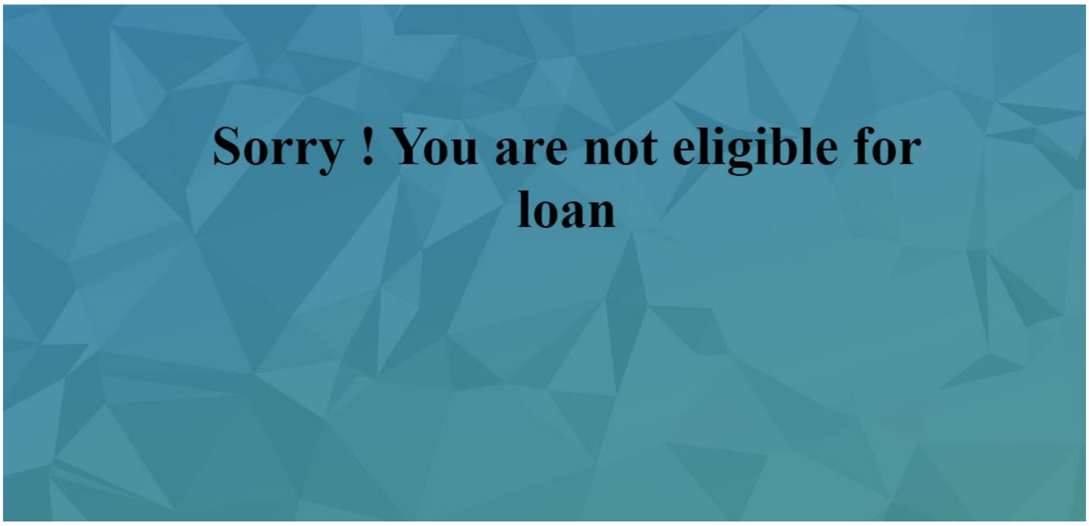
Number of Loans Taken:	10
Number of Loans Active:	10
Number of Accounts:	1
Principal Outstanding Amount:	1
Number of Loans Defaulted:	10
Average Loan Tenure:	100
Number of Inquiries:	011

9.2 Approval



Congratulations ! You are eligible for loan

9.3 Rejection



Sorry ! You are not eligible for loan