PREDICTING PERSONAL LOAN APPROVAL USING MACHINE LEARNING

INTRODUCTION

OVERVIEW

A loan is a sum of money that is borrowed and repaid over a period of time, typically with interest. There are various types of loans available to individuals and businesses, such as personal loans, mortgages, auto loans, student loans, business loans and many more.They are offered by banks, credit unions, and other financial institutions, and the terms of the loan, such as interest rate, repayment period, and fees, vary depending on the lender and the type of loan.

A personal loan is a type of unsecured loan that can be used for a variety of expenses such as home repairs, medical expenses, debt consolidation, and more. The loan amount, interest rate, and repayment period vary depending on the lender and the borrower's creditworthiness.To qualify for a personal loan, borrowers typically need to provide proof of income and have a good credit score.

PURPOSE

Predicting personal loan approval using machine learning analyses a borrower's financial data and credit history to determine the likelihood of loan approval. This can help financial institutions to make more informed decisions about which loan applications to approve and which to deny.

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result.

The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections:

(i)Data Collection

(ii) Comparison of machine learning models on collected data

(iii) Training of system on most promising model

(iv) Testing.

In this paper we are predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree and gradient boosting.

PROBLEM DEFINITION AND DESIGN THINKING

In this milestone, you will see the problem definition and design thinking using empathy map and brainstorm map with ideation.

**Specify The Business Problem**

A loan is a sum of money that is borrowed and repaid over a period of time, typically with interest. There are various types of loans available to individuals and businesses, such as personal loans, mortgages, auto loans, student loans, business loans and many more. They are offered by banks, credit unions, and other financial institutions, and the terms of the loan, such as interest rate, repayment period, and fees, vary depending on the lender and the type of loan.

A personal loan is a type of unsecured loan that can be used for a variety of expenses such as home repairs, medical expenses, debt consolidation, and more. The loan amount, interest rate, and repayment period vary depending on the lender and the borrower's creditworthiness. To qualify for a personal loan, borrowers typically need to provide proof of income and have a good credit score.

Predicting personal loan approval using machine learning analyses a borrower's financial data and credit history to determine the likelihood of loan approval. This can help financial institutions to make more informed decisions about which loan applications to approve and which to deny.

**Business Requirements**

The business requirements for a machine learning model to predict personal loan approval include the ability to accurately predict loan approval based on applicant information, Minimise the number of false positives (approved loans that default) and false negatives (rejected loans that would have been successful).Provide an explanation for the model’s decision, to comply with regulations and improve transparency.

### **Literature Survey**

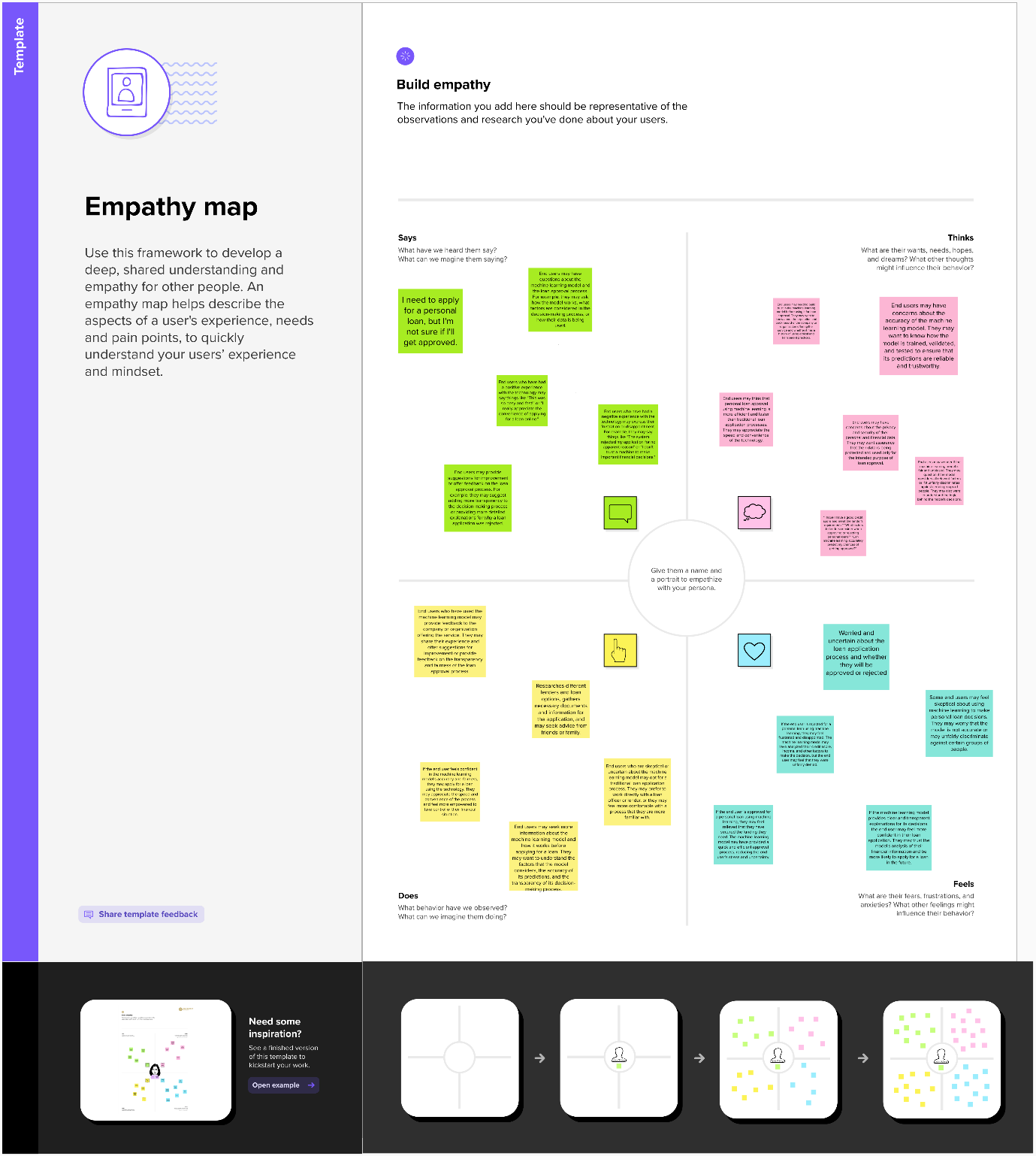
As the data is increasing daily due to digitization in the banking sector, people want to apply for loans through the internet. Machine Learning (ML), as a typical method for information investigation, has gotten more consideration increasingly. Individuals of various businesses are utilising ML calculations to take care of the issues dependent on their industry information. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high.Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications.There are various algorithms that have been used with varying levels of success. Logistic regression, decision tree, random forest, and neural networks have all been used and have been able to accurately predict loan defaults. Commonly used features in these studies include credit score, income, and employment history, sometimes also other features like age, occupation, and education level.

### **Social Or Business Impact.**

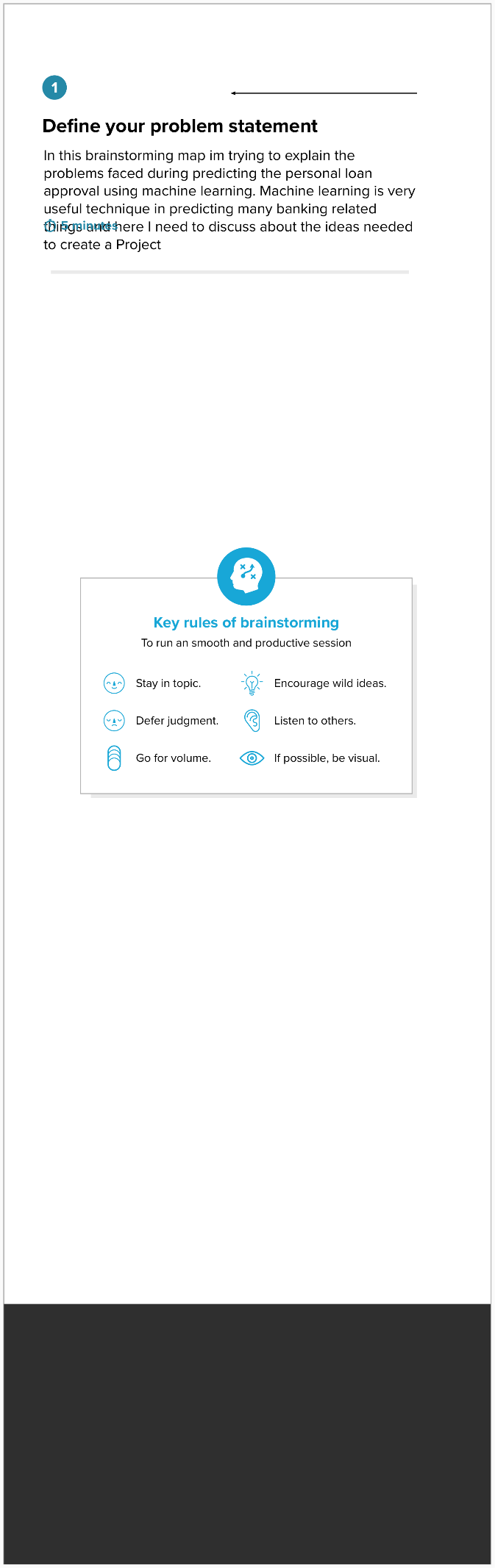
* Social Impact:- Personal loans can stimulate economic growth by providing individuals with the funds they need to make major purchases, start businesses, or invest in their education.

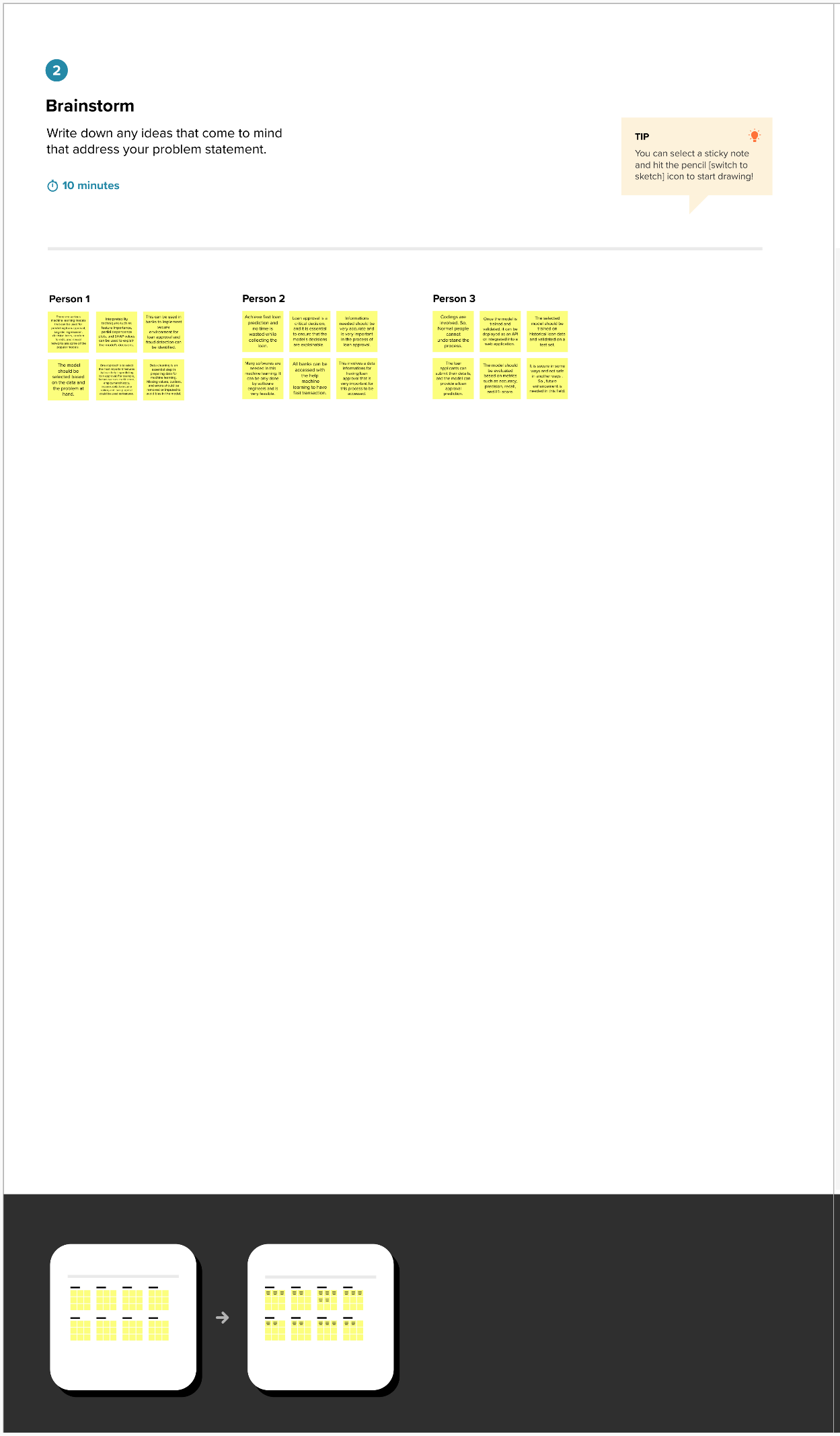
Business Model/Impact:- Personal loan providers may charge fees for services such as loan origination, processing, and late payments.Advertising the brand awareness and marketing to reach out to potential borrowers to generate revenue.

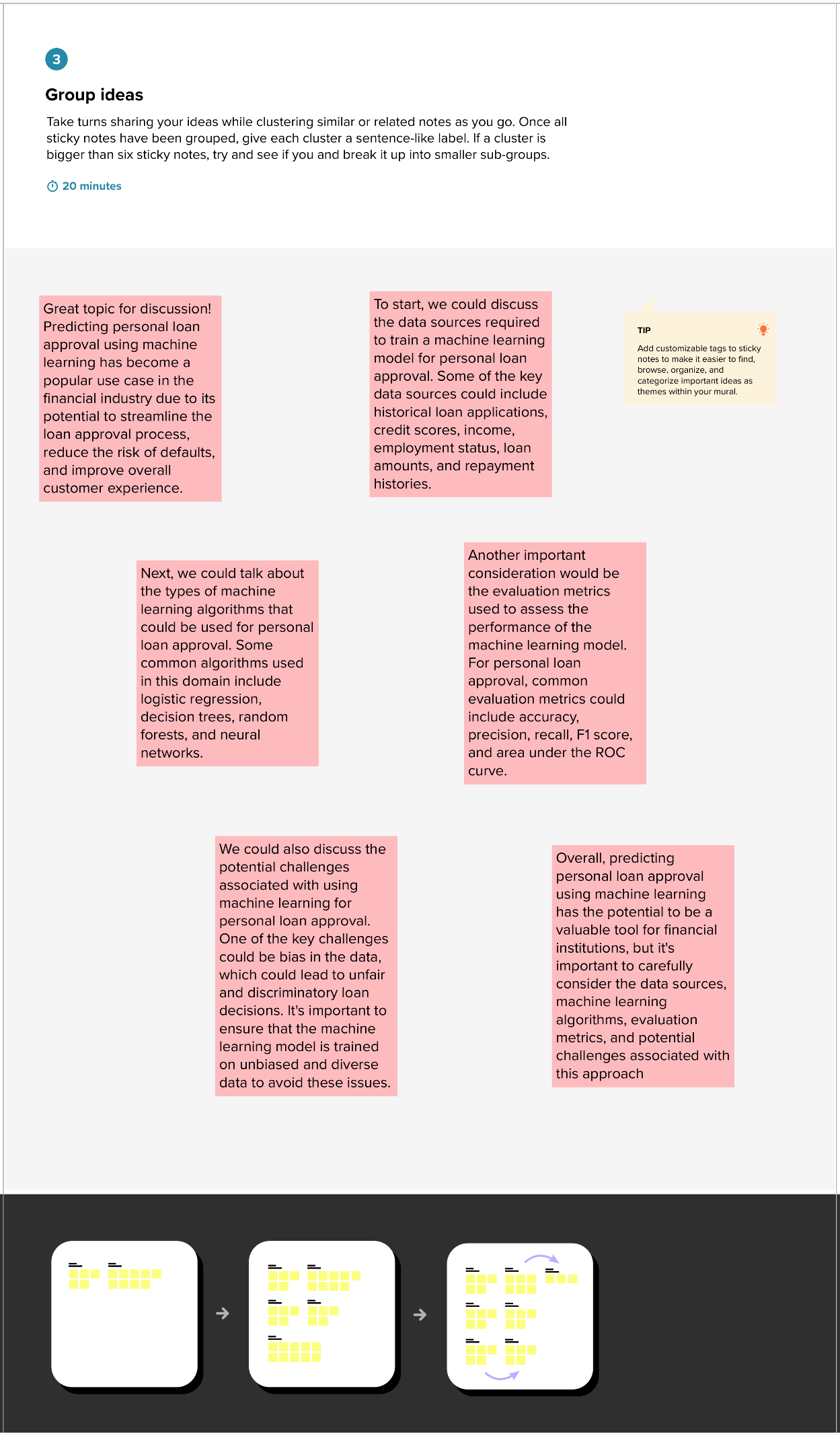
EMPATHY MAP



IDEATION AND BRAINSTORM MAP









RESULT

THE RESULTS FOR THE GIVEN PROJECT FOR PREDICTING LOAN APPROVAL USING MACHINE LEARNING IS GIVEN BELOW:

### **Collect The Dataset**

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

In this project we have used .csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset.

Link: <https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset>

As the dataset is downloaded. Let us read and understand the data properly with the help of some visualisation techniques and some analysing techniques.

### Importing The Libraries

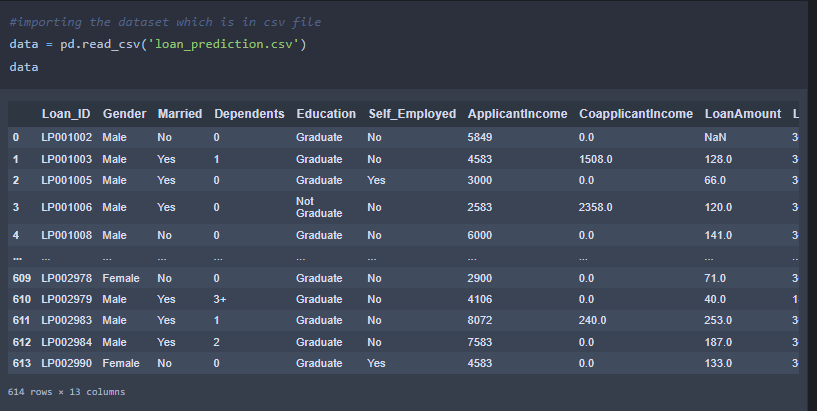
Import the necessary libraries as shown in the image. (optional) Here we have used visualisation style as five thirty eight.



**Read The Dataset**

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.

In pandas we have a function called read\_csv() to read the dataset. As a parameter we have to give the directory of the csv file.



### **Data Preparation**

As we have understood how the data is, let's pre-process the collected data.

The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

* Handling missing values
* Handling categorical data
* Handling Imbalance Data

### **Handling Missing Values**

* Let’s find the shape of our dataset first. To find the shape of our data, the df.shape method is used. To find the data type, df.info() function is used.
* For checking the null values, df.isnull() function is used. To sum those null values we use .sum() function. From the below image we found that there are no null values present in our dataset. So we can skip handling the missing values step.
* From the above code of analysis, we can infer that columns such as gender , married, dependents, self employed ,loan amount, loan amount term and credit history are having the missing values, we need to treat them in a required way.
* We will fill in the missing values in the numeric data type using the mean value of that particular column and categorical data type using the most repeated value.

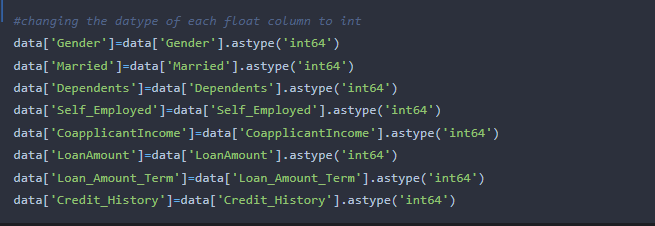
### **Handling Categorical Values**

As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding.

To convert the categorical features into numerical features we use encoding techniques.

There are several techniques but in our project we are using manual encoding with the help of list comprehension.

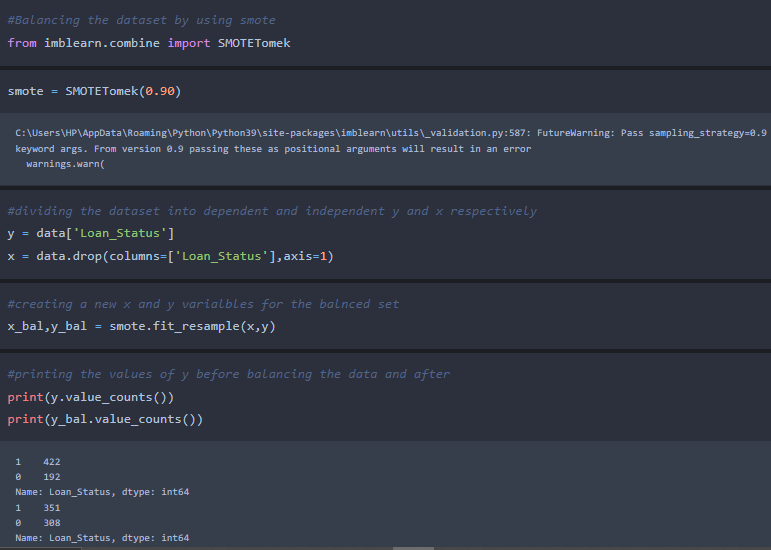
In our project, Gender ,married,dependents,self-employed,co-applicants income,loan amount ,loan amount term, credit history With list comprehension encoding is done.



### **Handling Imbalance Data**

Data Balancing is one of the most important step, which need to be performed for classification models, because when we train our model on imbalanced dataset ,we will get biassed results, which means our model is able to predict only one class element

For Balancing the data we are using the SMOTE Method.  
SMOTE: Synthetic minority over sampling technique, which will create new synthetic data points for under class as per the requirements given by us using KNN method.



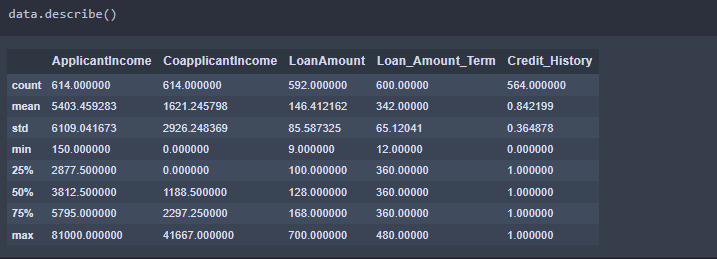
From the above picture, we can infer that ,previously our dataset had 492 class 1, and 192 class  items, after applying smote technique on the dataset the size has been changed for minority class.

**Exploratory Data Analysis**

In this milestone, we will see the exploratory data analysis

### **Descriptive Statistical**

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.



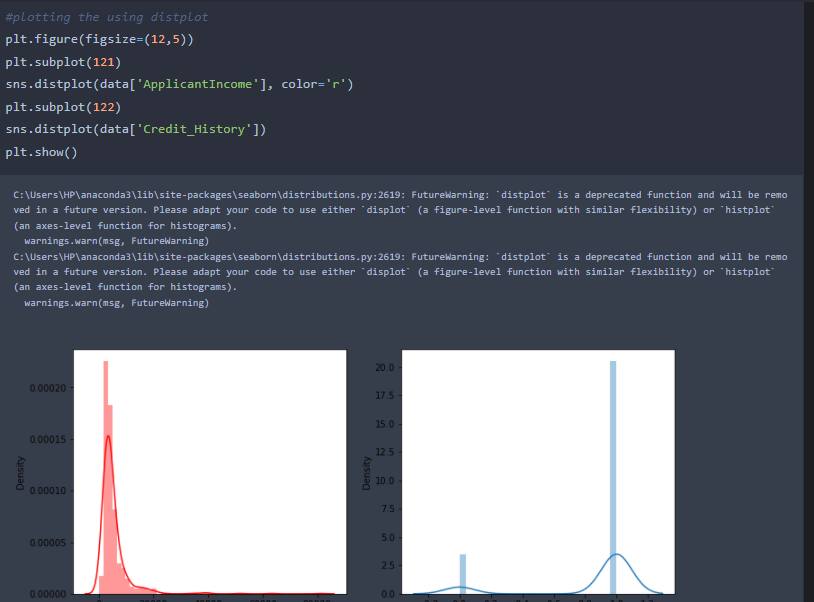
### **Visual Analysis**

Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

### **Univariate Analysis**

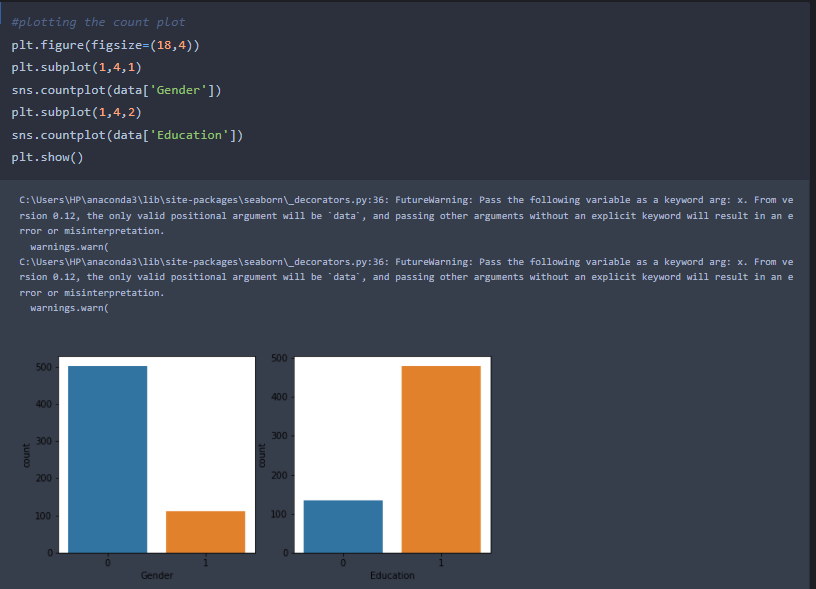
In simple words, univariate analysis is understanding the data with a single feature. Here we have displayed two different graphs such as distplot and countplot.

* The Seaborn package provides a wonderful function distplot. With the help of distplot, we can find the distribution of the feature. To make multiple graphs in a single plot, we use subplot.



* In our dataset we have some categorical features. With the count plot function, we are going to count the unique category in those features. We have created a dummy data frame with categorical features. With for loop and subplot we have plotted this below graph.
* From the plot we came to know, Applicants income is skewed towards left side, where as credit history is categorical with 1.0 and 0.0

### **Bivariate Analysis**





From the above graph we can infer the analysis such as

* Segmenting the gender column and married column based on bar graphs
* Segmenting the Education and Self-employed based on bar graphs ,for drawing insights such as educated people are employed.
* Loan amount term based on the property area of a person holding.

### **Multivariate Analysis**

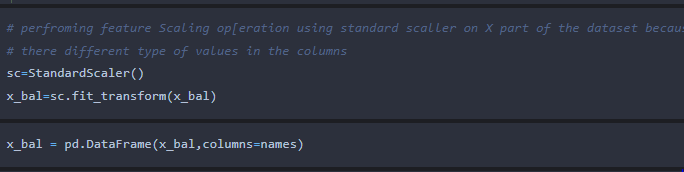
In simple words, multivariate analysis is to find the relation between multiple features. Here we have used a swarm plot from the seaborn package.



From the above graph we are plotting the relationship between the Gender, applicants income and loan status of the person.

Now, the code would be normalising the data by scaling it to have a similar range of values, and then splitting that data into a training set and a test set for training the model and testing its performance, respectively.

**Scaling the Data**  
Scaling is one the important process, we have to perform on the dataset, because of data measures in different ranges can leads to mislead in prediction.Models such as KNN, Logistic regression need scaled data, as they follow distance based method and Gradient Descent concept.

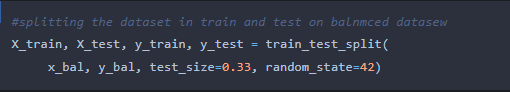


We will perform scaling only on the input values.Once the dataset is scaled, it will be converted into an array and we need to convert it back to a dataframe.

**Splitting data into train and test**Now let’s split the Dataset into train and test sets

Changes: first split the dataset into x and y and then split the data set.

Here x and y variables are created. On x variable, df is passed with dropping the target variable. And on y target variable is passed. For splitting training and testing data we are using the train\_test\_split() function from sklearn. As parameters, we are passing x, y, test\_size, random\_state.

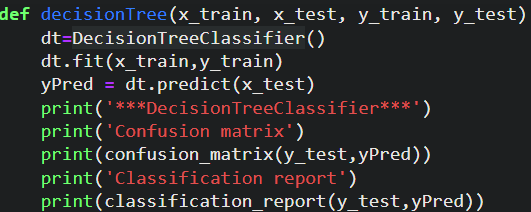


### **Training The Model In Multiple Algorithms**

Now our data is cleaned and it’s time to build the model. We can train our data on different algorithms. For this project we are applying four  classification algorithms. The best model is saved based on its performance.

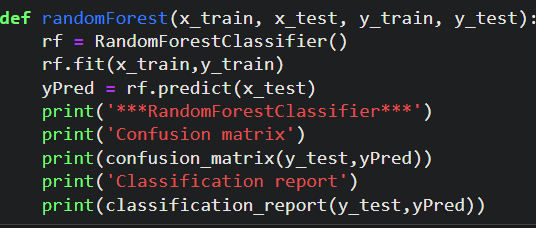
### **Decision Tree Model**

A function named decisionTree is created and train and test data are passed as the parameters. Inside the function, DecisionTreeClassifier algorithm is initialised and training data is passed to the model with the .fit() function. Test data is predicted with .predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.



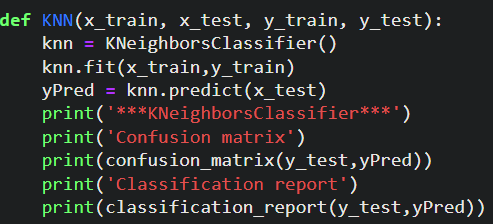
### **Random Forest Model**

A function named randomForest is created and train and test data are passed as the parameters. Inside the function, RandomForestClassifier algorithm is initialised and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.



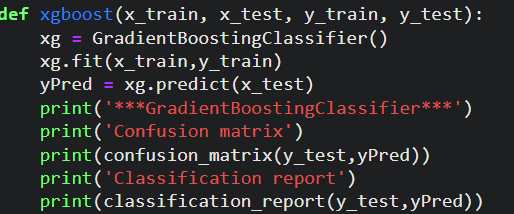
### **KNN Model**

A function named KNN is created and train and test data are passed as the parameters. Inside the function, KNeighborsClassifier algorithm is initialised and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.



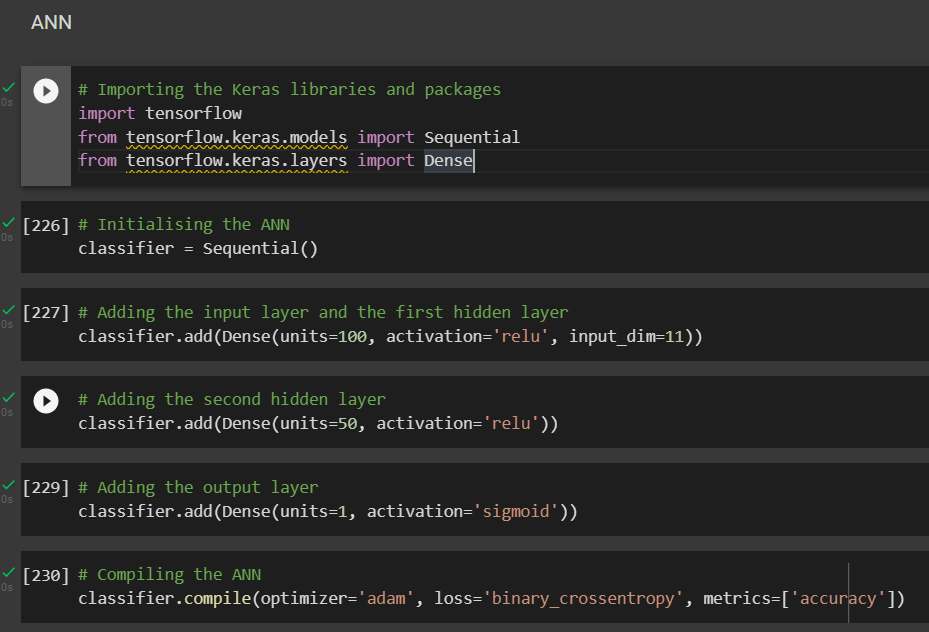
### **Xgboost Model**

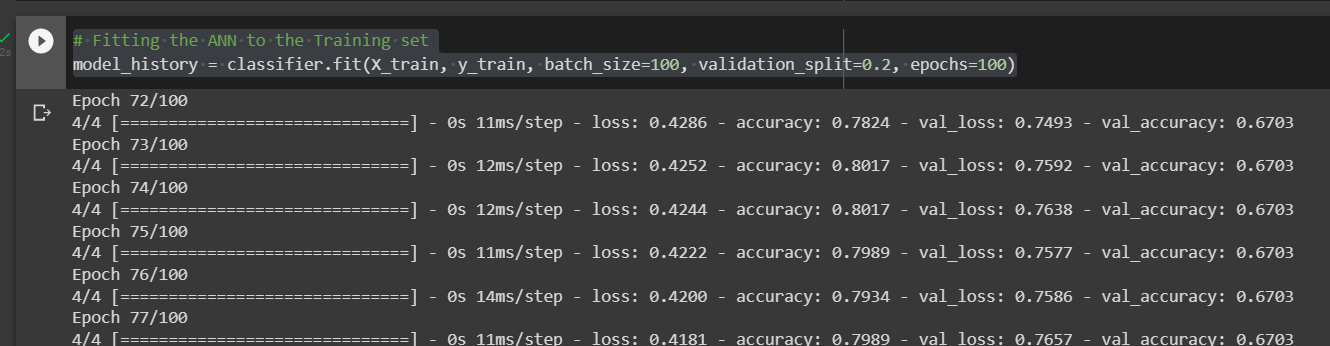
A function named xgboost is created and train and test data are passed as the parameters. Inside the function, GradientBoostingClassifier algorithm is initialised and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, confusion matrix and classification report is done.

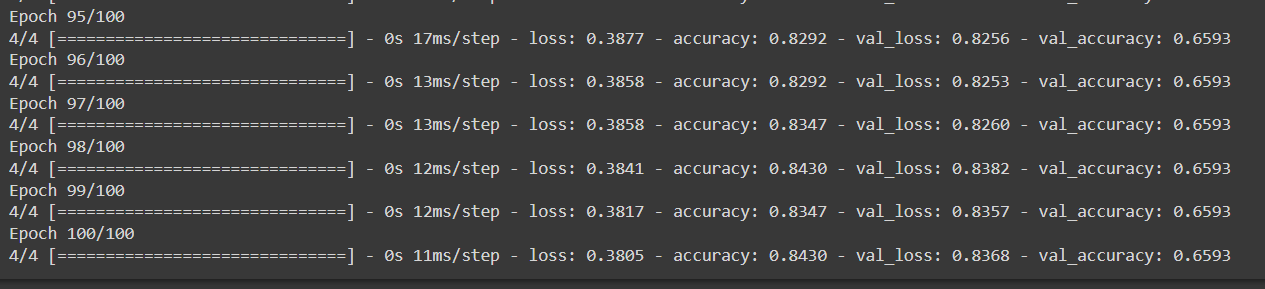


### **ANN Model**

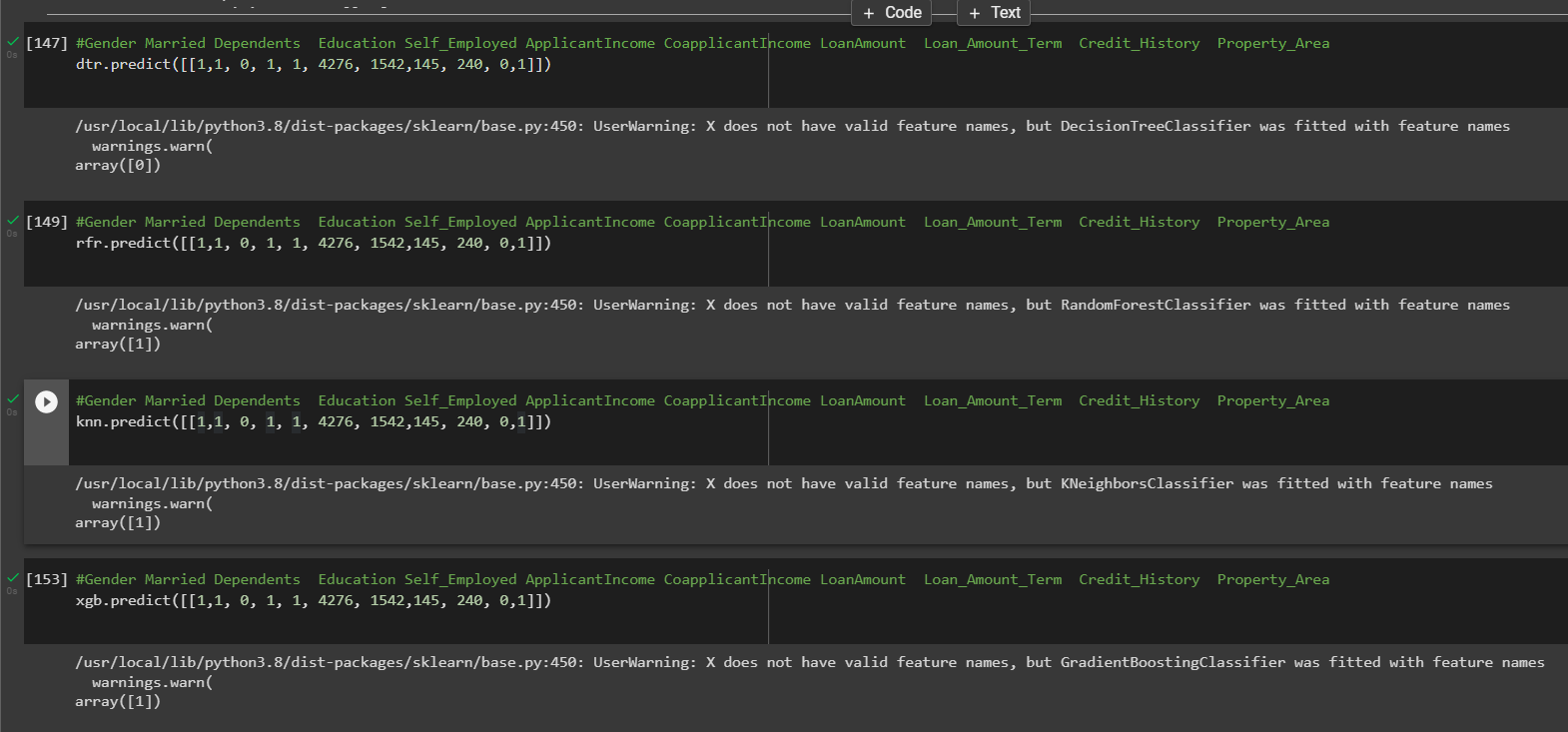
Building and training an Artificial Neural Network (ANN) using the Keras library with TensorFlow as the backend. The ANN is initialised as an instance of the Sequential class, which is a linear stack of layers. Then, the input layer and two hidden layers are added to the model using the Dense class, where the number of units and activation function are specified. The output layer is also added using the Dense class with a sigmoid activation function. The model is then compiled with the Adam optimizer, binary cross-entropy loss function, and accuracy metric. Finally, the model is fit to the training data with a batch size of 100, 20% validation split, and 100 epochs.



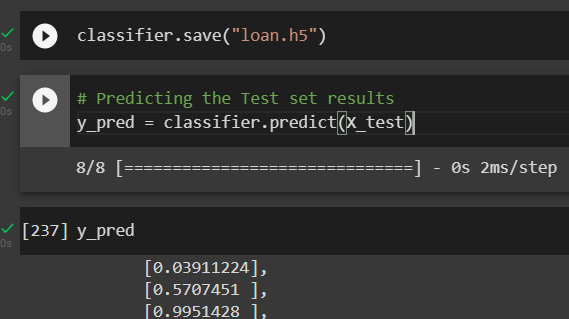


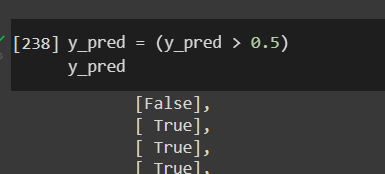


### **Testing The Model**

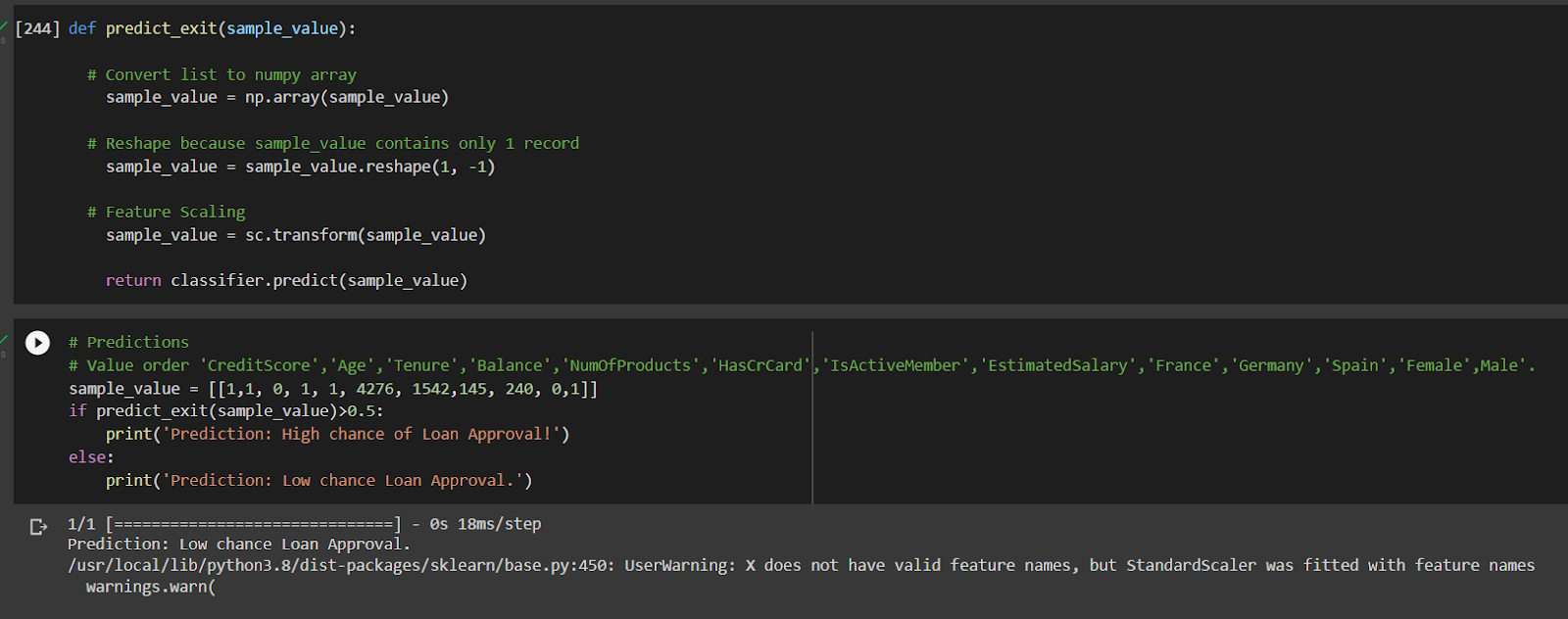
****

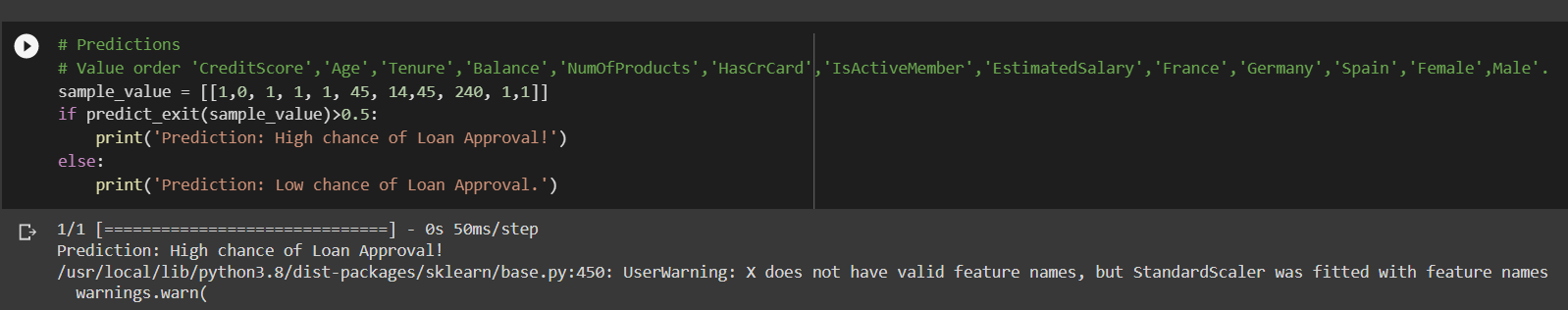
In ANN we first have to save the model to the test the inputs

****

****

This code defines a function named "predict\_exit" which takes in a sample\_value as an input. The function then converts the input sample\_value from a list to a numpy array. It reshapes the sample\_value array as it contains only one record. Then, it applies feature scaling to the reshaped sample\_value array using a scaler object 'sc' that should have been previously defined and fitted. Finally, the function returns the prediction of the classifier on the scaled sample\_value.

****

****

### **Model Building**

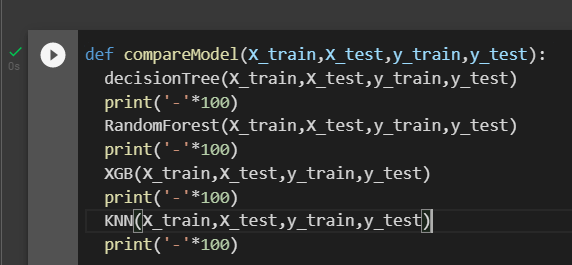
In this milestone, We will see the model building.

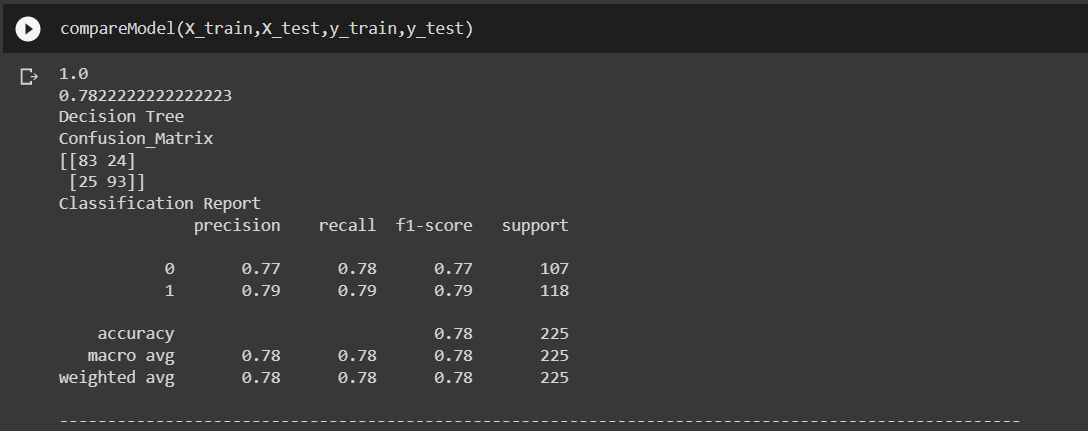
### **Testing Model With Multiple Evaluation Metrics**

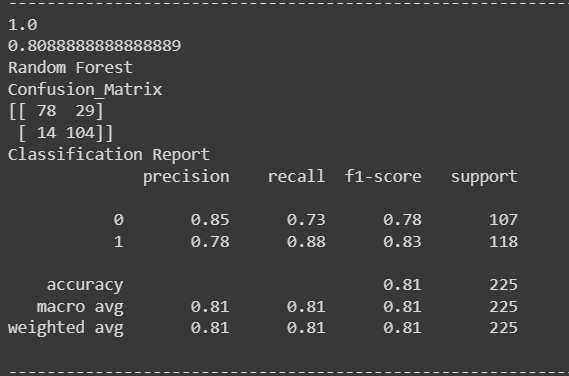
Multiple evaluation metrics means evaluating the model's performance on a test set using different performance measures. This can provide a more comprehensive understanding of the model's strengths and weaknesses. We are using evaluation metrics for classification tasks including accuracy, precision, recall, support and F1-score.

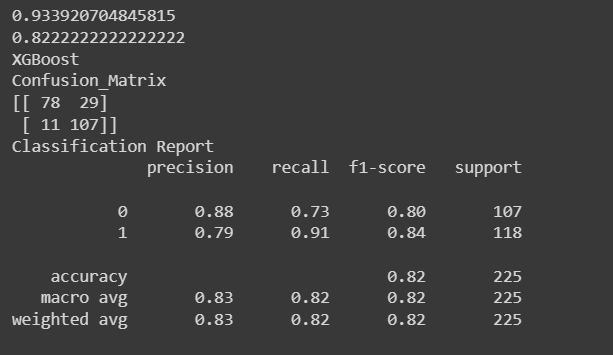
### **Compare The Model**

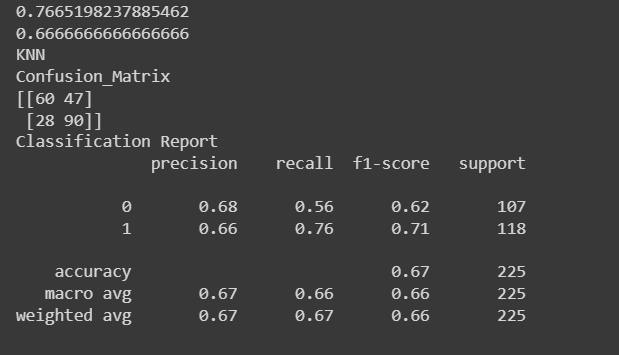
For comparing the above four models, the compareModel function is defined.

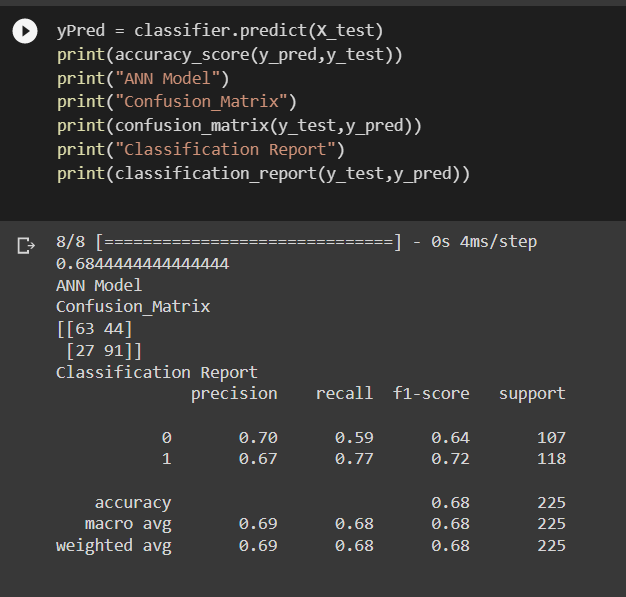










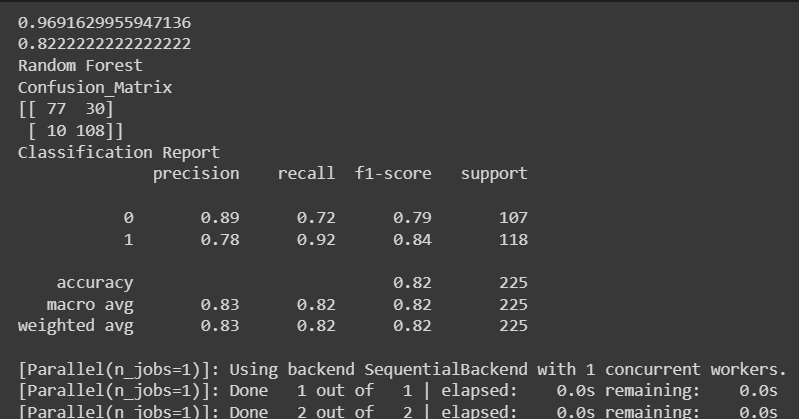


After calling the function, the results of models are displayed as output. From the five models Xgboost is performing well. From the below image, We can see the accuracy of the model. Xgboost is giving the accuracy of 93.39% with training data , 82.2% accuracy for the testing data.

### **Comparing Model Accuracy Before & After Applying Hyperparameter Tuning**

Evaluating performance of the modelFrom sklearn, cross\_val\_score is used to evaluate the score of the model. On the parameters, we have given rf (model name), x, y, cv (as 5 folds). Our model is performing well. So, we are saving the model by pickle.dump().



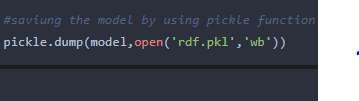


### **Performance Testing & Hyperparameter Tuning**

In this milestone, we will see the performance testing and hyperparameter Tuning

**Save The Best Model**

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its weights and configuration. This can be useful in avoiding the need to retrain the model every time it is needed and also to be able to use it in the future.



### **Integrate With Web Framework**

In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.

This section has the following tasks

* Building HTML Pages
* Building server side script
* Run the web application

### **Building Html Pages**

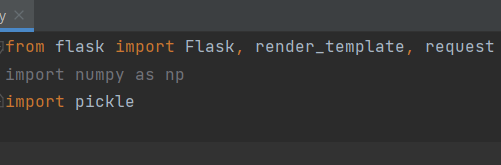
For this project create two HTML files namely

* home.html
* predict.html

and save them in the templates folder.

### **Build Python Code**

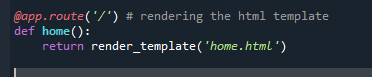
Import the libraries

****

Load the saved model. Importing the flask module in the project is mandatory. An object of Flask class is our WSGI application. Flask constructor takes the name of the current module (\_\_name\_\_) as argument.



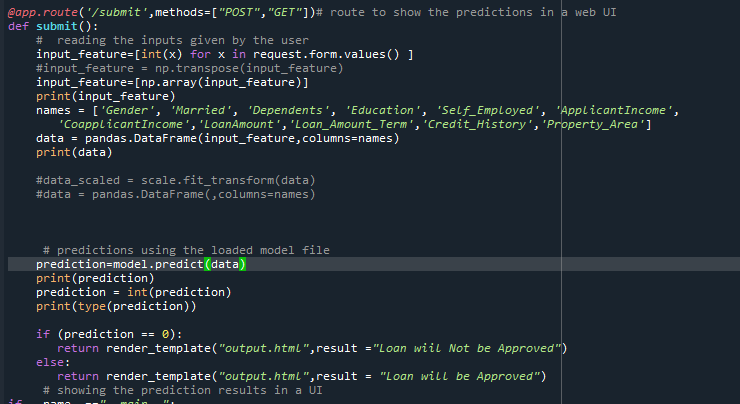
**Render HTML page:**



**Here we will be using a declared constructor to route to the HTML page which we have created earlier.**

In the above example, ‘/’ URL is bound with the home.html function. Hence, when the home page of the web server is opened in the browser, the html page will be rendered. Whenever you enter the values from the html page the values can be retrieved using POST Method.

Retrieves the value from UI:



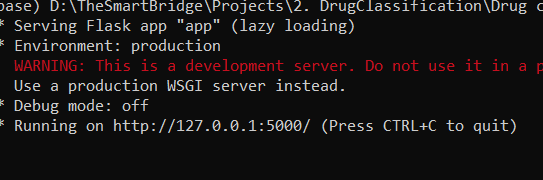
Here we are routing our app to predict() function. This function retrieves all the values from the HTML page using Post request. That is stored in an array. This array is passed to the model.predict() function. This function returns the prediction. And this prediction value will be rendered to the text that we have mentioned in the submit.html page earlier.

**Main Function:**

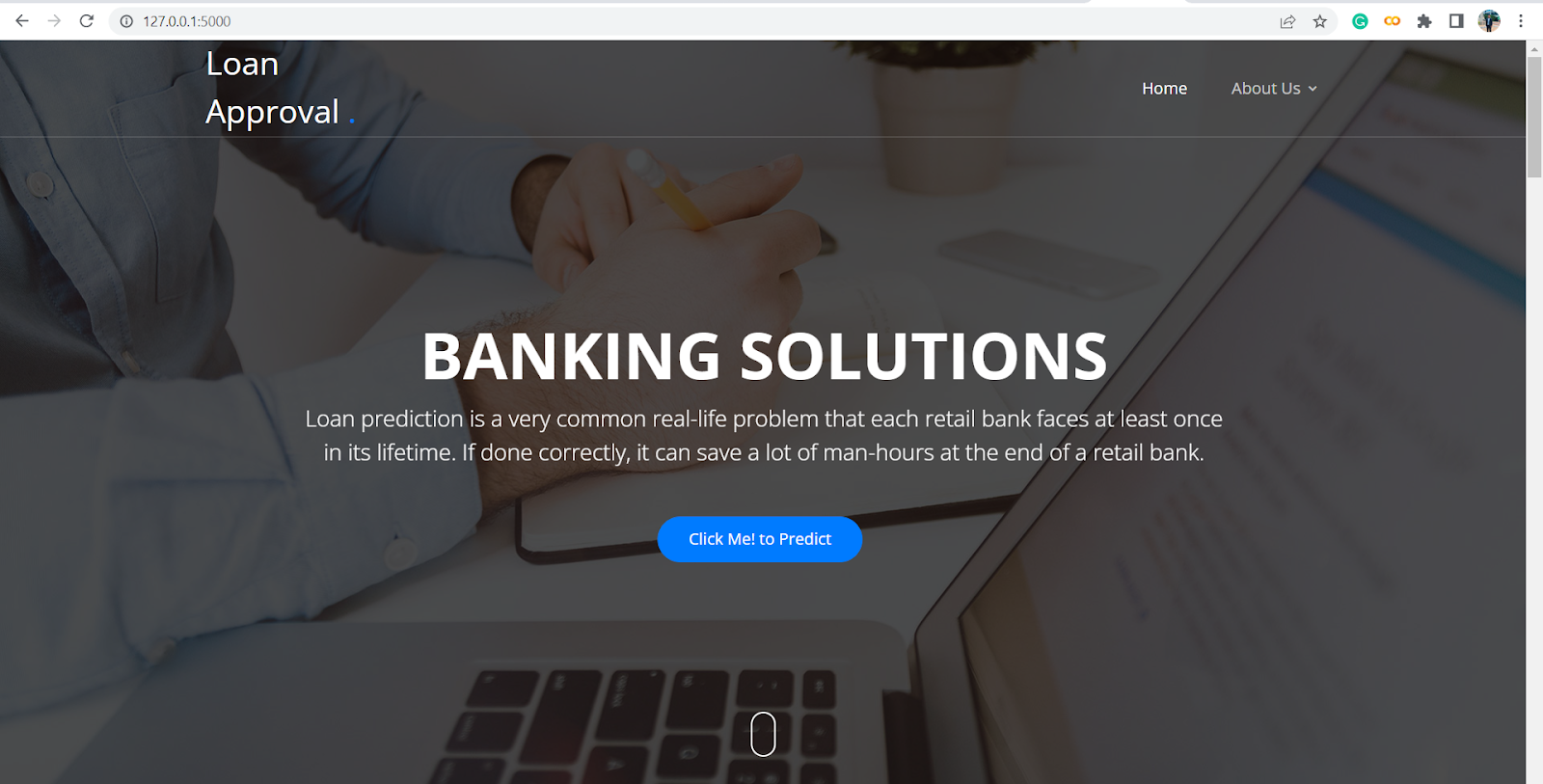


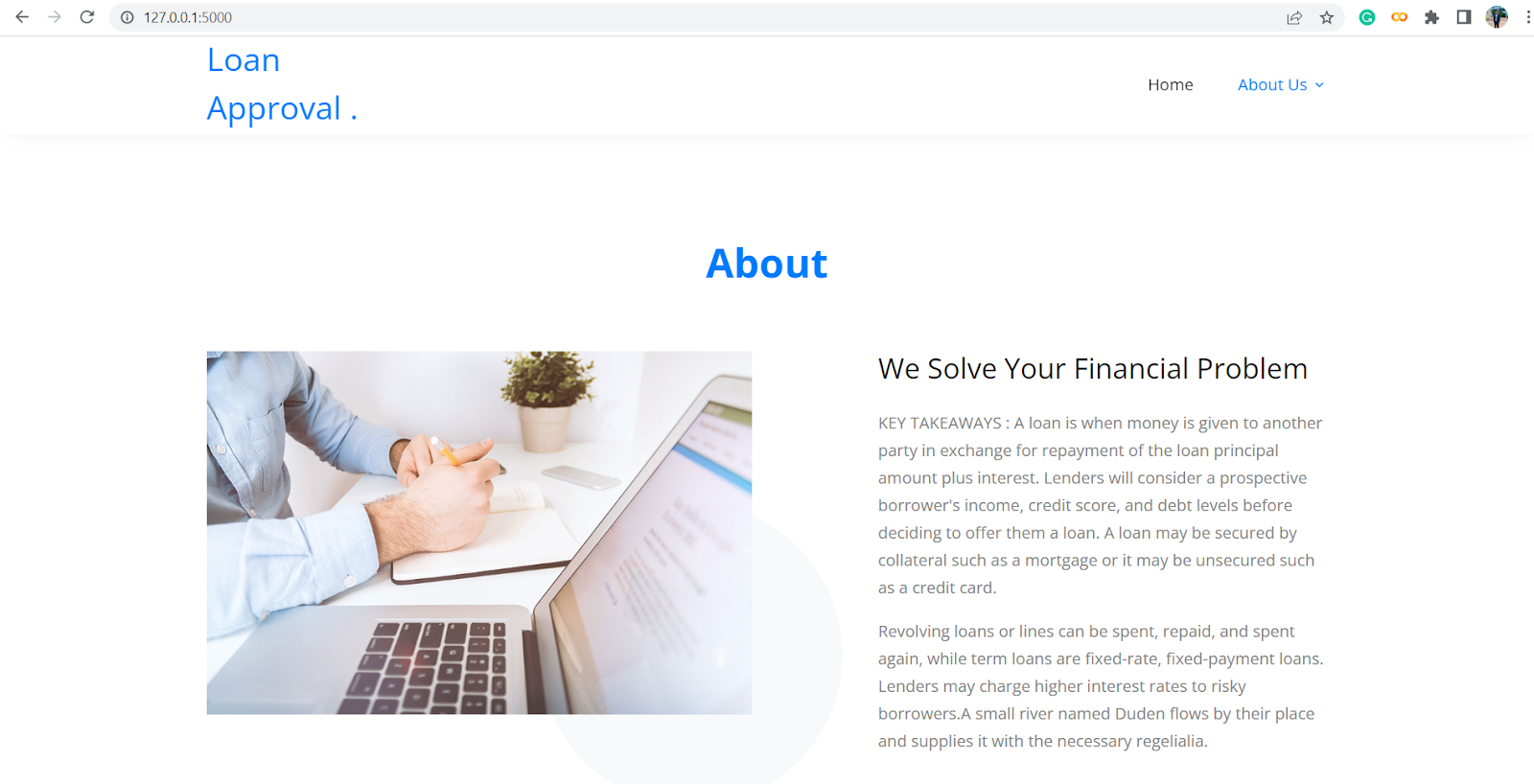
### **Run The Web Application**

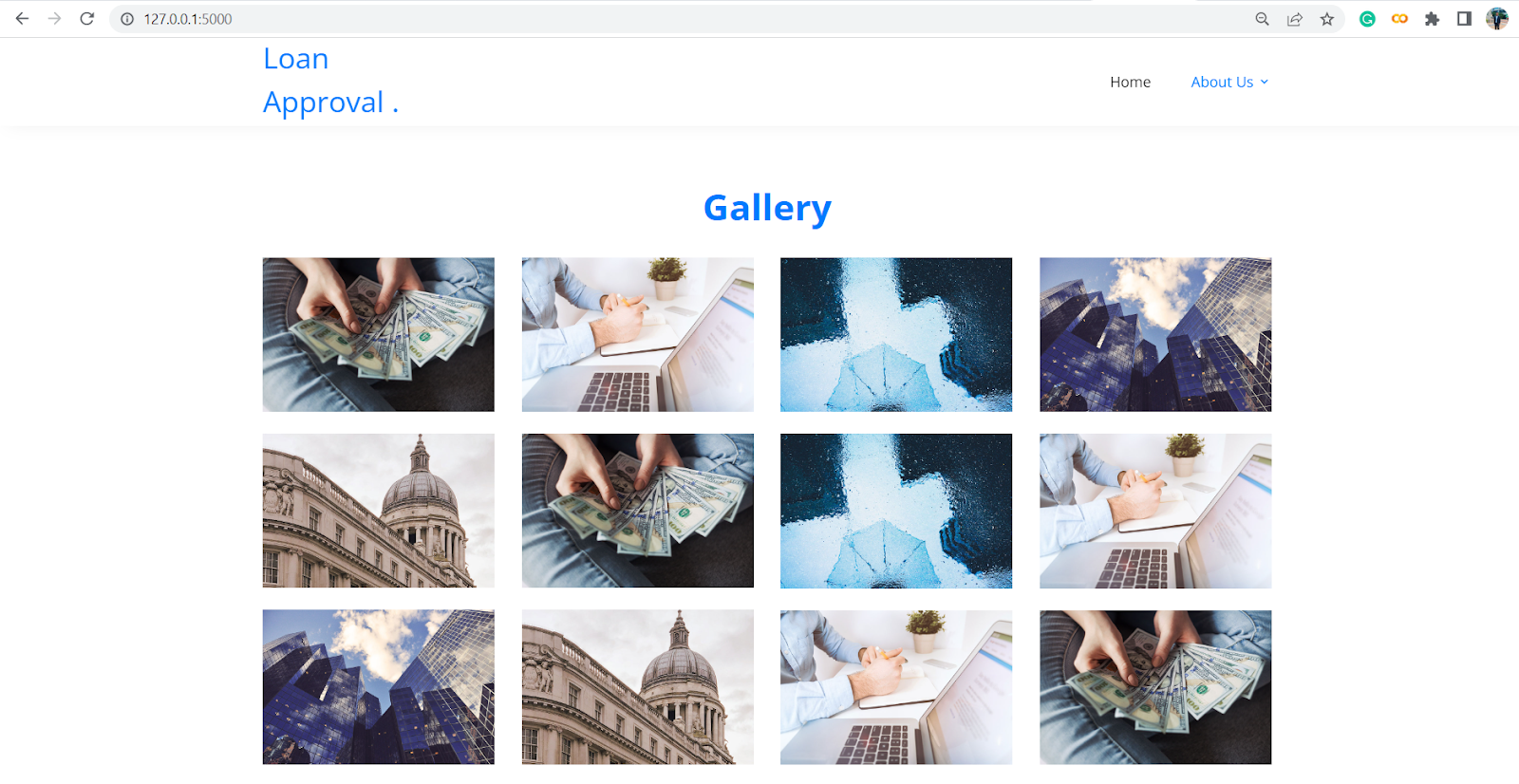
* Open anaconda prompt from the start menu
* avigate to the folder where your python script is.
* Now type “python app.py” command
* Navigate to the localhost where you can view your web page.
* Click on the predict button from the top left corner, enter the inputs, click on the submit button, and see the result/prediction on the web.

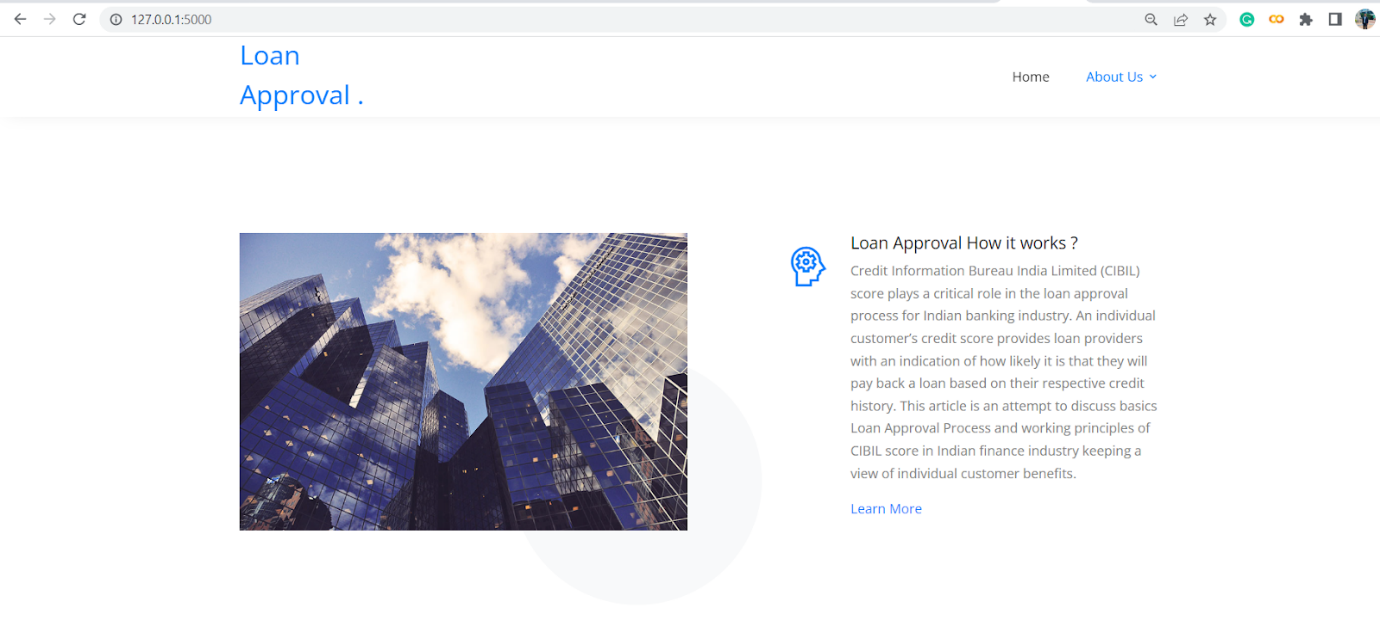


Now,Go the web browser and write the localhost url (http://127.0.0.1:5000) to get the below result

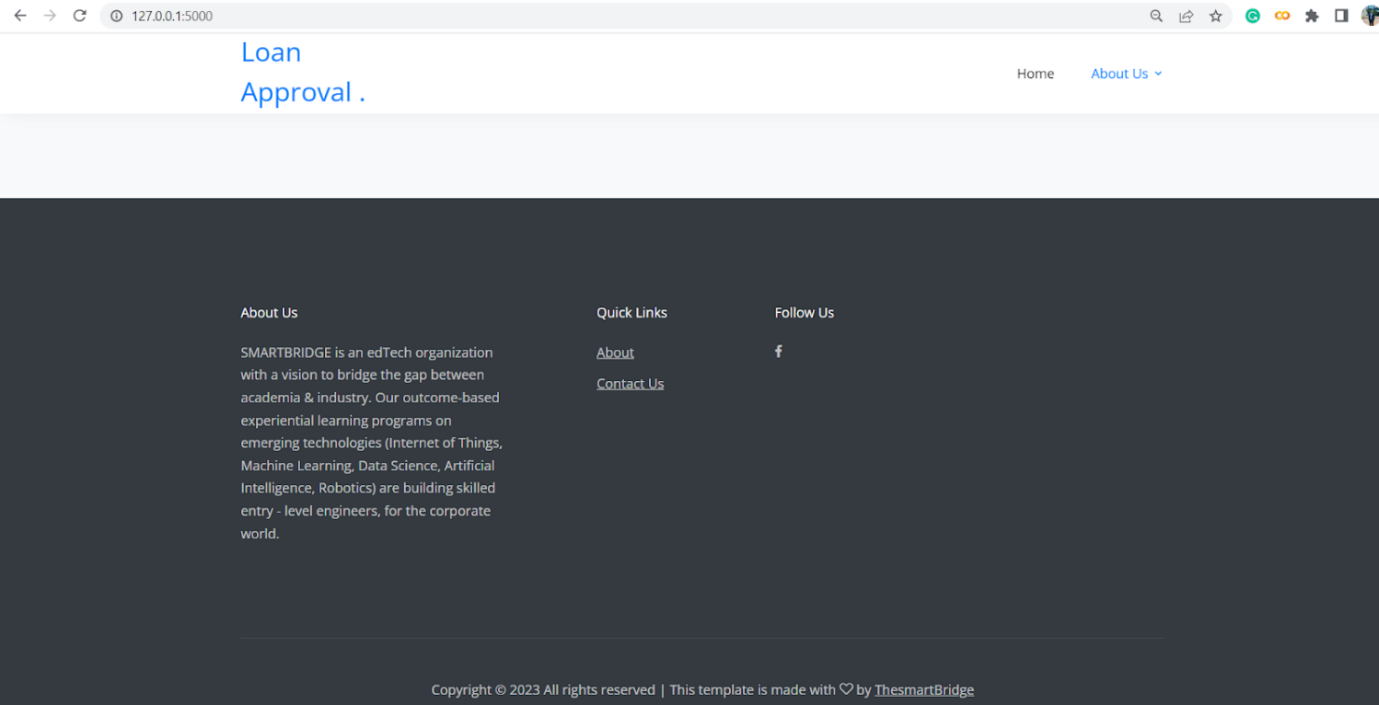




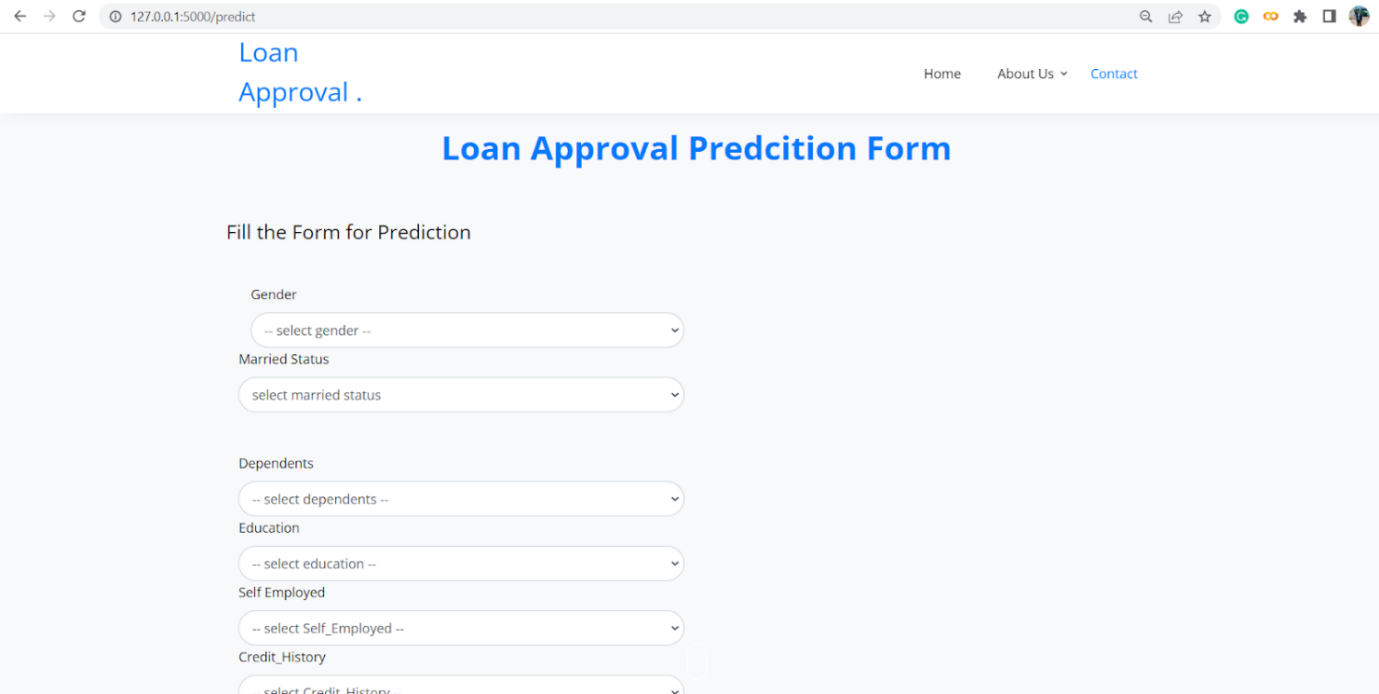


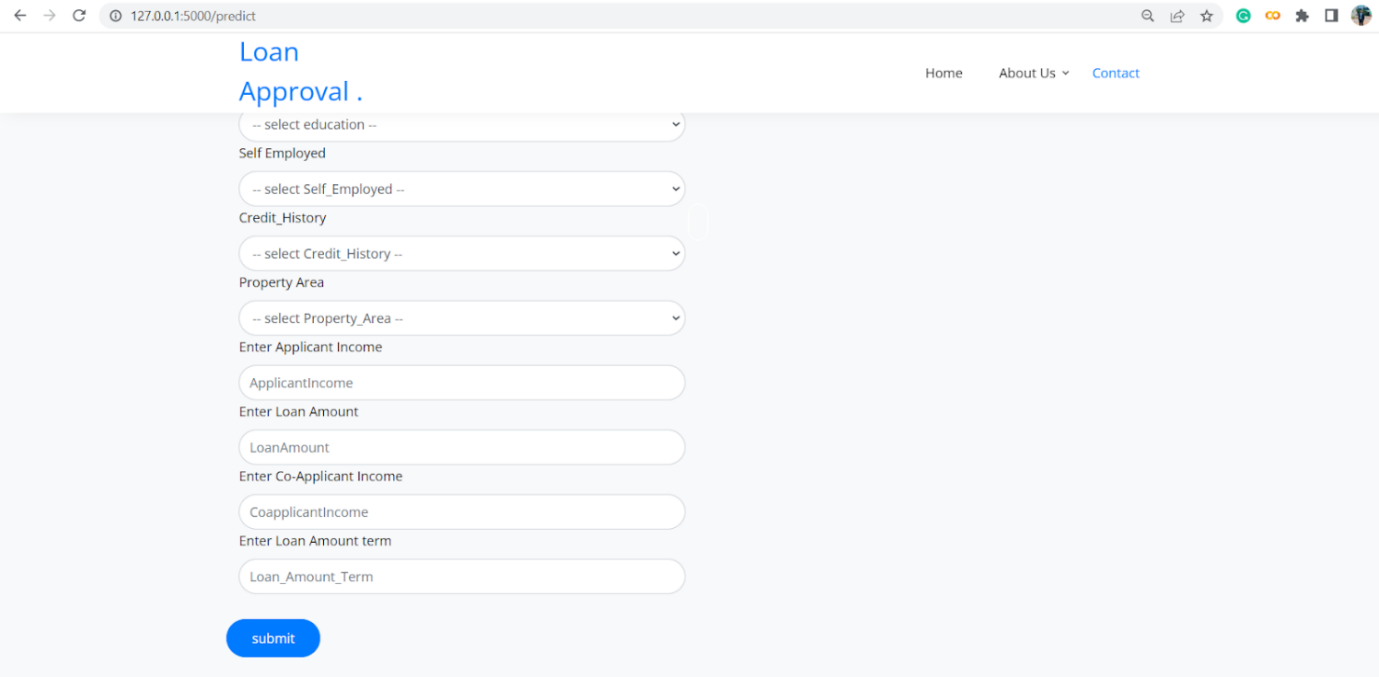




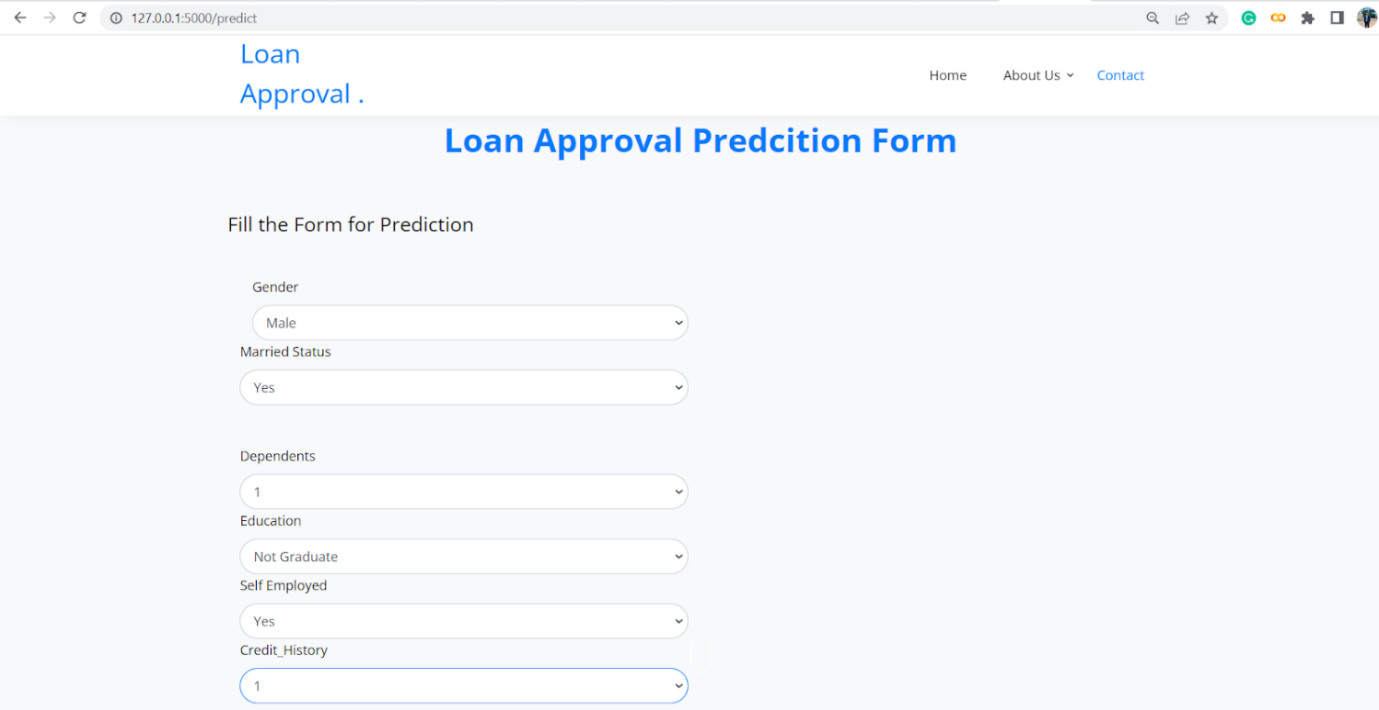


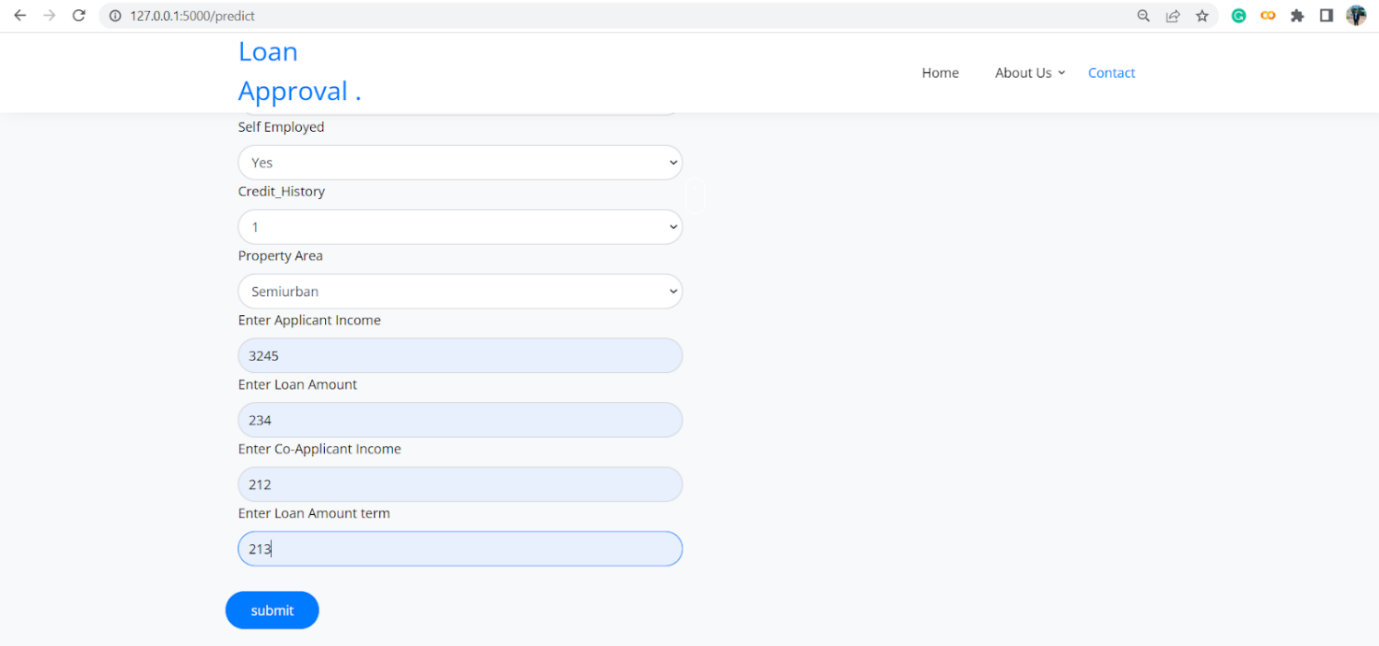
Now,when you click on click me to predict the button from the banner you will get redirected to the prediction page.



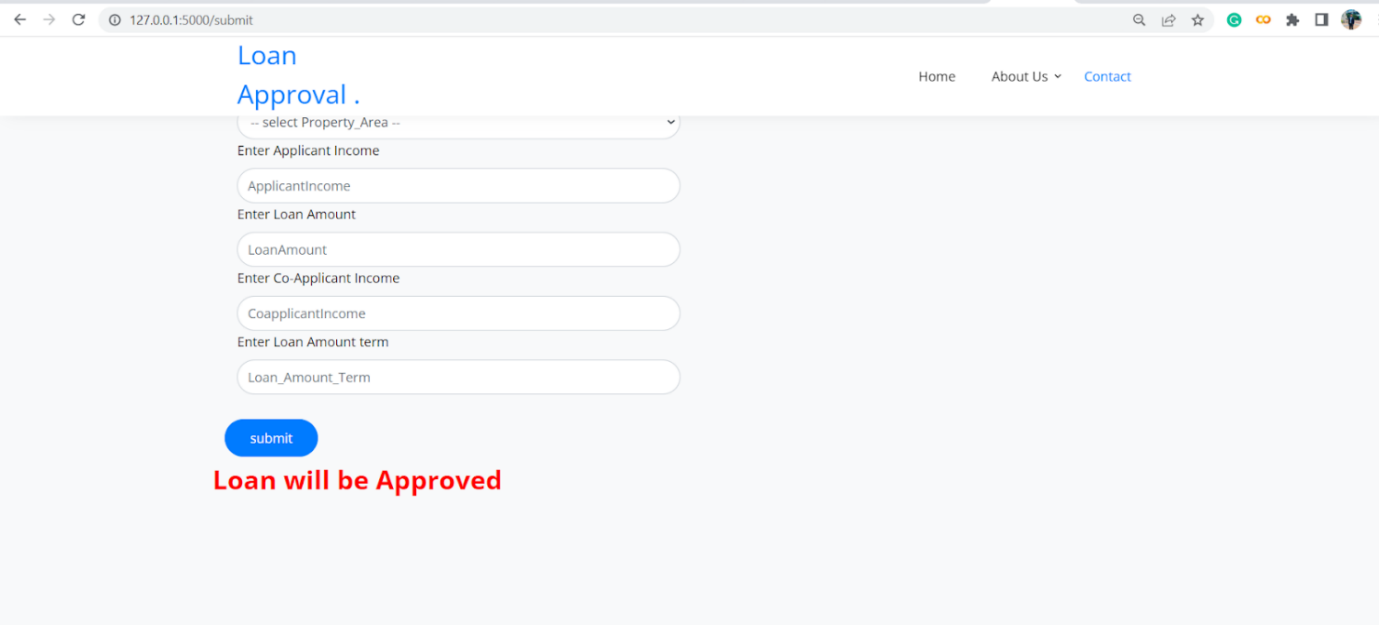


Input 1- Now, the user will give inputs to get the predicted result after clicking onto the submit button





Now when you click on the submit button you will get the result in the same page.



Now the loan is approved in regards to the prediction given by the machine learning process. Thus the result is shown above and hence play a very crucial role in the prediction of loan approval.

ADVANTAGES

The advantages of predicting personal loan using machine learning saves time in this unstoppable world. Thus it has a various advantages and can be seen below in this project.

Time saved:

* How much time could you potentially save your team when dealing with loan applications. By reducing the time spent evaluating credit scores using machine learning technology, you build your efficiency and can dedicate that time to increasing the number of loan applications processed and clients serviced.

Accuracy:

* One of the primary benefits of using machine learning for credit scoring is its accuracy. Unlike human manual processing, ML-based models are automated and less likely to make mistakes. This means that loan processing becomes not only faster but more accurate, too, cutting costs on the whole.

Ease of service:

* In the last decade, the lending industry has boomed with new competitors entering the market alongside traditional providers. Given a wider range of financial providers, borrowers are now choosing where to lend from. By creating ease-of-service and seamless onboarding, you increase your market competitiveness and make it more likely that clients will choose your service provider.

Personalization:

* This has long been growing as a trend, and no more so in the advancements of digital service use in the wake of the COVID-19 pandemic. Clients expect personalization and ML-based models are able to deliver. Not only can you deliver more tailored loans, but you can also establish an individual’s risk more closely than in previous years.

Growth potential:

* How much could your business potentially grow if more efficient methods of credit scoring were onboarded? The growth potential in terms of clients serviced or the variety of services offered helps to establish the value of intelligent bank underwriting software for your business.

DISADVANTAGES

With all those advantages to its powerfulness and popularity, Machine Learning isn’t perfect. The following factors serve to limit it:

Data Acquisition:

* Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

Time and Resources:

* ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

Interpretation of Results:

* Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

High error-susceptibility:

* Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

Thus the advantages and disadvantages of machine learning for predicting personal loan approval is defined properly.

APPLICATIONS

The applications of machine learning in loan approval are:

Customer service

* Just like other fields, the main driving force towards machine learning applications in banking is consumer demand. Customers want a secure and personalized approach to banking. Since banks rely heavily on customer loyalty, they have no other choice but to adopt technologies that will meet the ever-changing demands of customers and to use that technical competence for development further products.
* Machine learning and artificial intelligence improve customer service in so many ways. One of these is customer support. The only way a bank can secure a future with current clients is by offering top-notch customer support. The main problems banks face when it comes to customer service include:
* Lack of a personalized approach to customer complaints;
* Slow services;
* Inability to resolve issues;
* Limited channels.

Fraud detection

* Being used in banking, machine learning is helpful in detecting fraud. It is also applied to prevent fraudulent activities involving credit cards and insurance. Customers want to only work with banks that provide the very best cybersecurity. The level of credit card fraud is at its highest, with £112 million in banking losses in 2019 in the United Kingdom. So banks have to work extra hard to make sure that both customers and staff are protected from fraud.
* The best way to prevent fraud is early detection. This gives the bank ample time to block any activity on the affected account, thus preventing losses. Machine learning and AI are effective tools to create schemes that can detect and prevent fraud. Because every customer's data is unique, banks need experts who can provide practical knowledge on how to forecast, analyze, and classify information. For example, Citibank has invested in machine learning algorithms that provide real-time data that can alert the bank about imminent fraud.

Risk assessment

* ML in the banking sector reduces errors. So financial firms have more accurate reporting. Automation of credit risk testing limits the risks of losses for both banks and clients. By looking at history, ML and AI can provide more accurate foresting before banks can finance loan applications. This, in the long run, helps banks take the necessary steps to curb any potential problem.
* Algorithms can scan huge amounts of data within minutes. This is way faster as compared to humans. What's more, is that they do this with lower chances of making errors. This, combined with big data, helps banks to make more informed decisions about credit.

Marketing

* Success in trading is all about meeting the customer's needs. Any offer that a bank makes should meet the particular needs and concerns of a client. With ML in banking, it is possible to create personalized schemes to market a bank. A great way to identify customers for a new product is through data mining.
* Data analytics and machine learning projects require a training data set. This is what makes algorithm training possible. By looking at demographics, purchase behavior, and history, data scientists can determine the chances a customer will buy a new product. Sales prediction machine learning solutions analyses external and internal business data to provide accurate demand and sales forecasts of a bank product.
* Banks are more likely to make a more personalized scheme on attracting new customers with banking deposit prediction. This also allows banks to improve relationships with each customer.

CONCLUSION

From a proper analysis of positive points and constraints on the component, it can be safely concluded

that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems.

There have been numbers cases of computer glitches, errors in content and most important weight of

features is fixed in automated prediction system, So in the near future the so –called software could be made more secure, reliable and dynamic weight adjustment .In near future this module of prediction can be integrate with the module of automated processing system. the system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time.

FUTURE SCOPE

The future scope of machine learning in predicting personal loan mainly involves in banking and finance. Thus the scopes can be seen below. Here are some scope of usage:

Improved customer support

* How many times have you been irritated by the lack of real solutions when calling any bank helpline numbers? Machine learning is expected to change this picture with its intelligent solutions. Customer support will be aided with previous accounts details to help the system provide quick and helpful response.

Enhanced data quality

* Machine learning and AI will certainly enhance the quality of data. Right from collection to sorting and filtering complex and cumbersome calculations, a lot can be managed by modern day software and bots. It saves a large amount of manpower and yields higher accuracy in data.

Fraud prevention

* Artificial intelligence plays an important role in data security which is one of the greatest responsibilities and concerns of banking and financial services. Preventing financial frauds and data robberi­­es are two key areas where AI has been a game changer for financial institutions.

Digital financial assistants

* Allo, Siri and Cortana are pioneers in the space of digital assistants. For financial institutions, these assistants are going to be the new big thing and AI and machine learning are the key propellers of this huge change. Assistants will help the banks to maintain a personal relationship with each employee without having to spare their human resources. Digital assistants will provide quick and accurate info based on account holder’s transactional history and activity, improving responsiveness and TAT numbers.

Customized marketing strategies

* Deriving from marketing analytics, data coming in from various platforms and touch points like website, app, social media, and even offline media, marketing strategies can be tailored for a set of similar account profiles. The results can also be used for intuitive intelligence that can guide personal financial assistants.

APPENDIX

The given file is the source code and its result or output file.