|  |  |
| --- | --- |
|  | **Case Study** |
|  | Submitted To  Eastman Chemical India  Submitted by  Abin Singh R |

|  |
| --- |
| **[Predictive Model Building To identify The potential Customers]** |
|  |

**Table of Contents**

**01. Problem Statement**

**02. Data Sourcing and Basic Checks**

**03. Exploratory Data Analysis (for Categorical)**

**04. Outlier Detection, Correction and EDA (for Continuous)**

**05. Feature Engineering**

**06. Dummy Variable Creation and Sampling**

**07. Cost Sensitive Learning and Evaluation Metrics**

**08. Model Building and Cross-Validation**

**09. Finalizing and Saving the Model**

**10. Model Deployment**

**11. Conclusion**

**01. Problem Statement**

When Bank offers Personal Loans to customers, some customers may accept it and some of them may not accept it based on their requirements. In order to avoid this in-effective approach, Bank wants to identify the potential customers who can accept the loan in advance and then provide loan offers.

As a data scientist, we should look for the pattern in the given dataset and build the model which can help Bank to predict the potential customers and targeting them for offers.

**02. Data Sourcing and Basic Checks**

Missing value check, Unique features check and data type conversion has been carried out.

**03. Exploratory Data Analysis (for Categorical)**

* Uni-Variate analysis

**Key Findings:**

\* Classes are imbalanced (Accepted: 480 counts, Rejected: 4520 counts)

* Bi-Variate analysis

**Key Findings from Chi-Square Test (Personal Loan Vs Categorical predictors):**

\* Family, Education, CDAccount has a good association with Personal Loan (It can be a Significant Variable)\*.

SecuritiesAccount, Online, CreditCard has a bad association with Personal Loan (It can't be a Significant Variable)

**04. Outlier Detection, Correction and EDA (for Continuous)**

**Key Findings:**

Experience variable has negative entries, which is not possible. So, it has been changed to Zero.

Mortgage variable has an Outlier and it has been capped.

Two sample t-tests is used to measure the association

Income, CCAvg and Mortgage can be a significant Variable

Age and Experience does not show more association

**05. Feature Engineering**

Credit card spending to income ratio variable has been newly created.

**06. Dummy Variable Creation and Sampling**

Categorical variables are Dummy coded with 0’s and 1’s.

Train and test split is created with 80:20 ratio

**07. Cost Sensitive Learning and Evaluation Metrics**

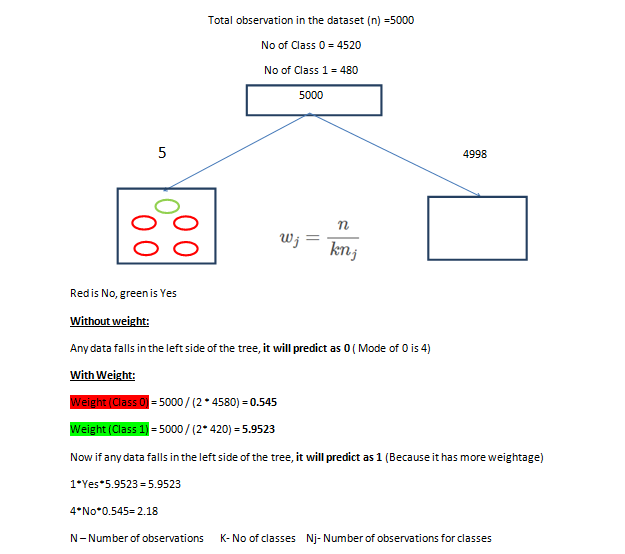


Figure 1: Hand Calculated

As our dependent variable PersonalLoan is heavily biased to the customer who does not accept the loan offers (0), dataset should be processed with undersampling, Oversampling or Cost Sensitive Learning methods.

As Cost Sensitive Learning (CSL) method is an effective method to tackle imbalanced datasets. So, CSL is considered.

In our case study, CSL method is applied by adding weights to the minority classes (In our case, minority class is the customer who accept the loan offers (1))

Class\_weight argument is added in the model for all the techniques (except XGB model)

**Evaluation Metrics**

Our business wants to correctly identify the potential customers (recall) and also to avoid false positive rate, f score (weighted recall & precision) is used for the entire model building process.

**08. Model Building and Cross-Validation**

* Random Forest (Used for Variable Selection)
* Logistic Regression
* Decision Tree
* Support Vector Machine + Feature Standardization
* Adaptive Boosting
* Gradient Boosting
* Xtreme Gradient Boosting

**Variable Selection**

Random Forest algorithm is used to identify the significant variables in the dataset.

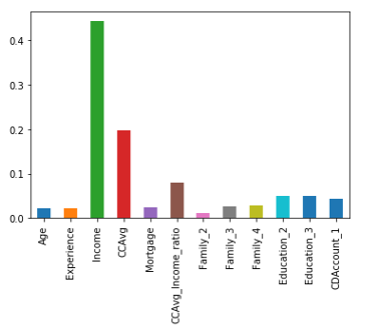


Figure 2: Extracted from Jupyter Notebook

**ROC Curve**

To improve the performance further for Logistic Regression, AUC metric and ROC curve is used.

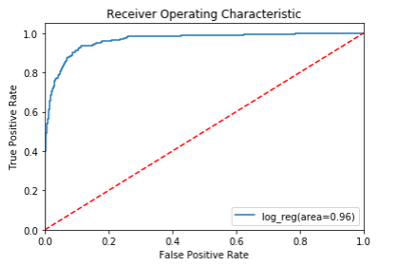


Figure 3: Extracted from Jupyter Notebook

**Grid Search Method**

To choose the optimal parameter, Grid Search method is used for the algorithm which has hyper parameter.

Parameter grid values are taken by considering the bias and variance trade-off.

**Cross-Validation**

To test the generalization of the model, cross validation technique is used for all the algorithms.

**Feature standardization:**

To tackle SVM Algorithm from convergence problem in training, continuous variable has been standardized.

**09. Finalizing and Saving the Model**

Gradient Boosting model is finalized by considering f\_score and Model generalization.

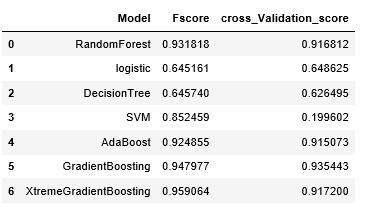


Table 1: Extracted from Jupyter Notebook

Actual business problem is to target the potential customers . If the Bank spend 1 hour in a day with non-potential customer, then averagely 20 buisness hour per month Bank has to convince the customer who would not buy. But if the Bank spend that 20 hour for the potential customer who can accept the loan offers, then Bank will get additional profit in a month.

So, we have to correctly identify potential customer and also avoid mis-classifying customers who are non-potential customers.

When compare with all other models, harmonic average of recall and precision (F1 Score) is high for Gradient Boosting model, So it has been considered as the final model

And also the model ‘s performance is better for out of sample which is tested using cross validation techniques.

There is no model assumptions behind this Gradient Boosting model.

And in future if we get more samples for prediction, it can do the same performance as it can handle large datasets

**Confusion Matrix for final model**

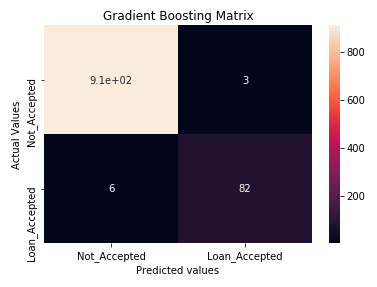


Figure 4: Extracted from Jupyter Notebook

**Saving the Final model**

Joblib module is used to save the final model. And for demonstrating deployment purpose, Logistic Regression model is also saved.

**10. Model Deployment**

**Web App Module:**  flask

**Project Structure:** Main folder is app

****

**Sub-folders:**

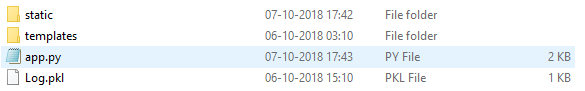
****

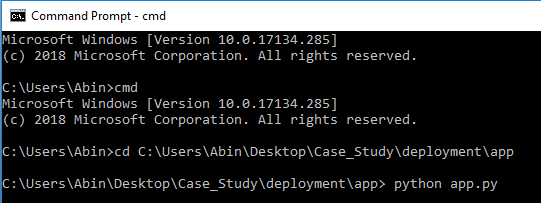
Image files which are used in the web page is placed in the **static folder**

Html home and result pages are placed in the **templates folder**

**app.py -** Script file for execution

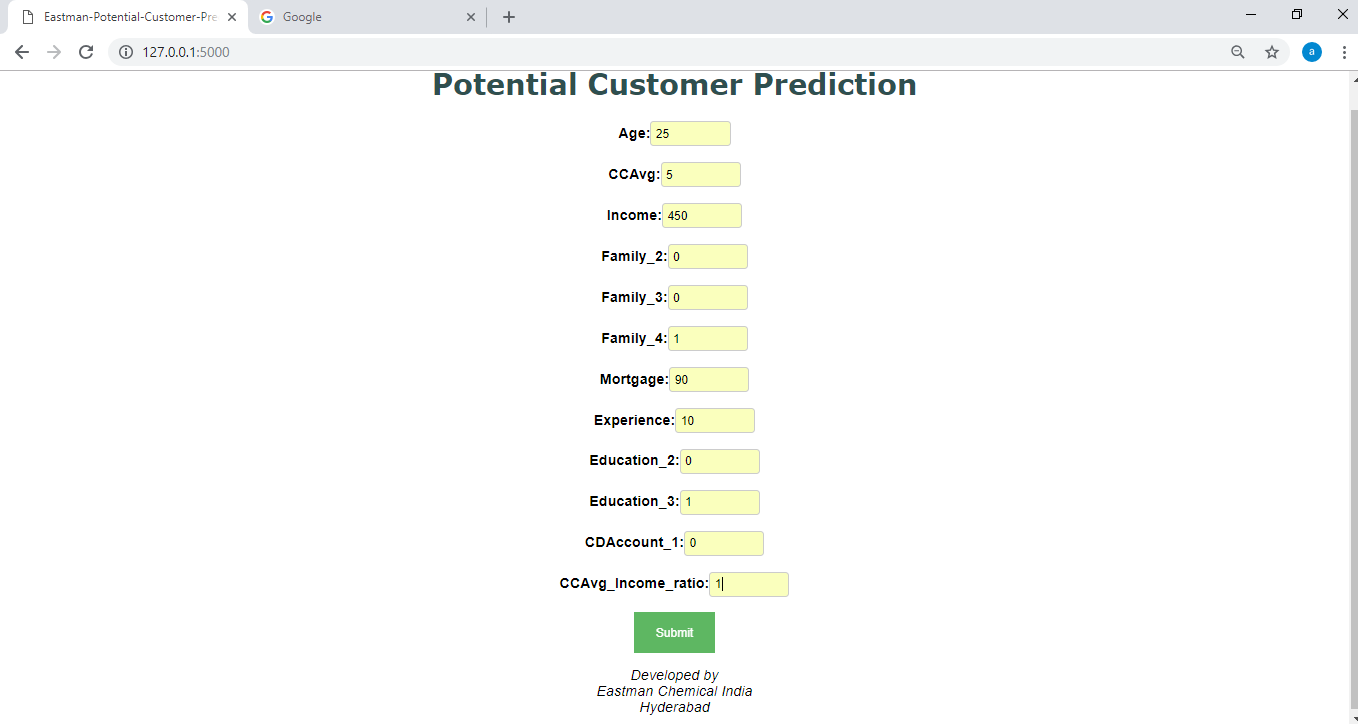
**Log.pkl** – Saved machine learning model

**Execution:**

****

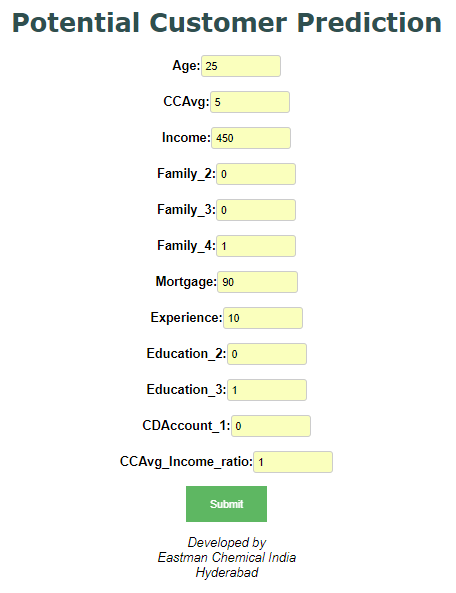
**Running on** [**http://127.0.0.1:5000/**](http://127.0.0.1:5000/)

**Home Page:**

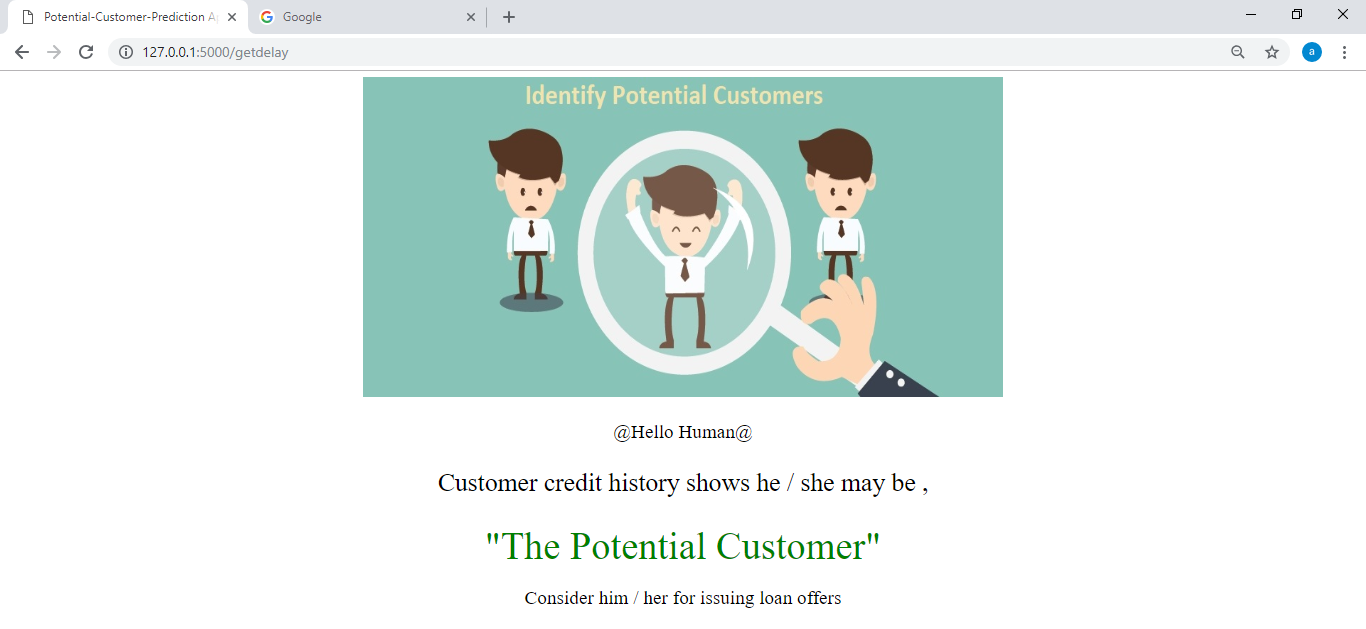
****

**Values Restriction for Categorical variable**

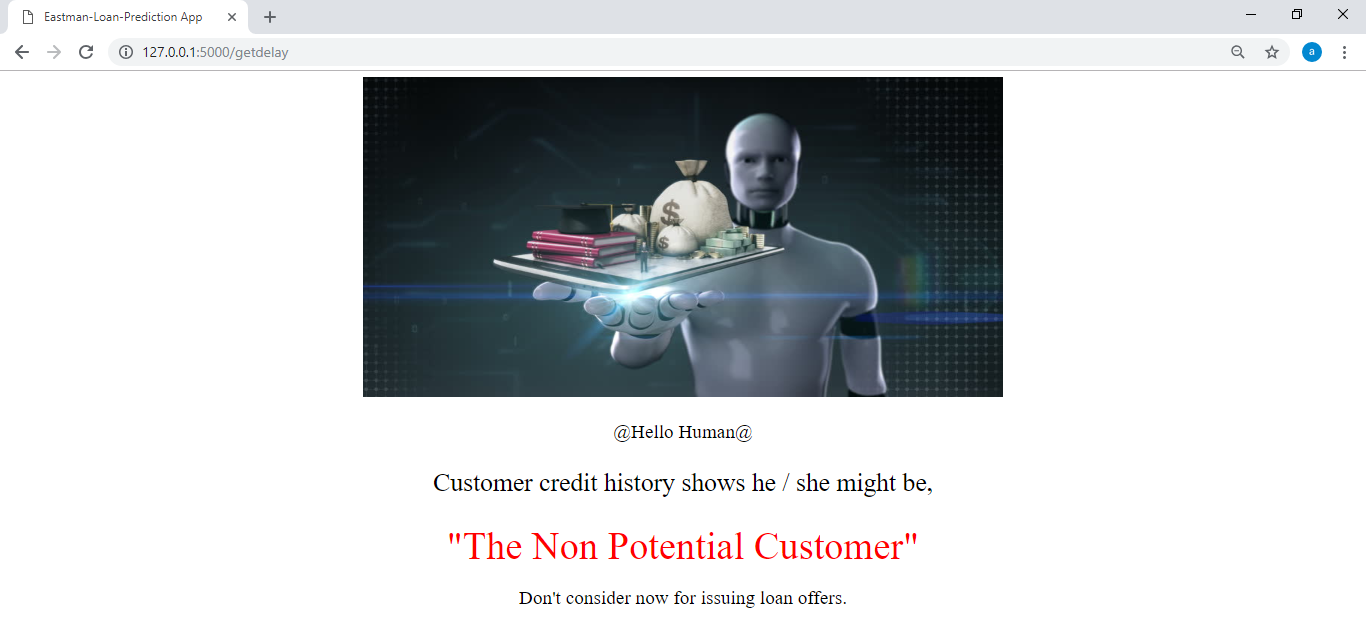
Family, Education and CDAccount\_1 should be either 0 or 1.

****

**For Potential Customer:**



**For Non-Potential Customer:**



**11. Conclusion**

Personal loan offers for the 5000 samples are thoroughly studied by various data analysis and explorations. Various machine learning techniques have been applied and the final model is finalized by considering business requirements.

And finally the model is deployed to make ease of the user who does not have practical knowledge about ML. Only thing he must do is to enter some inputs and get back the prediction of whether the person is the potential customer or not.