

**CS-401 Machine Learning and neural networks**

Module Lecturer: Dr. Barak Pearlmutter

2020-2021 Semester 1.

Student Name: Atikant Negi

Student ID No.: 19253717

Submission Date:

Word Count:

**­­Maynooth University**

**Declaration**

I declare that this project which I submit in fulfilment of the assessment requirements on cs-401 is my own work.

Atikant Negi

Msc. In Data science and analytics

**CS-401 Project (Loan Prediction) Report**

**Objective:** The objective of this Project is to train a classifier, attempting to get good generalization performance on the “loan\_data”set.

**Tools Used:** Python and Jupyter Notebook.

**Training Set:** loan\_data\_set (614 rows) has been divided into 2 set train set (80%) and validation set (20%) to check the accuracy of the classifier on the known outputs. The data set was taken from <https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset?select=train_u6lujuX_CVtuZ9i.csv> .The dataset contains 12 predictor variables (continuous) and one response variable (Y| N ).

**Steps taken to achieve the objective:**

1. Preprocessing of the data.
   * Checking if the dataset is balanced or unbalanced
   * Looking for null or empty data.
   * Transforming the data.
   * Relation between predictors and response variables
   * Splitting the train data into train and validation set
2. Training various classifier.
   * Logistic regression
   * Random forest
   * Artificial Neural Network
3. After the accuracy were observed for different models. The model with best accuracy was taken and various variants of that model was run to increase the accuracy.
4. Confusion metrics and various other visual aids were used to depict the results.
5. Output file (test-out.txt) was created for the best model.

**Elaboration on Steps taken:**

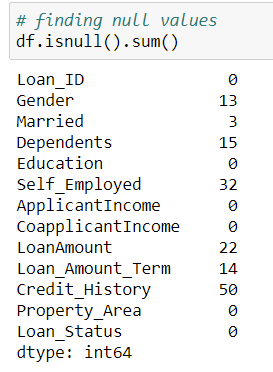
1. **Preprocessing of data**

* Dataset was balanced with the number of Y and N as below:

Y’s -422

N’s -192

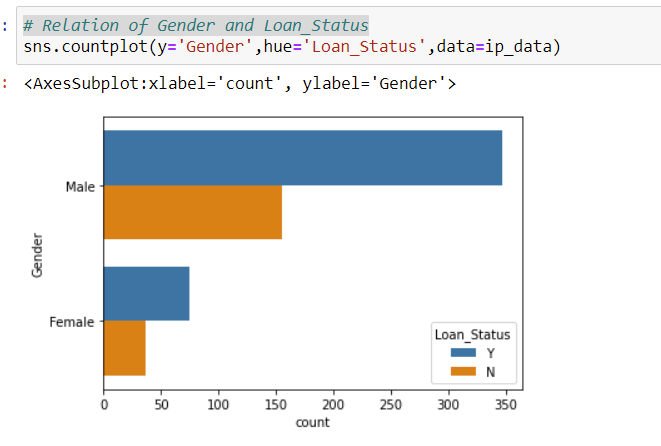
* The data set consist of 12 predictors as:
  + - Loan\_ID Unique ids
    - Gender M|F
    - Married M|Unmarried
    - Dependents continuous data
    - Education Graduate |Not Gradute
    - Self\_Employed Self\_Employed | not Self\_Employed
    - ApplicantIncome continuous data
    - CoapplicantIncome continuous data
    - LoanAmount continuous data
    - Loan\_Amount\_Term continuous data
    - Credit\_History 1|0
    - Property\_Area urban |Rural|semi urban
* The data set had a few missing values.



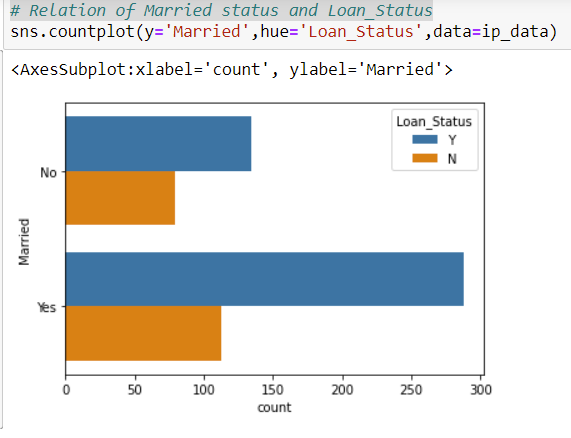
* The data missing values were corrected using the mode of the data of the columns.
* The range is more alike in almost every continuous variable hence scaling would not make any difference.
* No major relation between the variables in the dataset with each other or to the response.
* The categorical variables are converted to numeric variables such as below:
  + Gender ({'Male':1,'Female':0})
  + Loan\_Status ({'Y':1,'N':0})
  + Property\_Area ({'Urban':1,'Rural':2,'Semiurban':3})
  + Married ({'Yes':1,'No':0})
  + Self\_Employed ({'Yes':1,'No':0})
  + Education ({'Graduate':1,'Not Graduate':0})
  + Dependents ({'0':0,'1':1,'2':2,'3+':3})
* The train-set (614 rows) was split in train dataset (80%) and validation set (20%)

**1(b) Some General trends that can be seen are:**

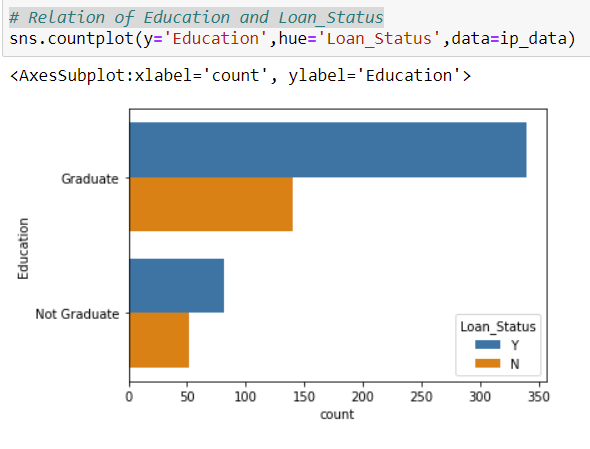
1. **It is seen that more males get loan than females.**



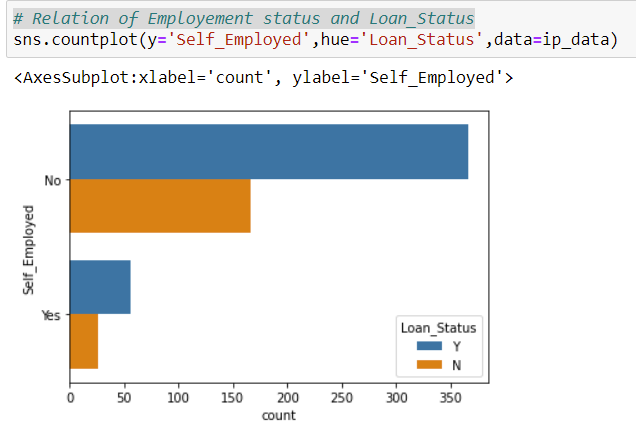
1. **It is seen that married couples collect more loans that unmarried people.**



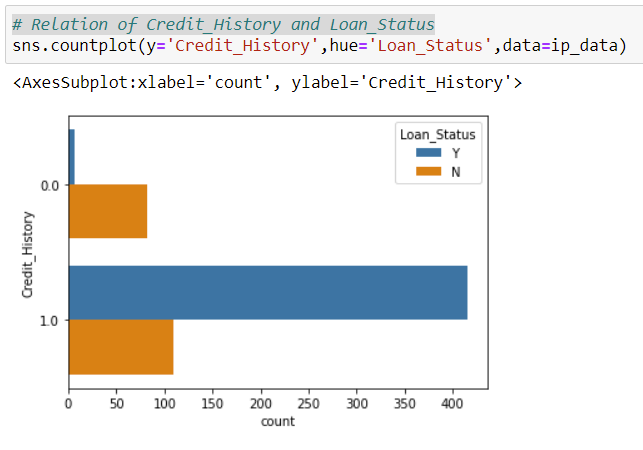
1. **Graduates collect more Loans than non- graduates.**



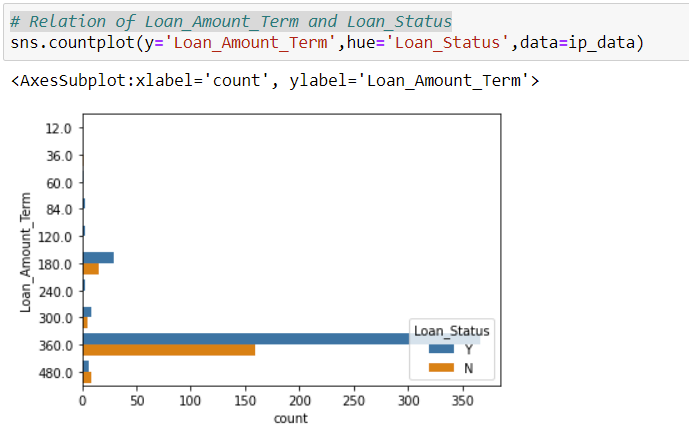
1. **Salary earners receives more loans than self employed people.**



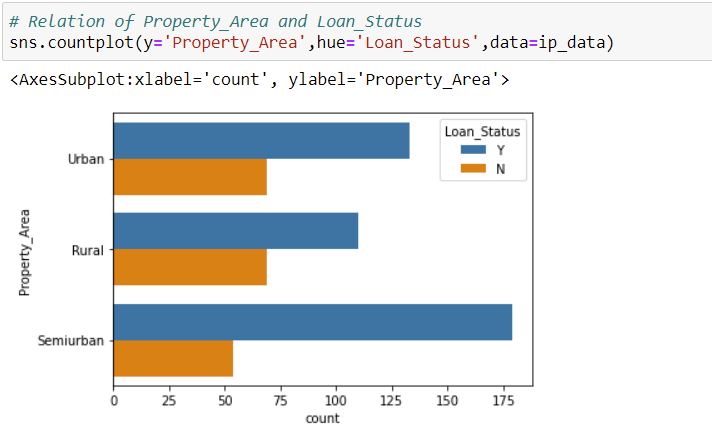
1. **The people with credit history receive more loans than people with no credit history.**



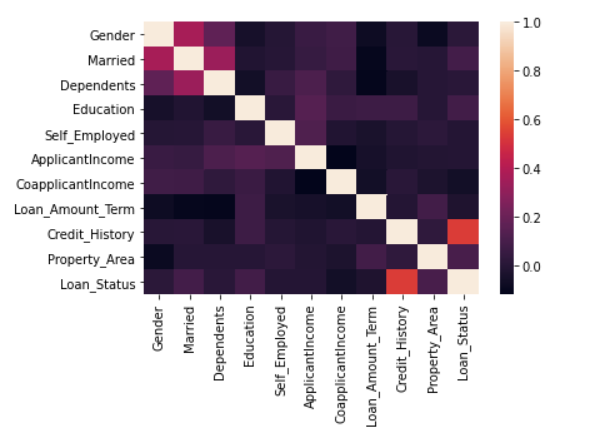
1. **More people receive loan if the loan amount term is 360.**



1. **Semiurban obtain more loan, folowed by Urban and then rural.**



**Correlation heat map:**



1. **Training various classifier**
   * **Logistic Regression**

The Logistic regression model was imported from SKlearn.

* + - * from sklearn.linear\_model import LogisticRegression

**Parameters:**

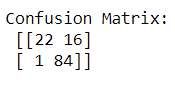
Penalty (default=l2)

dual (default=False)

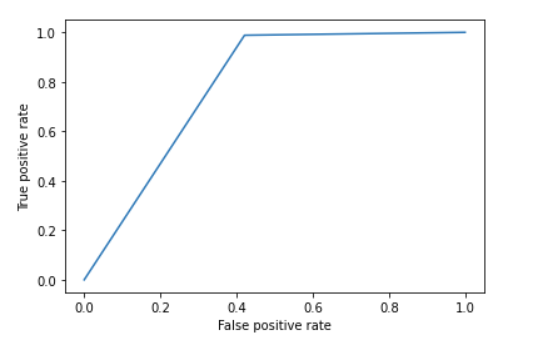
fit\_intercept (default=True)

intercept\_scaling (default=1)

**Confusion matrix:**



**Roc curve:**



The accuracy of the model was computed to be **90.81%**.

* + **Random forest**

The Random forest model was imported from SKlearn.ensemble.

* + - * from sklearn.ensemble import RandomForestClassifier

**Parameters:**

n\_estimators(default=100)

criterion ( default=”gini”)

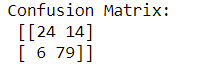
max\_depth (default=None)

min\_samples\_split (default=2)

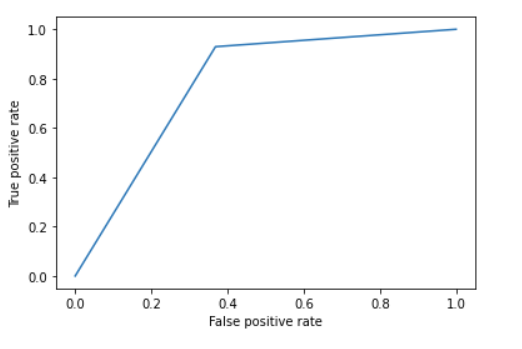
min\_samples\_leaf (default=1)

max\_features ( default=”auto”)

**Confusion matrix:**



**Roc curve:**



The accuracy was calculated to be **88.76%.**

* + **Artificial Neural Network**

The Artificial Neural Network model was imported from keras.models.

* + - * from keras.models import Sequential
      * from keras.layers import Dense

**Parameters:**

Layers: 3 (dense:4,2,1)

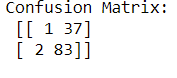
Activation =relu

Optimizer= ‘adam’

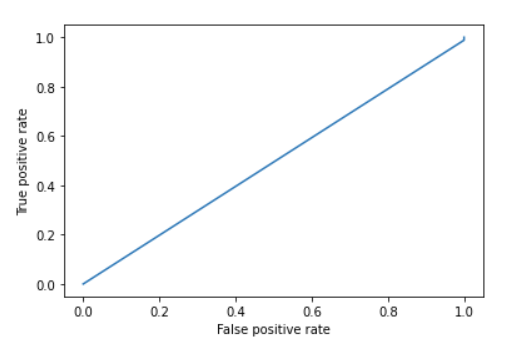
loss='binary\_crossentropy'

metrics=['accuracy']

**Confusion matrix:**



**Roc curve:**



The accuracy was calculated to be **68.29%.**

**Conclusion:**

As it was seen that the highest accuracy was achieved by Logistic regression. So we can clearly state that logistic regression performs better than the other models artificial neural network and an random forest in this case. Though all the models are more prone to errors while predicting the loans status where status is ‘N’ or 0, Artificial neural network performs the worst as it could identify only one case with ‘N’ | 0. This could have occurred due to the imbalance in the number of Y (422) and N (192) in the loan status.