**Content**

Finance and Risk Analytics

(Project Title: LOAN STATUS PREDICTION)

1. **Problem definition**
2. Problem statement
3. Goal
4. Motivation
5. **Data Understanding**
6. Feature understanding
7. **Exploratory Data Analysis**
8. Missing value treatment
9. Outlier treatment
10. Statistical tests
11. Categorical feature handling
12. Scaling the data
13. Handling imbalance of data
14. **Modelling**
15. Model used
16. Model performance
17. **Model deployment**
18. Model deployment in Heroku platform
19. **Problem Definition:**

Credit risk analysis is a form of analysis performed by a credit analyst on potential borrowers to determine their ability to meet debt obligations.

1. **Problem statement**:

* When a customer approaches a bank for loan, the bank has to look into several factors to decide whether the person is capable of getting a loan. The problem arises when the factors are not being analysed and loans that have been provided to such people are not repaid and it has its consequences like credit fraud.
* When a customer approaches to the bank for a loan, the bank has to look into several factors to decide whether the person is eligible for loan or not.
* But the actual problem arises when these factors are not being analysed and loans that have been given to those customers are not being repaid due to several consequences.
* Banks and Financial sectors face various types of risks. When we combine them together, we call it as ‘Financial Risk’. The risks can be classified as:

1. **Credit Risk:** A credit risk is a risk that arises due to default on a debt that may occur from a borrower failing to make required payments.
2. **Market Risk**: This risk arises because of the volatility of market, variation of different market behaviours like inflation, exchange rate.
3. **Goal**:

By understanding the relationship between the underlying factors that significantly affect a person’s credit score and therefore impacts the customer’s loan eligibility. An improved model that uses these significant features in order to predict the loan status and credit score which acts as a decision-making tool.

1. **Motivation**:

The single biggest reason for PSBs to post a 57,832-crore turnaround - from a loss of 26,015 crore in Y20 to a combined profit of 31,817 crore - was the end of their legacy bad loan problem. (2021)

**2) Data understanding**

* Data contain 100000 rows and 19 columns.
* Each feature and its description are given in **Table-01**.

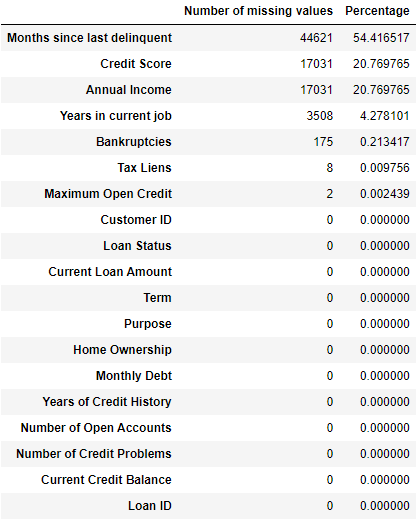
**Table-01. Description of features**

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| **Loan ID** | Id for the loan |
| **Customer ID** | Id for the customer |
| **Loan status** | Whether customer paid the loan completely or not **(Target variable)** |
| **Current Loan amount** | Loan amount borrowed by the customer |
| **Term** | Time duration of loan (Short duration, long duration) |
| **Credit score** | A 3-digit number that represents the credit worthiness of an individual |
| **Annual Income** | Annual income of customer (Rs) |
| **Years in current job** | Number of years the customer working in present job |
| **Home ownership** | Whether the customer has his own house or rent base mortgage (A mortgage is an agreement between you and a lender that gives the lender the right to take your property if you fail to repay the money you've borrowed plus interest. Mortgage loans are used to buy a home or to borrow money against the value of a home you already own.) |
| **Purpose** | Purpose of taking loan by the customer |
| **Monthly debt** | Monthly instalment to be paid by the customer once he accepts to get the loan |
| **Years of credit history** | A credit history is the record of how a person has managed his or her credit in the past, including total debt load, number of credit lines, and timeliness of payment. |
| **Months since last delinquent** | Delinquency means that you are behind on payments (Rs) |
| **Number of Open Accounts** | Number of accounts in the bank |
| **Number of Credit Problems** | A person is considered to have bad credit if they have a history of not paying their bills on time or owe too much money |
| **Current Credit Balance** | All charges and payments made to your account up to that day |
| **Maximum Open Credit** | Open credit refers to accounts that you can borrow from up to a maximum amount (like a credit card) but which must also be paid back in full each month. |
| **Bankruptcies** | customer liability |
| **Tax Liens** | A lien is a judgment or legal right in respect of properties that are usually used as collateral to pay a debt. A creditor or a legal opinion may create a lien. ... When the underlying duty is not fulfilled, the lender will be entitled to seize the asset, which is the subject of the lien. |

1. **Exploratory Data Analysis**

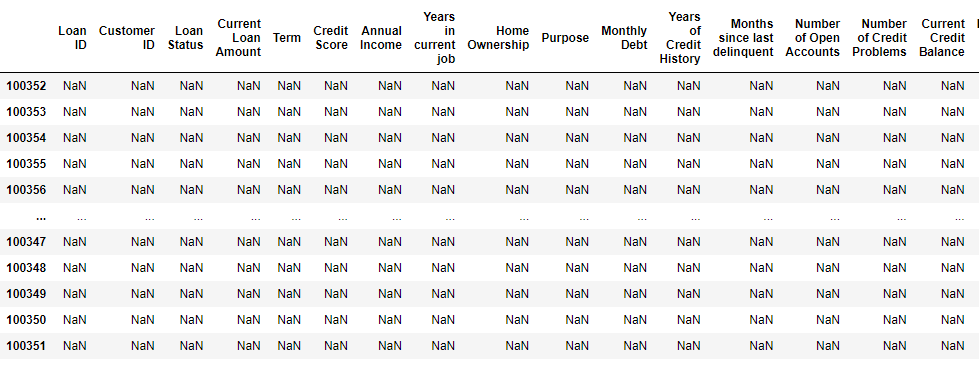
a) Missing value treatment

At initial stage the number of missing values in each column and their respective percentage with respect to the total length of data is given in table\_02

**Table-02. Missing value information**

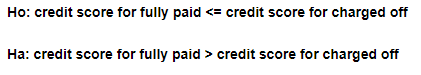
Apart from missing values in each column there are some rows which have completely null values. (Table-03)

**Table-03. Rows with zero information**

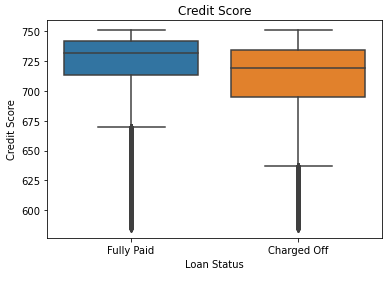


**How missing values are handled?**

* At first, all the rows which have missing value (zero information) are dropped.
* All features divided into categorical and numerical. Each features type treated separately.
* For each numerical feature we plot the distribution plots to visualize the normality of feature, if feature follows near normal distribution missing values in that column replaced by mean values, if feature is skewed then missing values are replaced with median
* **Special case:** In case of “credit score” and “annual income”, by statistical test (Two sample z-test) we found that mean values for these columns is significantly different for different loan status.



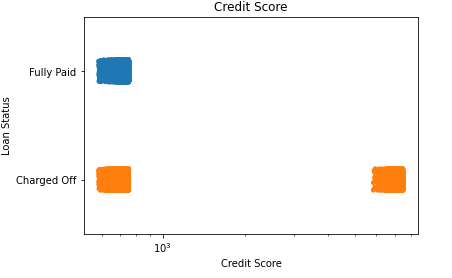
So, the missing values in these two columns replaced with mean values corresponding to loan status.



**Fig-01. Difference in mean of credit score for fully paid and charged of loan status**

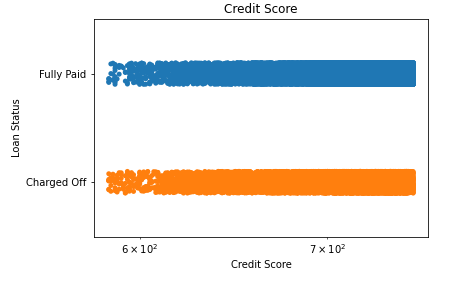
**Cleaning of Credit Score column**

Credit score is a 3-digit number that explains the customer liability (more the value more liable), but in our dataset we found credit score value contain 4-digit number also.



**Fig-02. Strip plot for credit score**

4-digit numbers are strange in case of credit score column, with keen observation we found that all digit number ends with ‘0’, so we think that this might be typo error so we just dropped the 0 in unit place for all 4-digit number.

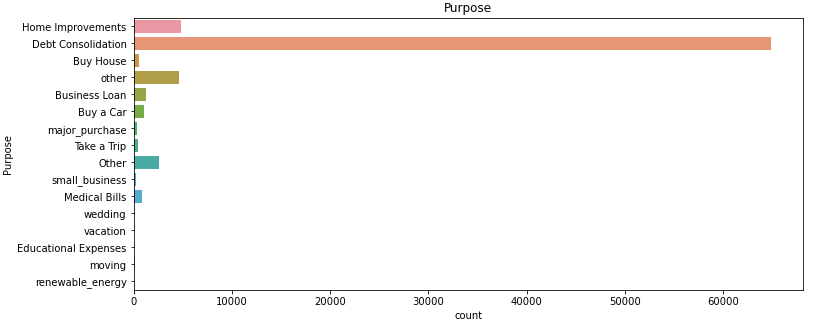


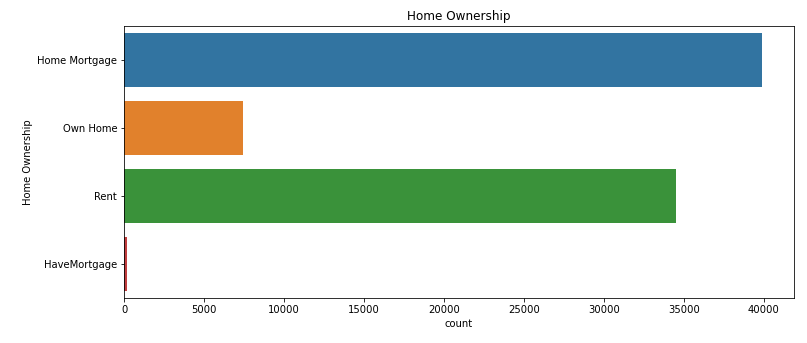
**after dropping ending zeroes in 4-digit numbers**

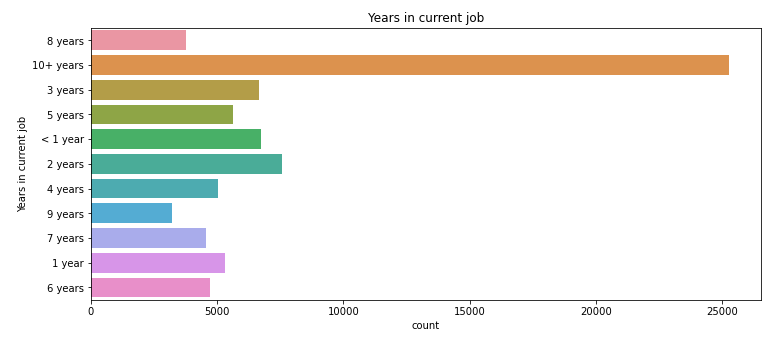
**Fig-03. Strip plot for credit score after cleaning**

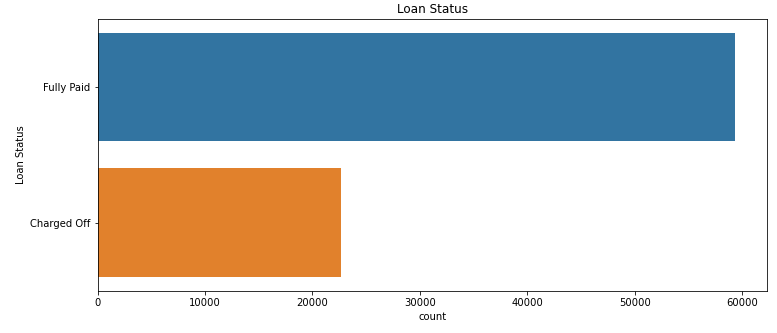
Missing values in Categorical features

For all categorical features missing values replaced with the mode value after observing the count plots of each feature.



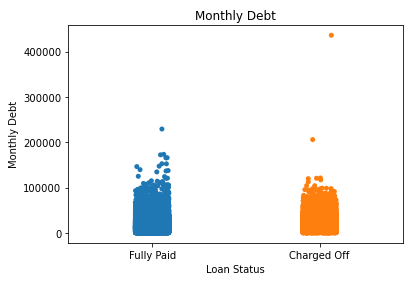






**b) Outlier treatment**

Here we did not take 3 sigma rules or IQR technique to remove the outliers, instead through visualization, if any data found completely apart from the group those observations are dropped thus handled the outliers.



**Fig-04. Outliers in monthly debt**

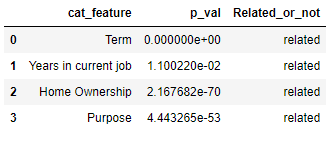
**C) Statistical tests**

* While treating missing values we did two sample z-test to check is mean value of any numerical feature is significantly different in different loan status.
* To check whether categorical features are related to target variable (Loan Status) we performed chisquare\_contingency test.

**Ho: Categorical features are not related**

**Ha: Categorical features are related**

Based on the result obtained from the test we found that each categorical variable related to the target variable.



**Fig-06. Relation between categorical features and target**

**d) Categorical feature handling**

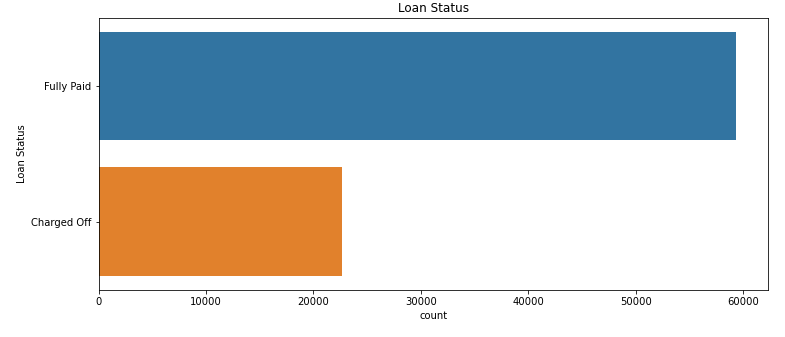
Since machine learning models do not understand the categorical variable, we need to convert each categorical variable into numerical values. In our case we used one hot encoding for the categorical variables by dropping the first one. Thus, each category converted into 1 if present else 0. After one hot encoding the numbers of columns have increased from 19 to 42.

**e) Scaling the data**

For some machine learning models data need to be scaled in order to compare the distances between features without having effect of scale. So, data is gone through StandardScaler wherever required to build the ML models.

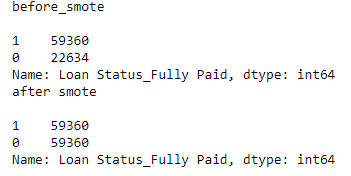
**f) Handling imbalance of data**

Whenever there is a huge difference between the observations in different categories of the target variable then the data is said to be imbalanced. In this work also dataset was imbalanced.



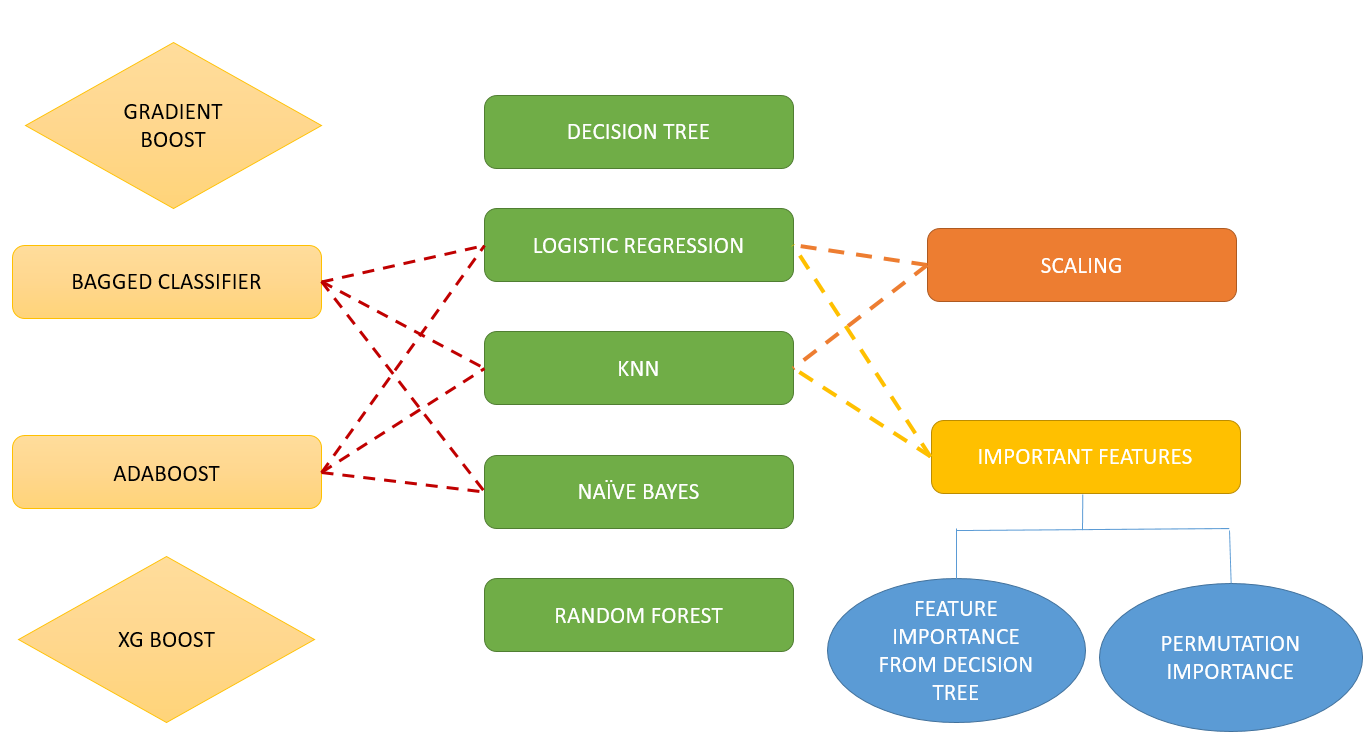
**Fig-07. Imbalance of the target variable**

This imbalance of the data is handled by using **SMOTE** technique available in imblean.oversampling module, which works on the basis of KMeans clustering thus adding data to the original data to make it balanced.



1. **Modelling**

Different machine learning models are fitted to check the performance of the models in order to classify different loan status. Models includes basic individual models such as Logistic regression, KNN, Decision tree. And also bagging and boosting techniques are also used such as RandomForest, AdaBoost, Grdadientboost and XgBoost. The pictorial representation of models used is given in below figure:



**Fig-08. ML models used to predict the loan status**

1. **Performance of each model**

Since the problem statement is of classification type, to evaluate the model performance one can use accuracy, precision, recall, f1\_score, roc\_auc\_score. At the very beginning the dataset was imbalance but to make the dataset balance we used smote technique. So, for this balanced dataset to measure the performance of models we used accuracy for test data and roc\_auc score. Performance of different models is given in terms of accuracy and roc\_auc score in below table. Darker the colour more is the accuracy.



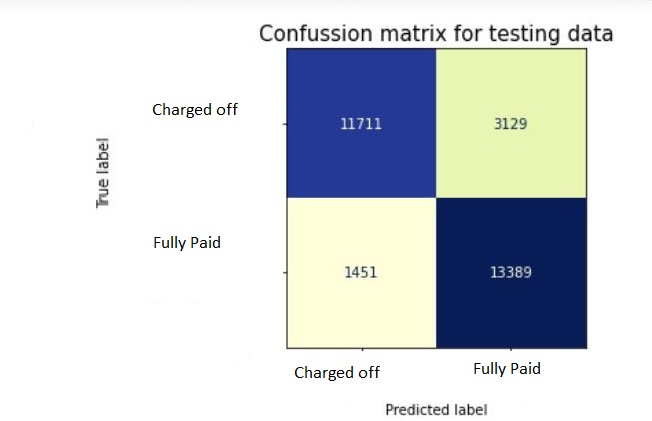
**Table-04. Performance of various models in terms of test accuracy and roc\_auc**

Among all models RandomForest is performing best. So, we will consider this two model for further hyperparameter tuning.

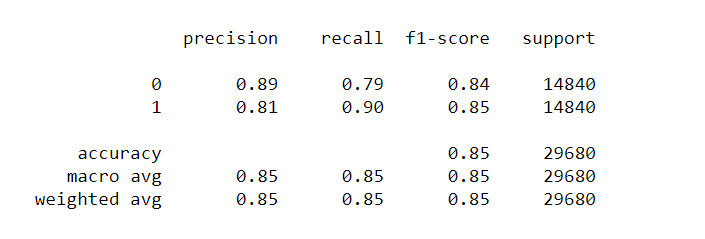
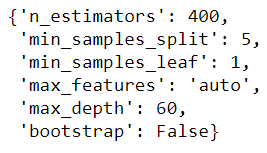
1. **Hyper Parameter Tuning for Best Models**

**c.1) Random Forest**

Below is the performance report for **Random Forest Classifier** model built using all features along with Hyper parameters based on Randomized search.

****

**Fig-09: Confusion matrix**

****

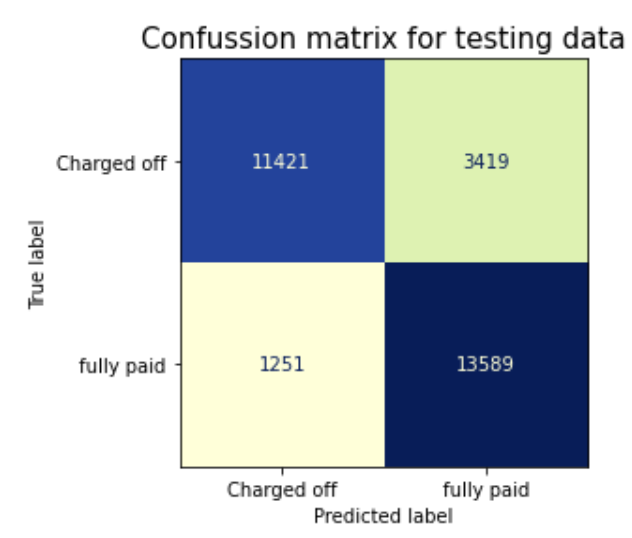
**Fig-10: Hyper parameters Fig-11: Classification Report**

Based on our observation, we found that Random Forest with all features, balanced data and hyper parameters is performing the best in terms of both ‘Accuracy score’ and ‘ROC\_AUC’.

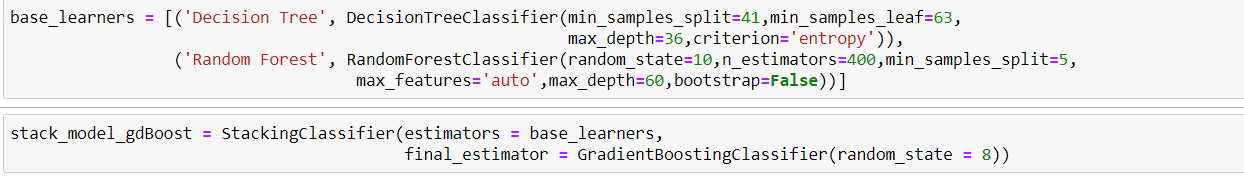
* **Pros:** Accuracy 85% and ROC\_AUC score 93%, which is well enough for taking into production.
* **Cons:** While building the model we utilized all the features and did not consider any feature selection.

**c.2) Stacking Classifier**

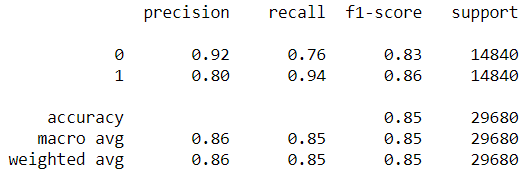
Below is the performance report for **Stacking Classifier model** using base\_learners as **Decision Tree** and **Random Forest** (Both with specific hyperparameter tuning), and final estimator as **GradientBoostingClassifier.**



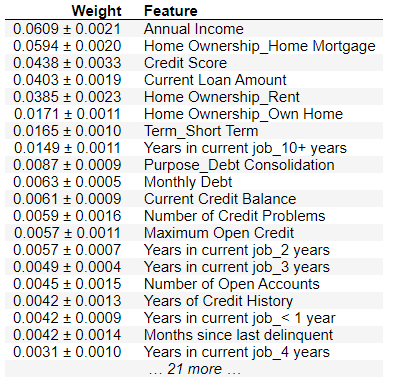
**Fig-12. Confusion matrix**



Classification Report



* **Pros:**
  + Accuracy 85% and ROC\_AUC score 93%, which is well enough for taking into production. The accuracy (85%) is achieved by using only 20(50% of total features) features instead of using all 42 features.
  + As a part of feature selection we used permutation importance technique (from **eli5**). And we did features selection based on this output.

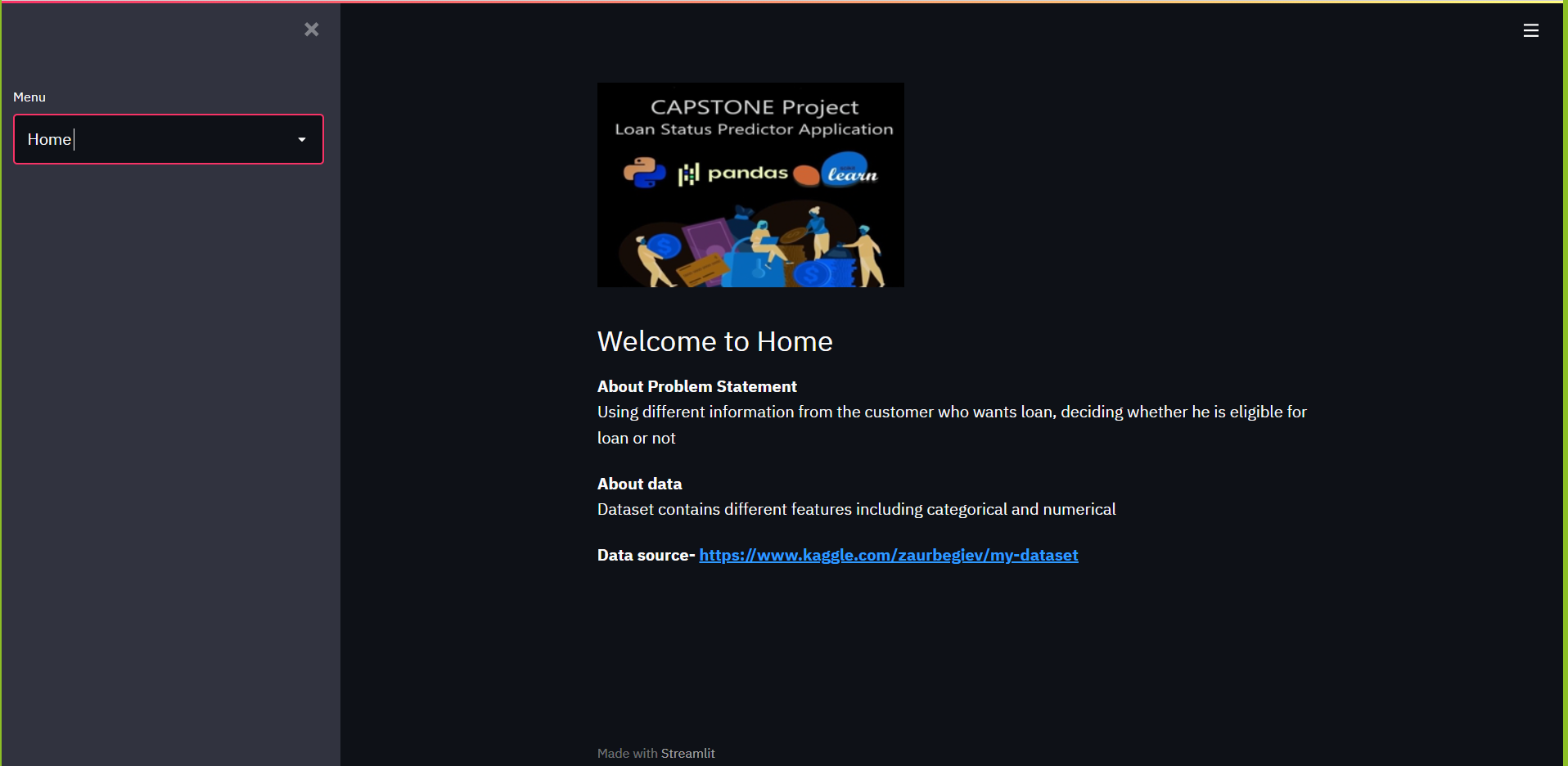


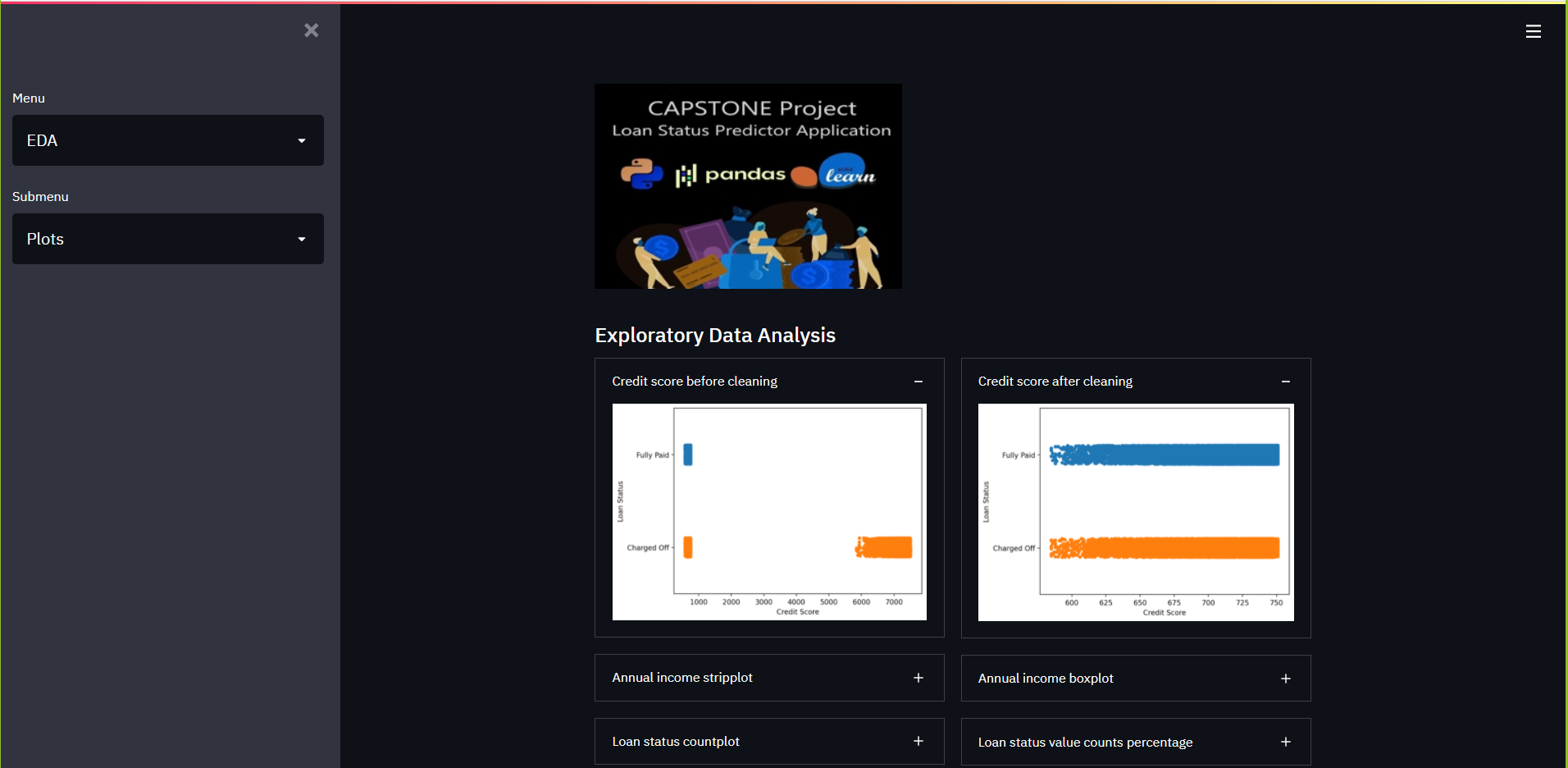
* **Cons:**
* Complexity of the model
* Excluded features may play important role if we had taken data from different sample from same population.

1. **Model deployment**

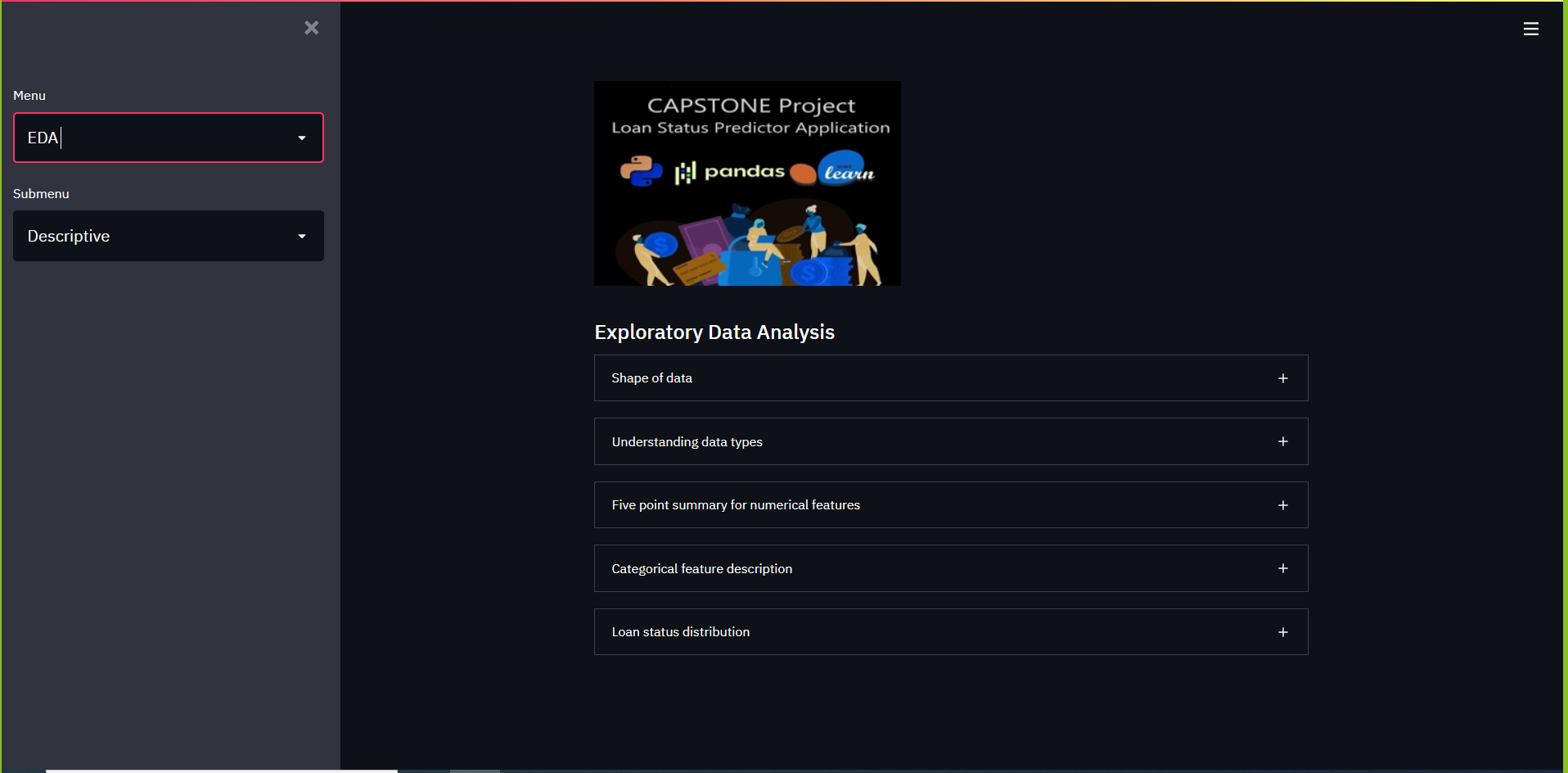
We have taken stacking classifier with base learners as Random Forest and Decision Tree, and final estimator as GradientBoostingClassifier. And deployed in local machine, using Streamlit.

**Home Page**

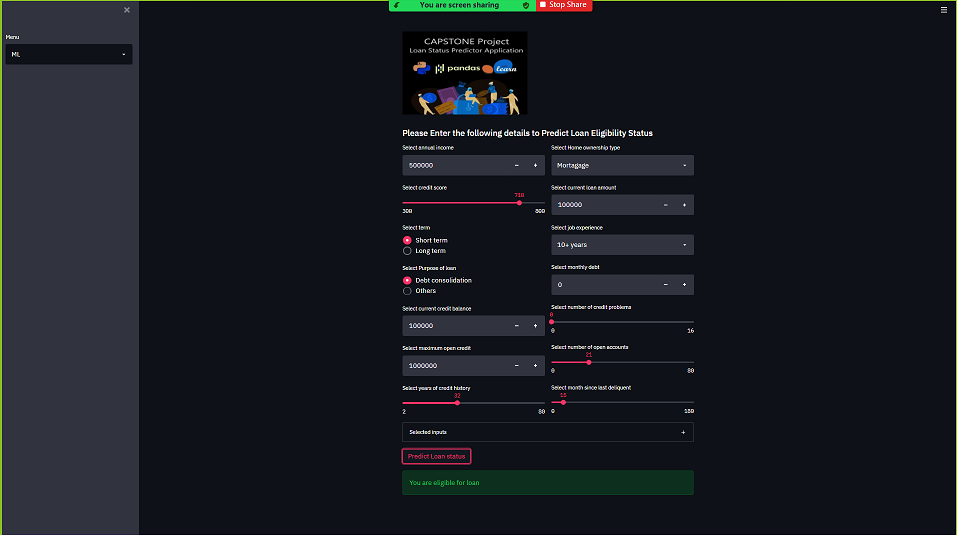


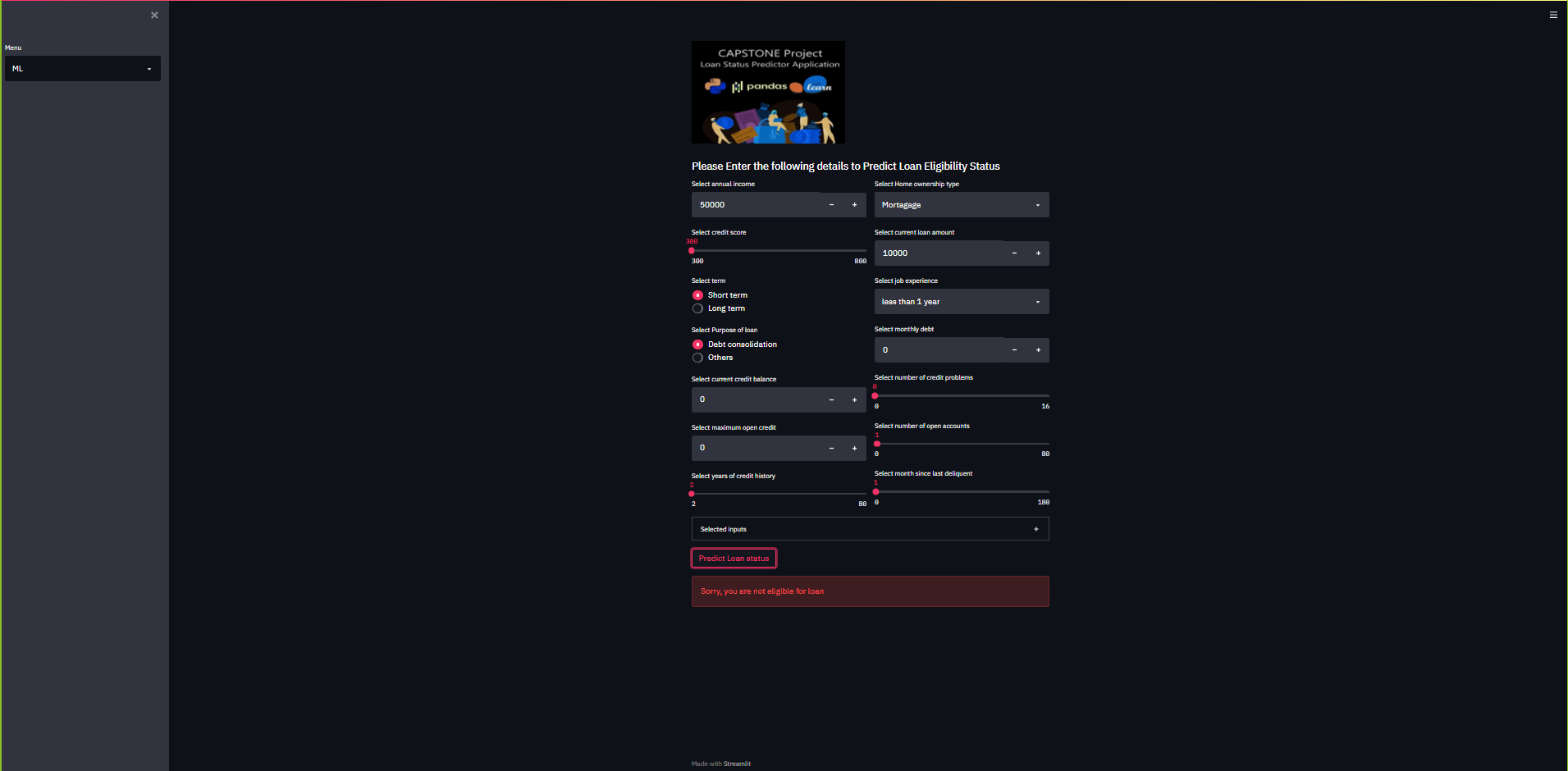


**EDA**



**ML**





**Conclusion**

* Artificial data injection due to SMOTE oversampling. (Biased data).
* Might perform poor in production due to unseen data.
* Extensive pre-processing is done before modelling. So, in real time, test-data needs to pass through data pre-processing filter before invoking prediction algorithm.
* Scope of prediction is limited to specific features used for modelling.
* As per our feature selection techniques (RFE), selective feature based models are not performing well.
* Using alternative feature selection technique (permutation importance) we successfully extracted 20 most important features out of 42 total features, which significantly improved model performance.
* Using Streamlit Python based framework model is deployed in local machine.