Imarticus Learning



**Project on**

**Credit Risk Analysis**

**Data Science Pro Degree Batch - 27**



Submitted By:

1. Nidhi Patel
2. Sanyukta Shenoy
3. Wasim Akram

**ACKNOWLEDGEMENT**

Apart from our efforts, the success of this project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

We would like to show our greatest appreciation to our project guide **Ms. Nikita Tandel**. We can’t thank her enough for the tremendous support and help. We feel motivated and encouraged every time we attend her Lectures. Without her encouragement and guidance, this project would not have materialised.

We express our deep thanks to the high authority of Data Science Pro Degree **Mr. Arun Upadhyay** and other faculty member for extending their support.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **SR.NO** | **TOPIC** | **PAGE NUMBER** |
| 1. | ABSTRACT | 3 |
| 2. | INTRODUCTION | 4 |
| 3. | PROPOSED ANALYSIS AND OBJECTIVE | 6 |
| 4. | PURPOSE OF THE STUDY | 7 |
| 5. | ORGANISATION DESCRIPTION | 8 |
| 6. | TECHNOLOGY USED FOR PREDICTION | 10 |
| 7. | VISUALISATION | 11 |
| 8. | MODEL DESCRIPTION | 17 |
| 9. | PREPROCESSING OF DATA | 18 |
| 10. | MODEL BUILDING | 23 |
| 11. | COMPARISION OF MODELS | 28 |
| 12. | CONCLUSION | 29 |

**ABSTRACT**

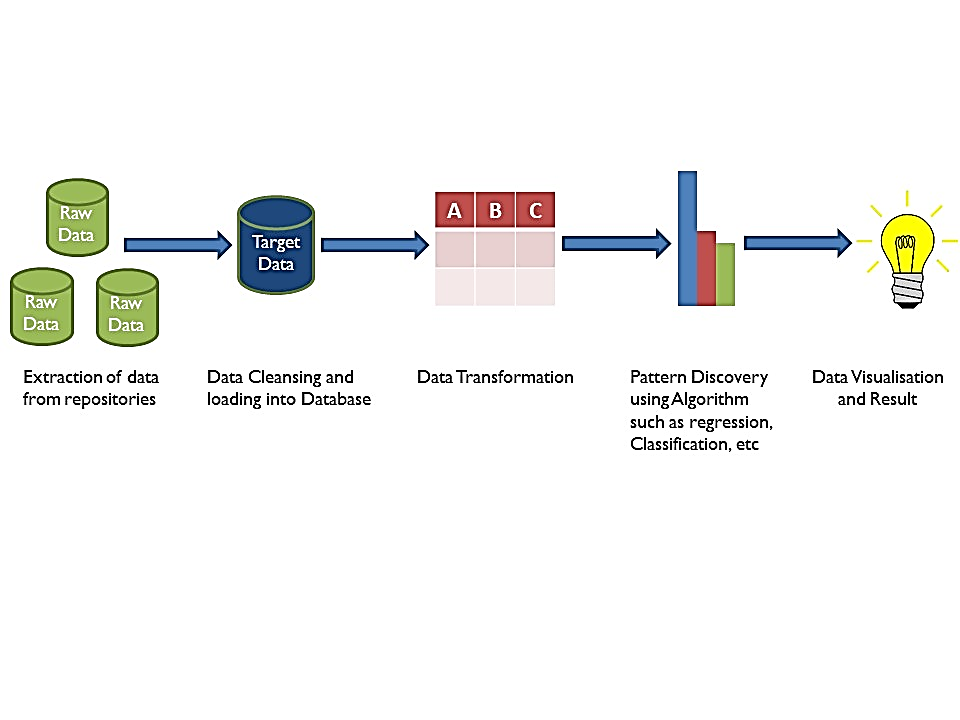
Bank databases is a very interesting field of research, which mainly focus on analysis and understand Credit Risk for bank.

All business including the business of banking requires top line growth in terms of volumes of business to increase the bottom line of profit growth. Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Credit risk also describes the risk that a bond issuer may fail to make payment when requested or that an insurance company will be unable to pay a claim. Hence, it is very important for any Businesses to quickly analyze data used to assess a customer's risk profile.

So, the goal of credit risk management in banks is to maintain credit risk exposure within proper and acceptable parameters to classify each borrower as defaulter or not, based on the loan issued by XYZ Co-operation through 2007-2015.

**INTRODUCTION**

In Data Science the amount of data being generated and stored is growing very rapidly, due in large part to the continuing advances in computer technology. This presents tremendous opportunities for those who can unlock the information embedded within this data, but also introduces new challenges. Data Science can be used to extract useful knowledge from the data that surround us. Those that can master this technology and its methods can derive great benefits and gain a competitive advantage.



**Proposed Architecture for Credit Risk Analysis**

Credit Risk is new trend in the data science and knowledge Discoveries in Databases field which focuses in mining and discovering the useful information such as our example consider for the Financial Domain to predict whether the borrower will repay loan or not.

The challenges faced by the bank today includes:

* Managing quality of loan assets.
* Boosting the credit flow to all productive sectors of the economy.
* To create an interest rate environment that supports revival of investment demand at the same time ensuring consistent growth in bank’s profitability.

The biggest and most difficult to manage for the Bank is Credit Risk, Credit risk can be described as the possibility of losses being incurred on account of deterioration in the quality of the borrowers in a portfolio. The basic function of commercial banks are accepting deposits and lending. So, to overcome this we can use the following details which is collected while opening an account such as:

1. Employment details such as job specifications, name and address of the employer, length of service, etc.
2. Provide details about source of income and annual income.
3. Details of assets owned such as house, vehicle, etc.
4. Other personal details such as qualification, marital status, etc.

Any Business including the business of banking requires top line growth in terms of volumes of business to increase the bottom line of profit growth, In this context, quality and value advances are of utmost importance to ensure sustainable business growth. Hence there is an urgent and immediate need for the bank to concentrate on quality of credit with profitability.

People often save their money in the banks which offer security but with lower interest rates. Lending Club operates an online lending platform that enables borrowers to obtain a loan, and investors to purchase notes backed by payments made on loans. It is transforming the banking system to make credit more affordable and investing more rewarding. But this comes with a high risk of borrowers defaulting the loans. Hence there is a need to classify each borrower as defaulter or not using the data collected from bank named as XYZ Co-operation through the year 2007-2015.

**PROPOSED ANALYSIS AND OBJECTIVE**

**1. TITLE**

**CREDIT RISK ANALYSIS**

**2. OBJECTIVE**

The objective of Credit Risk Analysis project is to put ourselves in the shoes of a loan issuer and manage credit risk by using the past data and deciding whom to give the loan to in the future. Model has to analyzing the XYZ Co-operation Data to analyze and detect defaulters by building models on data from June 2007 to May 2015 and testing it on data from June 2015 to December 2015. Based on the accuracy of each model we can determine how good a model is for predicting that a person applying for loan will default or not.

**PURPOSE OF THE STUDY**

In Credit Risk Analysis, The purpose of the project is to predict whether a borrower will default or not, so that investors can avoid those borrowers using manual investing feature provided by lending club. This, however, does not necessarily lead to highest return on investment because by completely avoiding potential defaults, one also avoid riskier loans that may lead to higher return on investment even though they default at some point in the future. In order to maximize return on investment, one needs to optimize return on investment instead. In this project, we work on the simpler problem that is to predict loan defaults.

0 represents – Not Defaulter

1 represents – Defaulter

**ORGANISATION DESCRIPTION**

**Name of the Organisation:** XYZ Co-operation Bank

XYZ Corporation Lending Data is used under the study. Data of Loans issued by XYZ Co-operation through the year 2007-2015 is used for analysis. The data contains the indicator of default, payment information, credit history and many more variables.

**Database Description:**

**Rows:** 855696 **Columns:** 73

|  |  |  |
| --- | --- | --- |
| **No** | **LoanStatNew** | **Description** |
| 1 | Id | A unique assigned ID for the loan listing. |
| 2 | member\_id | A unique Id for the borrower member. |
| 3 | loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| 4 | funded\_amnt | The total amount committed to that loan at that point in time. |
| 5 | funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| 6 | Term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| 7 | int\_rate | Interest Rate on the loan |
| 8 | Instalment | The monthly payment owed by the borrower if the loan originates. |
| 9 | Grade | XYZ corp. assigned loan grade |
| 10 | sub\_grade | XYZ assigned assigned loan subgrade |
| 11 | emp\_title | The job title supplied by the Borrower when applying for the loan. |
| 12 | emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| 13 | home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| 14 | annual\_inc | The self-reported annual income provided by the borrower during registration. |
| 15 | verification\_status | Was the income source verified |
| 16 | issue\_d | The month which the loan was funded |
| 17 | pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| 18 | Desc | Loan description provided by the borrower |
| 19 | Purpose | A category provided by the borrower for the loan request. |
| 20 | Title | The loan title provided by the borrower |
| 21 | zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| 22 | addr\_state | The state provided by the borrower in the loan application |
| 23 | Dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income. |
| 24 | delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| 25 | earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 26 | inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| 27 | mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| 28 | mths\_since\_last\_record | The number of months since the last public record. |
| 29 | open\_acc | The number of open credit lines in the borrower's credit file. |
| 30 | pub\_rec | Number of derogatory public records |
| 31 | revol\_bal | Total credit revolving balance |
|  |  |  |
| **No** | **LoanStatNew** | **Description** |
| 32 | revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| 33 | total\_acc | The total number of credit lines currently in the borrower's credit file |
| 34 | initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| 35 | out\_prncp | Remaining outstanding principal for total amount funded |
| 36 | out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| 37 | total\_pymnt | Payments received to date for total amount funded |
| 38 | total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| 39 | total\_rec\_prncp | Principal received to date |
| 40 | total\_rec\_int | Interest received to date |
| 41 | total\_rec\_late\_fee | Late fees received to date |
| 42 | Recoveries | post charge off gross recovery |
| 43 | collection\_recovery\_fee | post charge off collection fee |
| 44 | last\_pymnt\_d | Last month payment was received |
| 45 | last\_pymnt\_amnt | Last total payment amount received |
| 46 | next\_pymnt\_d | Next scheduled payment date |
| 47 | last\_credit\_pull\_d | The most recent month XYZ corp. pulled credit for this loan |
| 48 | collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| 49 | mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| 50 | policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| 51 | application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| 52 | annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| 53 | dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested loan, divided by the co-borrowers' combined self-reported monthly income |
| 54 | verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by XYZ corp., not verified, or if the income source was verified |
| 55 | acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| 56 | tot\_coll\_amt | Total collection amounts ever owed |
| 57 | tot\_cur\_bal | Total current balance of all accounts |
| 58 | open\_acc\_6m | Number of open trades in last 6 months |
| 59 | open\_il\_6m | Number of currently active installment trades |
| 60 | open\_il\_12m | Number of installment accounts opened in past 12 months |
| 61 | open\_il\_24m | Number of installment accounts opened in past 24 months |
| 62 | mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| 63 | total\_bal\_il | Total current balance of all installment accounts |
| 64 | il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| 65 | open\_rv\_12m | Number of revolving trades opened in past 12 months |
| 66 | open\_rv\_24m | Number of revolving trades opened in past 24 months |
| 67 | max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| 68 | all\_util | Balance to credit limit on all trades |
| 69 | total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| 70 | inq\_fi | Number of personal finance inquiries |
| 71 | total\_cu\_tl | Number of finance trades |
| 72 | inq\_last\_12m | Number of credit inquiries in past 12 months |
| 73 | Default\_ind | Current status of the loan |

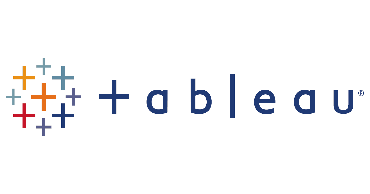
**TECHNOLOGY USED FOR PREDICTION**

**For Coding:**

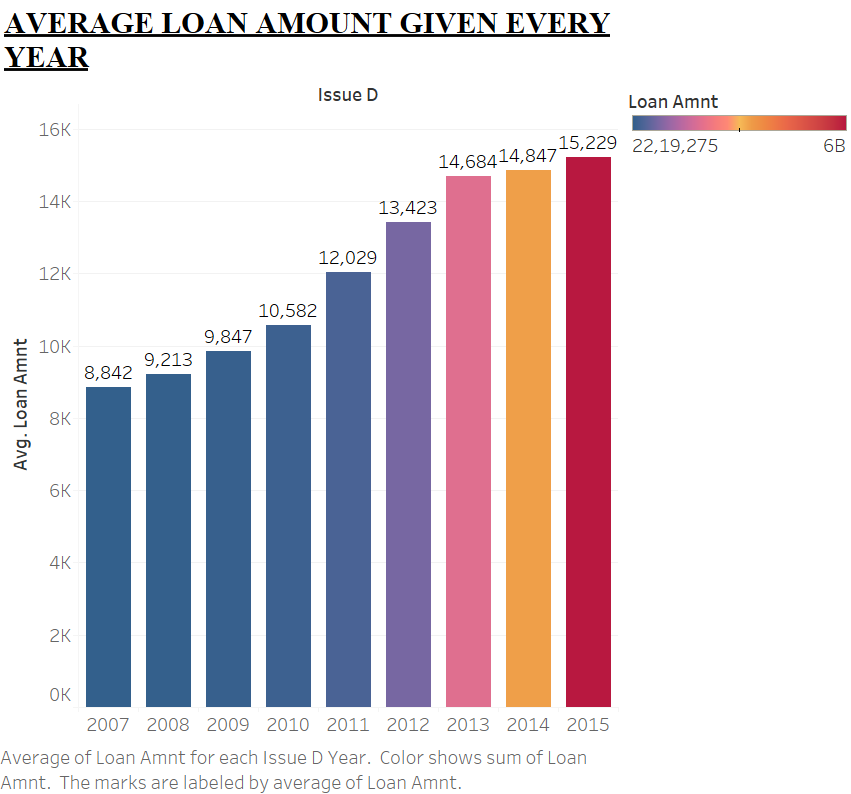
**Anaconda Navigator Jupyter Notebook:**

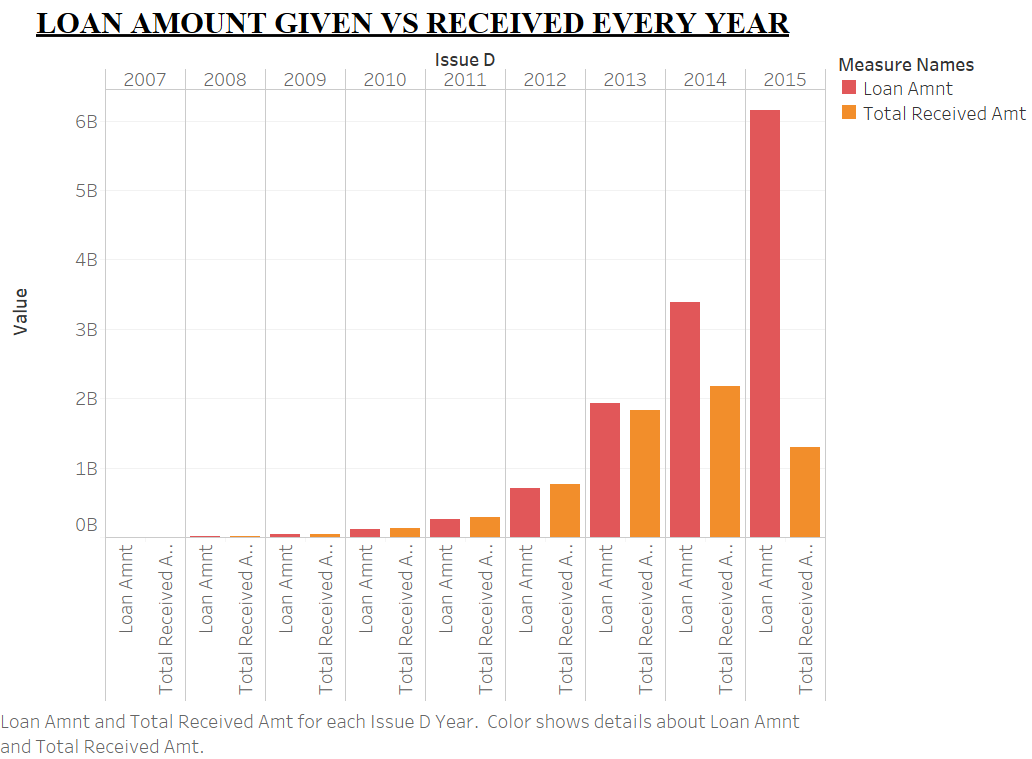
  

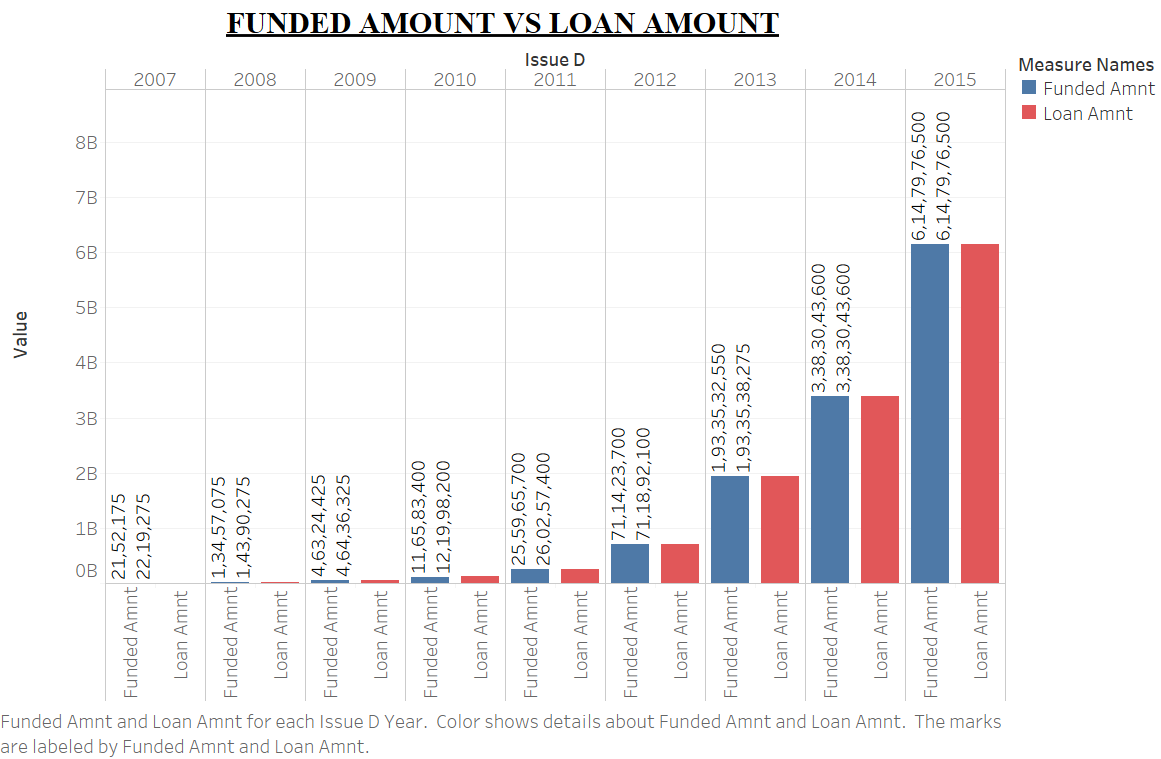
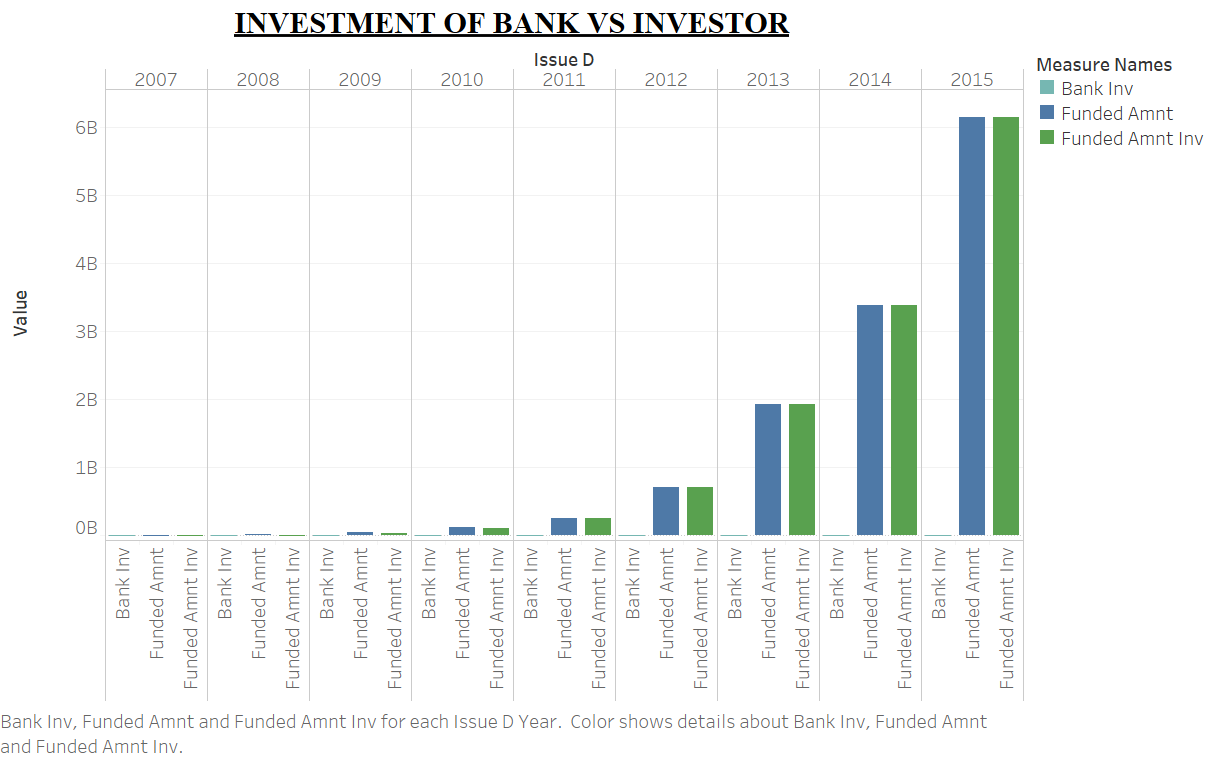
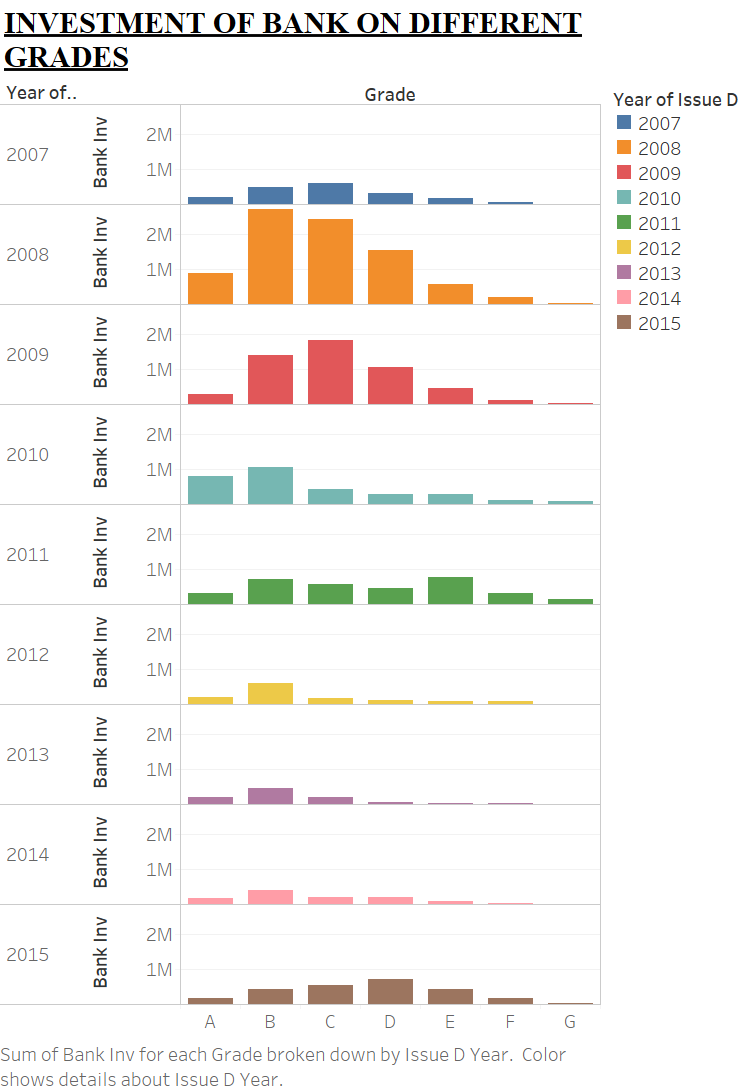
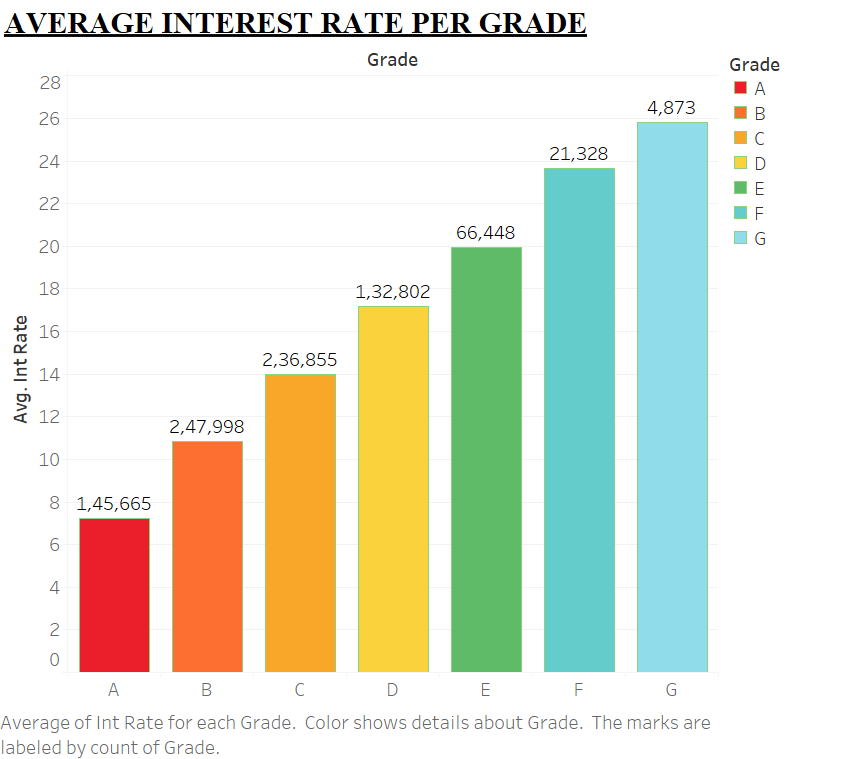
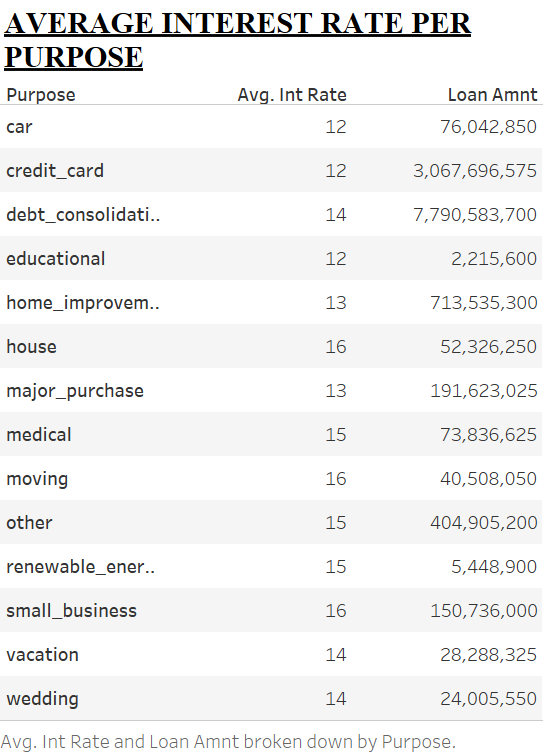
**For Visualisation:**

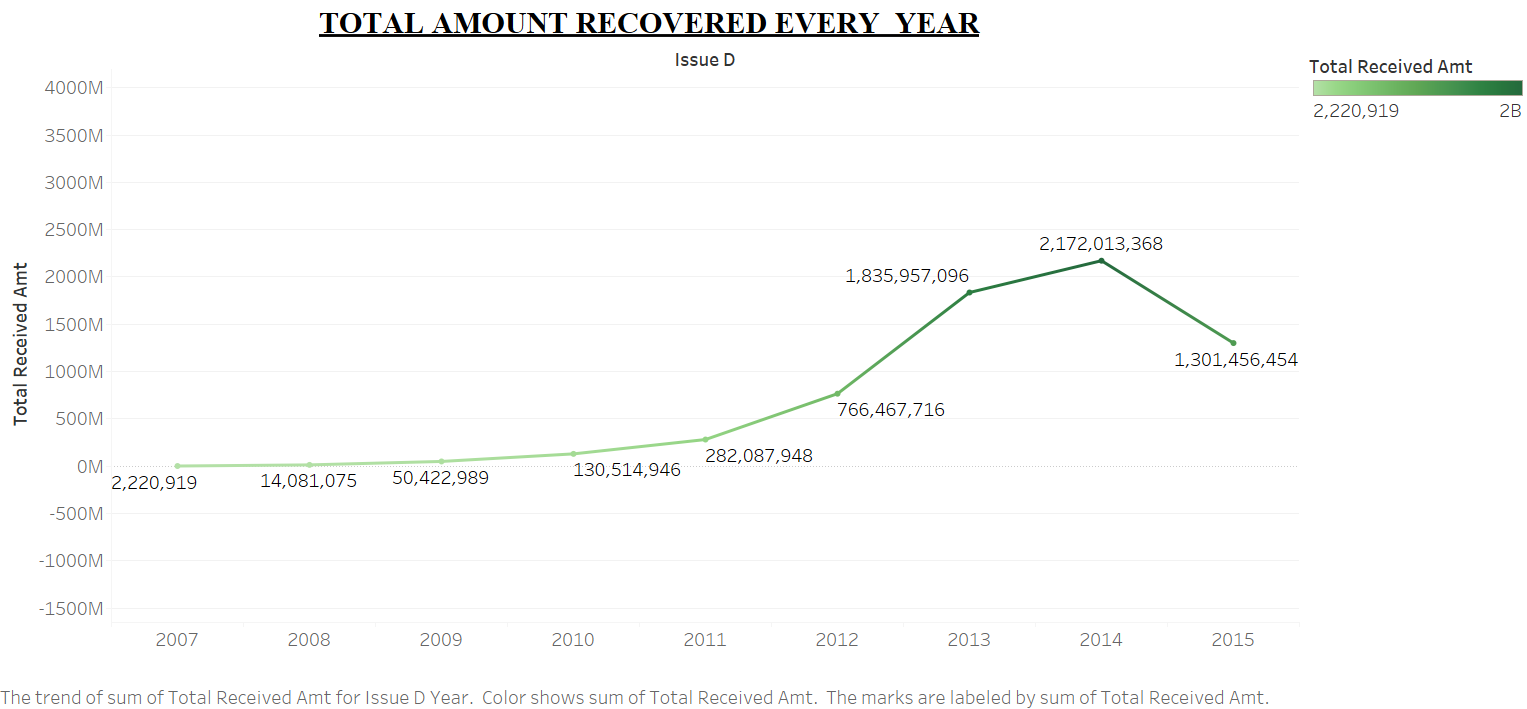
 

**VISUALISATION**











**MODELS DESCRIPTION**

**MODEL 1 – LOGISTIC REGRESSION**

*Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).*

**MODEL 2 – DECISION TREE**

*Decision Tree Classifier is a simple and widely used classification technique. It applies a straight forward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. The decision tree classifiers organized a series of test questions and conditions in a tree structure.*

**MODEL 3 – ADAPTIVE BOOST CLASSIFIER (ADABOOST)**

*Boosting is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers.* [*AdaBoost*](https://en.wikipedia.org/wiki/AdaBoost)*was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting.*

**MODEL 4 – EXTREMELY RANDOMIZED TREES (EXTRA TREES CLASSIFIER)**

*An “extra trees” classifier, otherwise known as an “Extremely randomized trees” classifier, is a variant of a random forest. Unlike a random forest, at each step the entire sample is used and decision boundaries are picked at random, rather than the best one. In real world cases, performance is comparable to an ordinary random forest, sometimes a bit better.*

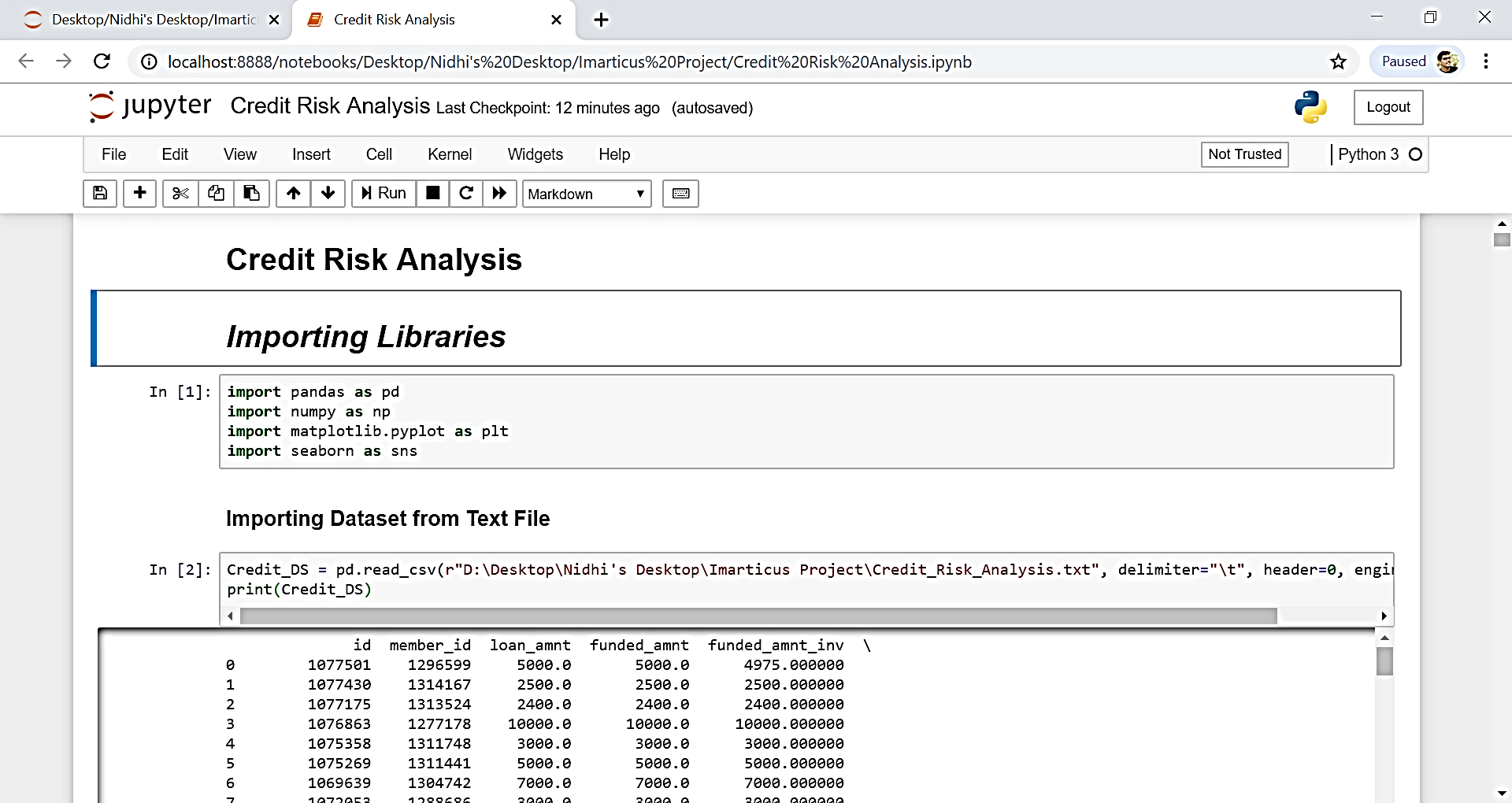
**CROSS VALIDATION TECHNIQUE**

**K-FOLD CROSS VALIDATION TECHNIQUE**

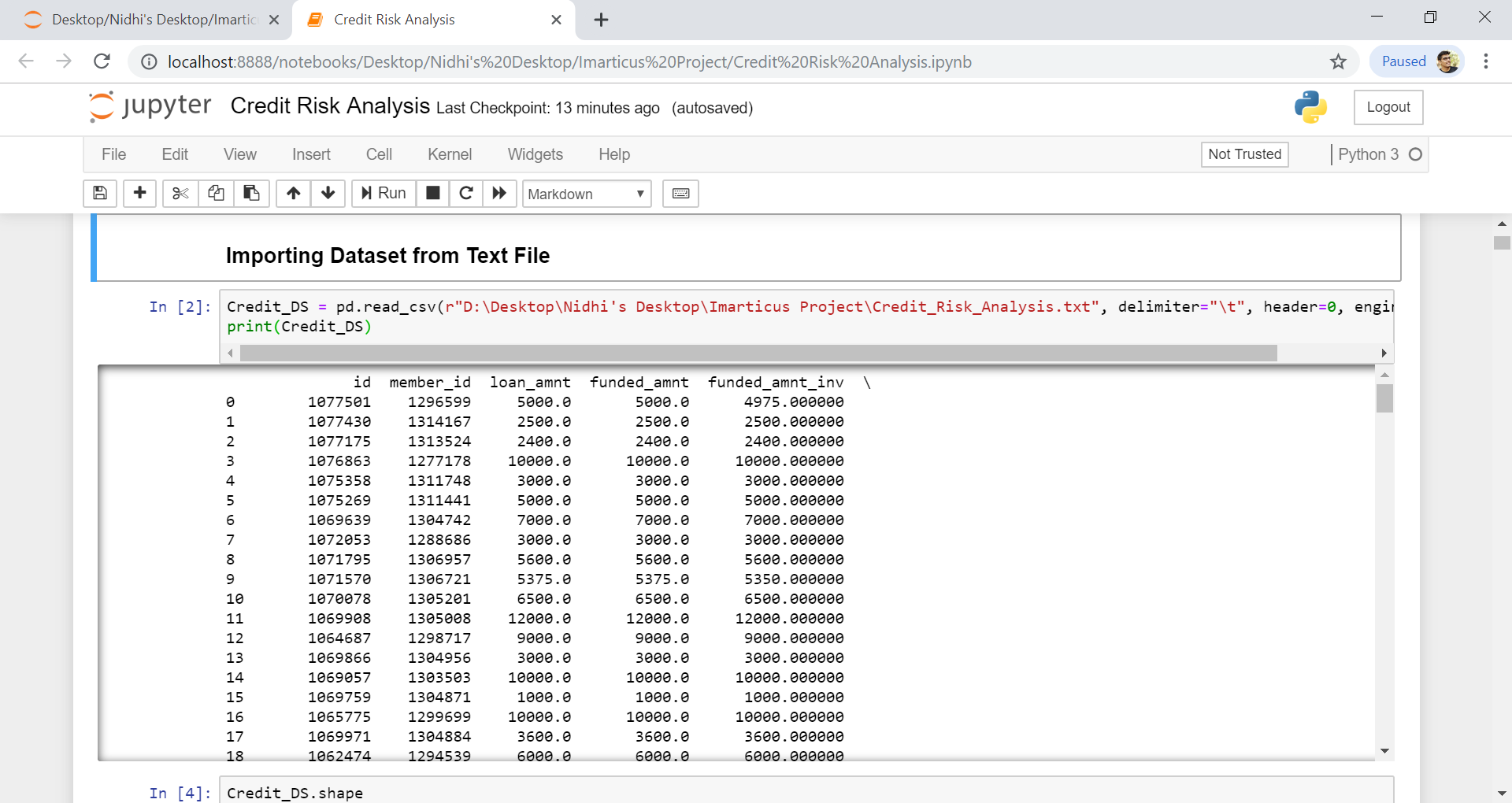
*Cross-validation is a statistical method used to estimate the skill of machine learning models .It is commonly used in applied machine learning to compare and select a model for a given predictive modelling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.*

**DATA EXTRACTION AND CLEANING**

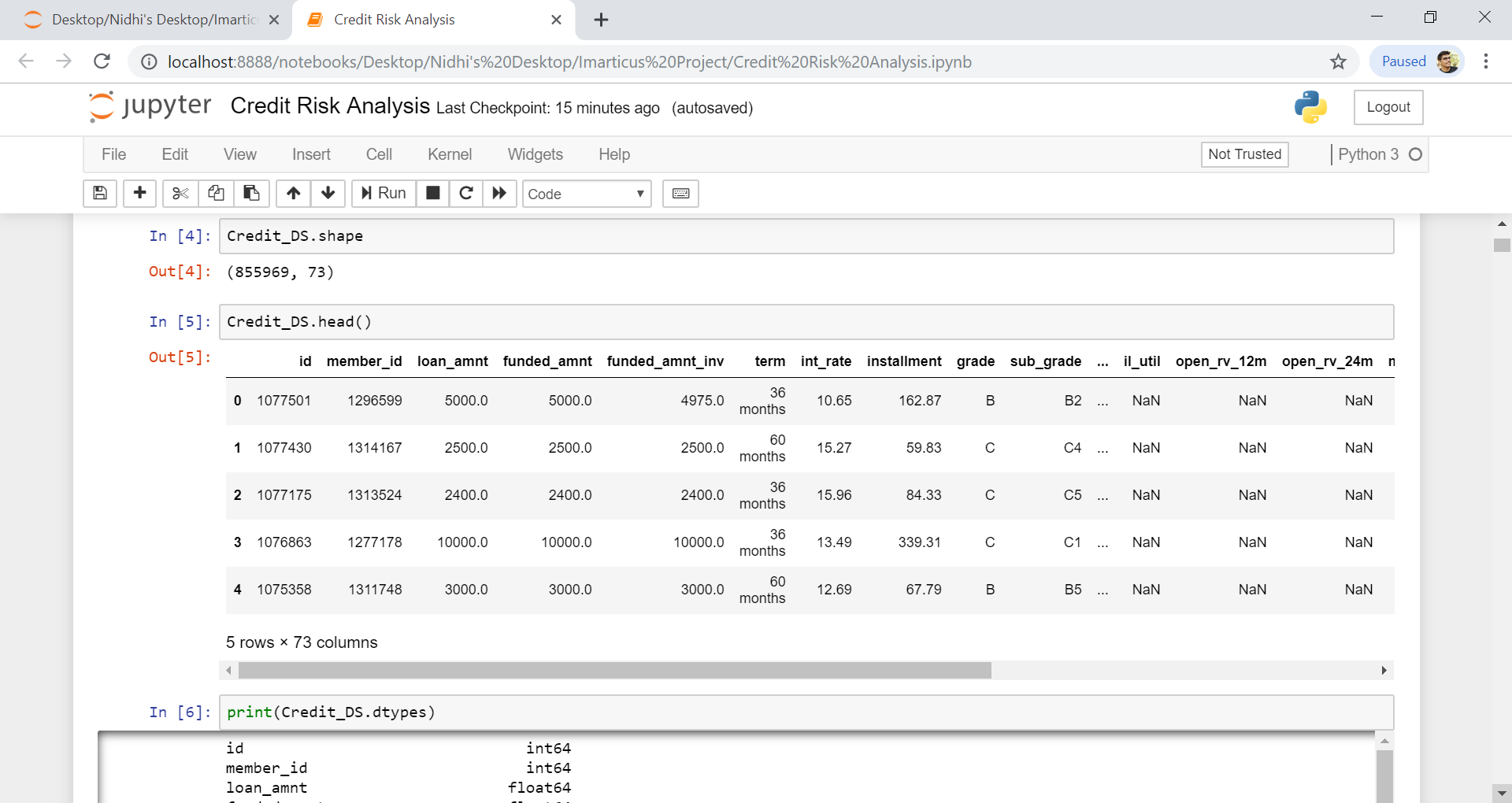
1. Importing the Libraries



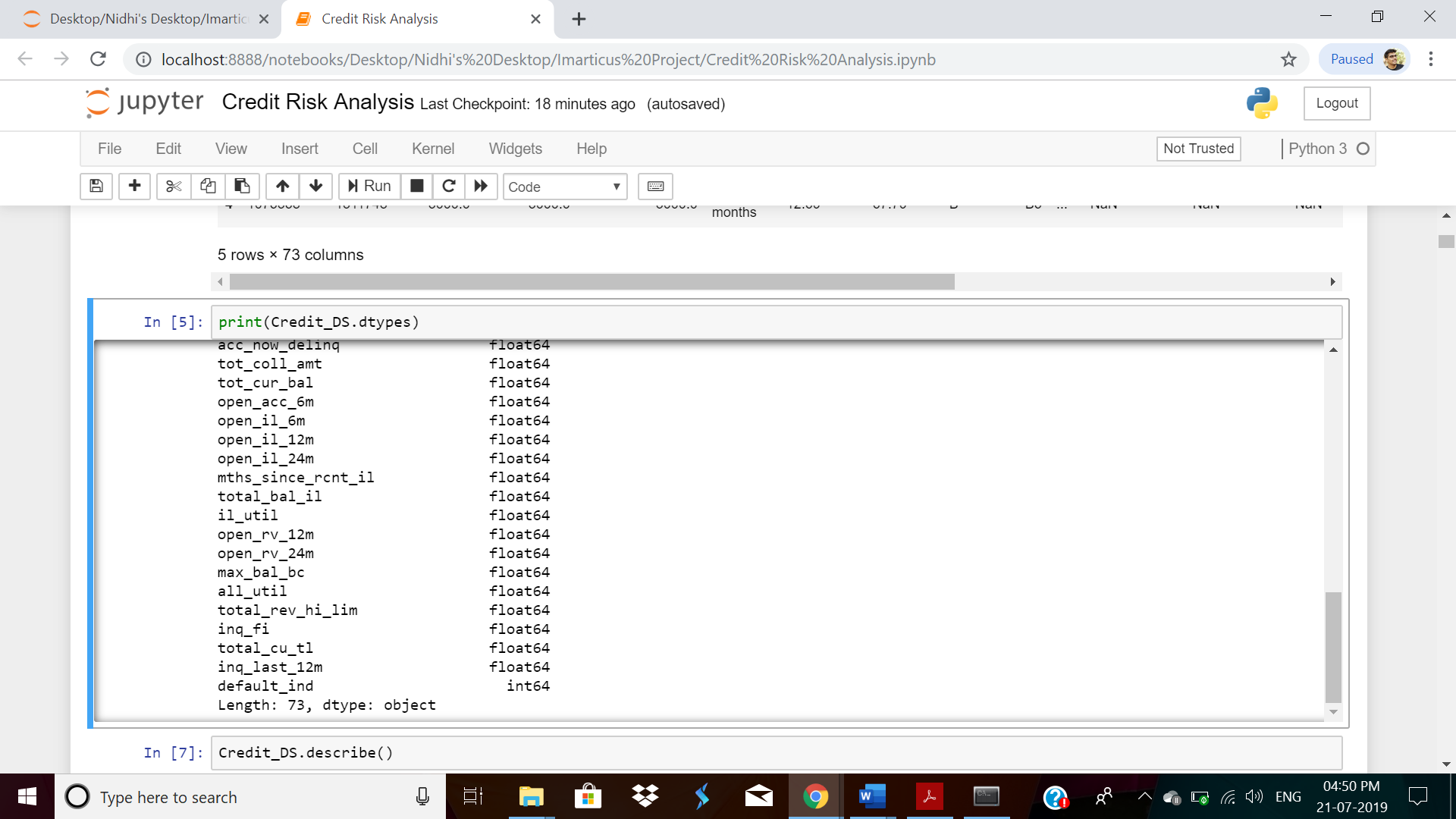
2. Importing the Dataset



3. View the Shape and head of the Data

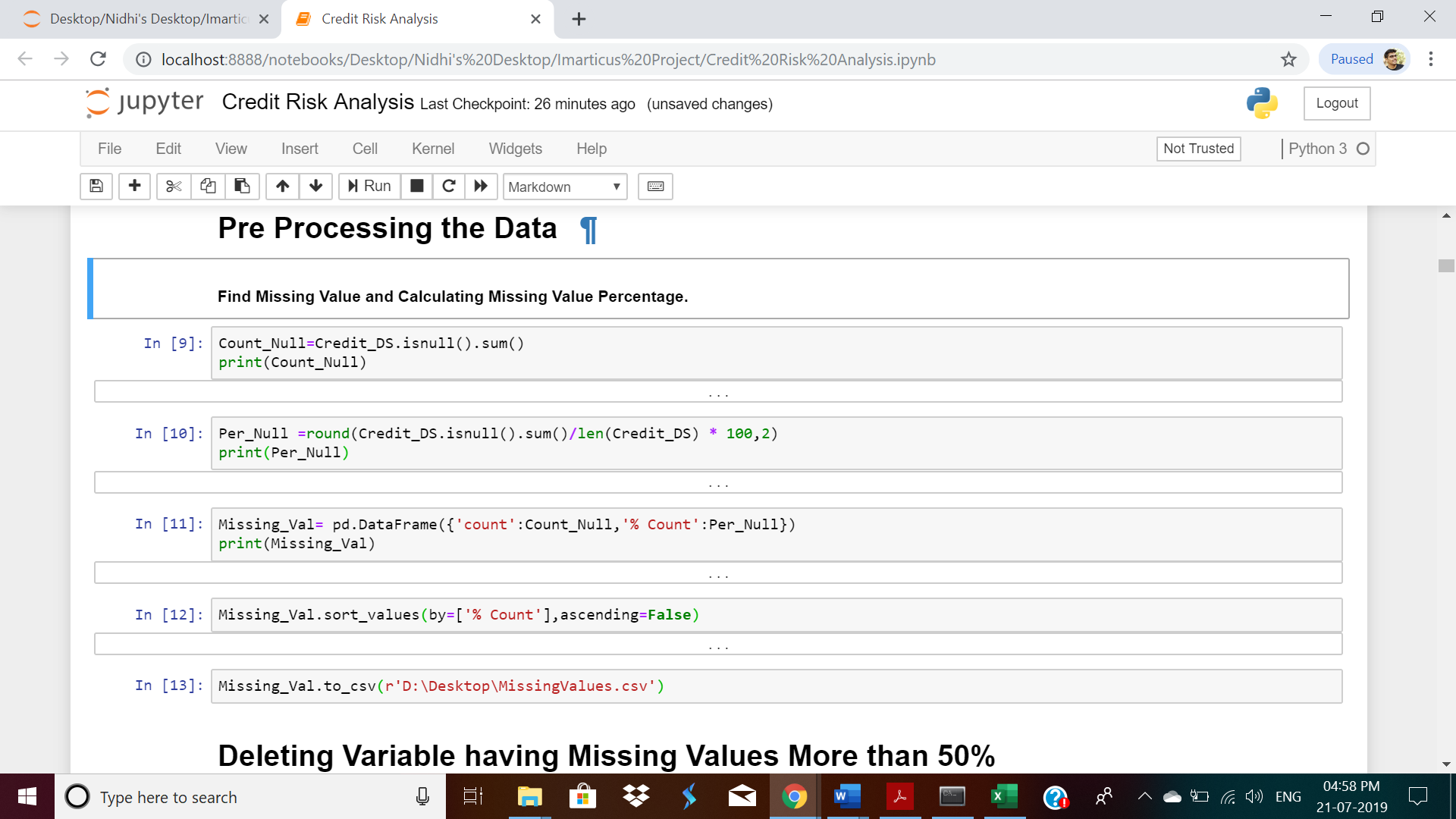


4. Datatype of Datasets



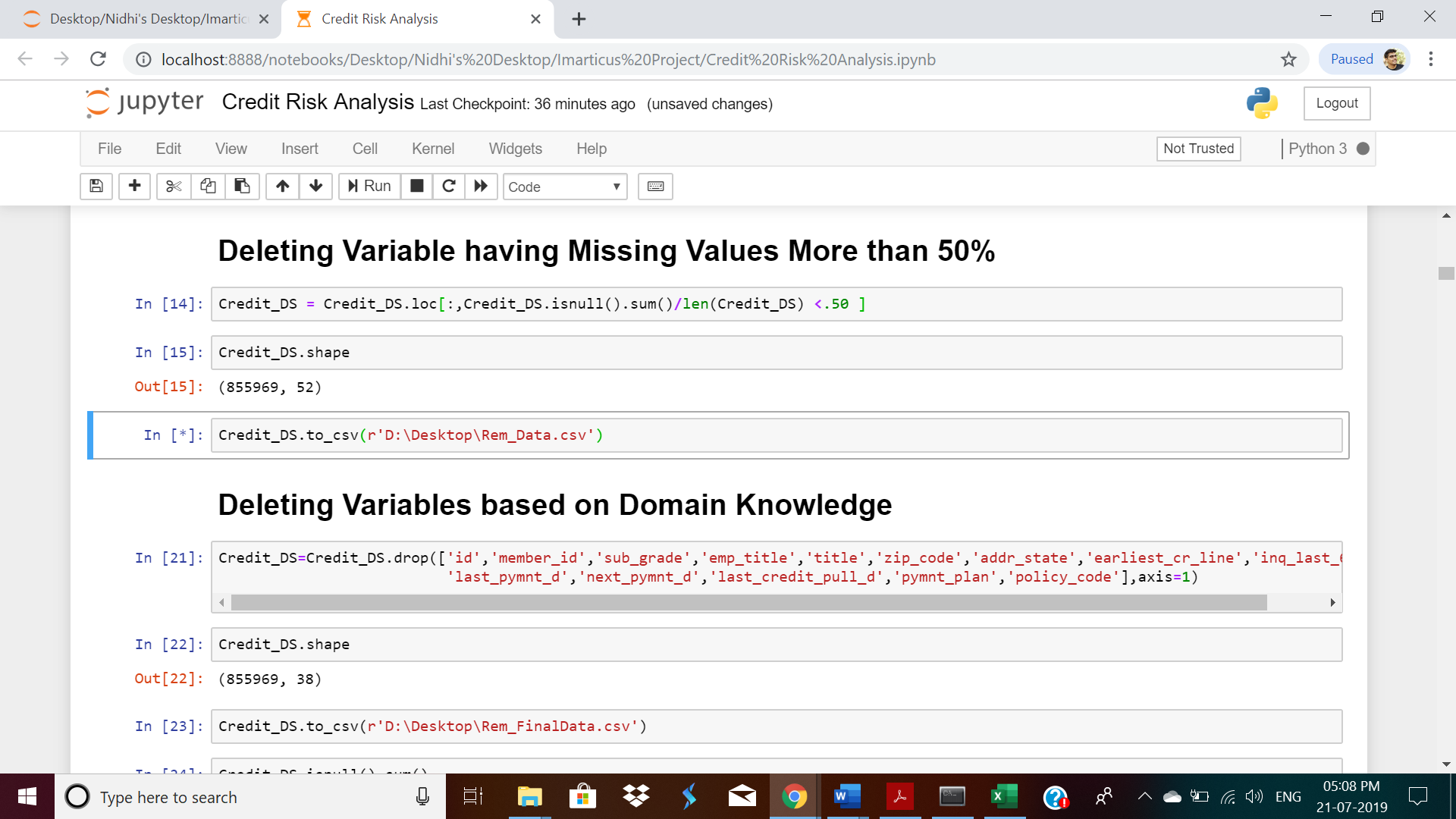
**Treating Missing Values**

5. Checking Missing Value and Deleting all the Column having Missing Value more than 50%

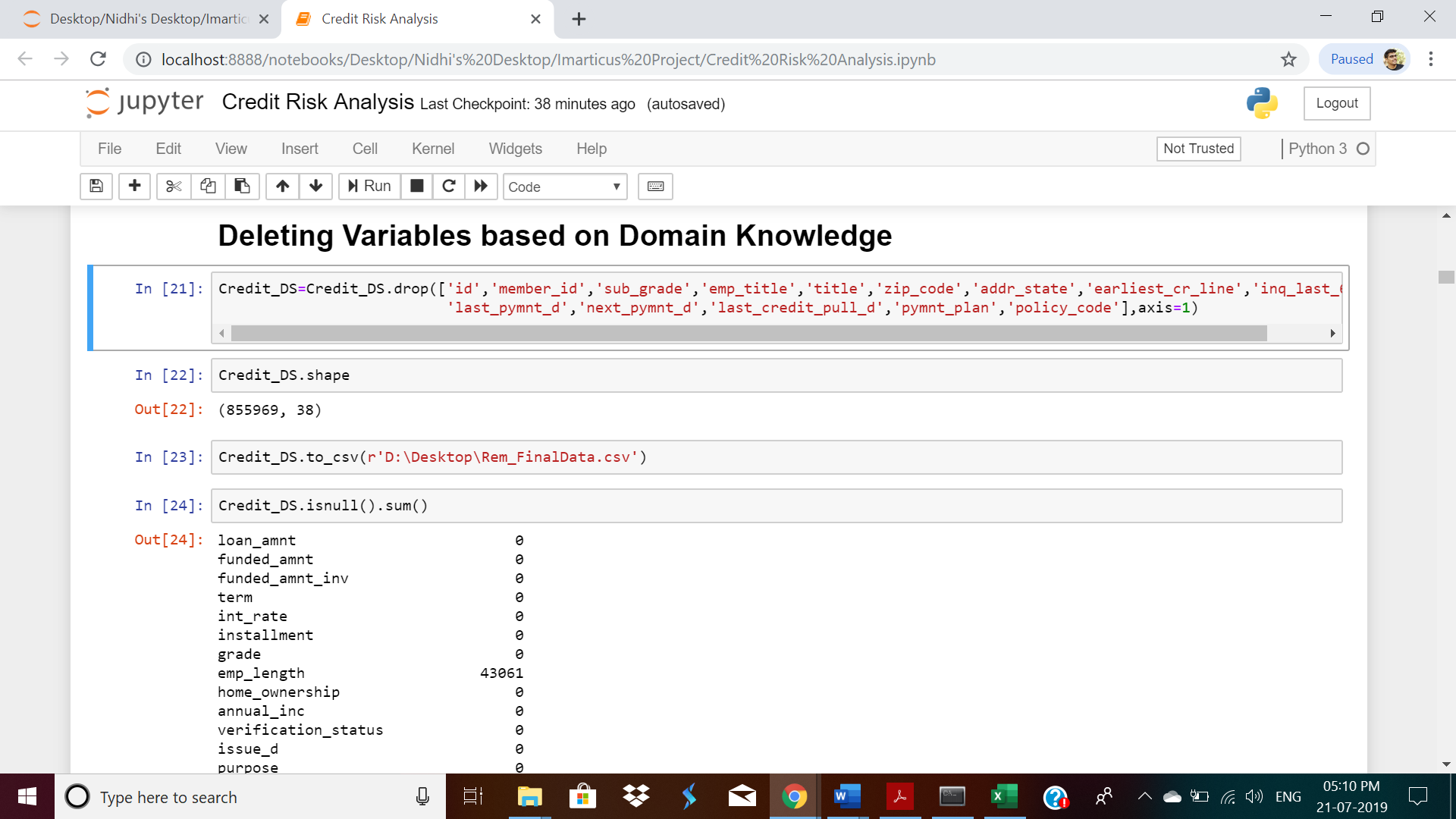


**Output:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **count** | **% Count** |
| annual\_inc\_joint | 855527 | 99.95 |
| dti\_joint | 855529 | 99.95 |
| verification\_status\_joint | 855527 | 99.95 |
| il\_util | 844360 | 98.64 |
| mths\_since\_rcnt\_il | 843035 | 98.49 |
| open\_acc\_6m | 842681 | 98.45 |
| open\_il\_6m | 842681 | 98.45 |
| open\_il\_12m | 842681 | 98.45 |
| open\_il\_24m | 842681 | 98.45 |
| total\_bal\_il | 842681 | 98.45 |
| open\_rv\_12m | 842681 | 98.45 |
| open\_rv\_24m | 842681 | 98.45 |
| max\_bal\_bc | 842681 | 98.45 |
| all\_util | 842681 | 98.45 |
| inq\_fi | 842681 | 98.45 |
| total\_cu\_tl | 842681 | 98.45 |
| inq\_last\_12m | 842681 | 98.45 |
| Desc | 734157 | 85.77 |
| mths\_since\_last\_record | 724785 | 84.67 |
| mths\_since\_last\_major\_derog | 642830 | 75.1 |
| mths\_since\_last\_delinq | 439812 | 51.38 |
| next\_pymnt\_d | 252971 | 29.55 |
| tot\_coll\_amt | 67313 | 7.86 |
| tot\_cur\_bal | 67313 | 7.86 |
| total\_rev\_hi\_lim | 67313 | 7.86 |
| emp\_title | 49443 | 5.78 |
| emp\_length | 43061 | 5.03 |
| last\_pymnt\_d | 8862 | 1.04 |
| revol\_util | 446 | 0.05 |
| last\_credit\_pull\_d | 50 | 0.01 |
| collections\_12\_mths\_ex\_med | 56 | 0.01 |
| Id | 0 | 0 |
| member\_id | 0 | 0 |
| loan\_amnt | 0 | 0 |
| funded\_amnt | 0 | 0 |
| funded\_amnt\_inv | 0 | 0 |
| term | 0 | 0 |
| int\_rate | 0 | 0 |
| installment | 0 | 0 |
| grade | 0 | 0 |
| sub\_grade | 0 | 0 |
| home\_ownership | 0 | 0 |
| annual\_inc | 0 | 0 |
| verification\_status | 0 | 0 |
| issue\_d | 0 | 0 |
| pymnt\_plan | 0 | 0 |
| purpose | 0 | 0 |
| title | 33 | 0 |
| zip\_code | 0 | 0 |
| addr\_state | 0 | 0 |
| dti | 0 | 0 |
| delinq\_2yrs | 0 | 0 |
| earliest\_cr\_line | 0 | 0 |
| inq\_last\_6mths | 0 | 0 |
| open\_acc | 0 | 0 |
| pub\_rec | 0 | 0 |
| revol\_bal | 0 | 0 |
| total\_acc | 0 | 0 |
| initial\_list\_status | 0 | 0 |
| out\_prncp | 0 | 0 |
| out\_prncp\_inv | 0 | 0 |
| total\_pymnt | 0 | 0 |
| total\_pymnt\_inv | 0 | 0 |
| total\_rec\_prncp | 0 | 0 |
| total\_rec\_int | 0 | 0 |
| total\_rec\_late\_fee | 0 | 0 |
| recoveries | 0 | 0 |
| collection\_recovery\_fee | 0 | 0 |
| last\_pymnt\_amnt | 0 | 0 |
| policy\_code | 0 | 0 |
| application\_type | 0 | 0 |
| acc\_now\_delinq | 0 | 0 |
| default\_ind | 0 | 0 |



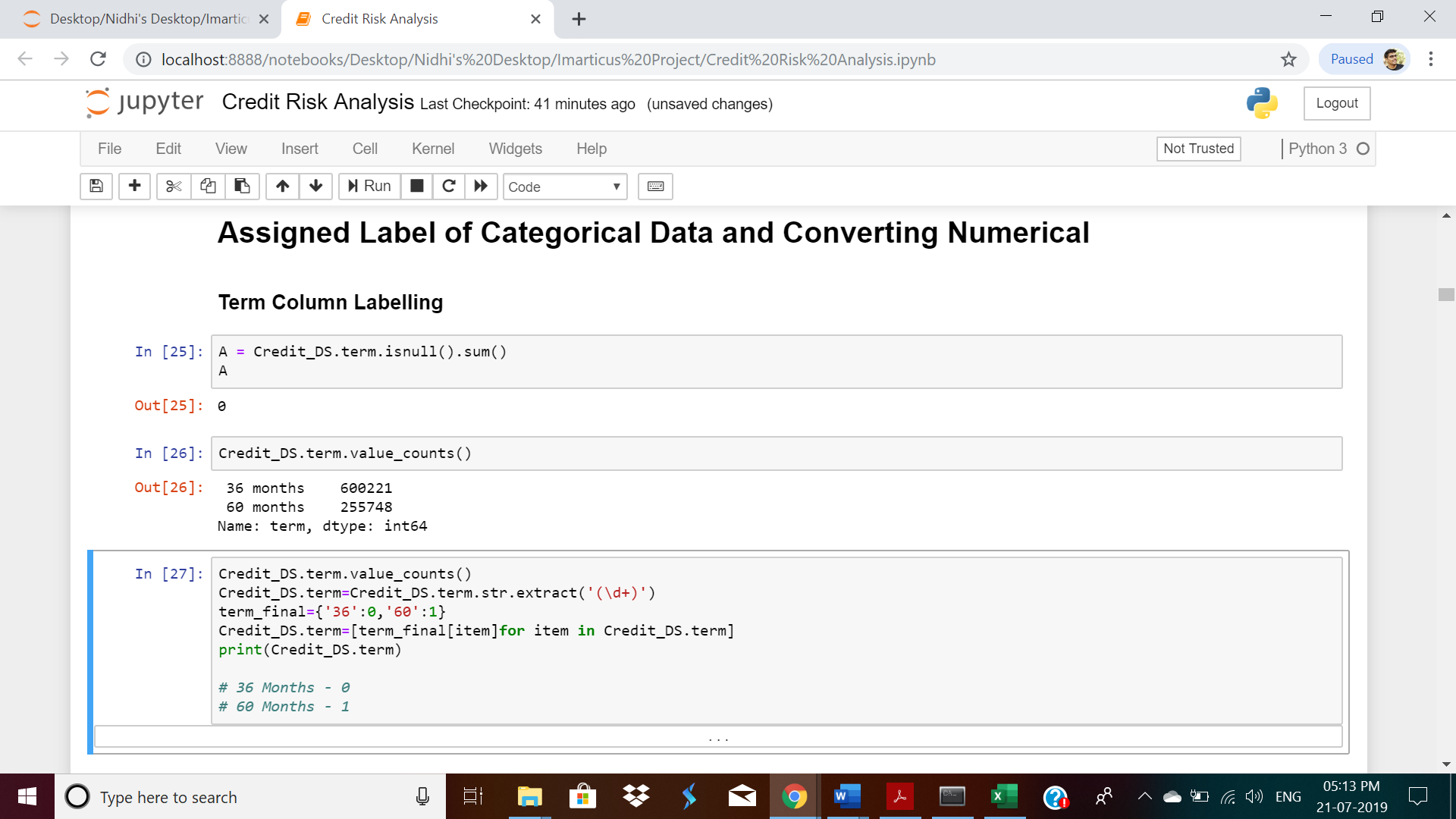
6. Deleting Variable based on Domain Knowledge



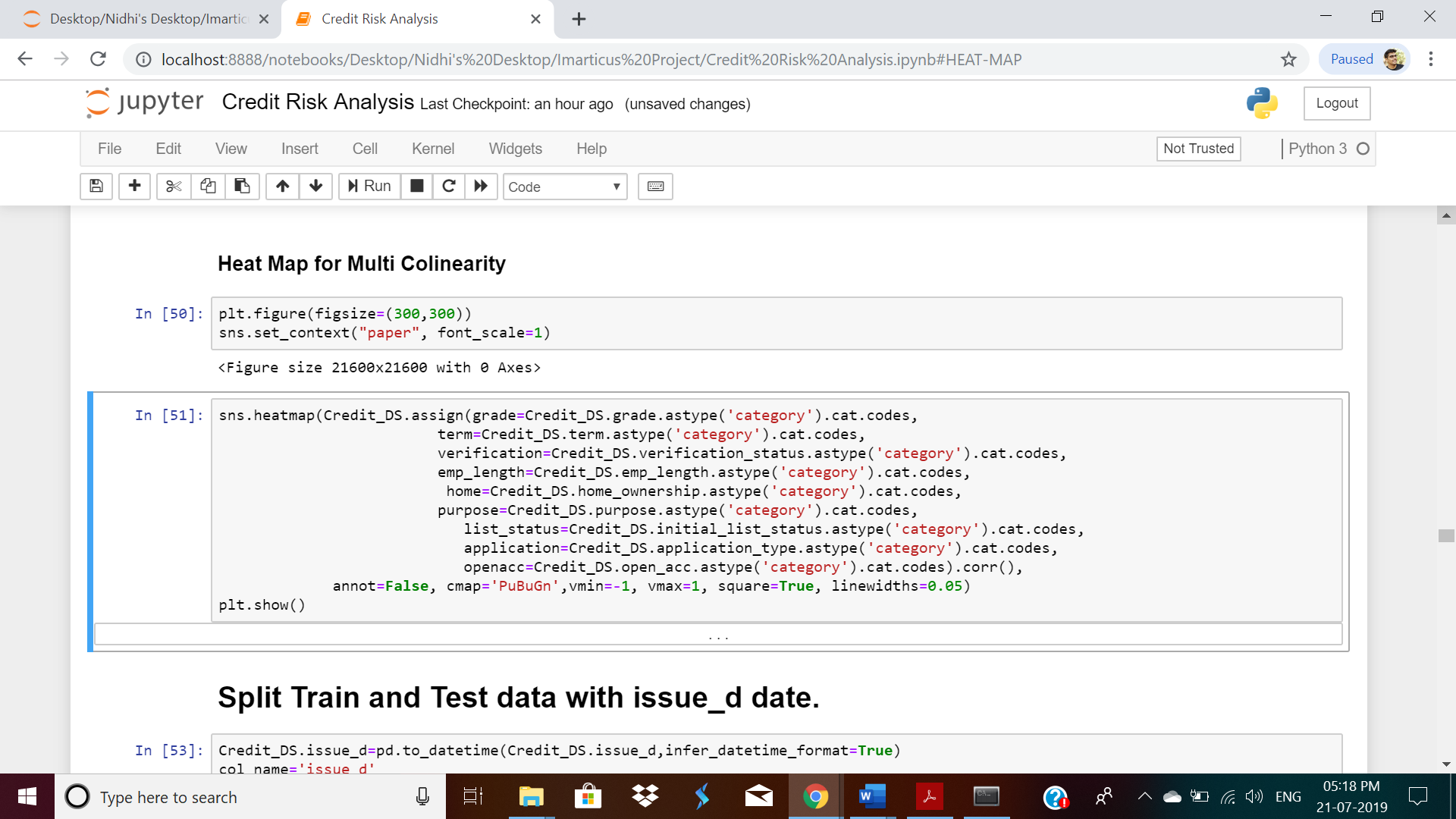
7. Treating Missing Value of Remaining variable and Converting Categorial Variable to

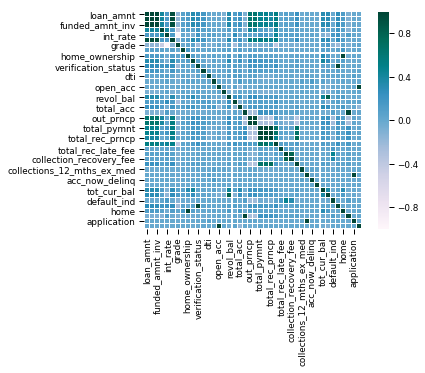
Numeric by Label Encoding Method.

For Example:



8. To Check there is no Much Multi – Collinearity using Heat Map

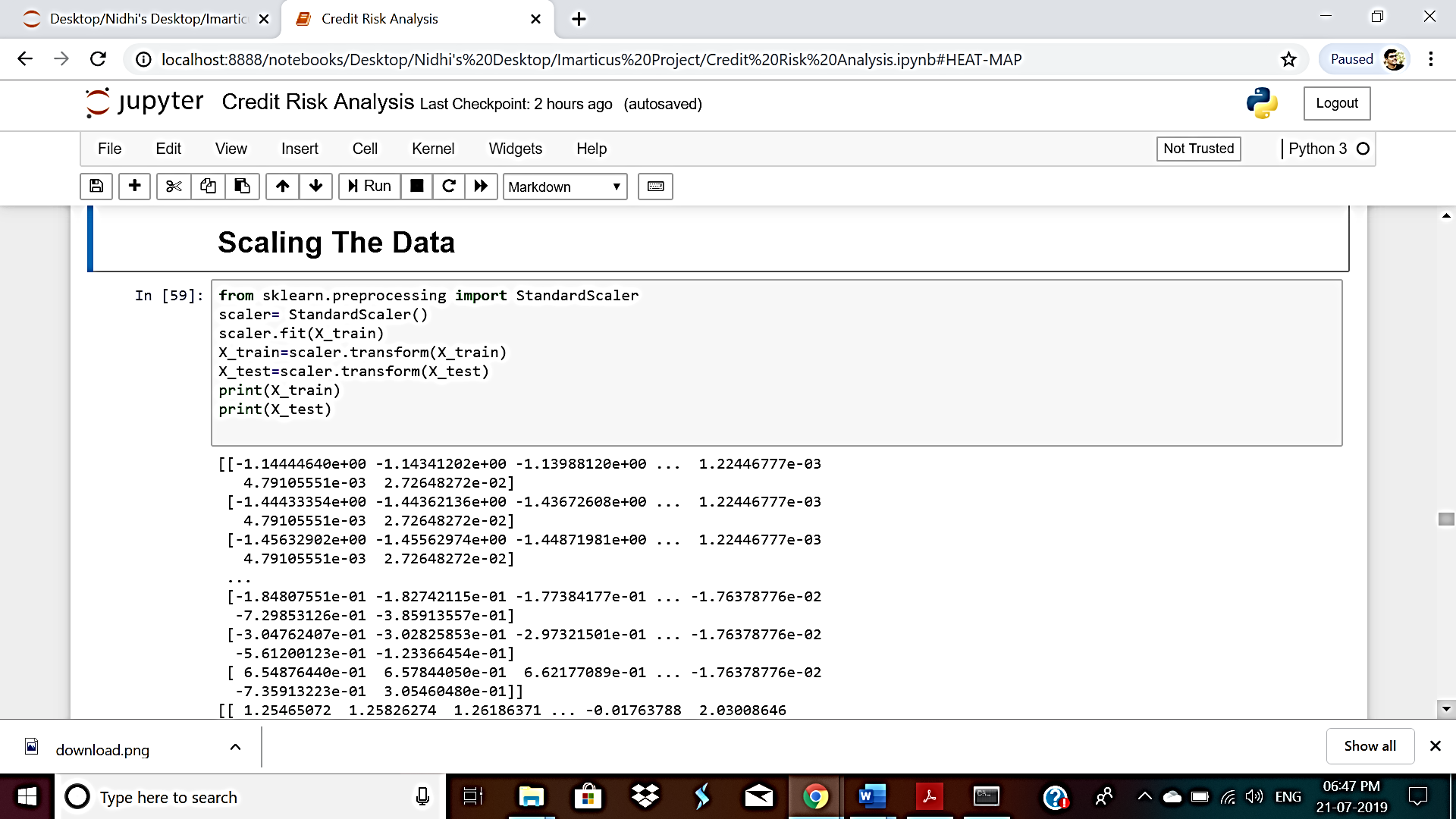




9. Splitting the Data into Testing and Training Data using issue\_d Column

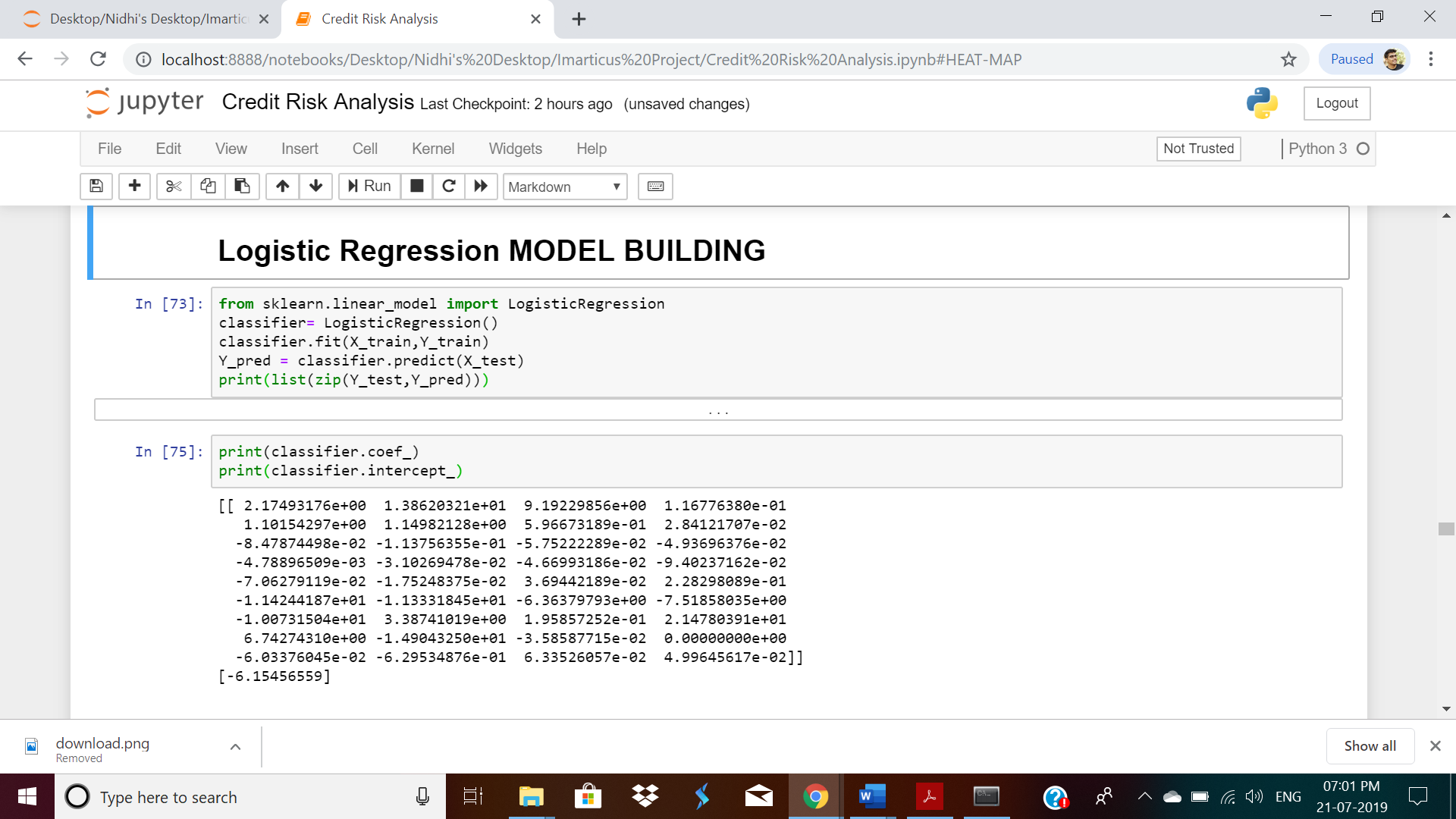


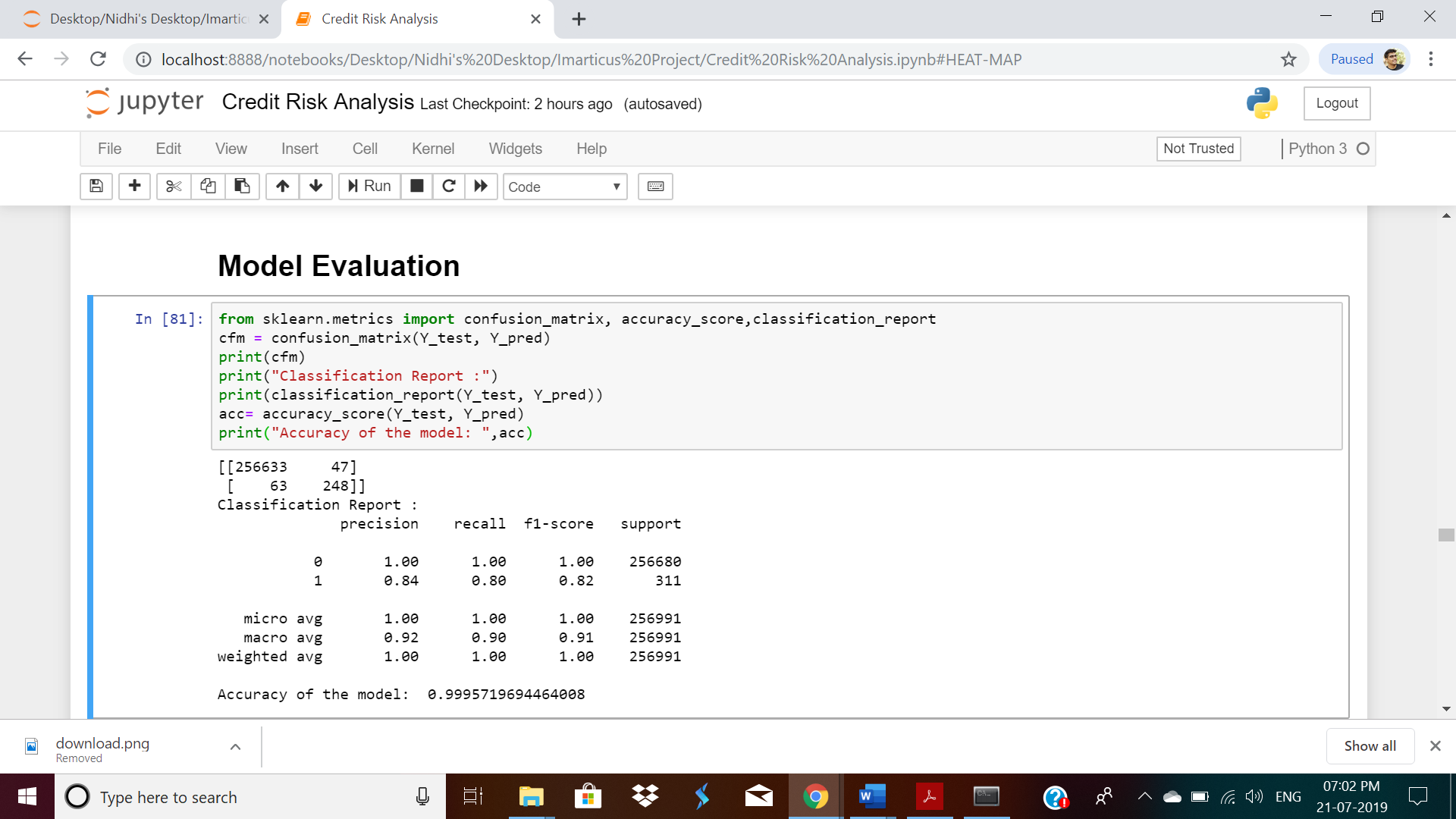
10. Scaling the Data

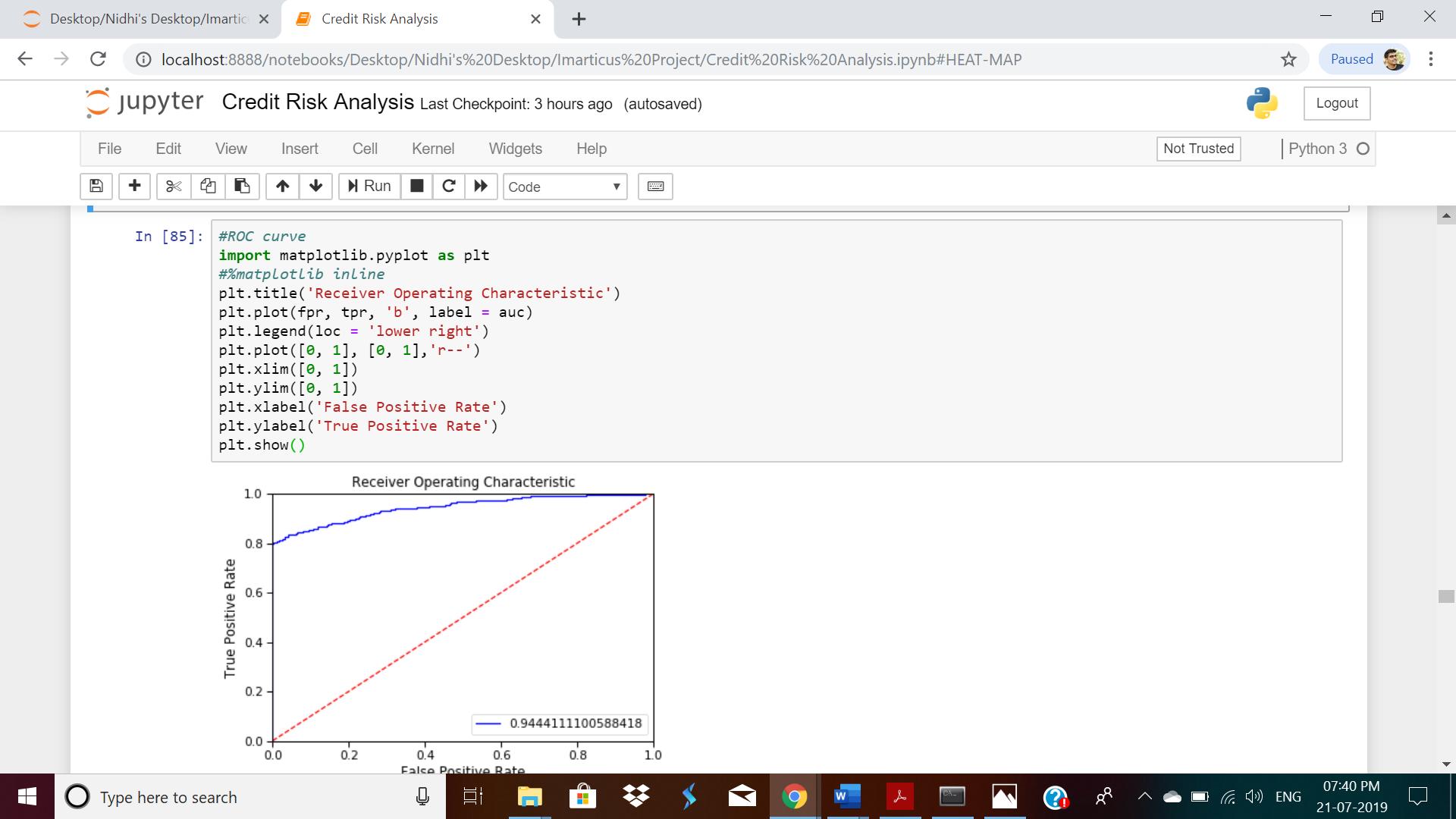


**BUILDING MULTIPLE MODELS**

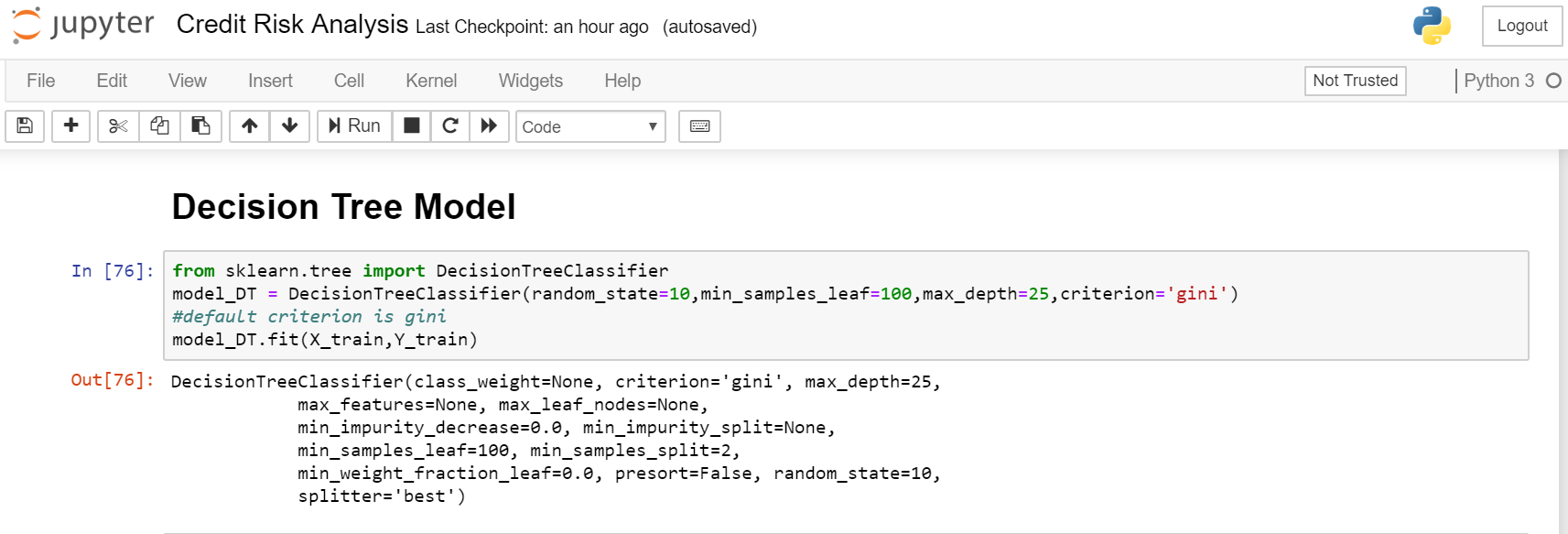
**Model – 1 – Logistic Regression**

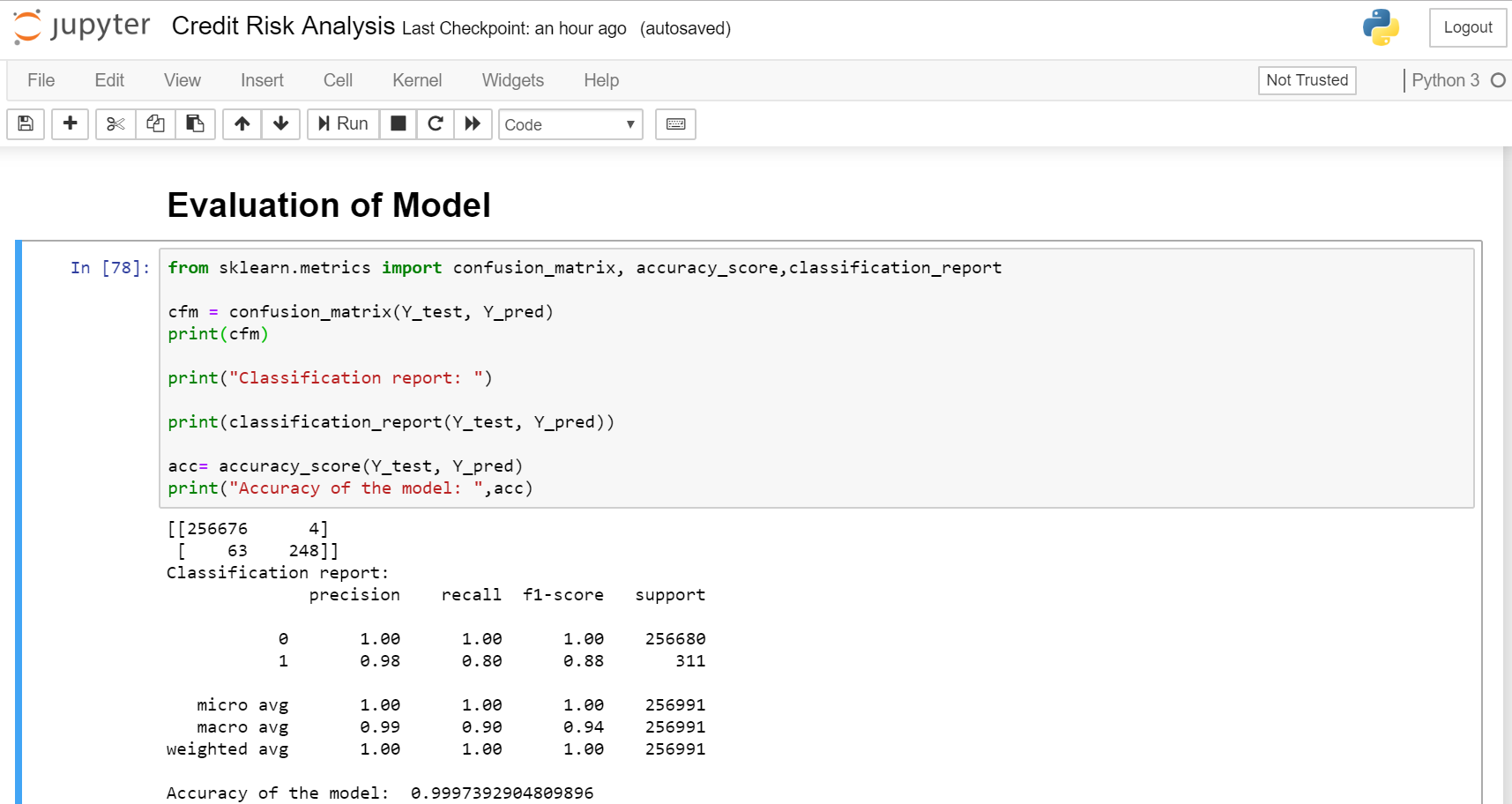


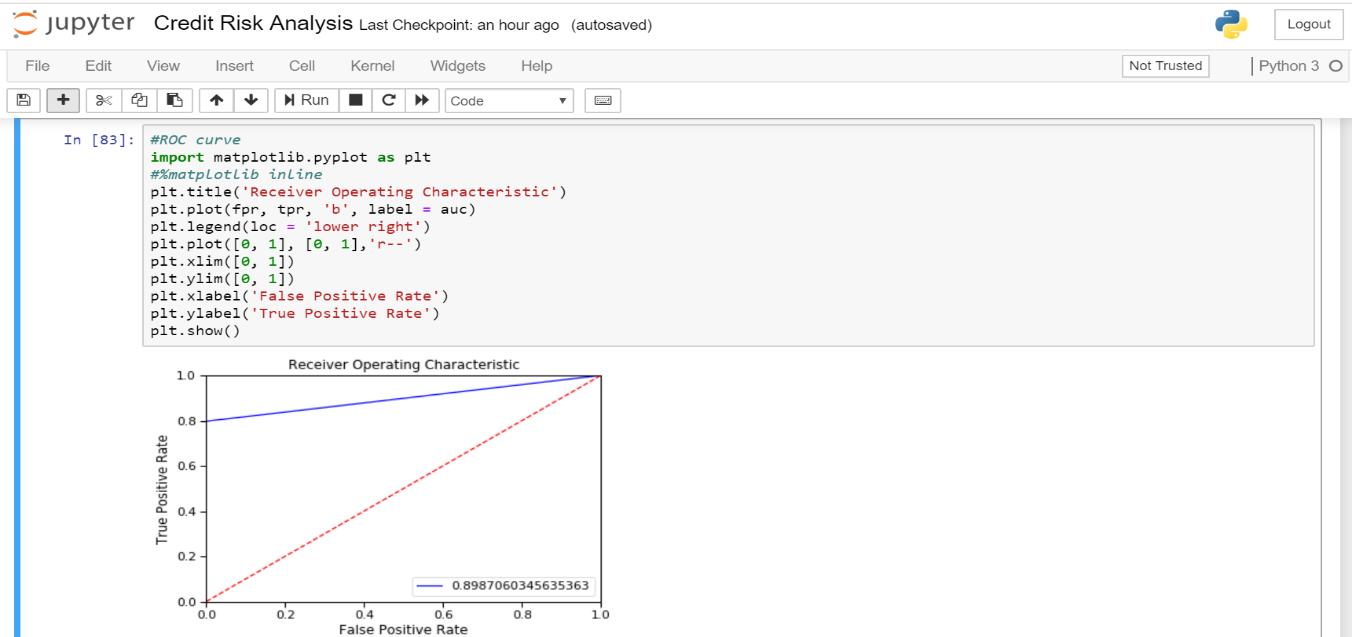




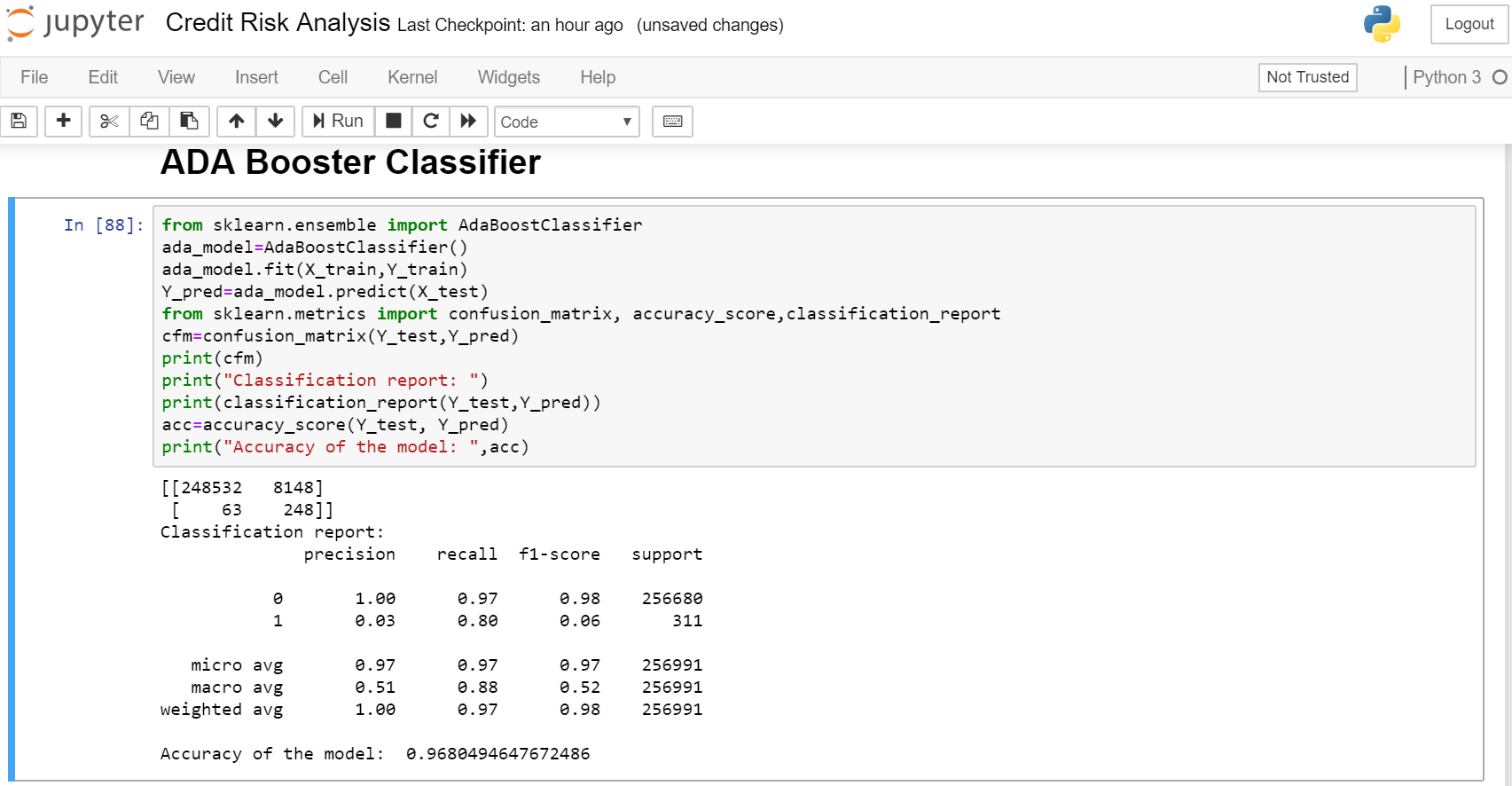
**Model – 2 – Decision Tree**

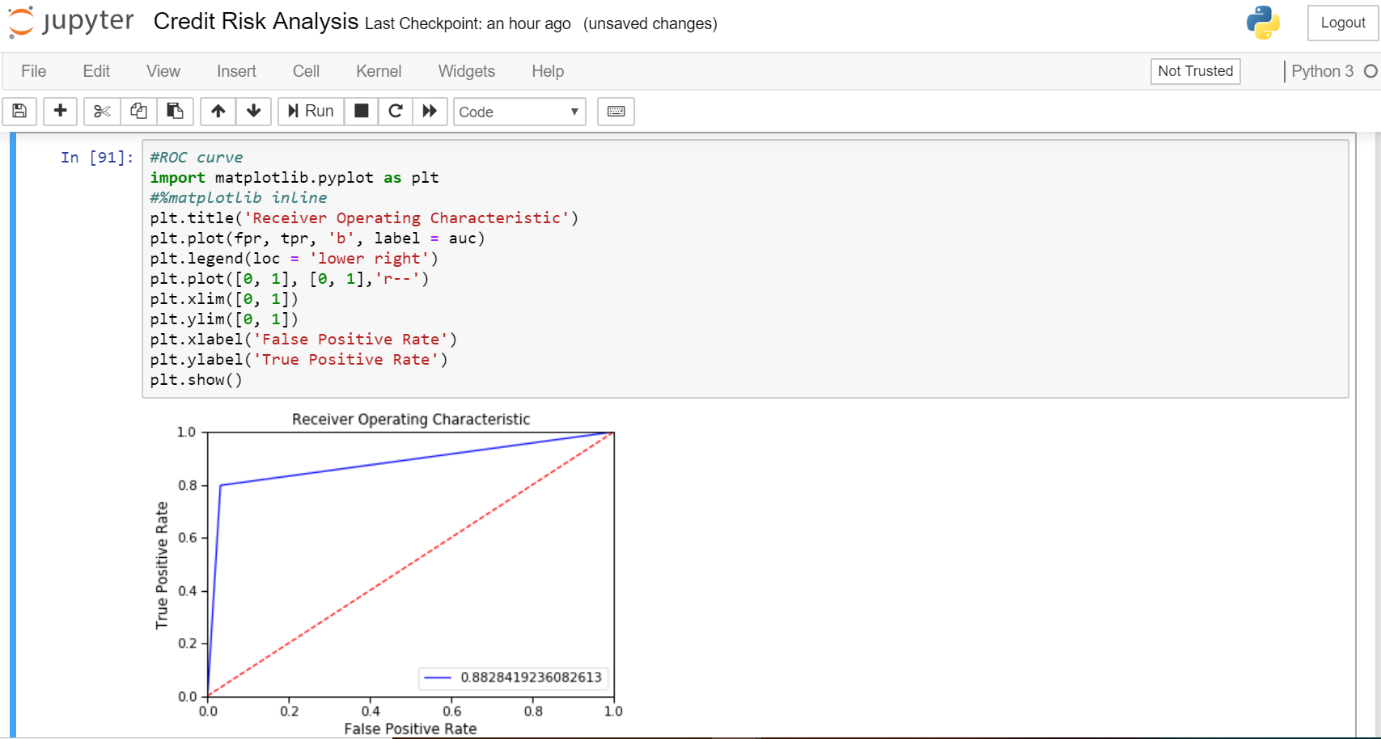






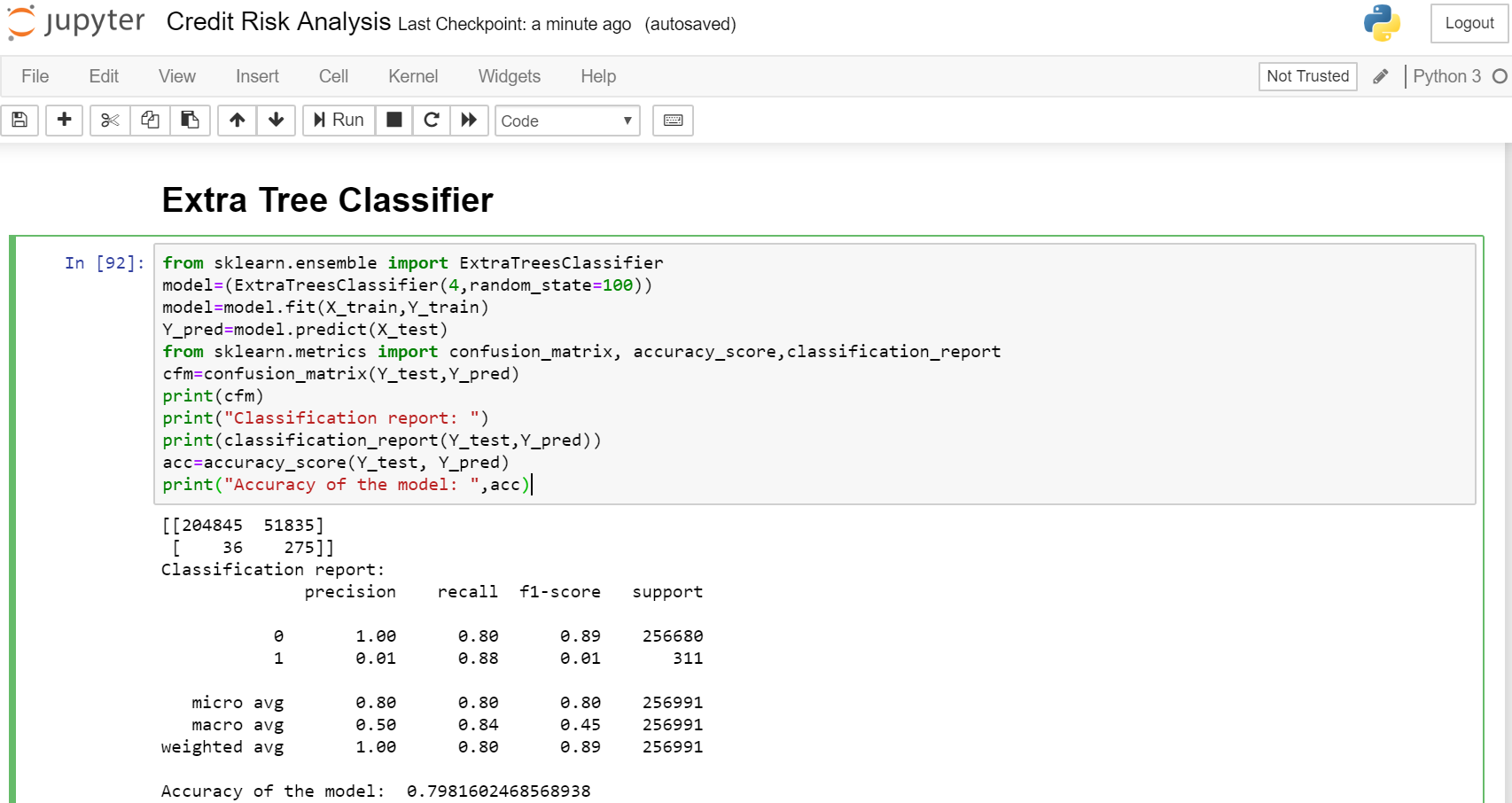
**MODEL 3 – ADAPTIVE BOOST (ADABOOST)**

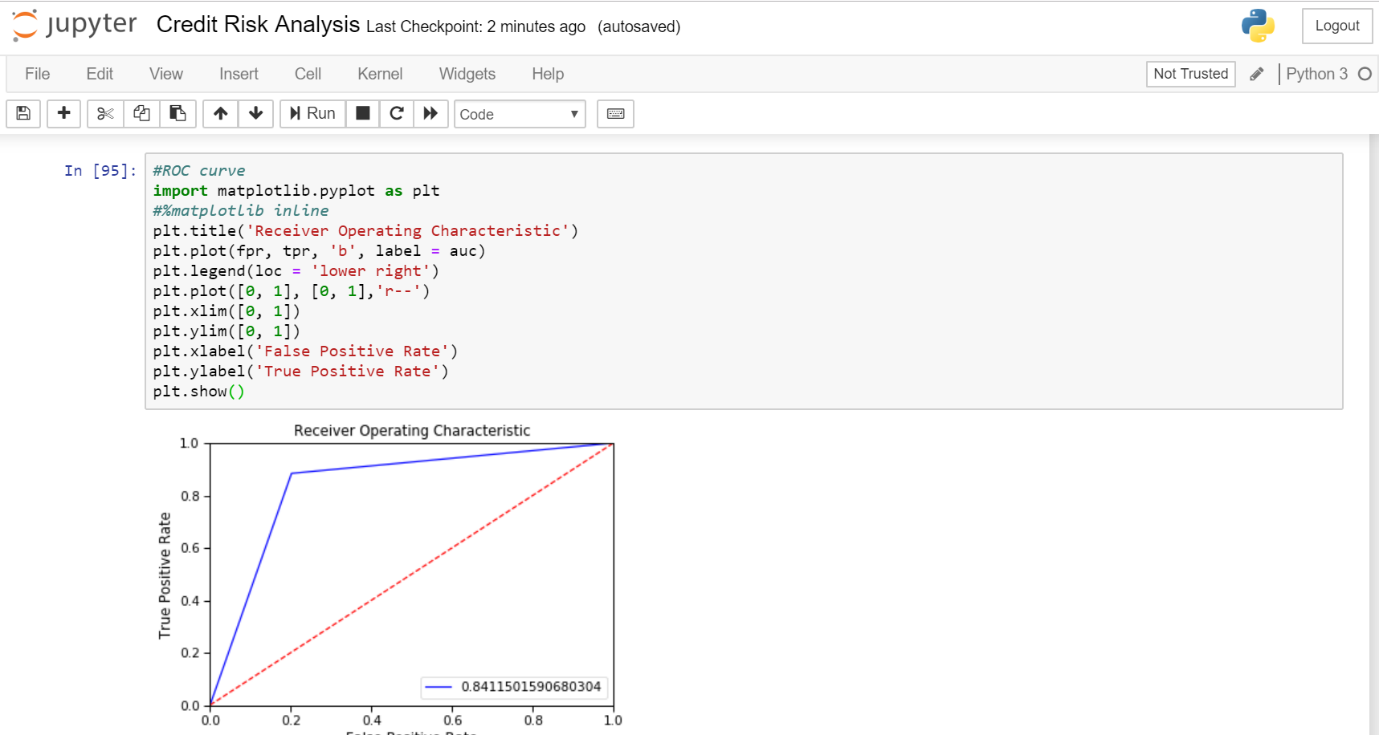




**MODEL 4 – EXTREMELY RANDOMIZED TREES**

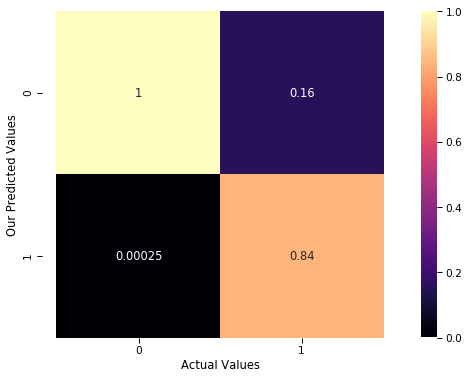
**(EXTRA TREES CLASSIFIER)**



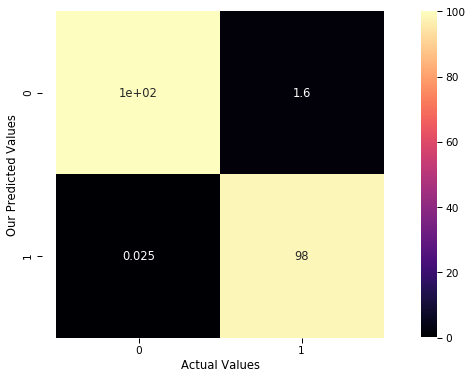


CONFUSION MATRIX

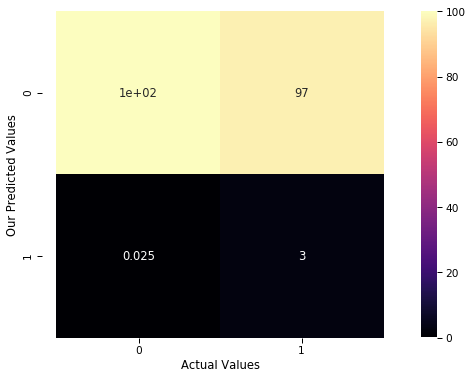
Model 1 – Logistic Regression



Model 2 – Decision Tree



Model 3 – Ada Boost Classifier



Model 4 – Extra Trees Classifier



**COMPARISION OF AUC, FPR, TPR**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr.no | Model Name | AUC | Sensitivity | Specificity | Accuracy |
| 1 | Logistic Regression | 94.44% | 0.797428 | 0.999817 | 99.96% |
| 2 | Decision Tree | 89.87% | 0.797428 | 0.999984 | 99.97% |
| 3 | Ada Boost Classifier | 88.28% | 0.797428 | 0.968256 | 96.80% |
| 4 | Extra Trees Classifier | 84.11% | 0.884244 | 0.798056 | 79.82% |

**CONCLUSION**

Credit Risk Analysis is a very crucial part of the banking sector and it plays an important role in the growth of the bank’s profit. Using analysing techniques one can predict or analyse that a person applying for loan will repay the loan or not. So, multiples algorithms have been implemented to analyse a defaulter. The best model selected out of all models that have been tested is Logistic Regression model with an accuracy of 99.957%, which is cross verified with K-fold cross validation technique having the same accuracy of 99.956%. Further insights gained using visualizations from Tableau is that the bank is really working hard to come up from crisis of loss of revenue. To overcome this crisis, credit risk analysis will help them to grow and earn profit.