Loan Prediction Problem

A project report submitted for the partial fulfillment of the
Bachelor of Technology Degree
in
IT under
Maulana Abul Kalam Azad University of Technology
BY
AMIT KUMAR JHA
ROLL NO: 10400217129

Under the Guidance of:

Prof Swagatam Basu

Department of Information Technology

For the Academic Year 2020-2021

Institute of Engineering & Management

Management House, D-1, Sector-V, Salt Lake Electronics Complex,
Kolkata-700091 Affiliated To:

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL
(FORMERLY KNOWN AS WEST BENGAL UNIVERSITY OF TECHNOLOGY)
Main Campus : Haringhata, Nadia, Pin-741249
Kolkata Campus : BF-142, SECTOR-I, SALTLAKE CITY, KOLKATA-700 064, (INDIA)
Website : www.wbut.ac.in

Maulana Abul Kalam Azad University of Technology

BF-142, Salt Lake,Sector I,Kolkata-700064

JUNE 2021

INSTITUTE
OF ENGINEERING & MANAGEMENT
श्रद्धावान लभते ज्ञानम् Salt Lake Electronics Complex, Kolkata - 700091, WB, INDIA

Phone : (033) 2357 2969/2059/2995
(033) 2357 8189/8908/5389
Fax : 91 33 2357 8302
E-mail : director@iemcal.com
Website : www.iemcal.com

CERTIFICATE

TO WHOM IT MAY CONCERN

This is to certify that the project report entitled **"Loan Prediction Problem"**, submitted by

AMIT KUMAR JHA Roll no-10400217129.

Students of **INSTITUTE OF ENGINEERING & MANAGEMENT,** in partial fulfilment of requirements for the award of the degree of **Bachelor of Technology in Information Technology ,**is a bonafide work carried out under the supervision and guidance of **Prof. Swagatam Basu** during the final year of the academic session of 2017-2021.The content of this report has not been submitted to any other University or Institute for the award of any other degree.It is further certified that work is entirely original and its performance has been found to be quite satisfactory.

_____

Prof. Swagatam Basu

Project Mentor

Dept. of Information Technology

Institute of Engineering & Management

_____

Prof. Indraneel Mukhopadhyay

H.O.D (IT)

Dept. of Information Technology

Institute of Engineering &Management

ACKNOWLEDGEMENT

We should like to take this opportunity to extend our gratitude to the following revered persons without whose immense support, completion of this project wouldn't have been possible.

We are sincerely grateful to our advisor and mentor **Prof. Swagatam Basu**, **Information Technology**, IEM Kolkata, for his constant support, significant insights and for generating in us a profound interest for this subject that kept us motivated during the entire duration of this project.

We would also like to express our sincere gratitude to **Prof. Indraneel Mukhopadhyay**, HOD of Information Technology **, Satyajit Chakrabarti** (Director**,** IEM) and other faculties of Institute of Engineering & Management, for their assistance and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Amit Kumar Jha
Roll No -10400217129.
Dept. of IT
Institute of Engineering & Management, Kolkata

ABSTRACT

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 op>on for the bank is a typical process. So in this paper we try to reduce this risk factor behind selec>ng 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 objec>ve of this paper is to predict whether assigning the loan to par>cular person will be safe or not. This paper is divided into four sec>ons (i)Data Collec>on (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing

Keywords:

Loan, Machine Learning, Training, Testing, Prediction

TABLE OF CONTENTS

TOPIC **PAGE NUMBER**

Chapter-1: Introduction

Motivation

Predicting the outcome of a loan is a recurrent, crucial and difficult issue in insurance and banking.

The objective of our project is to predict whether a loan will default or not based on past information of the people.

We used a dataset provided by one of the nationalised banks, concerning almost 1000 loans issued.

Using a very structured pipeline to load and test algorithm, we reviewed two of the classification Machine Learning strategies to extract information from the data provided by banks.

Need for this Project

Distribu>on of the loans is the core business part of almost every banks. The main por>on the bank's assets is directly came from the profit earned from the loans distributed by the banks. Today many banks/financial companies approves loan aOer a regress process of verifica>on and valida>on but s>ll there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that par>cular applicant is safe or not and the whole process of valida>on of features is automated by machine learning technique. The disadvantage of this model is that it emphasise different weights to each factor but in real life some>me loan can be approved on the basis of single strong factor only, which is not possible
through this system.

Scope of Project

Loan prediction can provide special advantages to the bank. The Loan Predic>on System can can automa>cally calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A >me limit can be set for the applicant to check whether his/her loan can be sanc>oned or not. Loan Predic>on System allows jumping to specific applica>on so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process

of predic>on is done privately no stakeholders would be able to alter the processing. Result against par>cular Loan Id can be send to various department of banks so that they can take appropriate ac>on on applica>on. This helps all others department too carried out other formali>es. In near future this module of predic>on can be integrate with the automated processing system. Whenever you will fill the details of applicant the applica>on will automa>cally predict the guarantee of Loan returning.

Chapter 2: Problem review & Formulation

Loan Predic>on is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants.The Loan Predic>on System can can automa>cally calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A >me limit can be set for the applicant to check whether his/her loan can be sanc>oned or not. Loan Predic>on System allows jumping to specific applica>on so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of predic>on is done privately no stakeholders would be able to alter the processing.In near future this module of predic>on can be integrate with the automated processing system. Whenever you will fill the details of applicant the applica>on will automa>cally predict the guarantee of Loan returning.

Chapter 3: Proposed Methodology

3.1 Collec>on of Data Sets

It consists into approximately 10,000 samples of loans granted by the bank, with the full set of informa>ons about the borrower, the history of payments and the outcome of the loan. The dataset is quite clean and the figures can be considered as ground truth, but lots of columns are either irrelevant, very sparse or non informa>ve. Moreover, the dataset is very unbalanced, with approximately 17% of loans considered as defaulted. Since the objec>ve is to predict the outcome from the informa>ons gathered at the signature of the loan, we cannot use the data concerning the history of payments or the current situa>on of a loan. Excluding features for which the informa>on is incomplete, or uninforma>ve, we get a total of 19 features, that cover personal informa>on (credit grade, income, housing status, ...) and credit informa>on (amount, interest rates, ...). Accuracy is not well-suited for our problem. The unbalance of the classes would lead an algorithm to never predict a default. Predic>on score allows us to quan>fy a good predic>on on both precision and recall.[4]

3.2 Classifica>on of Different Variables

Loan_ID	614 non-null object
Gender	601 non-null object
Married	611 non-null object

Dependents 599 non-null object

Educa>on 614 non-null object

Self_Employed 582 non-null object

ApplicantIncome 614 non-null int64

CoapplicantIncome 614 non-null float64

LoanAmount 592 non-null float64

Loan_Amount_Term 600 non-null float64

Credit_History 564 non-null float64

Property_Area 614 non-null object

Loan_Status 614 non-null object

dtypes: float64(4), int64(1), object(8)

memory usage: 62.5+ KB

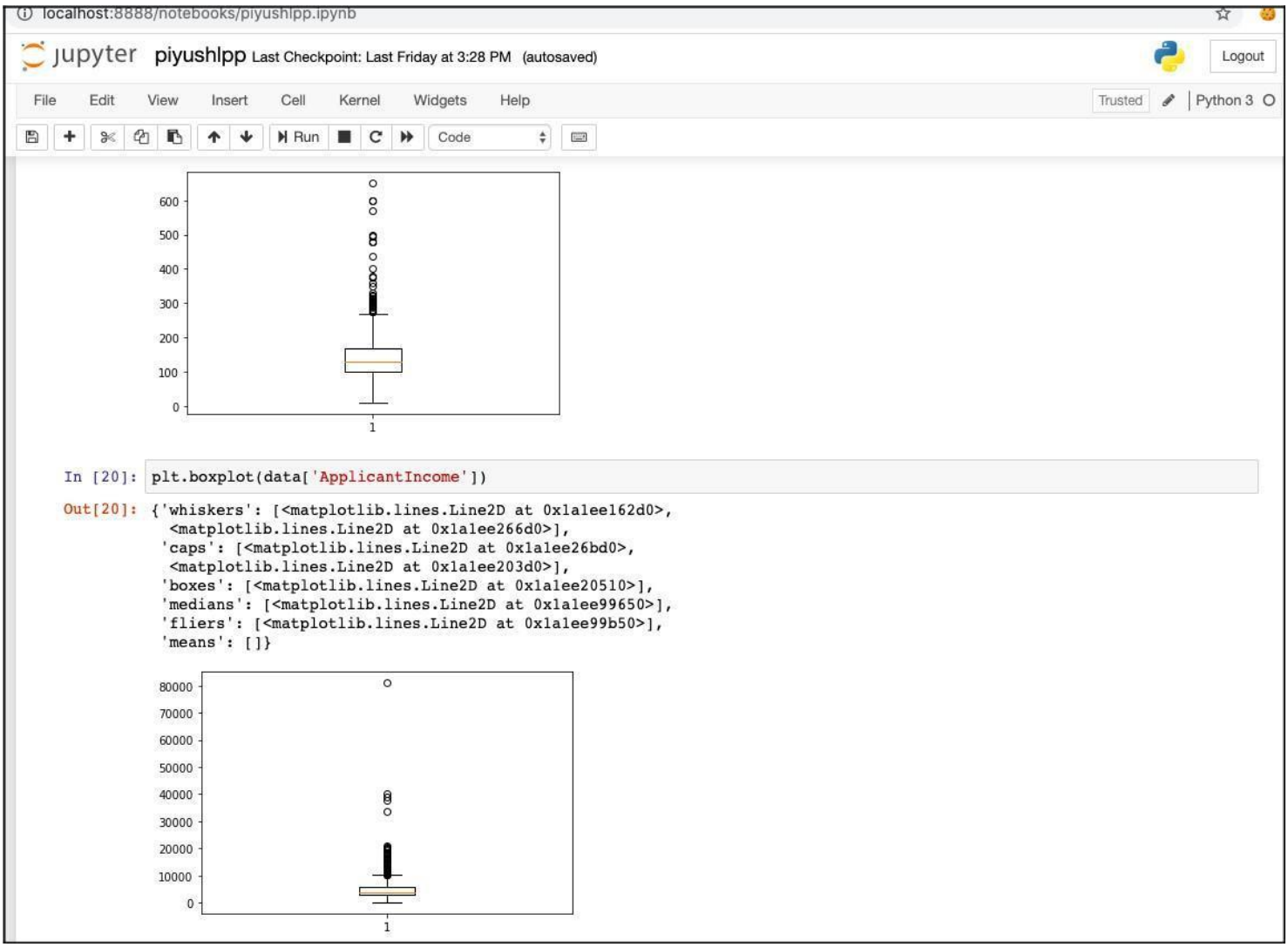
3.3 Graphs plohng

3.3.1.Univariate Graph

Use of one variable to plot a graph(Box-plot).

Applicant income

Fig.1.Box Plot Applicant Income



Loan Amount Term

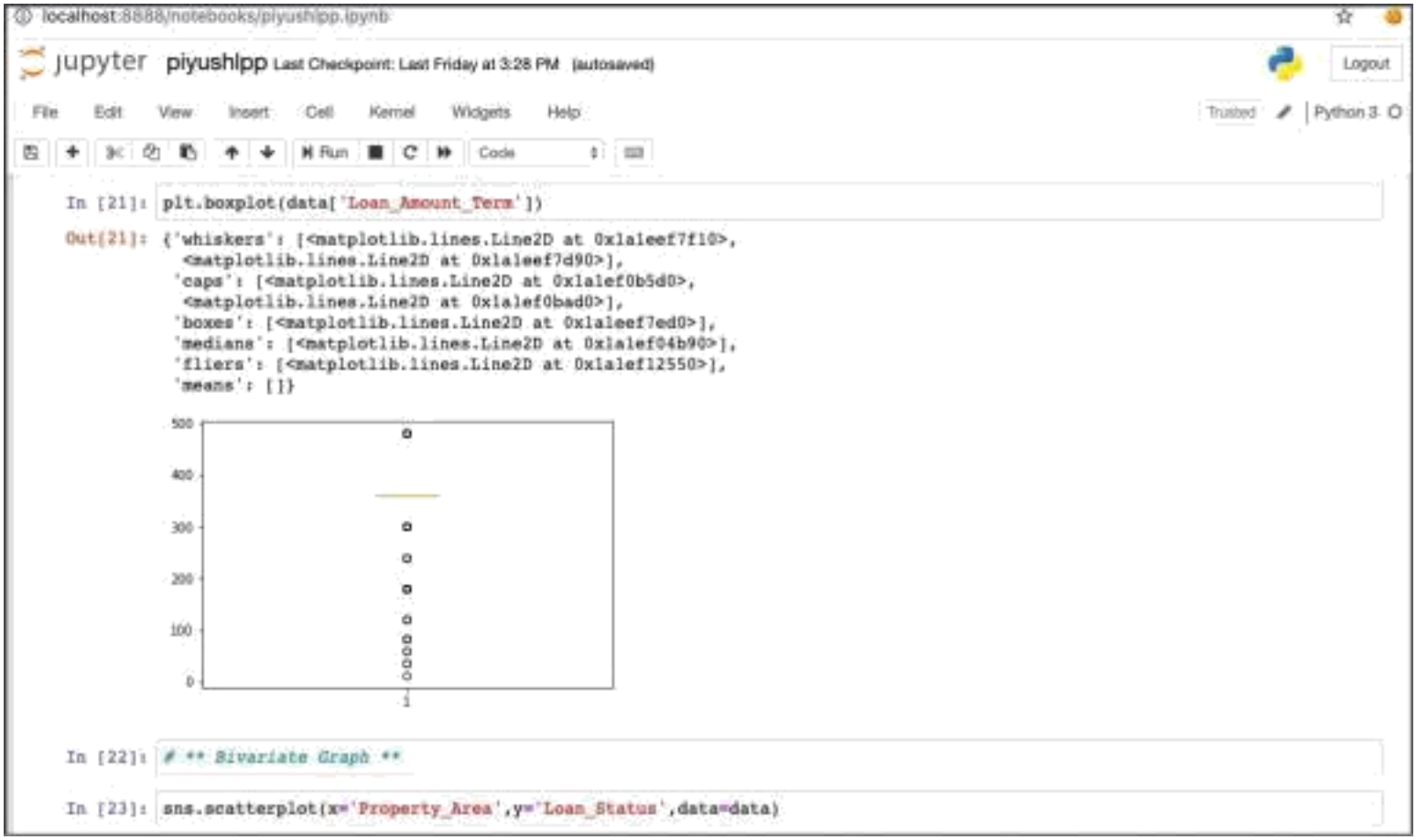


Fig.2.Loan Amount Term(Box Plot)

3.3.2. Bivariate Graph

1.Property Area Vs Loan Status(Box Plot) 2.Self Employed vs Loan Status(Box Plot)

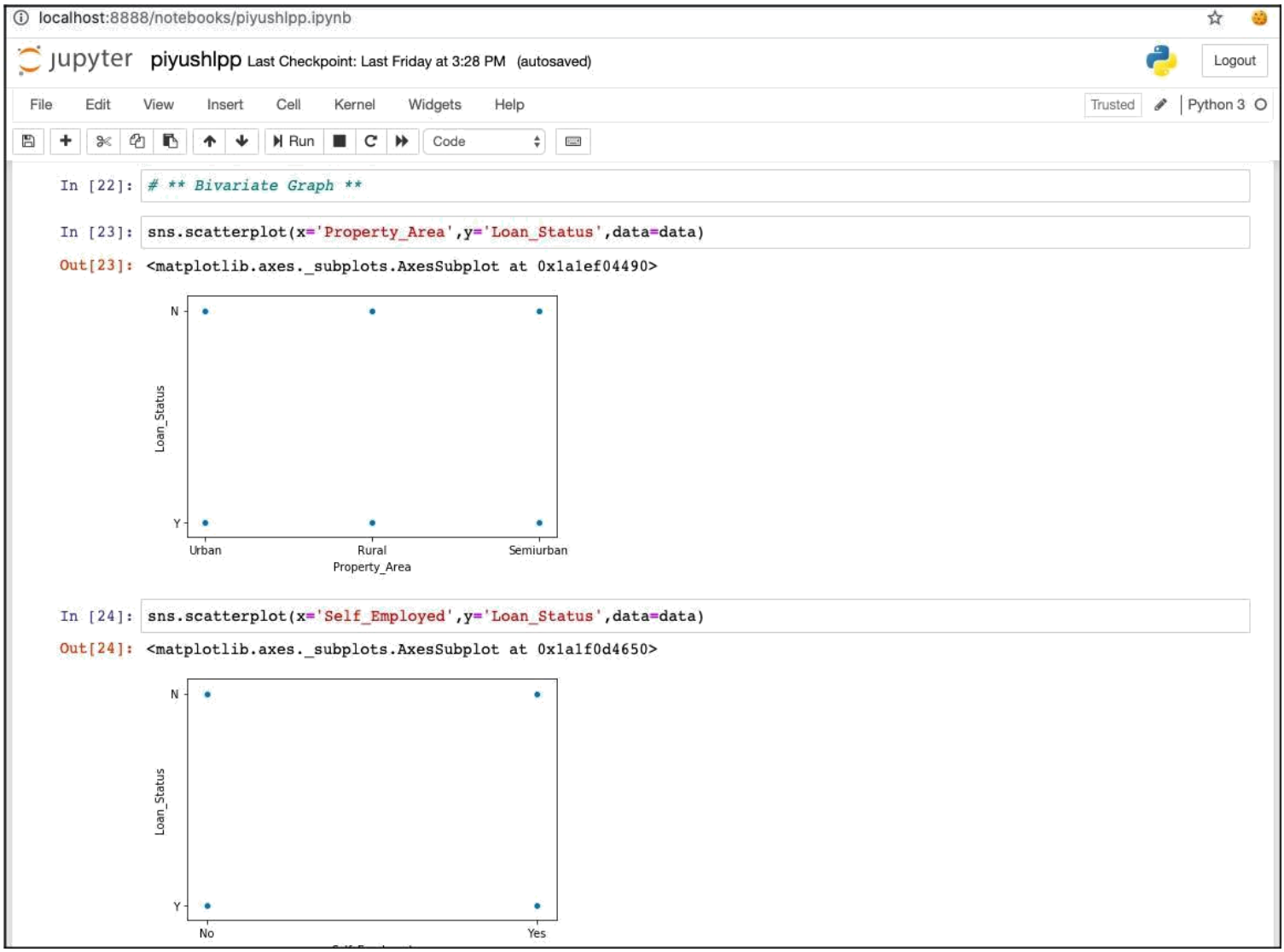


Fig.3.Property Area Vs Loan Status(Box Plot)

Property Area Vs Loan Status(Bar Graph)

1.Urban

2.Rural

3.Semiurban

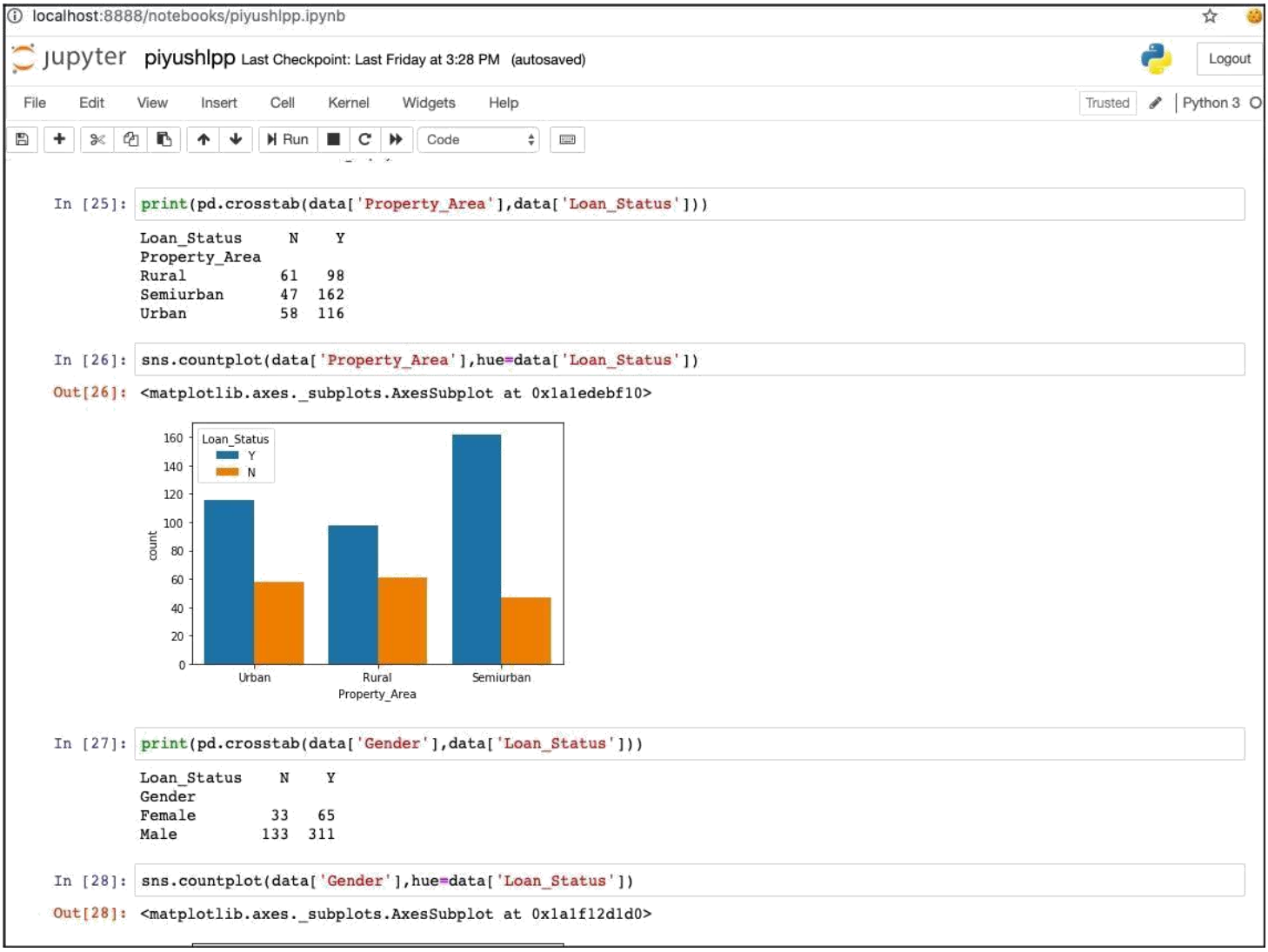


Fig.4.Property Area Vs Loan Status(Bar Graph)

Gender Vs Loan Status(Bar Graph)

1.Male

2.Female

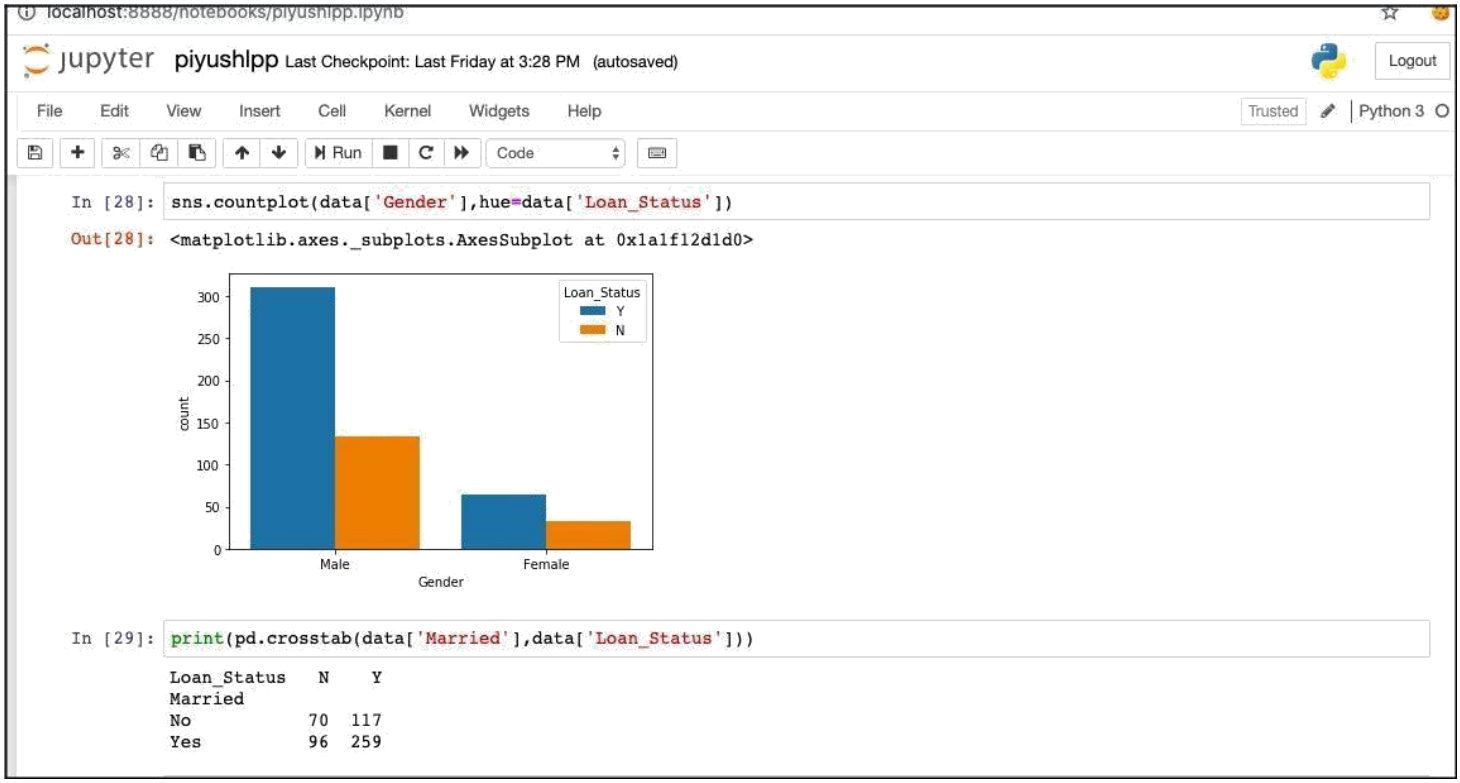


Fig.5.Gender Vs Loan Status(Box Plot)

Married Vs Loan Status(Bar Graph)

1.Yes(Married)

2.No(Not Married)

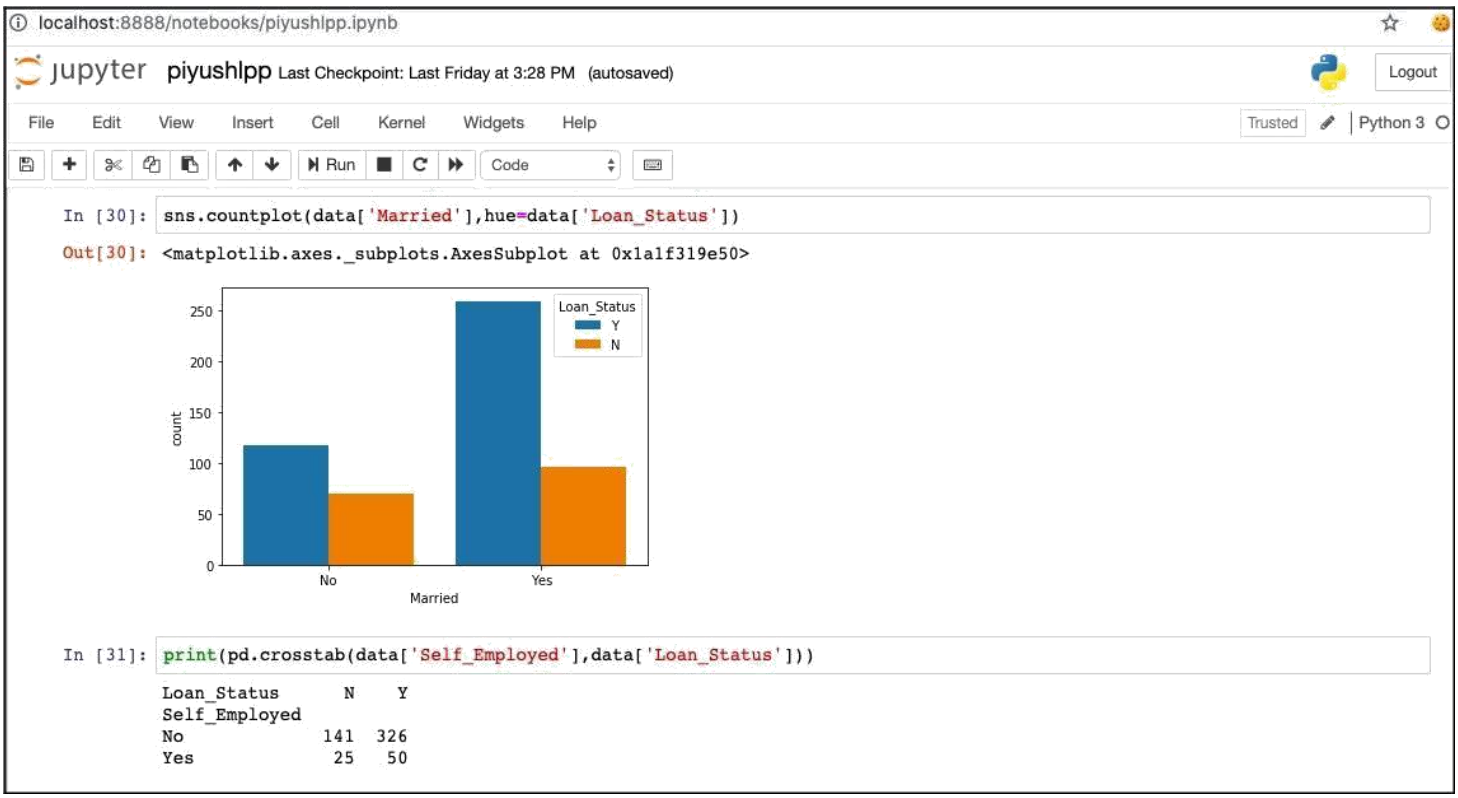


Fig.6.Marital Status Vs Loan Status(Bar Graph)

Education Vs Loan Status(Bar Graph)

1.Graduate

2.Not Graduate

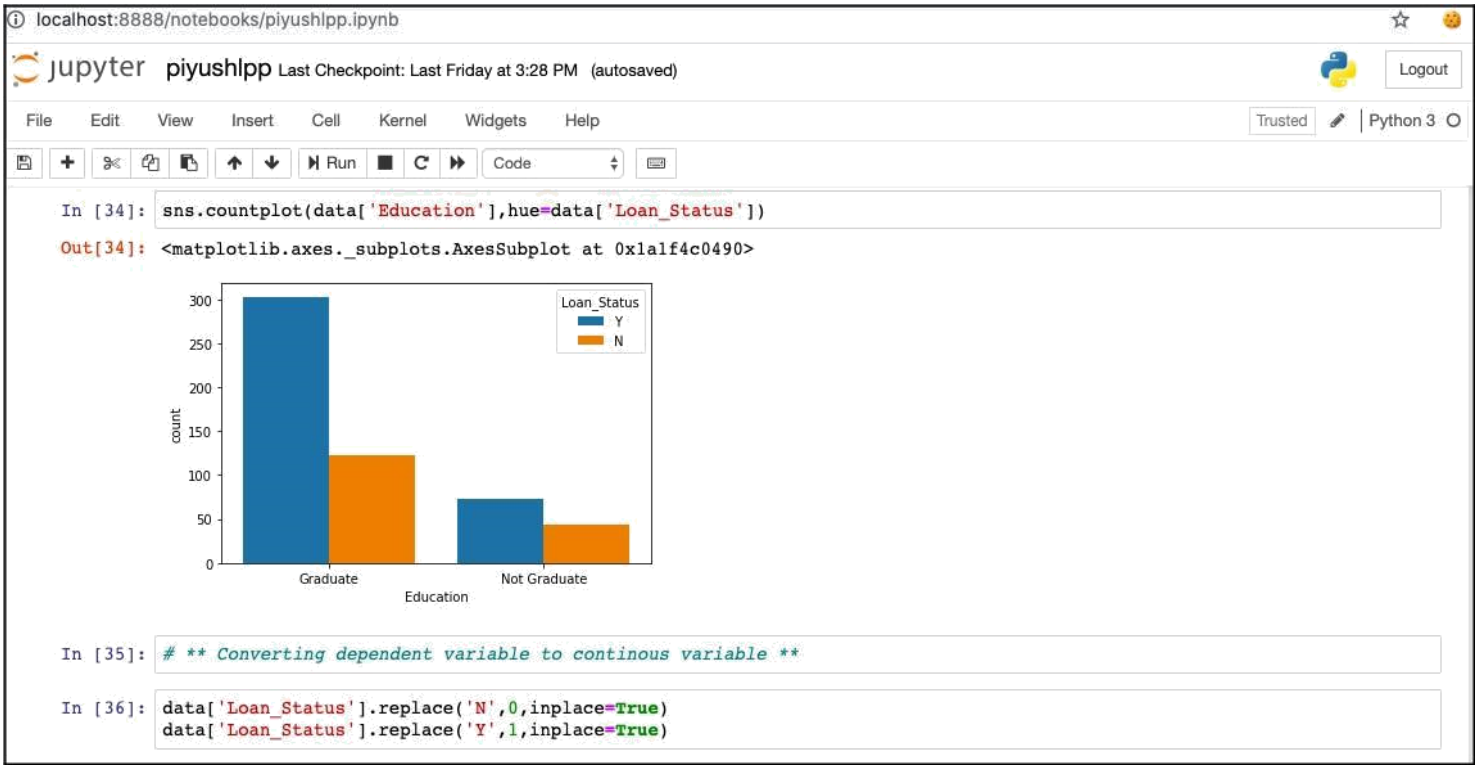


Fig.7.Education Vs Loan Status(Bar Graph)

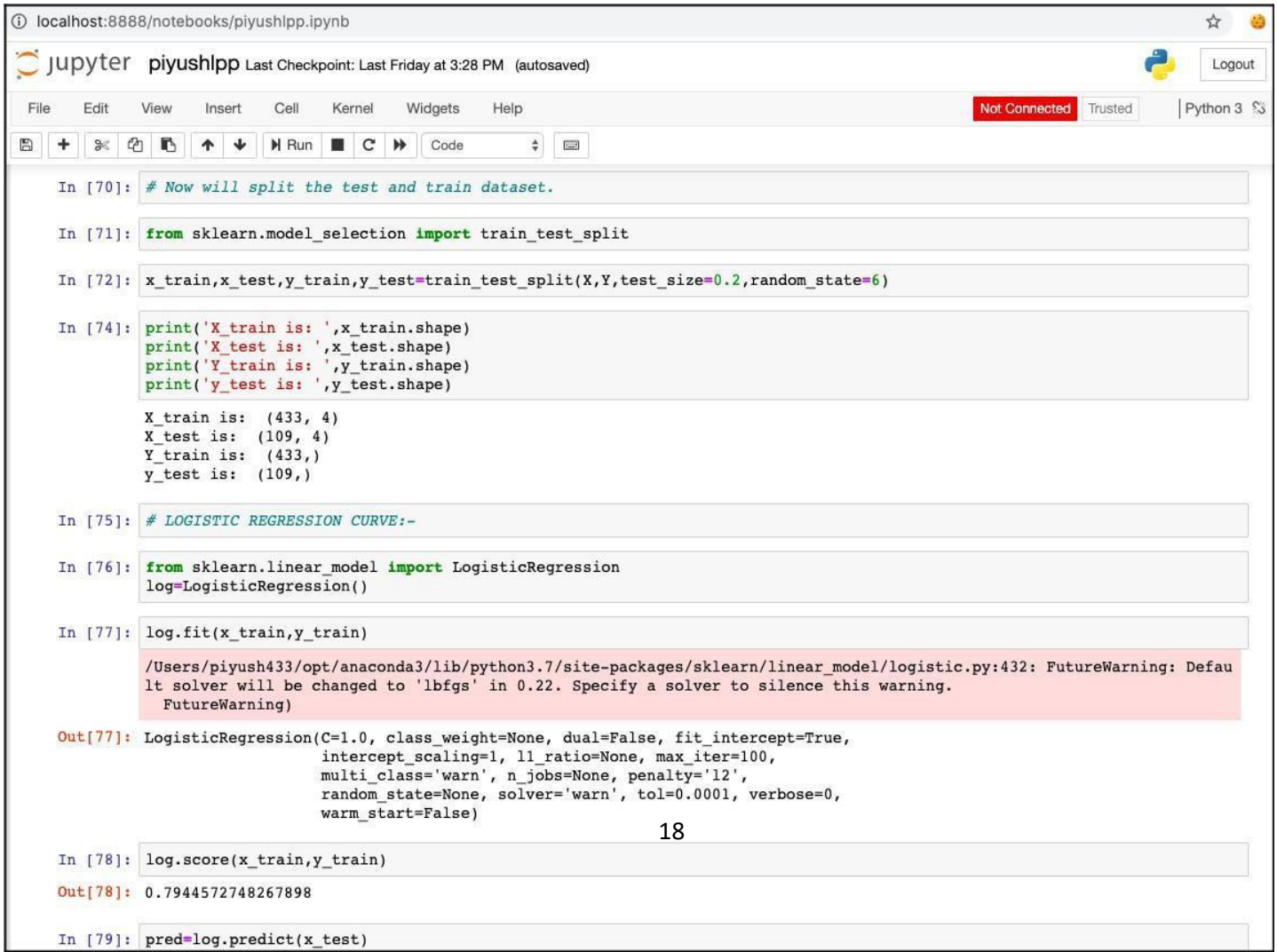
3.4Test the model on Algorithm

3.4.1.Logistic Regression Curve

Logis>c Regression is a Machine Learning algorithm which is used for the classifica>on problems, it is a predic>ve analysis algorithm and based on the

concept of probability.The hypothesis of logis>c regression tends it to limit the cost func>on between 0 and 1 .[1]

Fig.8.Splihng Test and Train Dataset



Logis>c Regression Algorithm

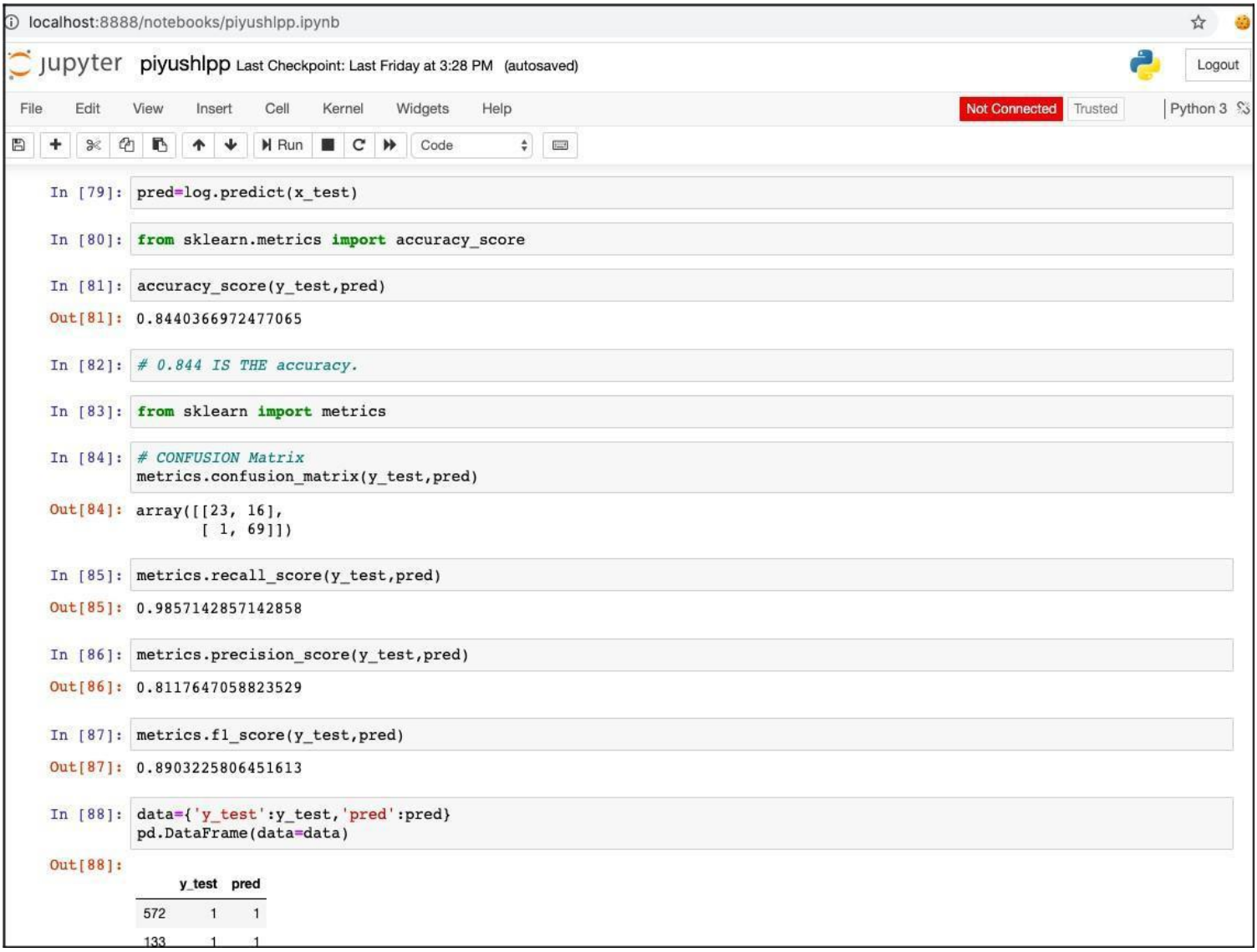


Fig.9.Logistic Regression Algorithm

3.4.2. Decision Tree

The basic algorithm of decision tree requires all attributes or features should be discretised. Feature selection is based on greatest information gain of features. The knowledge depicted in decision tree can represented in the form of IF-THEN rules. This model is an extension of C4.5 classification algorithms described by Quinlan.[2]

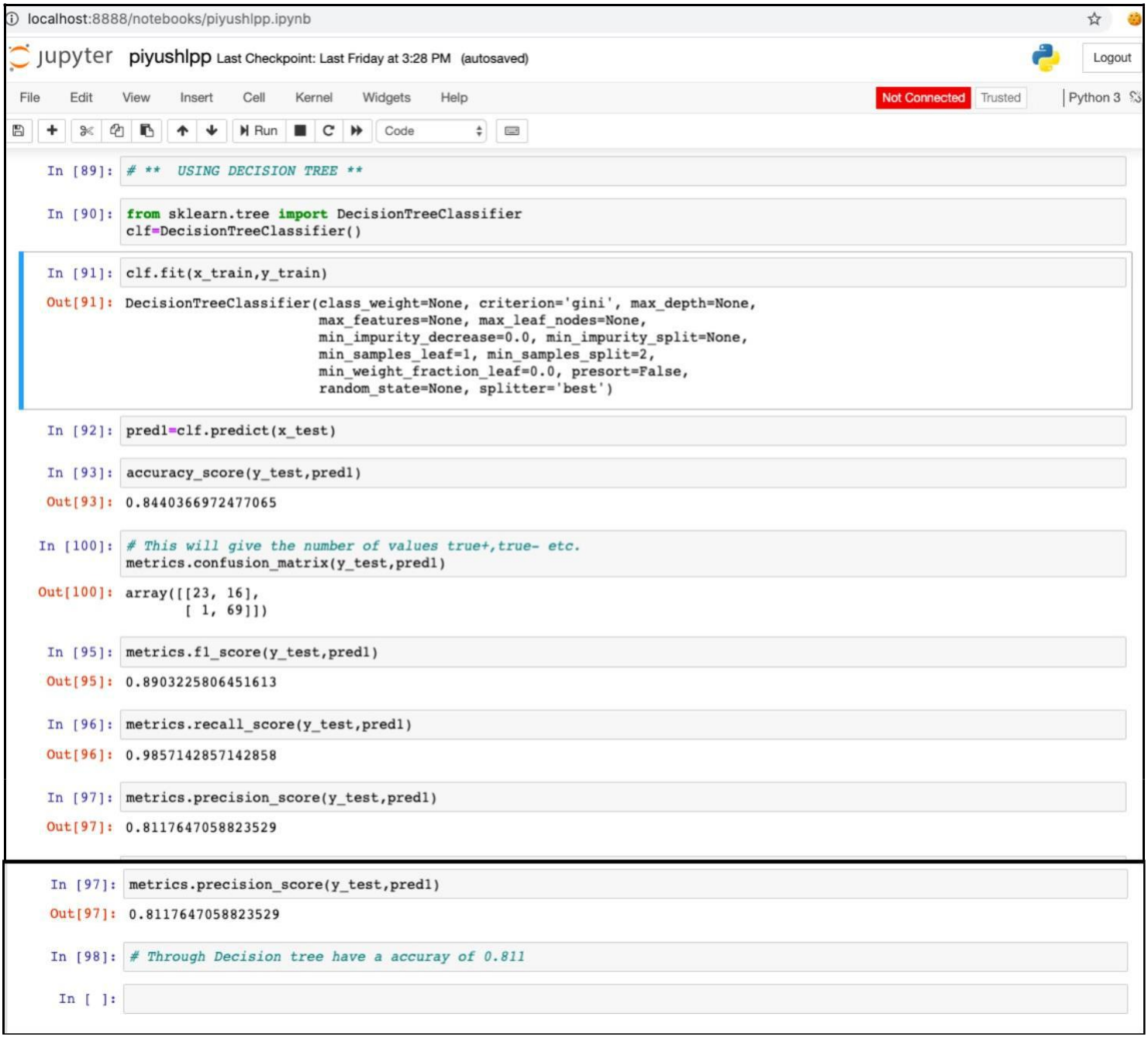


Fig.10.Decision Tree Algorithm

3. 5.Result analysis

We will check for the accuracy of project with both the algorithm.As the dataset is less the accuracy of the project more than expected from the algorithm.

Chapter 4: Project Implementation

Platform used

1. Python 3
2. Anaconda Jupyter notebook
3. Web browser Chrome

Fetching Dataset

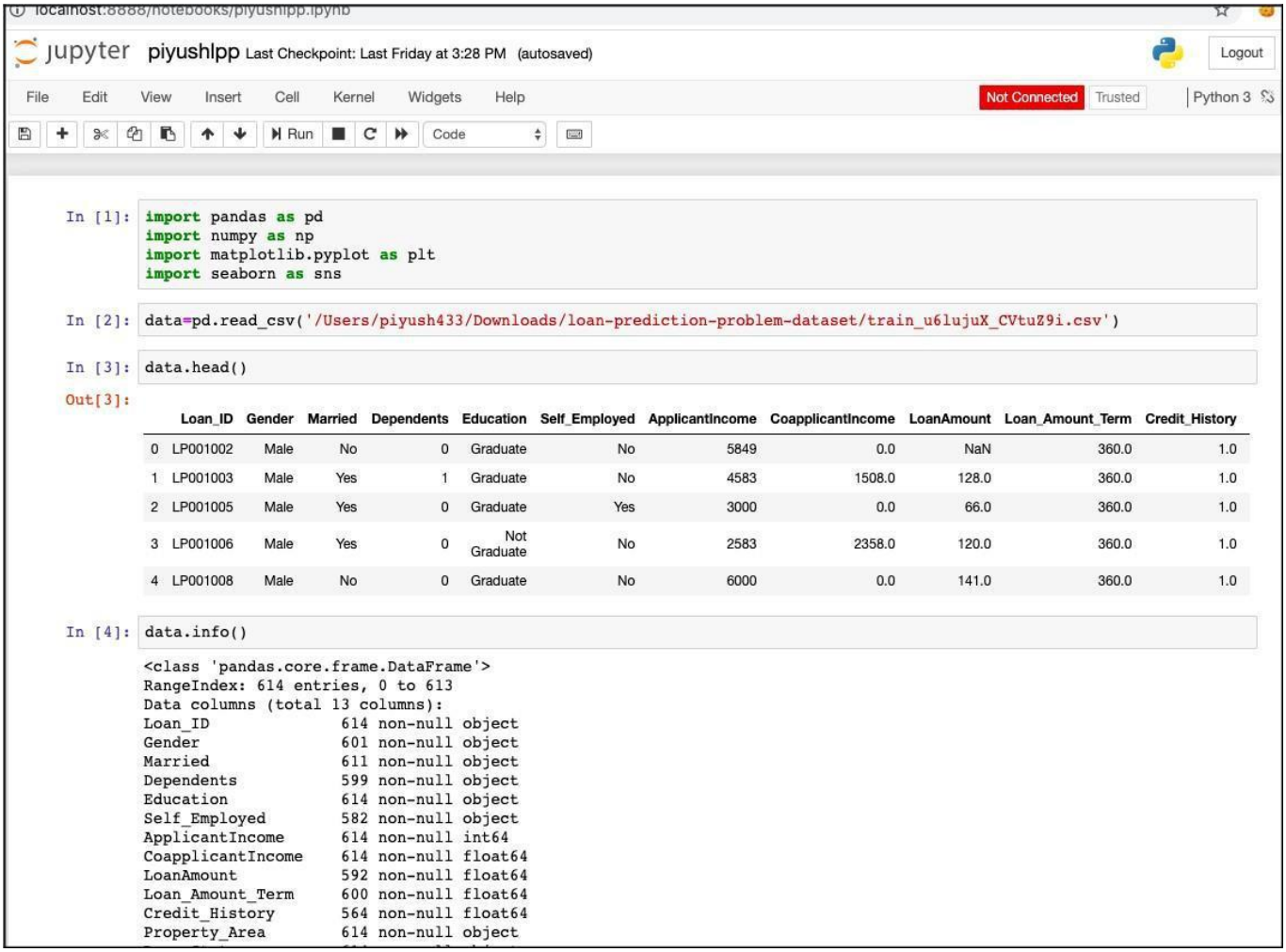
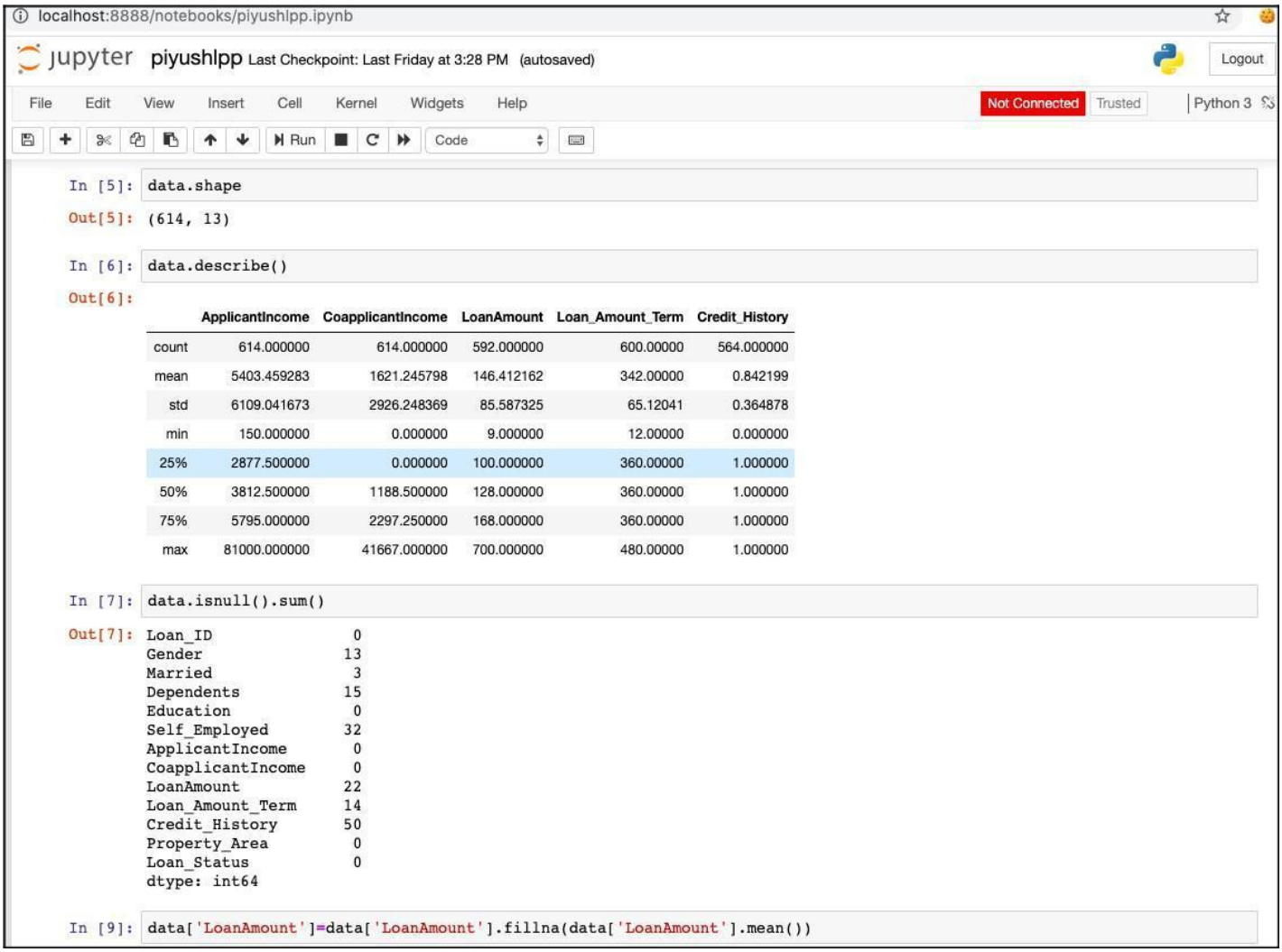


Fig.11.Dataset Fetching

Dividing the different Variables

Fig.12.Differentiating Independent and Dependent Variable



Correlation Matrix

For finding the variable dependent on Loan Status

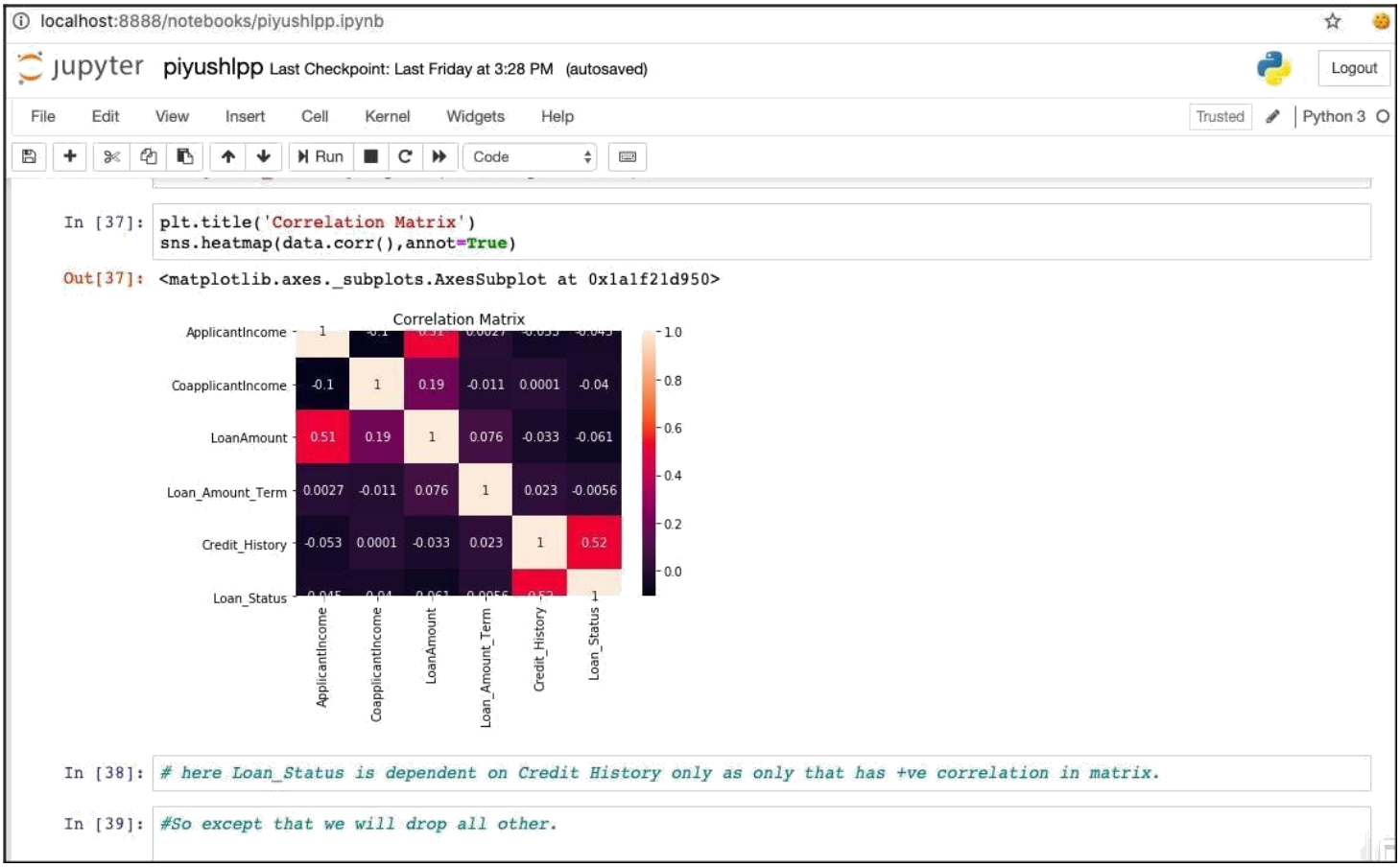


Fig.13.Correlation Matrix

Splitting the Dataset for Getting training dataset

It is done by dropping the variable on which Loan Status doesn't depend.

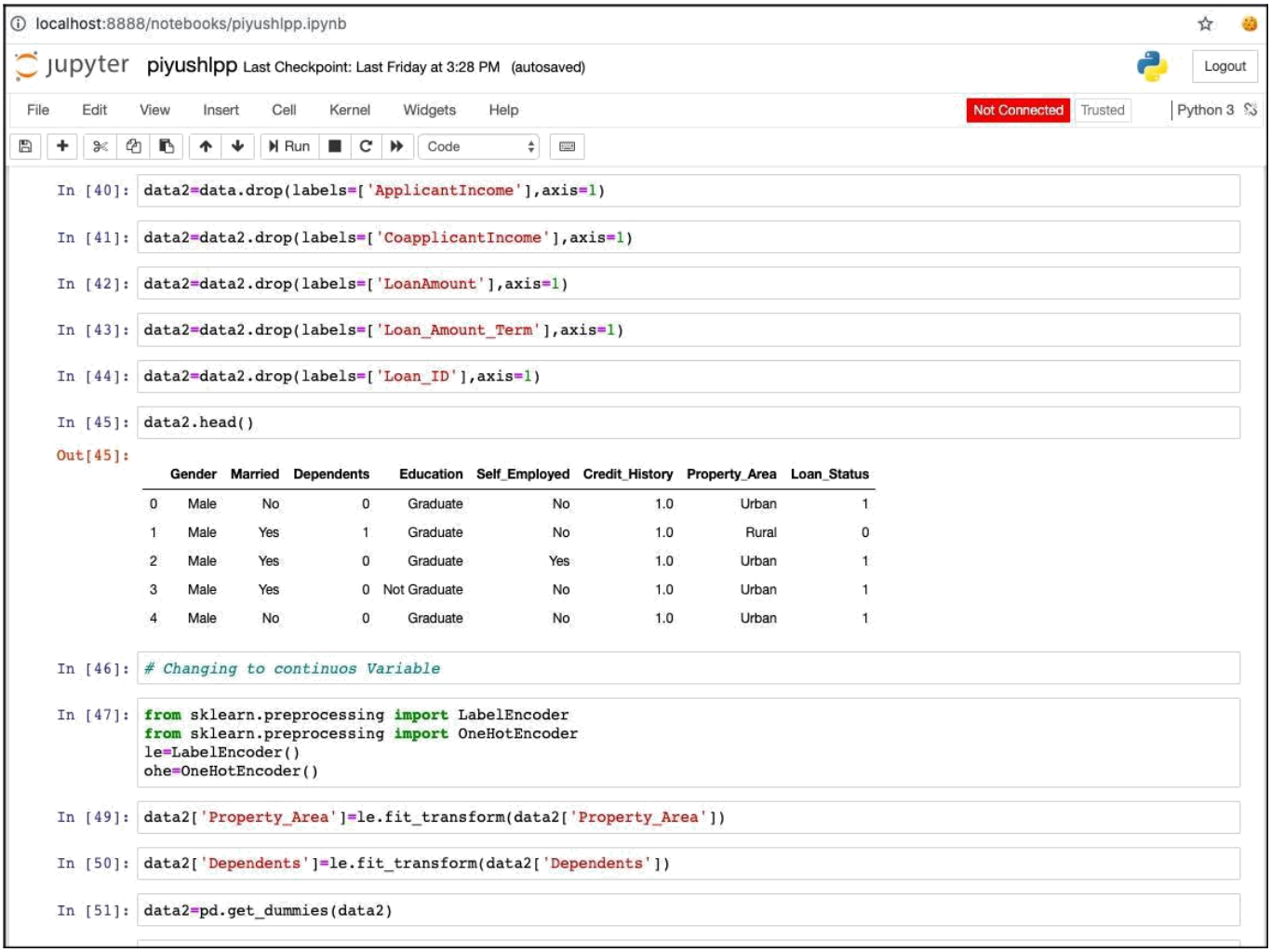
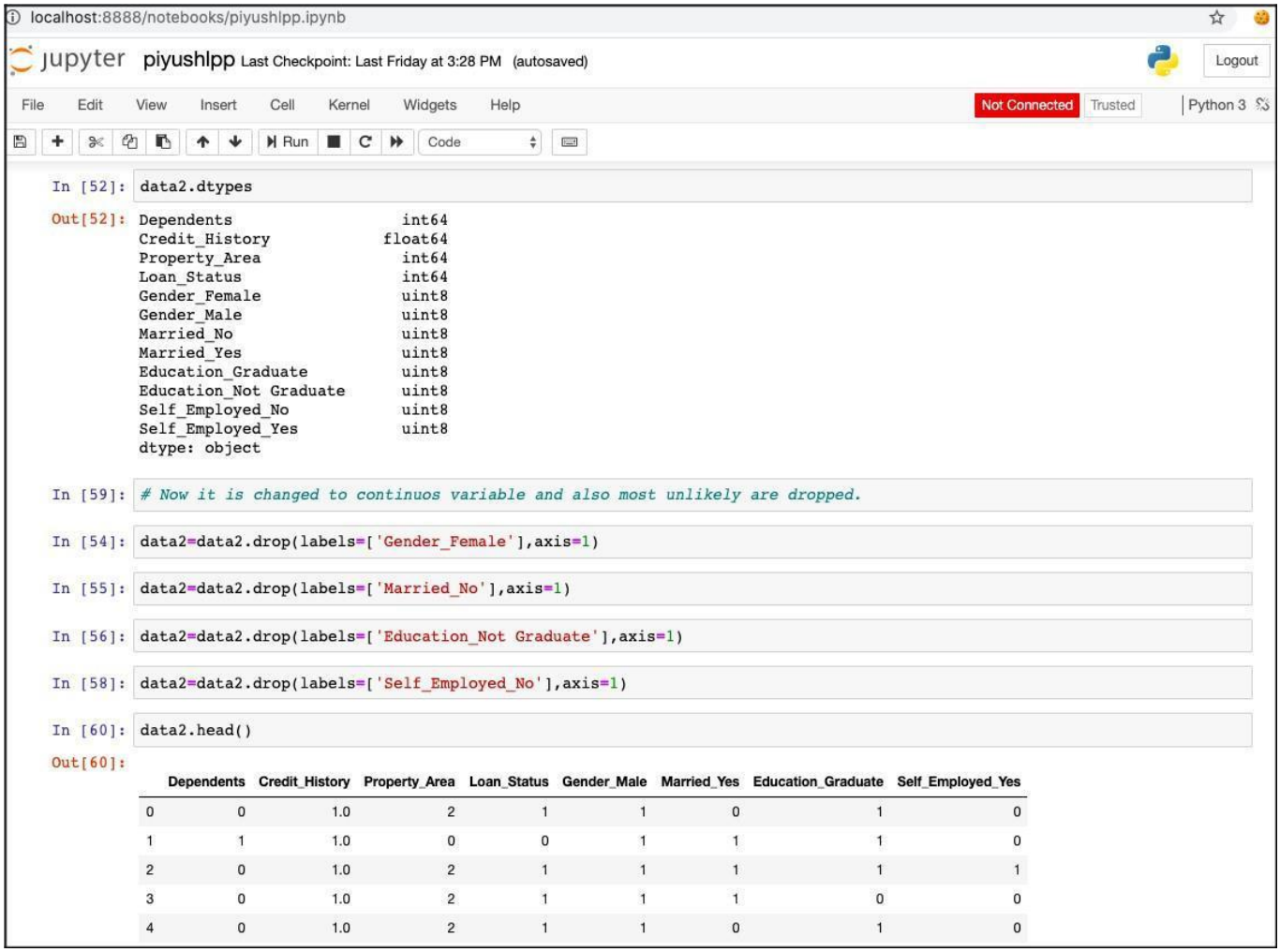


Fig.14.Variable Dropping

Dropping the Variable and making continuous the dataset

Fig.15.Making Continuos Dataset Dropping Null Set



Dividing the test and trained dataset and seeing the output of it

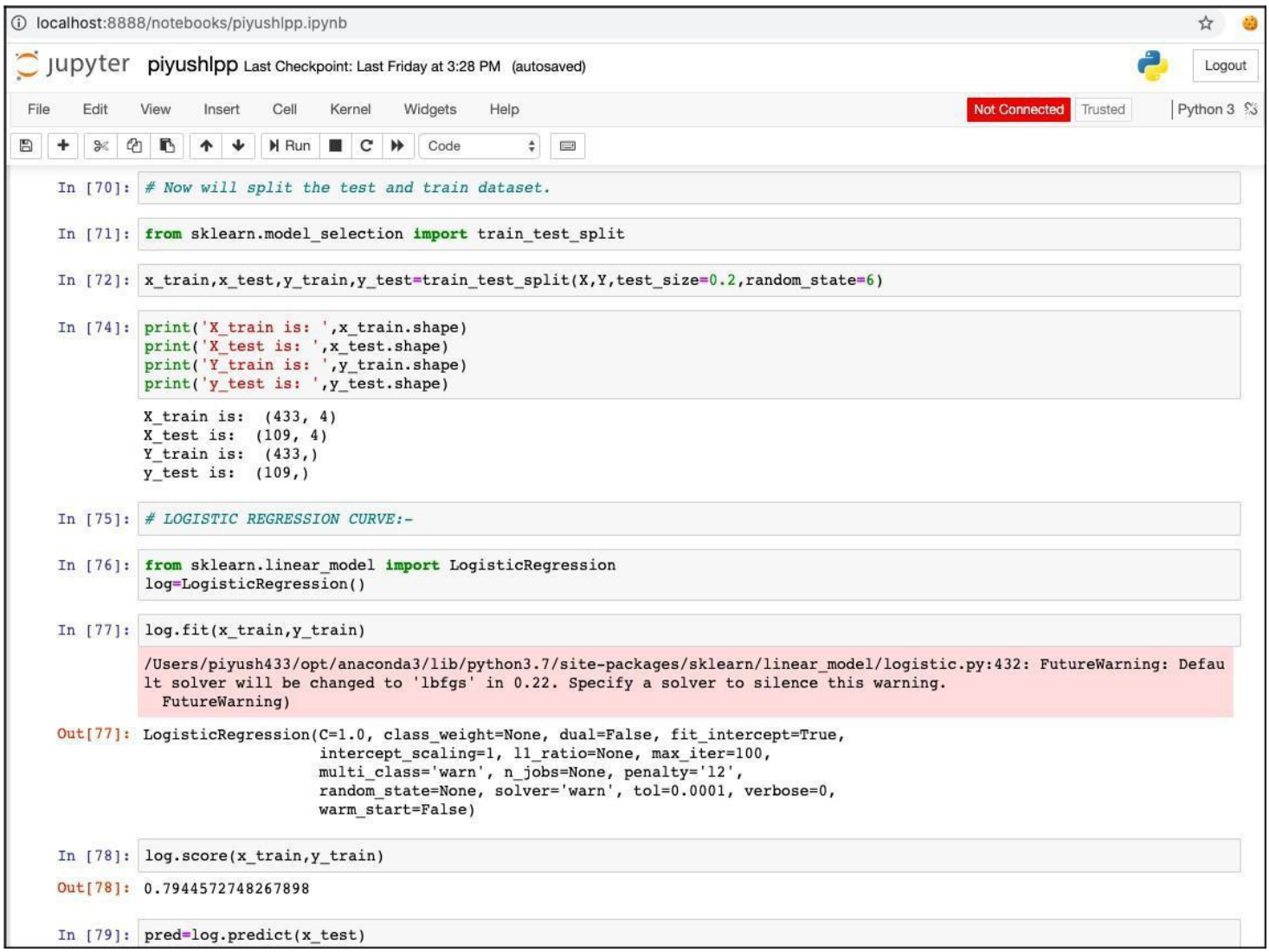


Fig.16.Split test and Train Dataset

Logistic Regression Curve

Confusion Matrix is used for finding the probability.

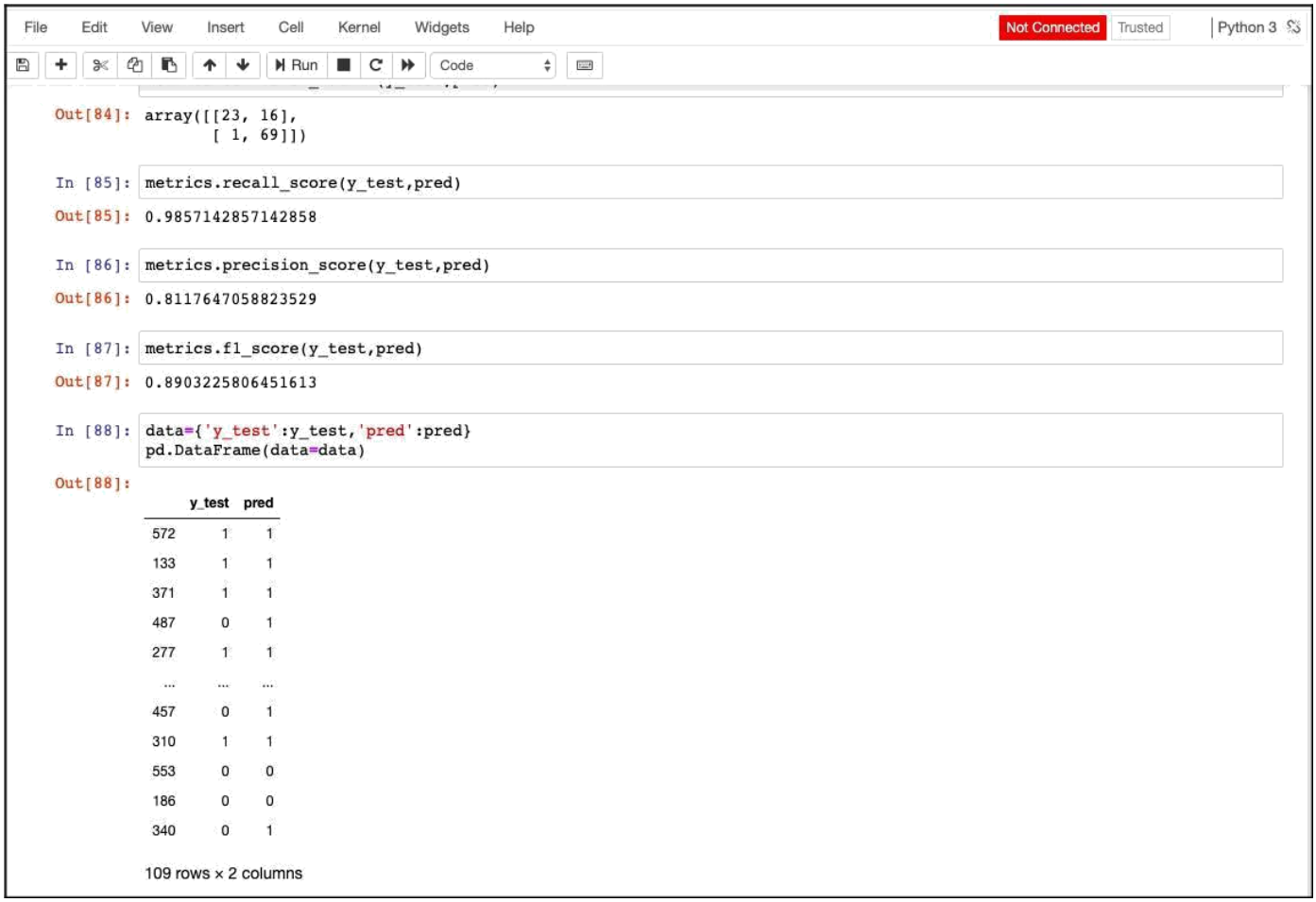


Fig.17.Confusion Matrix

Decision Tree

Using weighted variable for plotting the graph and finding the probability

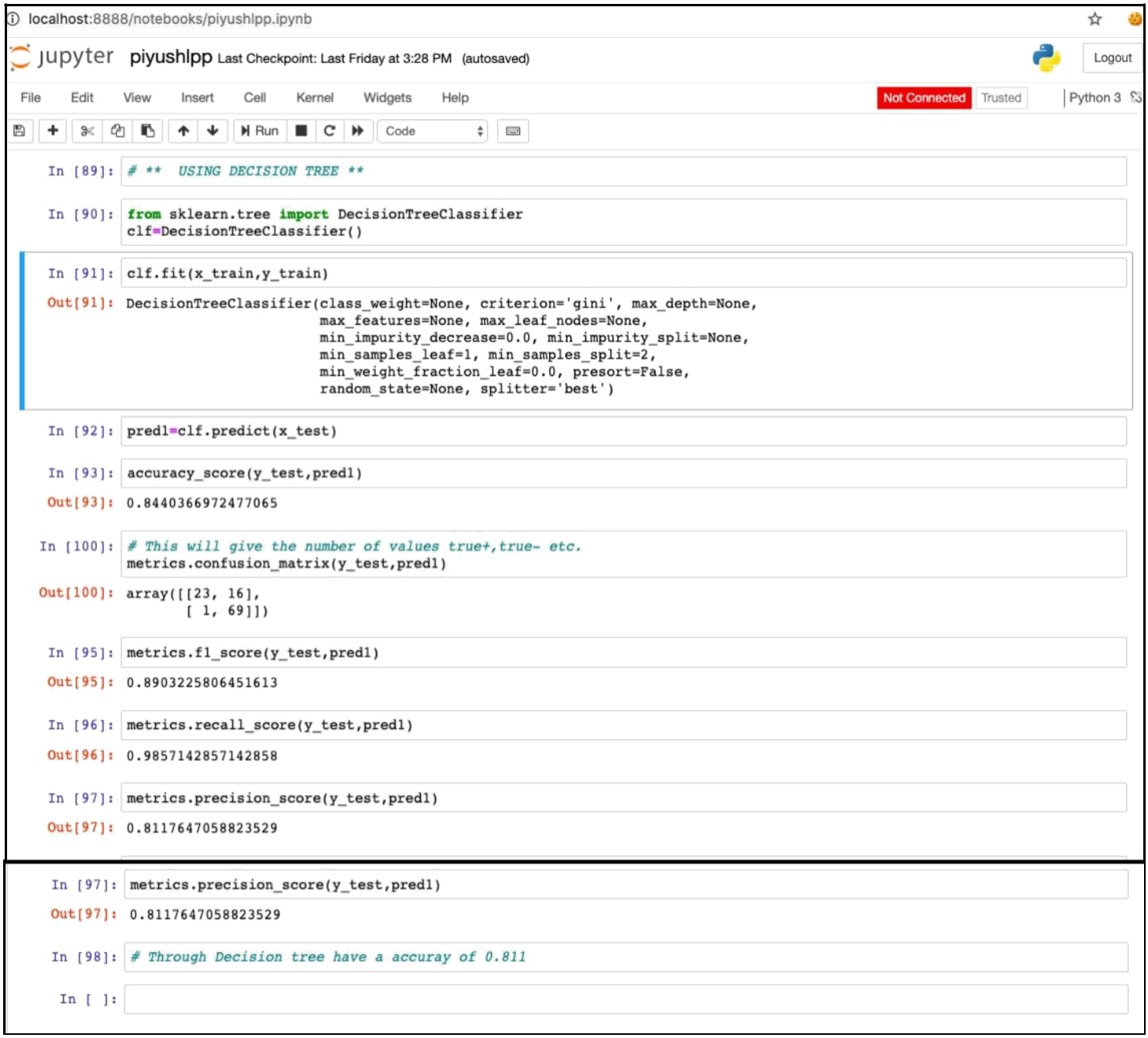


Fig.18.Decision Tree

Chapter 5: Results

From Logistic Curve We get

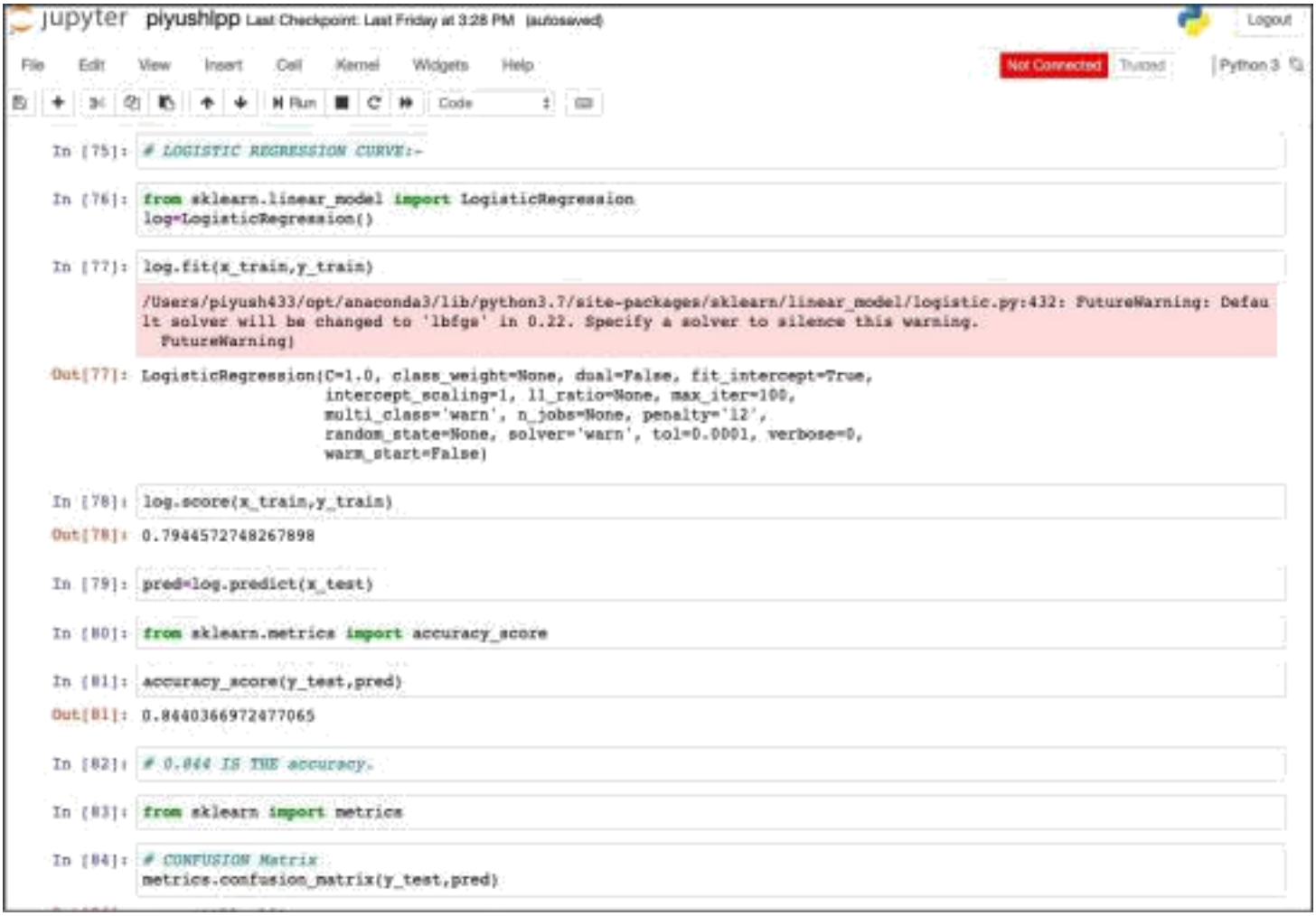


Fig.19.Logistic Regression Accuracy

Confusion Matrix

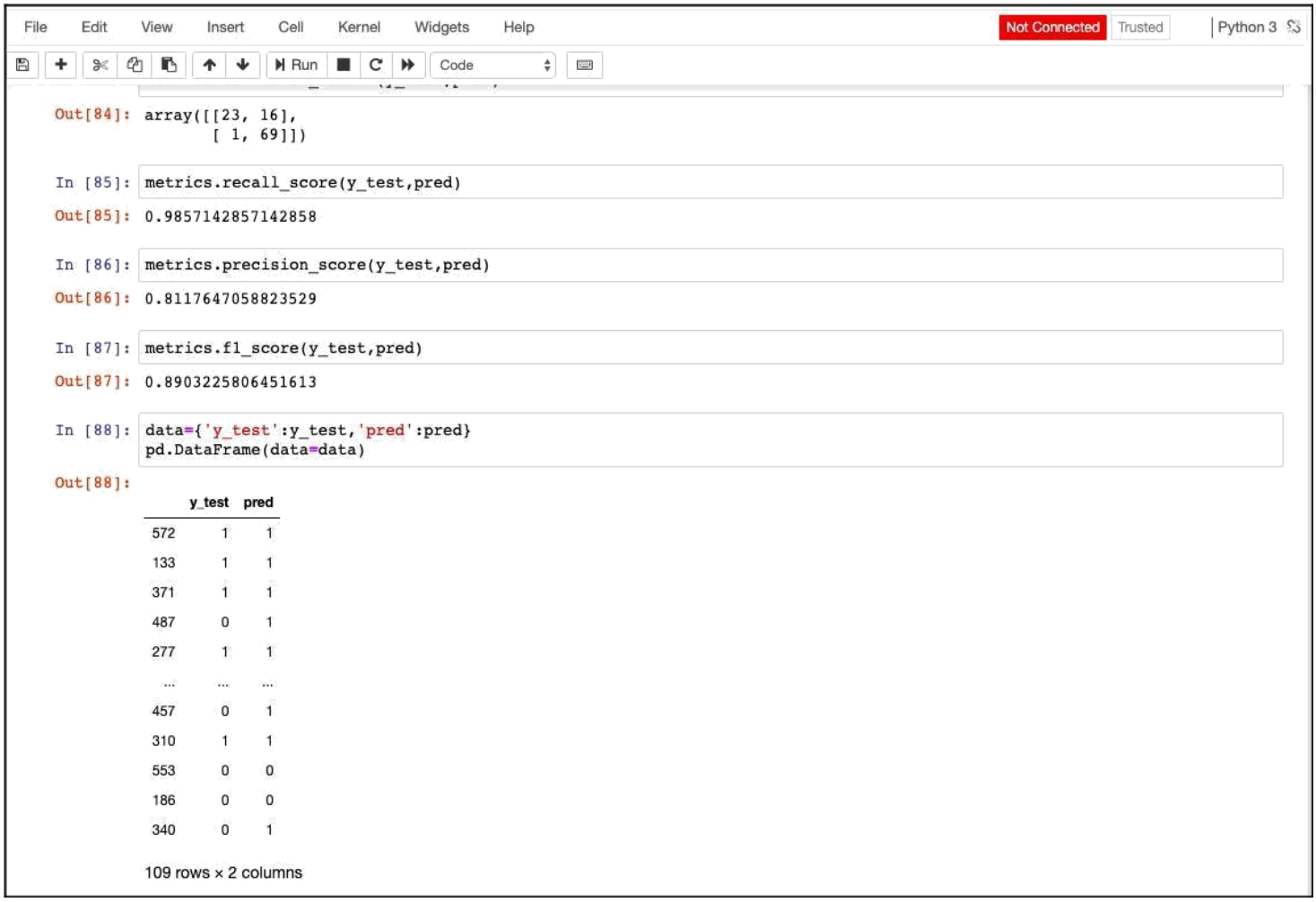


Fig.20.Logistic Regression(Confusion Matrix)

We compared the training and testing dataset and also through confusion matrix we compared and get an accuracy of 0.84.

Using Decision Tree:-

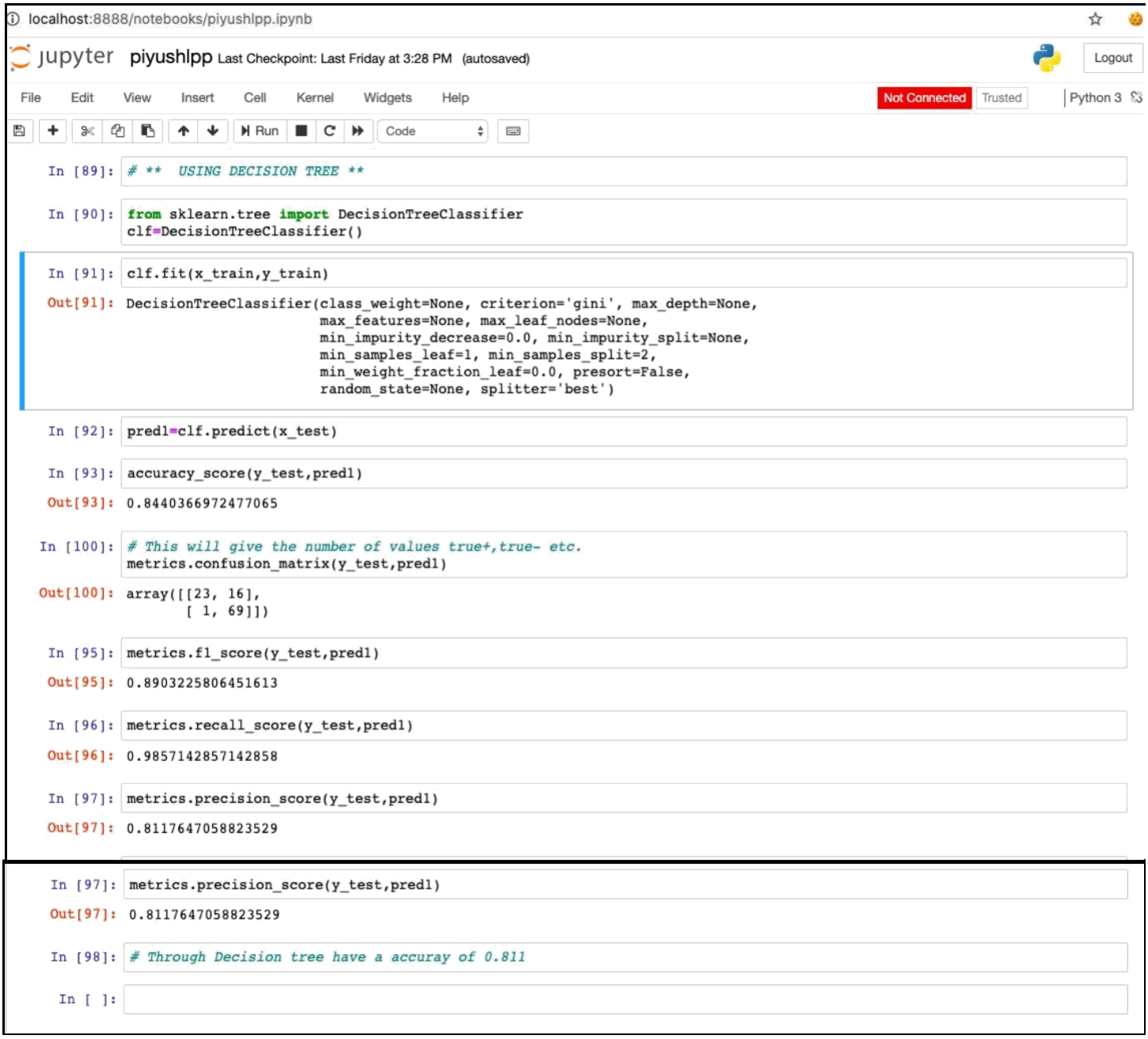


Fig.21.Decision Tree Accuracy

We used weighted variable to compare with the testing and training dataset and got an accuracy of 0.811.[3]

Chapter 6: Conclusion

From a proper analysis of posi>ve points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This applica>on is working properly and mee>ng 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 predic>on system, So in the near future the so –called soOware could be made more secure, reliable and dynamic weight adjustment .In near future this module of predic>on can be integrate with the module of automated processing system. the system is trained on old training dataset in future soOware can be made such that new tes>ng date should also take part in training data aOer some fix >me.

Chapter 7: References

[1]**Machine Learning Algorithms: Interna,nal Journal of Computer Science and Telecommunica,ons (Volume2, Issue3, June 2011).**

[2] ***J. R. Quinlan. Induc0on of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.*** **[3]** ***J.R. Quinlan. Induc0on*** **of decision trees. MachinelearningSpringer, 1(1):81 –106, 1086.**

[4] kaggle.com **for dataset.**