**PROJECT REPORT**

**Project Name**

**XYZ Corp Loan Default Prediction**

*Submitted towards the partial fulfillment of the criteria for award of PGA by Imarticus*

*Submitted By:*

*Kavya. A*

*Sarika. B*

*Vengam S Harathi*

*Course and Batch: DSP-04 May 2019*



# Abstract

**Keywords**

*Disclaimer: \*Data shared by the customer is confidential and sensitive, it should not be used for any purposes apart from capstone project submission for PGA. The Name and demographic details of the enterprise is kept confidential as per their owners’ request and binding.*

# Acknowledgement

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have **Priyansu Panda** as our mentor. He has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the PGA program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: November 02, 2019 Member 1 : Kavya A

Place:Bangalore Member 2 : Sarika B

Member 3 : Vengam S Harathi

# Certificate of Completion

I hereby certify that the project titled **“ Loan Default Prediction”** was undertaken and completed under my supervision by **Kavya. A, Sarika. B** and **Vengam S Harathi** from the batch of DSP(May 2019)

Mentor: Priyansu Panda

Date: November 06, 2019

Place – Bangalore

Table of Contents

[Abstract 2](#_Toc510332824)

[Acknowledgements 2](#_Toc510332825)

[Certificate of Completion 3](#_Toc510332826)

[CHAPTER 1: INTRODUCTION](#_Toc510332827) 5

[1.1 Title & Objective of the study 5](#_Toc510332828)

[1.2 Need of the Study 5](#_Toc510332829)

[1.3 Business Model of Enterprise 5](#_Toc510332831)

[1.4 Data Sources 6](#_Toc510332832)

[1.5 Tools & Techniques 6](#_Toc510332833)

[CHAPTER 2: DATA PREPARATION AND UNDERSTANDING 6](#_Toc510332835)

[2.1 Phase I – Data Extraction and Cleaning: 7](#_Toc510332836)

[2.2 Phase II - Feature Engineering 8](#_Toc510332837)

[2.3 Data Dictionary: 9](#_Toc510332838)

[2.4 Exploratory Data Analysis: 1](#_Toc510332839)2

[CHAPTER 3: FITTING MODELS TO DATA 1](#_Toc510332840)4

[CHAPTER 4: MODEL SELECTION 1](#_Toc510332841)5

[4.1LOGISTIC REGRESSION 1](#_Toc510332842)6

[4.2 RANDOM FOREST 1](#_Toc510332845)6

[CHAPTER 5: KEY FINDINGS 1](#_Toc510332847)7

[CHAPTER 6: RECOMMENDATIONS AND CONCLUSION 1](#_Toc510332848)8

[CHAPTER 7: REFERENCES 1](#_Toc510332849)8

# CHAPTER 1: INTRODUCTION

## Scope & Objective of the study

The objective of this project is to build a predictive model based on the available historic data and predict the probability of loan default and deciding a cut-off.

The scope of this project is to show how predictive modeling can be used to predict bad loans, loans that are to default, vital information that can be used to reject the loan at the time of application, and, thereby saving the bank from loses incurred.

## Need of the Study

Prevail loans from financial organization has become a very common phenomenon. Every day many people apply for loans, for a various purpose. But not all the applicants are trustworthy, and not everyone can be approved. Every year, there are cases where people do not refund the bulge of the loan amount to the bank which results in huge financial loss. The risk associated with making a decision on a loan approval is immense. Hence, the idea of this project is to gather loan data from the Kaggel website and use machine learning techniques on this data to extract important information and predict if a customer would be able to repay the loan or not.

## 1.3 Business Model of Enterprise

People often save their money in banks which offers security but with lower interest rates. XYZ Corporation operates an online lending platform that enables borrowers to obtain loan, and investors to get money back by payments made on loans. It is transforming the bank system to make credit more affordable and investing more profitable. But this comes with high risk of borrowers defaulting loans. Hence there is a need to classify each borrower as defaulter or not using the collected data.

## 

## 1.4 Data Sources

The provided dataset corresponds to all loans issued to individuals in the past from 2007-2015. The dataset has 855969 observations and 73 features. The data contains the indicator of default, payment information, credit history, etc. Customers under 'current' status have been considered as non-defaulters in the dataset. We have also been provided with a Data dictionary that describes the features.

## 1.5 Tools & Techniques

Tools: Python 3.7.2, Jupyter Notebook, Pandas, Matplotlib, Seaborn, Sklearn, Numpy

Techniques: Logistic regression, Random Forest Classifier, Confusion Matrix, Classification Report, Decision Tree Classifier, K Neighbors Classifier.

# CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below

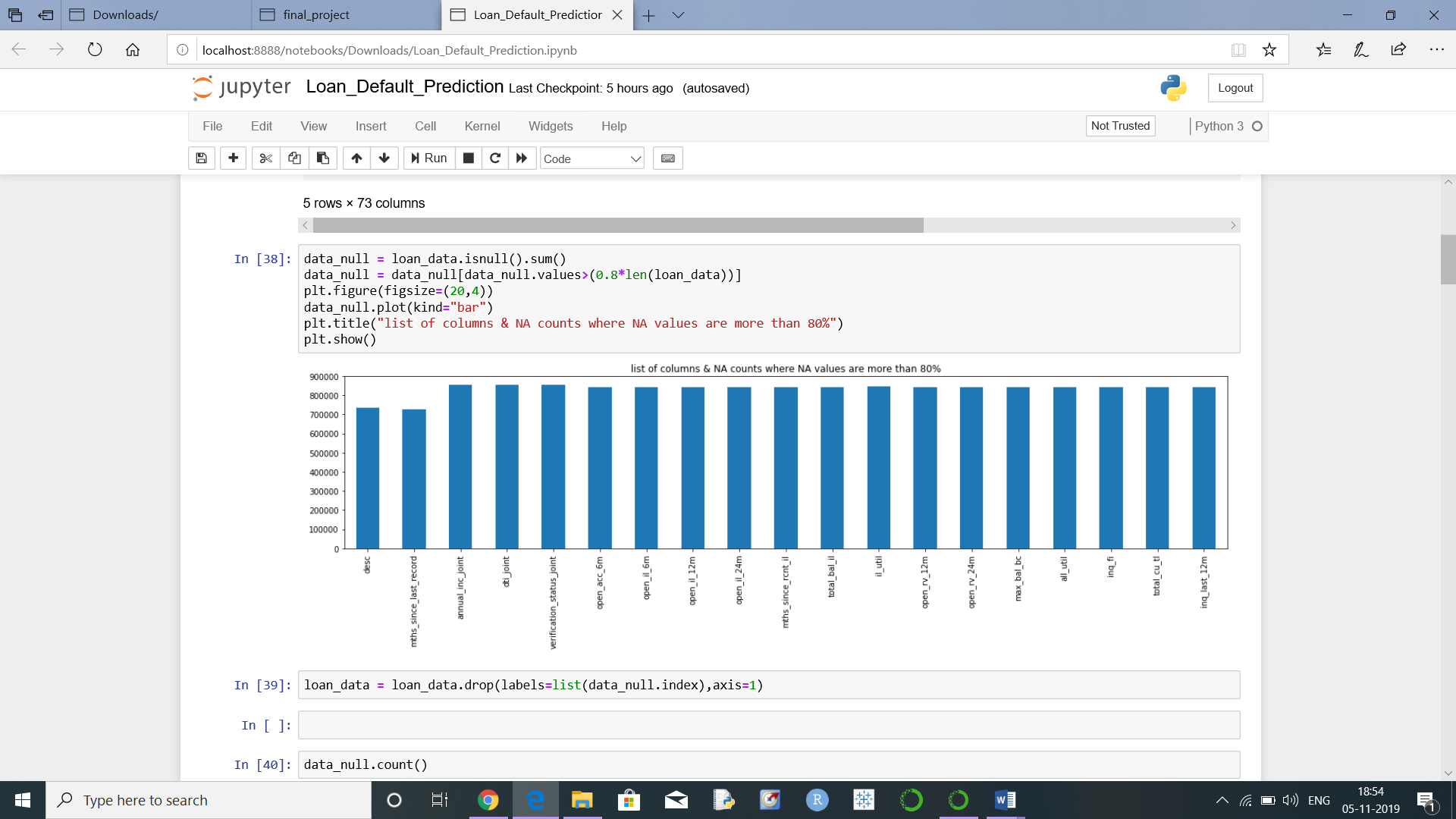
## 2.1 Phase I – Data Extraction and Cleaning:

* **Missing Value Analysis and Treatment**

In the dataset the target variable shows that 94% of features are not defaulted and 6% are defaulters or charged off. So we can clearly say this is an unbalanced dataset.

The first step is to know whether the columns are filled with useful information or empty. Data exploration exposed many empty or almost empty columns which were removed from the dataset because it would be a difficult task to go back and try to answer for each data factor that is not necessary at the time of the loan application.

The dataset has 73 features including the target in which 19 features have missing values or NAN. Below we will look at a plot and get some insights.



Insights: So, we can see from the above plot that there are 19 columns which has complete NAN values.

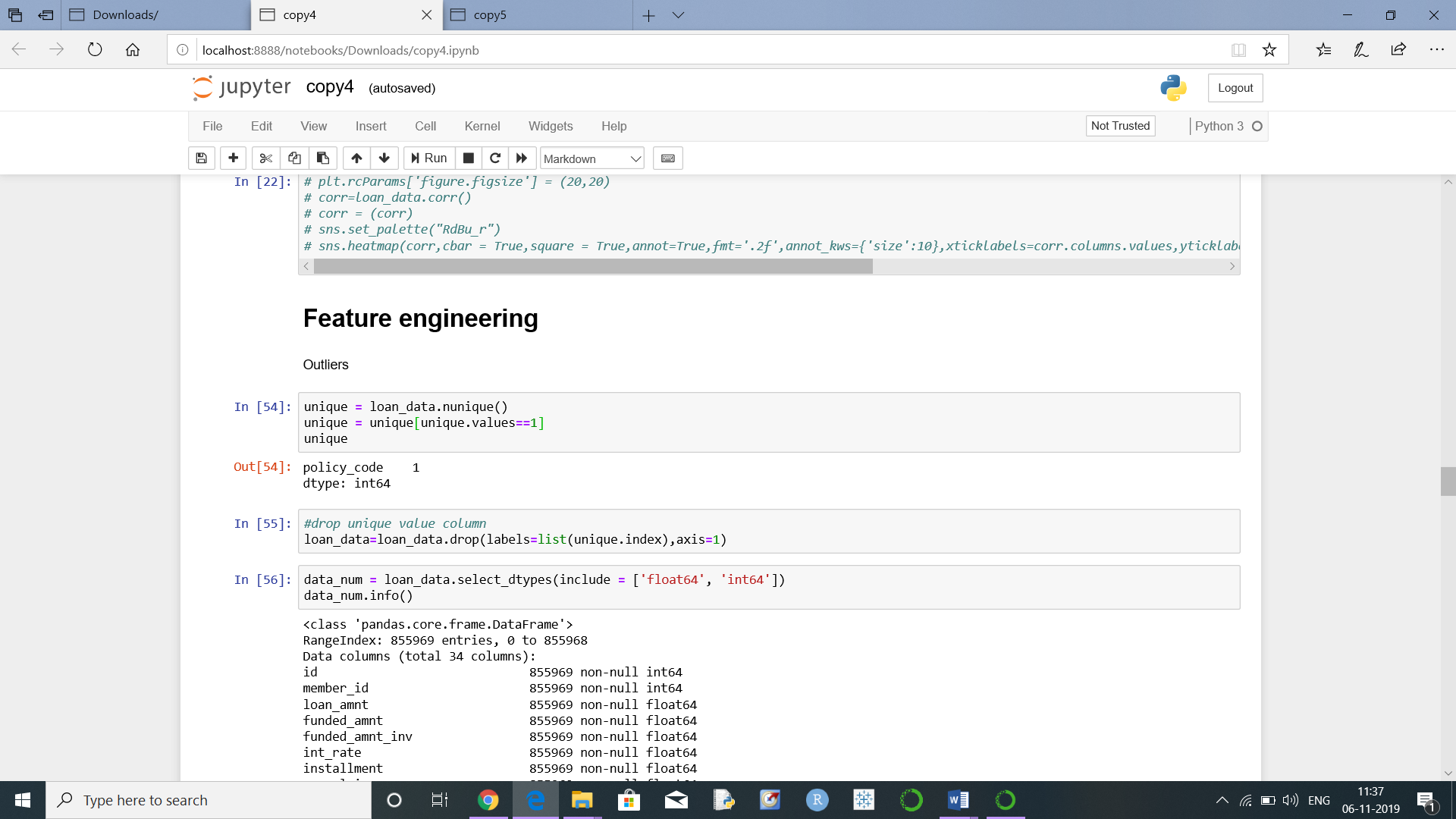
As we can see there are 855969 observations & 73 columns in the dataset, it will be very difficult to look at each column one by one & find the NA or missing values. So let's find out all columns where missing values are more than certain percentage, let's say 80%. We will remove those columns as it is not feasible to impute missing values for those columns.

Out of 73 features we only kept 54. So we removed about 19 features that had more than 40% missing values, since it doesn’t make any sense in further exploration.

Some irrelevant columns Unique ID's such as "id","member\_id" because they did not provide any useful information about the customer. As last 2 digits of zip code is masked 'xx', we can remove that as well.

* **Handling Outliers**

We explored the columns where number of unique value is only 1. We have removed columns where number of unique value is only 1 because that will not make any sense in the analysis. That unique value is policy\_code so we have removed it.



After removing the variables with low variance and all irrelevant columns we are left with (855969, 37) rows & columns.

* **Feature Extraction**

### Deciding Target Column

Has we learned from the description of columns in the Data Dictionary that **default\_ind** is the only field in the main dataset that describe a loan status, so let’s use this column as the target column.

**Transformation**

We have transformed term, emp\_length, grade, homeownership, to integer values using transformation technique so that they provide some information to our model.

We are dropping columns which has more number of categories since it is not feasible to encode them. The variable which we removed are sub\_grade', 'emp\_title', 'title', 'verification\_status',’purpose','addr\_state','earliest\_cr\_line','last\_pymnt\_d', 'last\_credit\_pull\_d'. We're not removing issue\_d because the variable need to be used for splitting.

## 2.2 Phase II - Feature Engineering

**Mapping:**

We are mapping the issue date from “June 2007 - May 2015” as Train set and "Jun-2015" to "Dec-2015" as Test set for the easiness of splitting.

**Filling missing values:**

We are fill the numerical missing values with the median value using the fillna function.

**One Hot Encoding**

Since we have some categorical variables for the analysis and the machine learning algorithms doesn't take categorical and string variables directly, we have encode them using label encoder .

Finally we have selected 37 features that are necessary for our model to make the prediction.

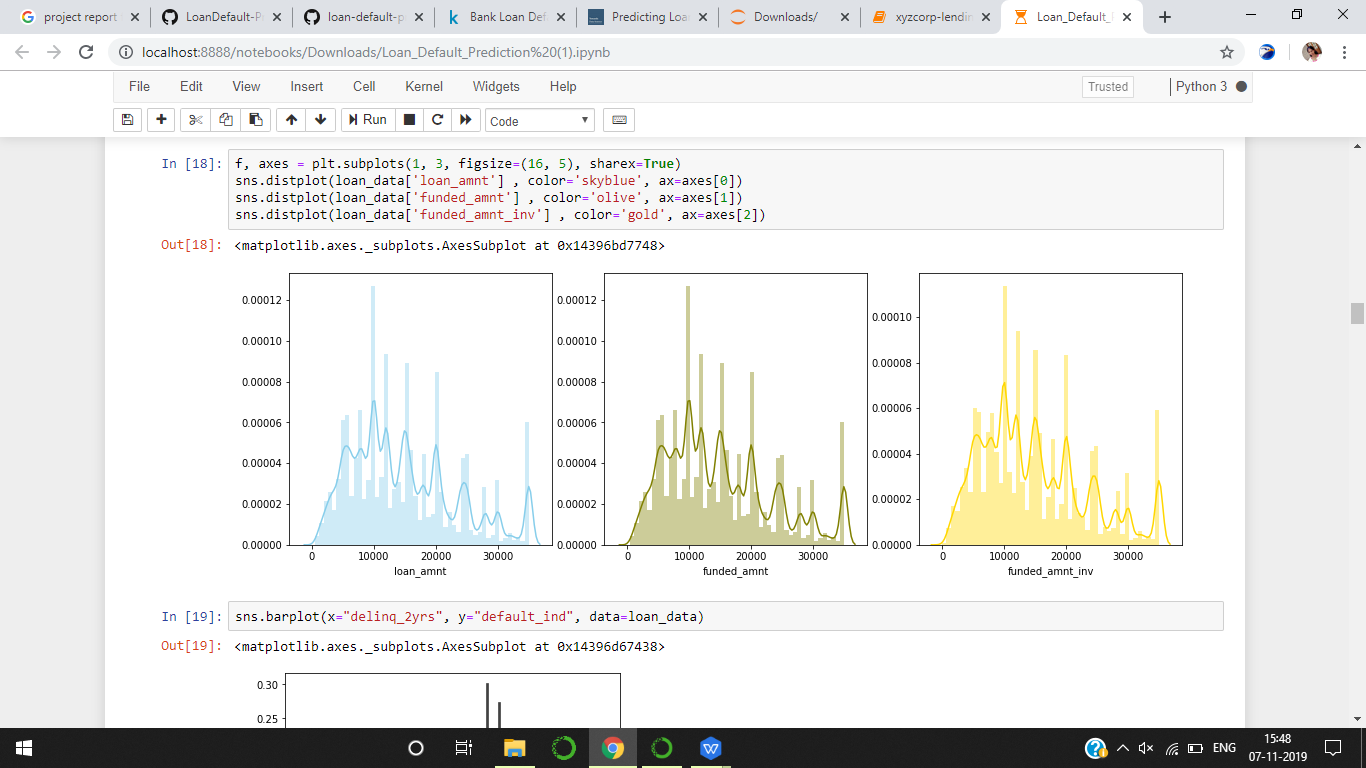
**2.3 Data Dictionary:**

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| addr\_state | The state provided by the borrower in the loan application |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| desc | Loan description provided by the borrower |
| 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. |
| 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 |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 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. |
| emp\_title | The job title supplied by the Borrower when applying for the loan. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| grade | XYZ corp. assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| id | A unique assigned ID for the loan listing. |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month XYZ corp. pulled credit for this loan |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| 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. |
| **loan status** | Current status of the loan |
| member\_id | A unique Id for the borrower member. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| next\_pymnt\_d | Next scheduled payment date |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| sub\_grade | XYZ assigned assigned loan subgrade |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | Principal received to date |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by XYZ corp., not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| total\_bal\_il | Total current balance of all installment accounts |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| all\_util | Balance to credit limit on all trades |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| inq\_fi | Number of personal finance inquiries |
| total\_cu\_tl | Number of finance trades |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| verification\_status | Was the income source verified |

## 2.4 Exploratory Data Analysis:

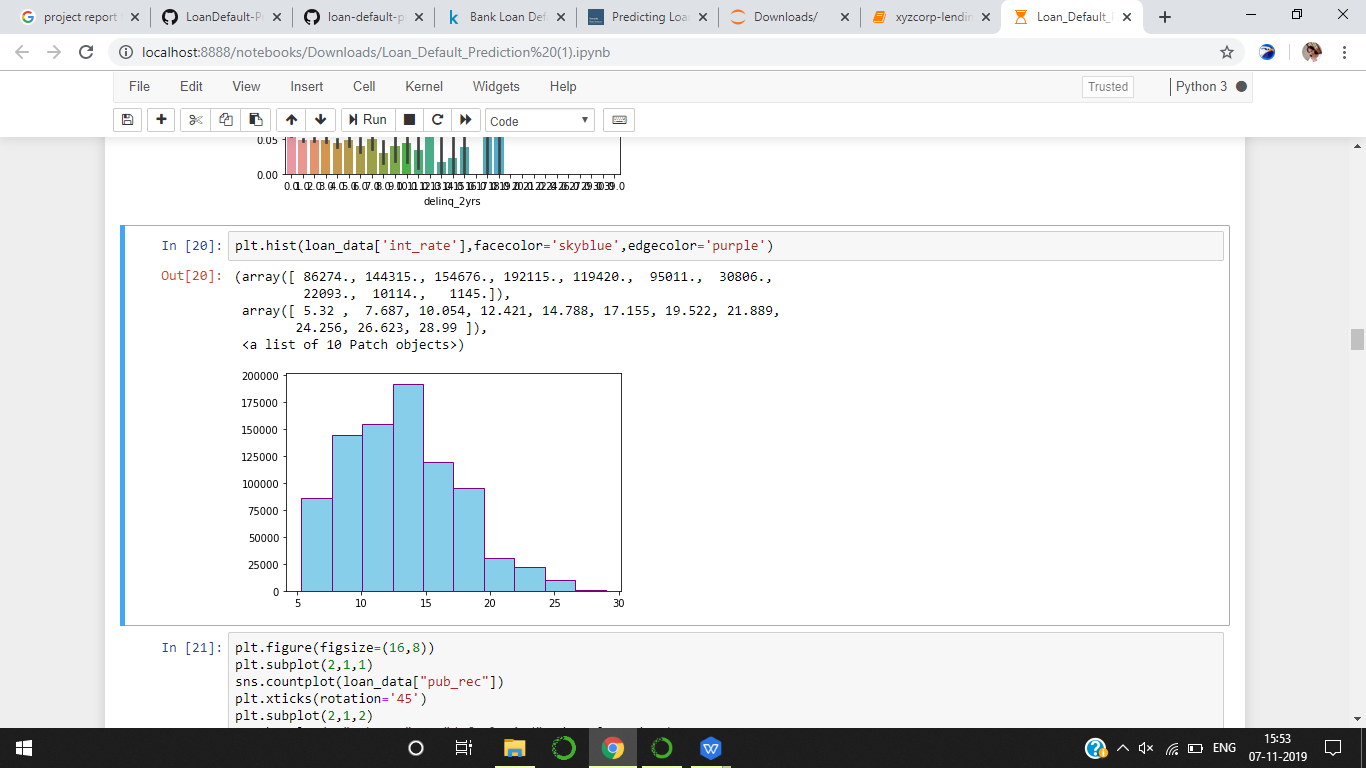
**Derived Features For plotting:**

#### **Loan Amount**



**Insights**: Most of the loan amounts are distributed between 8000 to 20000

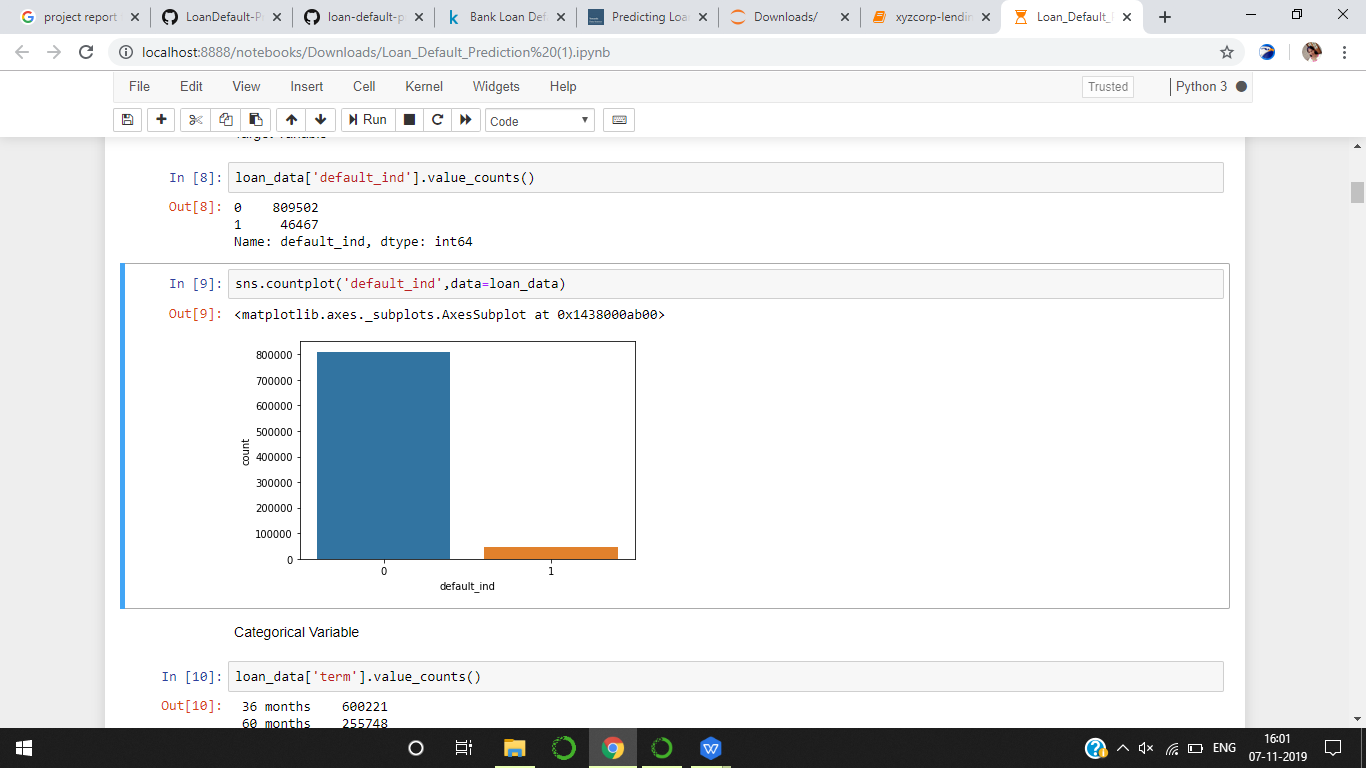
**2.Interest Rate**



**Insights**: Most of the loan amounts are distributed with the interest rate between 10% to 16%.

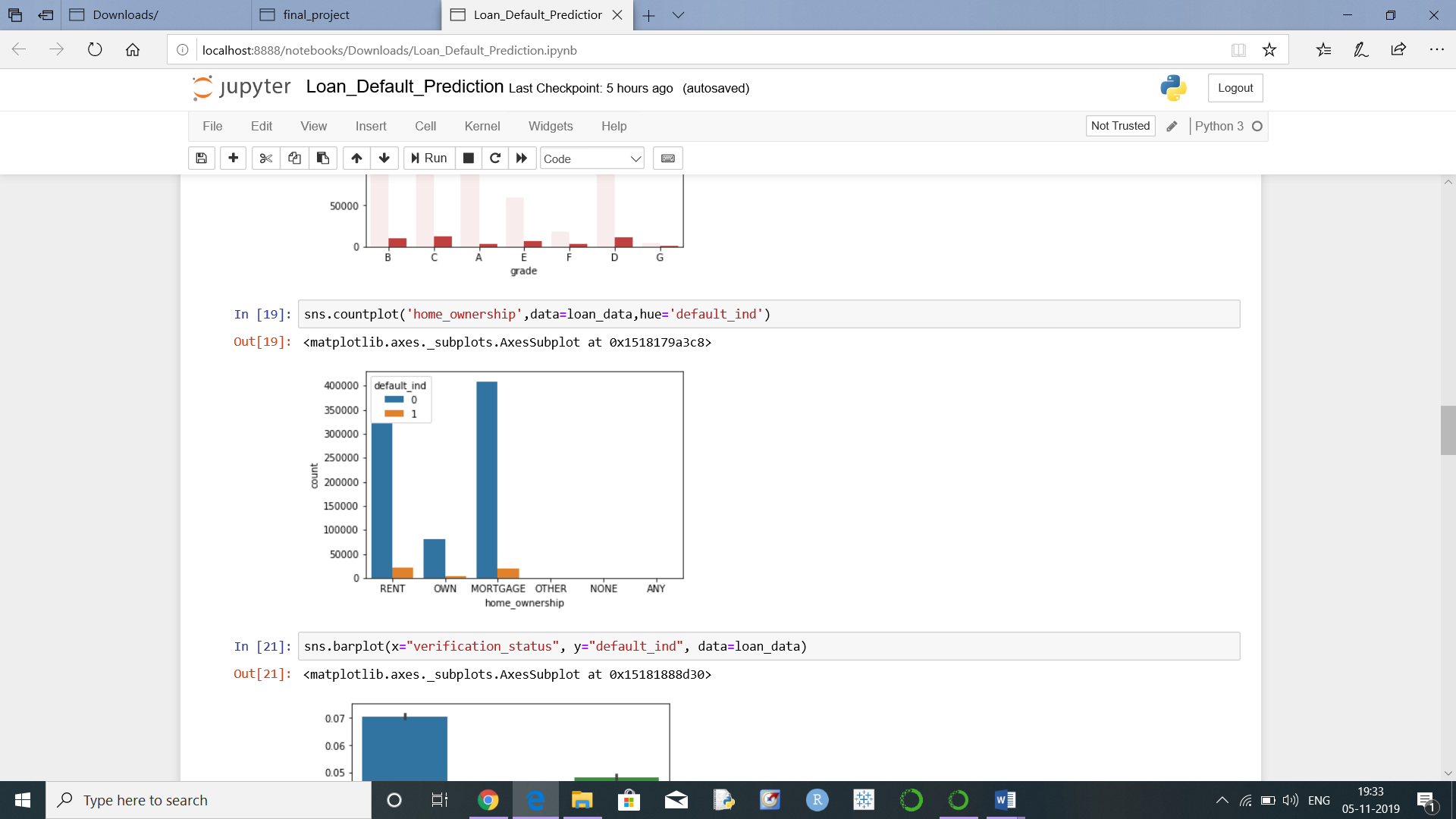
### Categorical Variables:

#### **3. Default Ind**



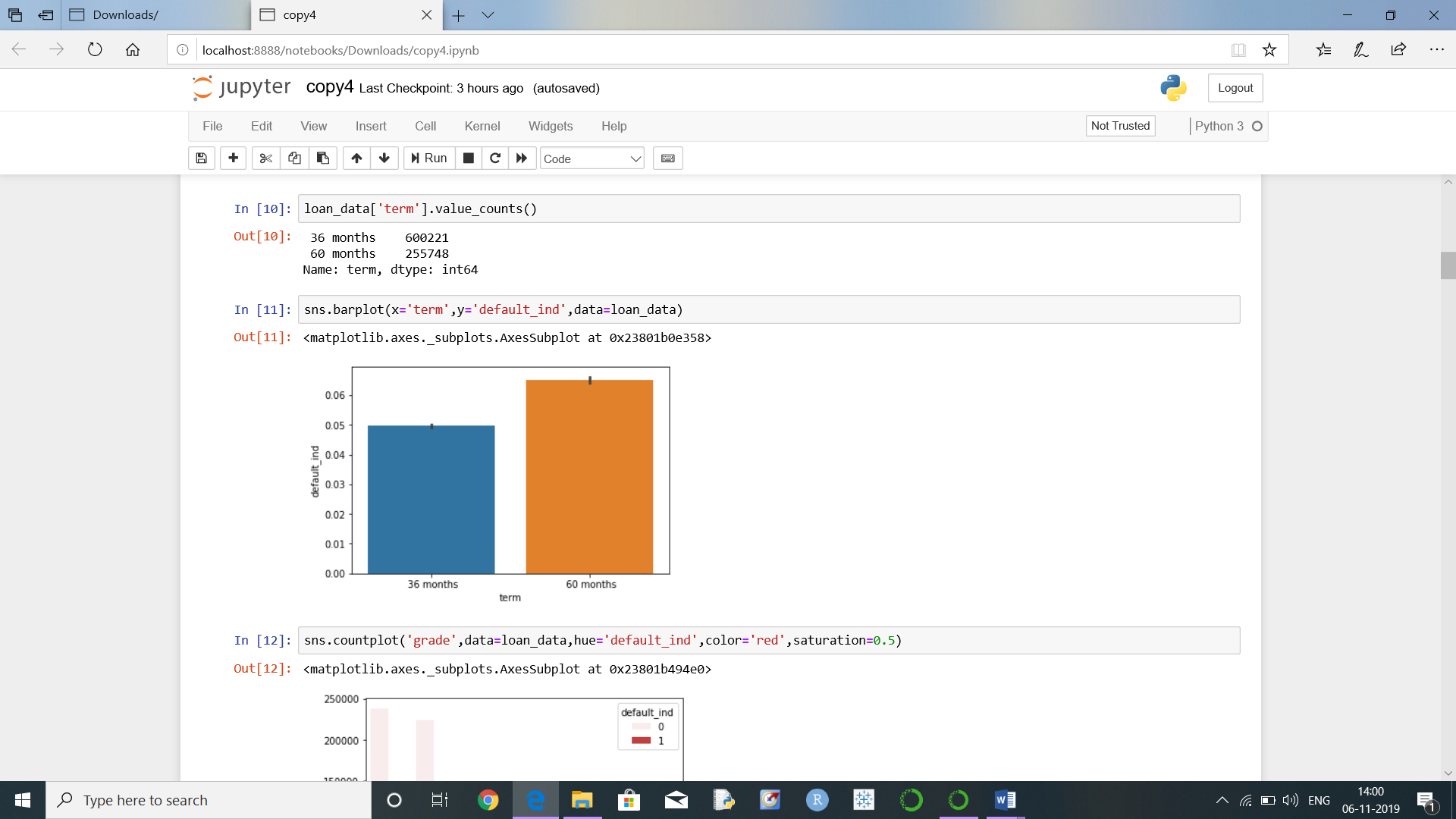
**Insights**: About 6% of loans are charged off.

#### **4. Home Ownership wise Loan**



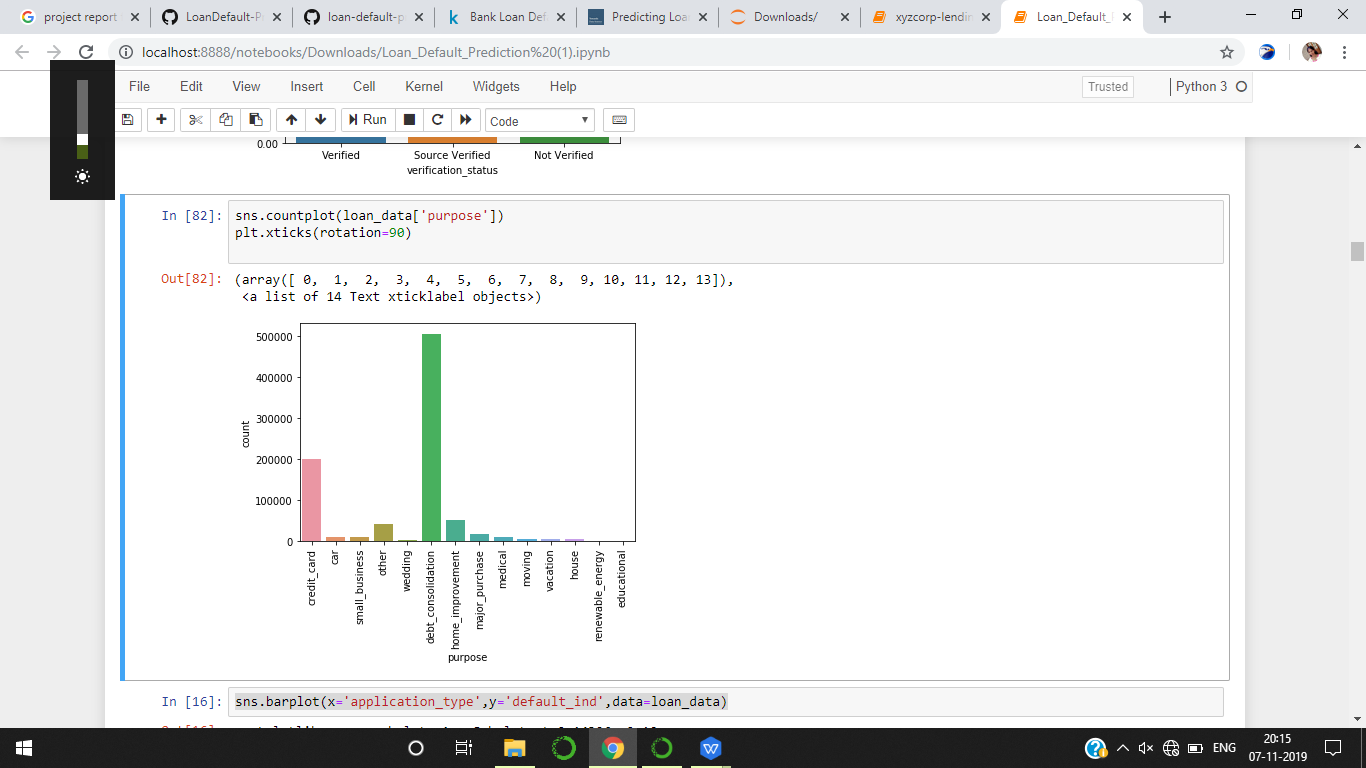
**Insights**: 30% of applicants are living in rented home and 50% of applicants were mortgaged their home.

**5.Term**



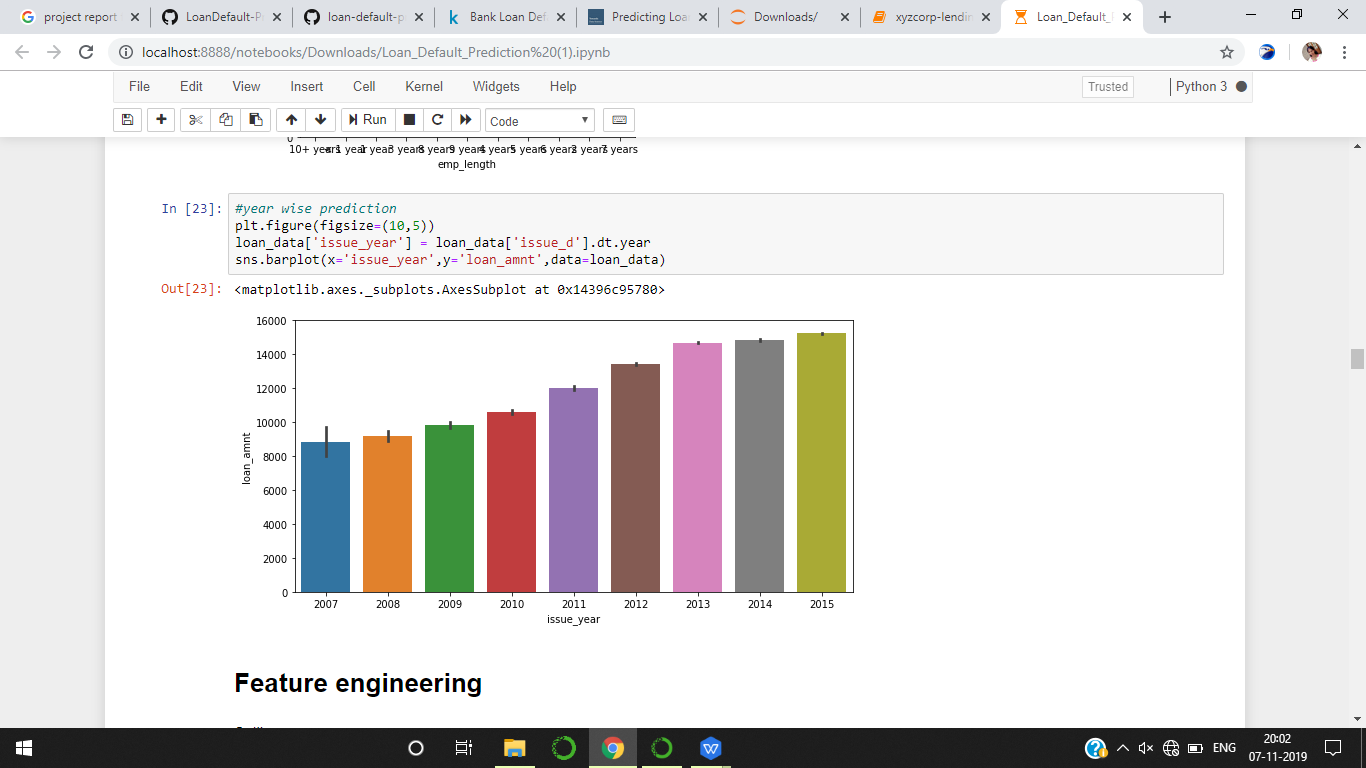
**Insight**: All the defaulters fall under the duration of 60 months.

**6.Purpose of Loan**



**Insights**: Approx. 60% of the applicants applied loan for paying their other loans (Debt Consolidation).

**7.Year wise Loan**



**Insights**: Loan applicants are increasing year on year, most of loan applicants received loans in 2015.

**CHAPTER 3: FITTING MODELS TO DATA**

We have used the below models for our classification:

**Logistic regression**:

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a LogisticRegression Function.

**Random forest Classifier:**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

**Decision Tree:**

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

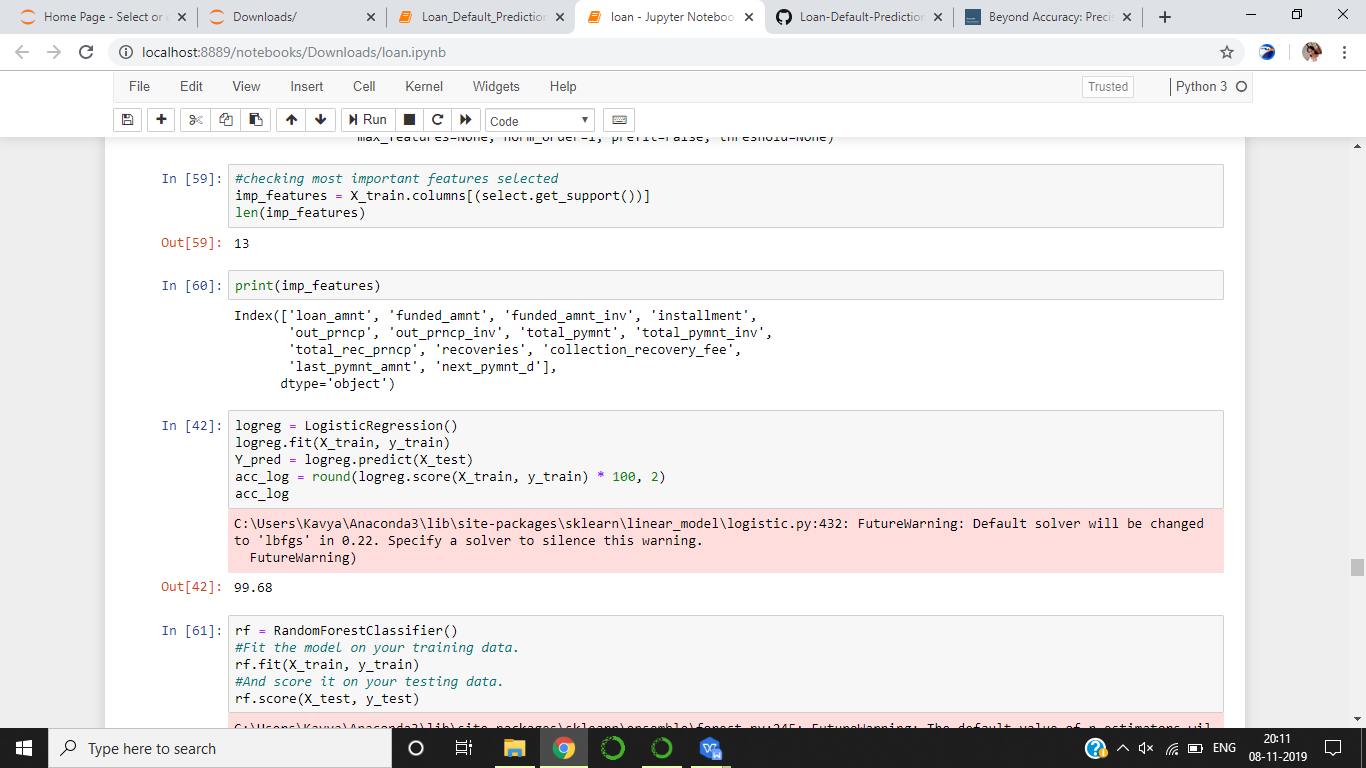
**K Neighbors Classifier:**

KNN is a non-parametric, lazy learning algorithm. When we say a technique is non-parametric, it means that it does not make any assumptions about the underlying data.

The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required

## 4.MODEL SELECTION

Before fitting the model with machine learning algorithms we are finding the important features for the model.



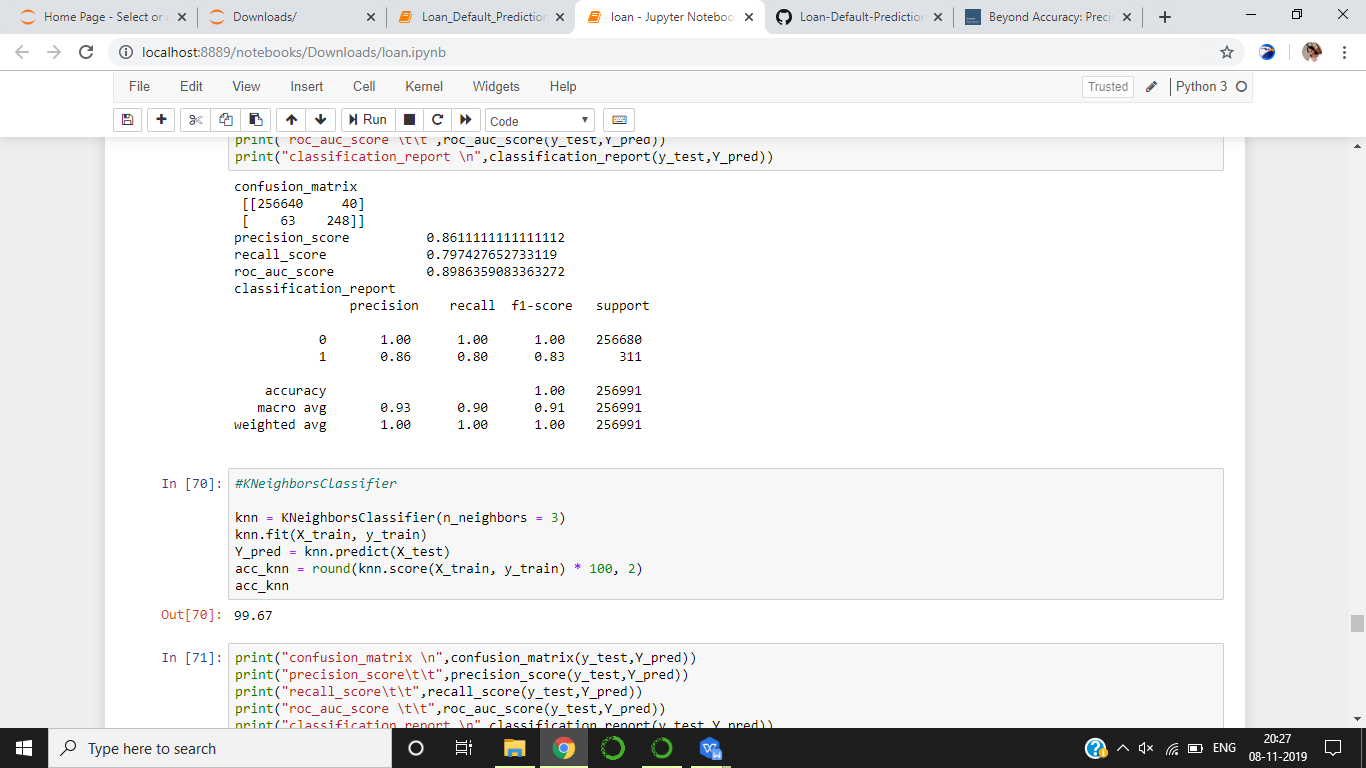
The above displayed variables are the important features selected for the model.

Based on this we are removing the columns which has least importance those are ['open\_acc','total\_acc','revol\_bal','revol\_util','collections\_12\_mths\_ex\_med','total\_rec\_late\_fe',

'acc\_now\_delinq','tot\_coll\_amt','tot\_cur\_bal', 'total\_rev\_hi\_lim']

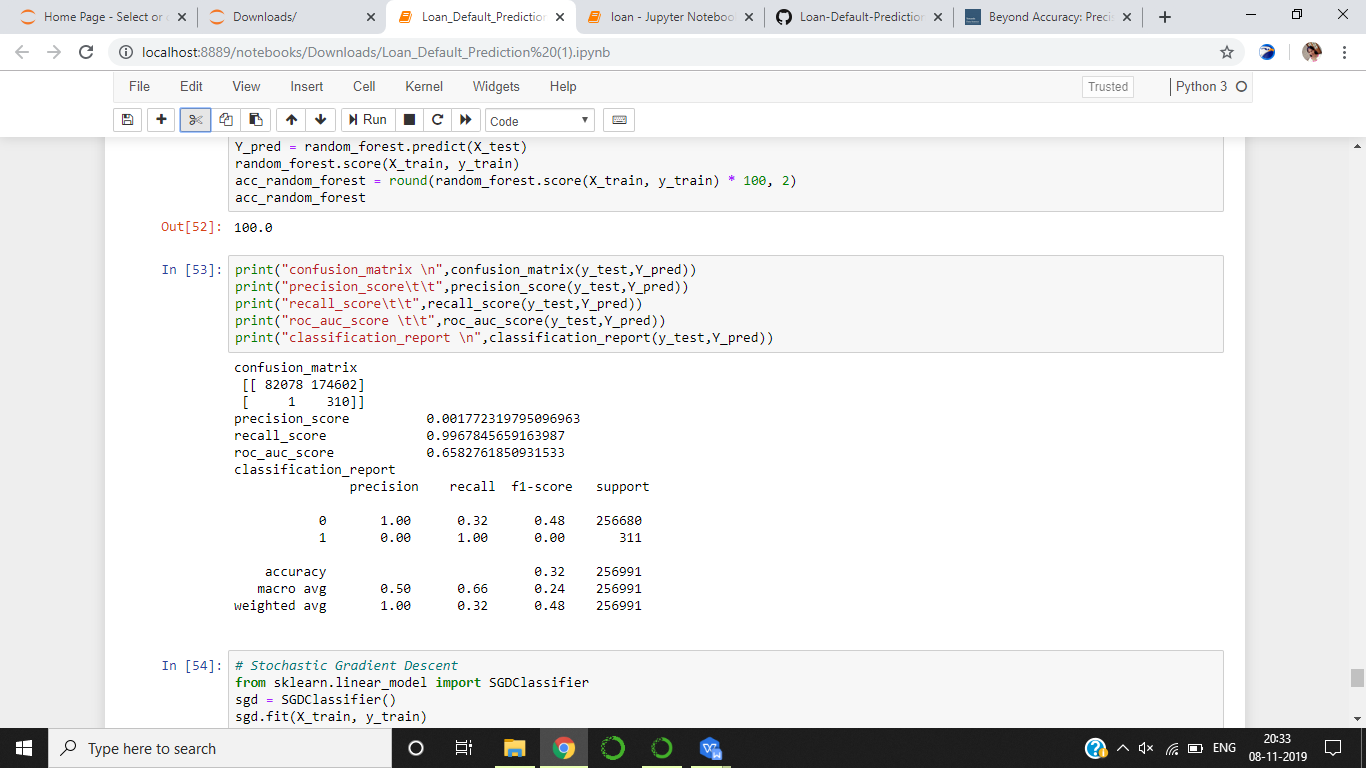
**4.1 LOGISTIC REGRESSION**

This model proving 99.69% of accuracy score and 0.86% of precision score. Confusion matrix also providing quite good values.



4.2 RANDOM FOREST

We applied Random Forest on the Training data set to validate if any further improvement of the model can be performed post the Logistic regression. Below were the parameters which were applied for Random Forest.



The model is providing 100% of accuracy but the confusion matrix is not looking good.

CHAPTER 5: KEY FINDINGS

Significant Variables identified in linear models are also used in Random forest

Below table provides a snapshot of the various models which the business can choose from based on the pros and cons of each model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No** | **Model Name** | **Accuracy** | **precision** |
| **1** | Logistic\_Regression\_Model | 99.69% | 83% |
| **2** | Random\_Forest\_Model | 100% | 0% |
| **3** | Decision\_Tree\_Model | 100% | 0% |
| **4** | K Nearest Neighbour\_Model | 99.67% | 10% |

**Below are some of the key findings:**

All the models can predict who are going to pay off the loan with a good accuracy of 99% but we cannot consider all the models has best model for prediction. We should consider other factors like precision score, F1 score and confusion matrix. By looking after all the factors we are considering Logistic Regression as a good model.

CHAPTER 6: RECOMMENDATIONS AND CONCLUSION

Conclusion:

We have successfully built a machine learning algorithm to predict the applicants who can charged off or default their loans.

As the dataset contains only 6% of people might default their loan, the model may not be efficient to predict if we pass the updated dataframe with increasing number of defaulters. So we have to improve the model by applying some other techniques and selecting the variables.

**Business Insights and Recommendations:**

The facts from our analysis shows that Applicants who has taken the Loan for 'small business' has the highest probability of charge off. Hence, bank should take extra caution like take some asset or guarantee while approving the loan for purpose of 'small business'.

Banks should consider "Grade" as a major variable while providing loans.

Also, As the annual income is decreasing the probability that person will default is increasing. The banks should either start with less principal loan amount and check the credibility.

CHAPTER 7: REFERENCES

1. <https://www.kaggle.com/deepanshu08/prediction-of-lendingclub-loan-defaulters>
2. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html>
3. [https://scikit-learn.org/stable/model\_selection.html#model-selection](https://scikit-learn.org/stable/model_selection.html" \l "model-selection)
4. [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html" \l "sklearn.ensemble.RandomForestClassifier)
5. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
6. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>