

**PROJECT REPORT**

ON

**Home Loan Prediction**

FOR

**MINI PROJECT 2022**

**SUBMITTED TO:                             SUBMITTED BY:**

Mr. ADITYA PAI                         JAHNAVI MAHARA

                                     SEMESTER: 6

                                   SECTION : D

                                                                       ROLL NO: 2014690

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| CHAPTER NO. | **TITLE** | PAGE NO. |
| **1** | ACKNOWLEDGEMENT | 3 |
| **2** | OVERVIEW | 4 |
| **3** | DATA EXPLORATION | 4 |
| **4** | DATA IMPUTATION | 6 |
| **5** | DATA CLEANING | 7 |
| **6** | MODEL BUILD | 8 |
| **7** | RESULT and CONCLUSION | 9 |

**ACKNOWLEDGEMENT**

I have carried out this project during the period February 2022 – June 2022

. I have taken time and efforts for the completion of this project however, it

would not have been possible without the kind support and help of many

individuals and organizations. I would like to extend my sincere thanks to all of

them.

I would like to express my gratitude towards my parents and members of

Graphic Era University for their kind cooperation and encouragement which

helped in completion of this project. My thanks and appreciation also goes to

my friends in developing the project and people who have willingly helped me

out with their abilities.

**JAHNAVI MAHARA**

                                                                    CSE D

                                                                                              2014690

**Overview**

This project’s main goal is to see if a system can be built to help automate the loan approval process.

An applicant will fill out an online form for variables such as income, number of dependents, loan amount, and more. The system will take that data, run it through a model, and predict whether the applicant’s loan will be approved. Any applicant predicted to be approved can be targeted by representatives to expedite the loan process. This project focuses on building the predictive model.

**Data**

**Exploration**

The dataset was taken from <https://www.kaggle.com/> .

The dependents variable was provided as a string, however, it is better to use it as a numerical

variable.

Education was one variable hypothesized to be an important indicator, however, while being a

graduate was more favorable for loan approval, it was not as large of a difference as expected.

360.0 terms was the most common value for Loan\_Amount\_Term, and most of the other value

counts were too small to keep on their own, so grouping based on standard loan terms was

implemented.

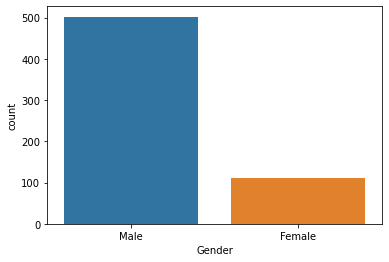
Figure 5 shows the value counts before and after grouping.

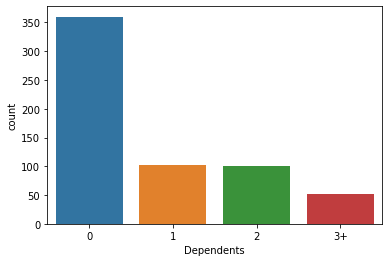
Credit History was another variable hypothesized to be an important indicator, however, the actual

importance was not expected. After seeing the breakdown based on category, this will be the most

explanatory variable. It also has the most missing values, so imputation here will be vital. Figure 6

illustrates the breakdown for Credit\_History.





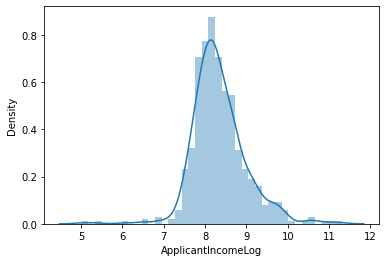
A higher family income was hypothesized as an indicator for a better chance of loan approval,

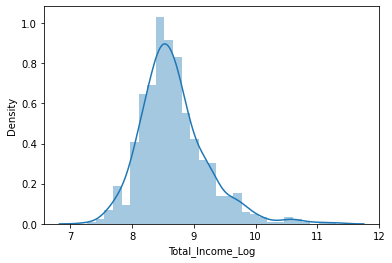
however, breaking down the income into quartiles, and deciles did not confirm that thought. It’s

possible that income isn’t an end all signal of financial health. The following variables and their formulas were created to help paint a better picture of the

applicants financial standing:

● FamilyIncome - ApplicantIncome + CoapplicantIncome



****

**Imputation**

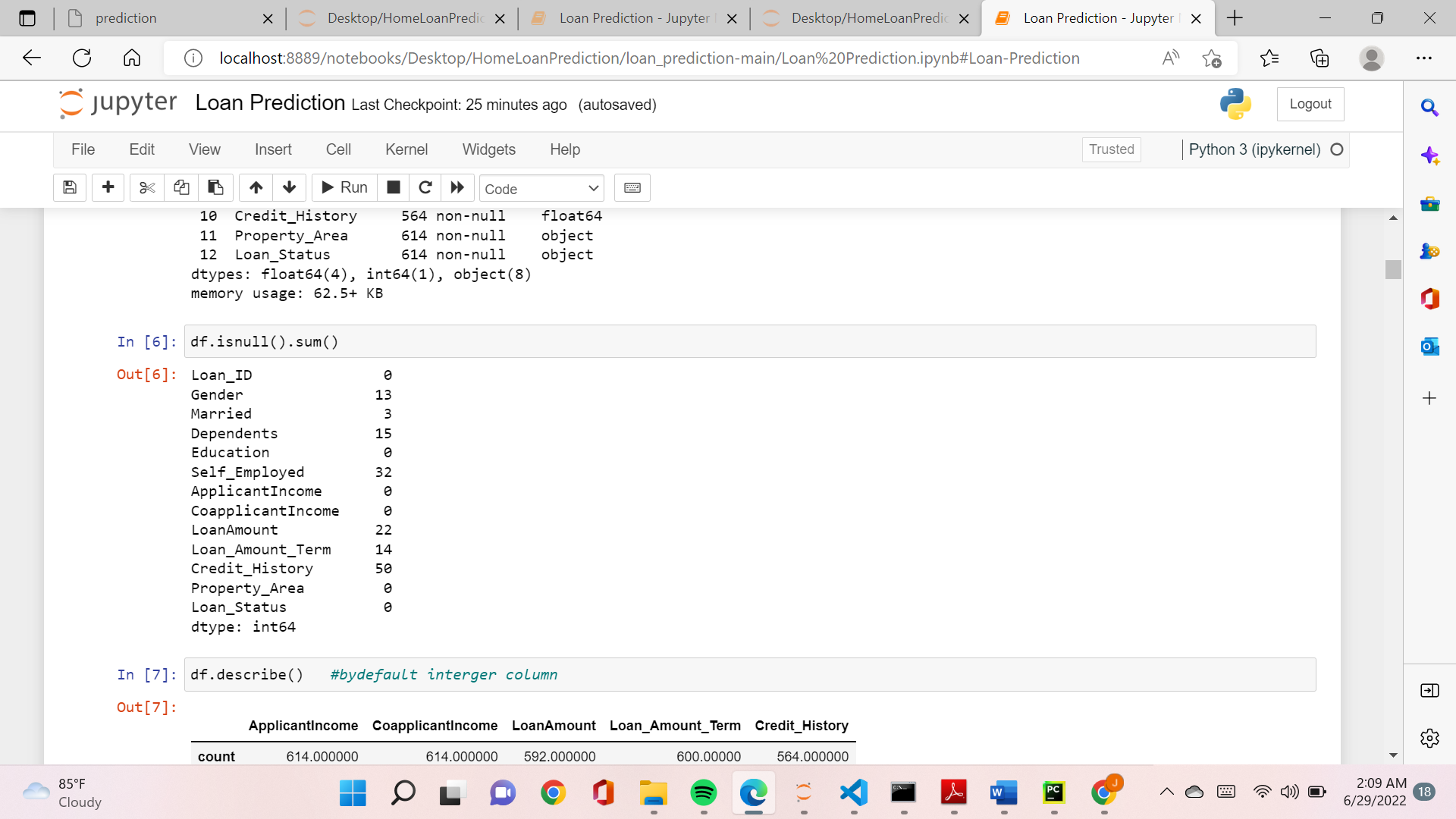
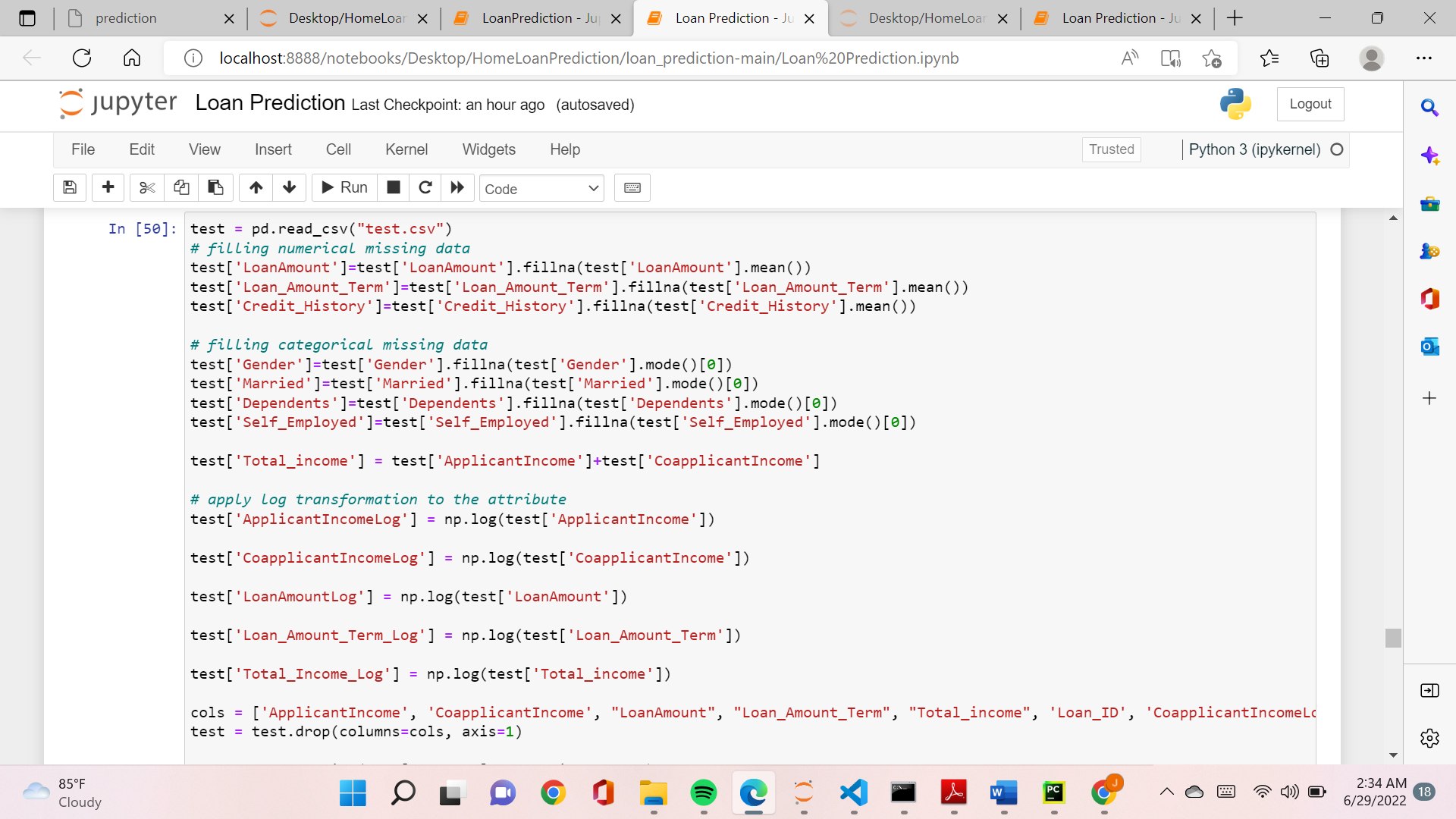
****

Figure shows which variables needed to be imputed and how many values were

missing. All the categorical and numerical data are handeled by taking mean and mode respectively.



**Cleaning**

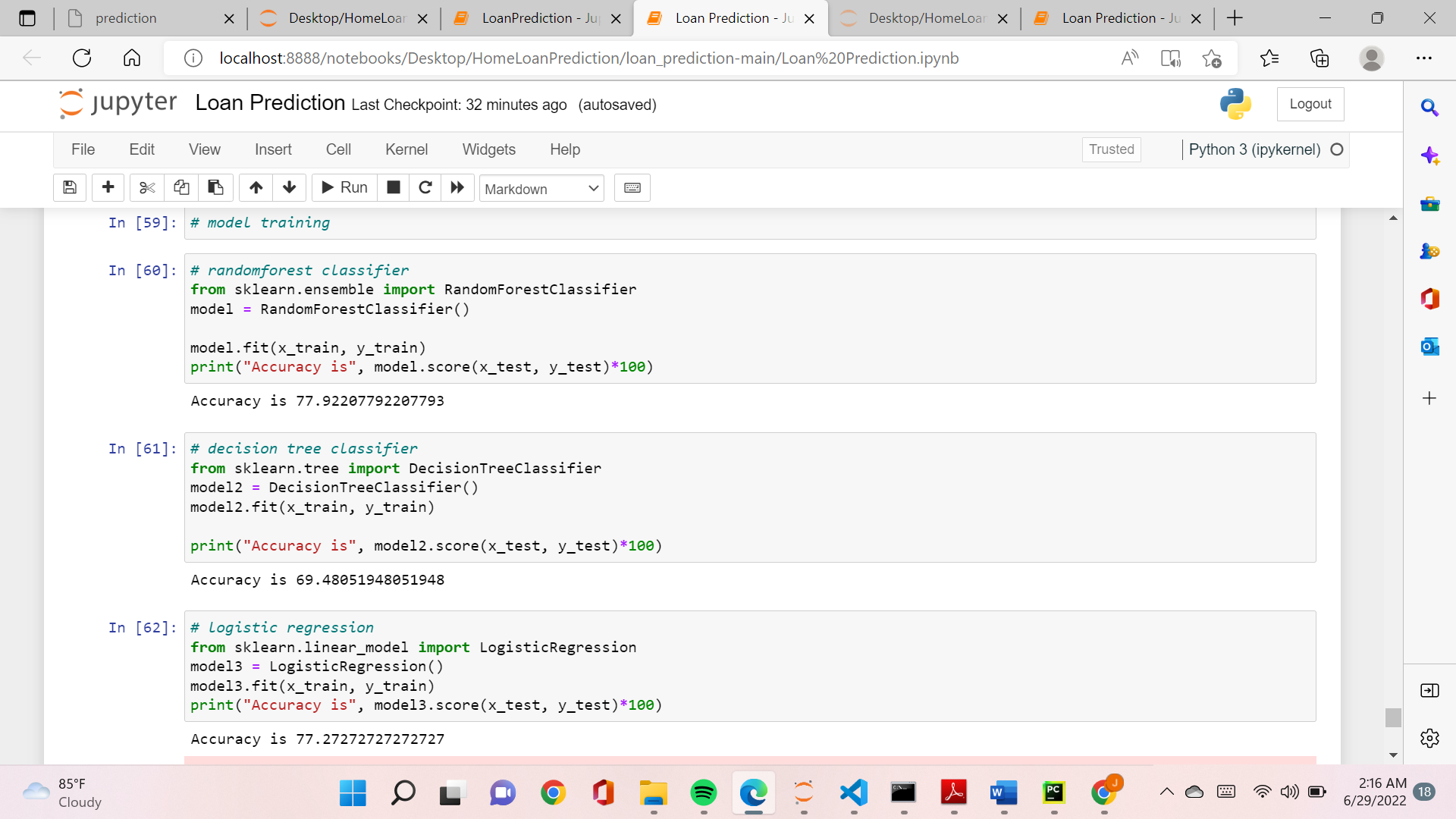
Any variable that was Yes or No, or anything of that sort, was converted to an indicator, 1 in place

of Yes (Male, Graduate), and 0 in place of No. Property Area was converted to a dummy variable.

Dependents were converted to a numeric variable (3+ converted to 3).

**Model Build**

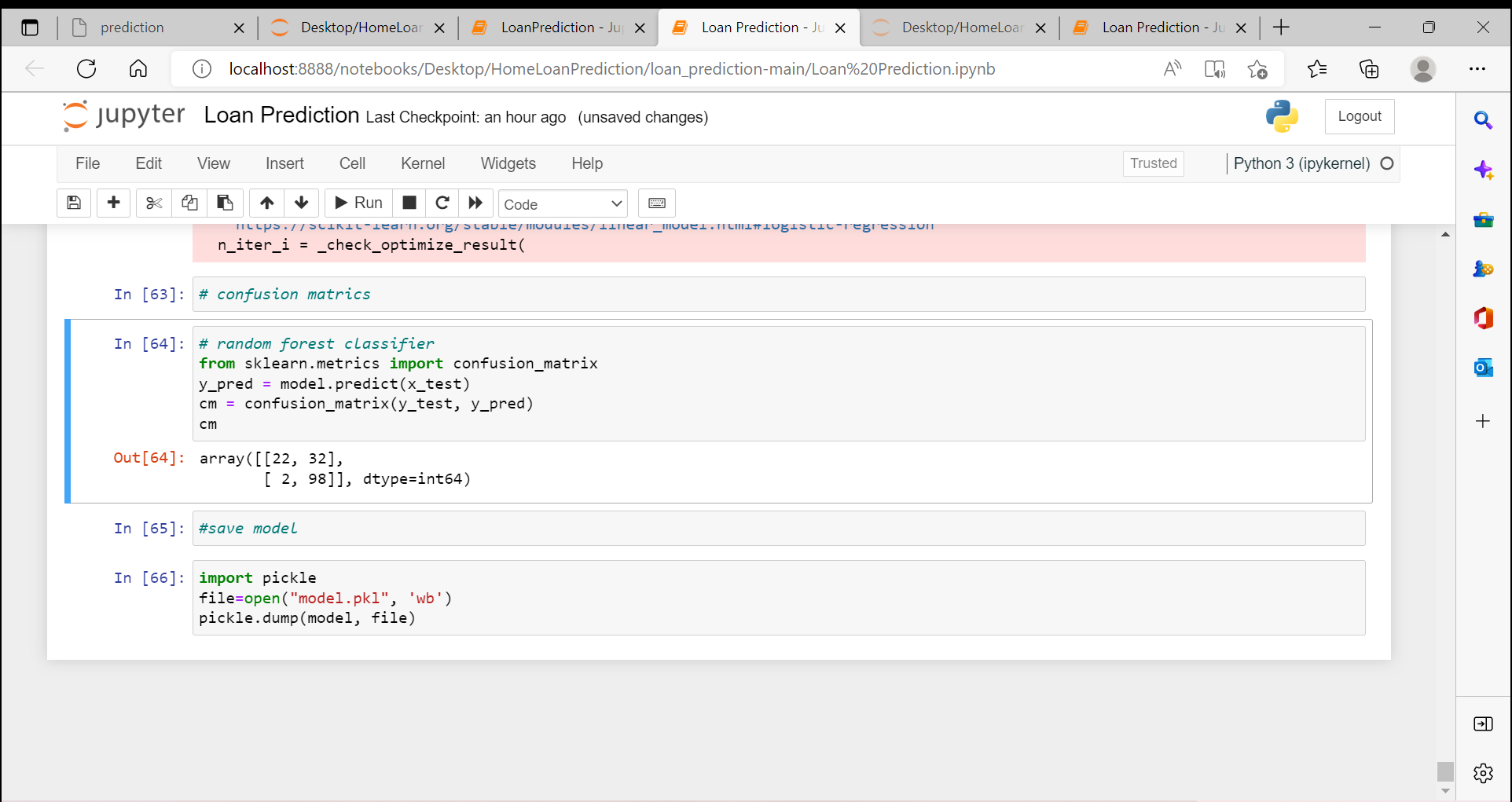
Many models were built using multiple methods, such as logistic regression and random forests. A logistic regression model performed well, and can be seen later, however, ultimately, a random forest model was chosen as the best performing model. It was built using hyperparameter optimization and cross validation, as well as feature selection. Number of trees, maximum features per tree, maximum depth per tree, minimum samples for a split, and minimum samples for a leaf were all optimized .



**Results**

The benchmark rate, based on the percentage of loans approved from the training dataset, is

77.9%.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Conclusion**

The random forest model predictions were submitted to the html website where the RandomForest model was taken as a binary file and was connected to the website using “flask” library of python . the website used a form which has all the required parameters for Loan Approval Prediction . On submitting the form we get the result displayed on the top of the form.

