A Project Report on

**Bank Loan Default Case**

Submitted by :

**Abhishek Darekar**

## TABLE OF CONTENTS

**CHAPTER TITLE PAGE NO.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  |  |
| 01 | | Introduction | 3 |
|  | 1.1 | Problem Statement | 3 |
|  | 1.2 | Data | 3 |
| 02 | | Load dataset and Data Pre Processing | 4 |
|  | 2.1 | Understand datatype of required Variables | 4 |
|  |  |  |  |
| 03 | | Missing Value Analysis | 5 |
|  | 3.1 | Find Missing value percentage Column wise | 5 |
| 04 | | Splitting the data set into two sets - existing customers and new customers | 6 |
| 05 | | Checking for Outliers | 7 |
|  | 5.1 | Dealing with outliers | 9 |
| 06 | | Data Exploratory Analysis | 10 |
|  | 6.1 | Bivariate Analysis - (TTest) | 10 |
|  | 6.2 | Bivariate Analysis Visualization | 11 |
| 07 | | Modelling | 12 |
|  | 7.1 | Logistic Regression Model | 14 |
|  | 7.2 | Random Forest Model | 15 |
|  | 7.3 | Decision Tree Classifier | 17 |
| 08 | | Finalize the model | 18 |
|  | 8.1 | Model Evaluation | 18 |
|  | 8.2 | Model Selection and Business Insights | 18 |
|  | |  |  |

Chapter 1 : Introduction

### 1.1 Problem Statement

The loan default dataset has 8 variables and 850 records, each record being loan default status for each customer. Each Applicant was rated as “Defaulted” or “Not-Defaulted”. New applicants for loan application can also be evaluated on these 8 predictor variables and classified as a default or non-default based on predictor variables.

### 1.2 Data

Here we have a data set with following features, we need to go through each and every variable of it to understand and for better functioning.

Size of Dataset Provided: -Rows : 850, Columns : 9 (includes 1 dependent variable)

Missing Values: Yes

Outliers Presented: Yes

Below mentioned is a list of all the variable names and what they stand for:

Attributes: ·

* **Age** : Age of each customer
* **Education** : Education categories.
* **Employment** : Employment status Corresponds to job status and being converted to numeric format.
* **Address** : Geographic area -Converted to numeric values.
* **Income** : Gross Income of each customer
* **debtinc** : Individual’s debt payment to his or her gross income.
* **creddebt** : debt-to-credit ratio is a measurement of how much you owe your creditors as a percentage of your available credit (credit limits)
* **othdebt** : Any other debts

Chapter 2 :  **Load dataset and Data Pre Processing**

It is a good practice to understand the data first and try to gather as many insights from it. Data Pre processing is all about making sense of data in hand.

### 2.1 Understand datatype of required Variables

Attribute Name Data Type.

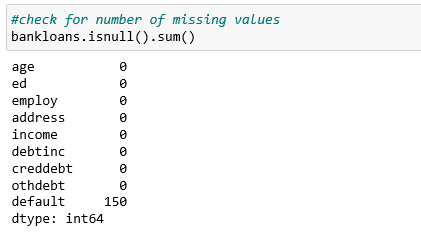
|  |  |  |
| --- | --- | --- |
| **1** | age | Int |
| **2** | ed | Int |
| 3 | employ | Int |
| 4 | address | Int |
| **5** | income | Int |
| **6** | debtinc | Float |
| **7** | creddebt | Float |
| **8** | othdebt | Float |
| **9** | default | Float |
|  |  |  |

Before proceeding we check the data types of variable, Change the data type of variables which are not appropriate so they can be processed correctly moving forward.

Chapter 3 : Missing value Analysis.

There are missing values in the dataset. We need to deal with the missing values Before moving forward.

### 3.1 Find Missing value Number Column wise



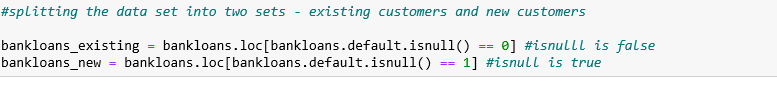
The table indicates the number of missing values in a attribute in relation to the whole dataset.

Now we will decide how to deal with these missing values in order to go forward..

Chapter 4 : **Splitting the data set into two sets - existing customers and new customers.**

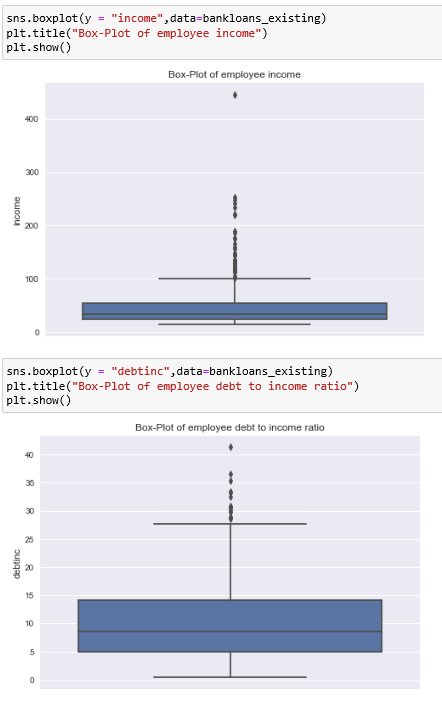
As we found out in the Missing value stage there were missing values only for “Default” column.

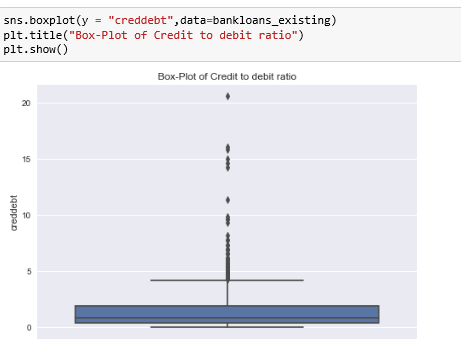
Those were not the missing values but the entries of new customers. So we split the data so we can predict the data later.



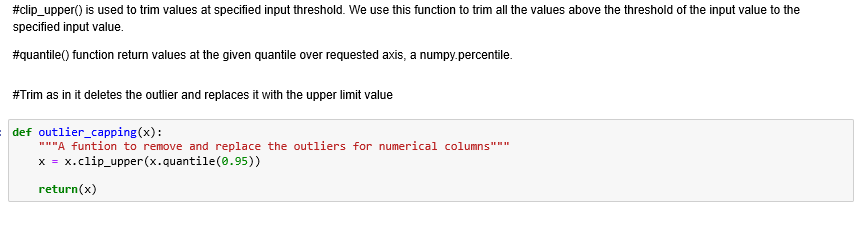
Chapter 5 : **Checking for Outliers.**

|  |
| --- |
| * An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. * Such incorrect values will mean we will be feeding wrong data to our model. * To avoid this we will be removing all the values which are incorrect. * The following are the parameters for removing values from each attribute : |
|  |





### 5.1 Dealing with Outliers



* We have used boxplot to find the outliers.
* So we will use 95th percentile and clip the value so all items above that value will be purned down to 95th.

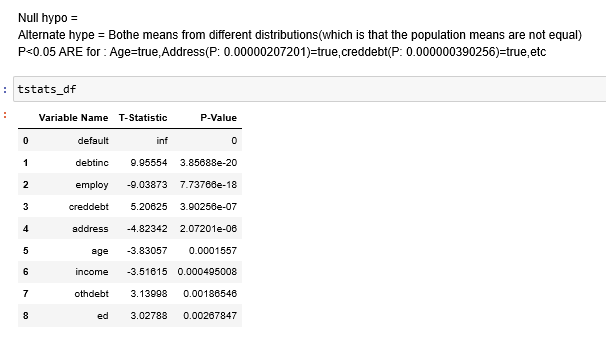
Chapter 6 : **Data Exploratory Analysis.**

* The EDA is an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.
* It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand.

### Bivariate Analysis - (TTest)

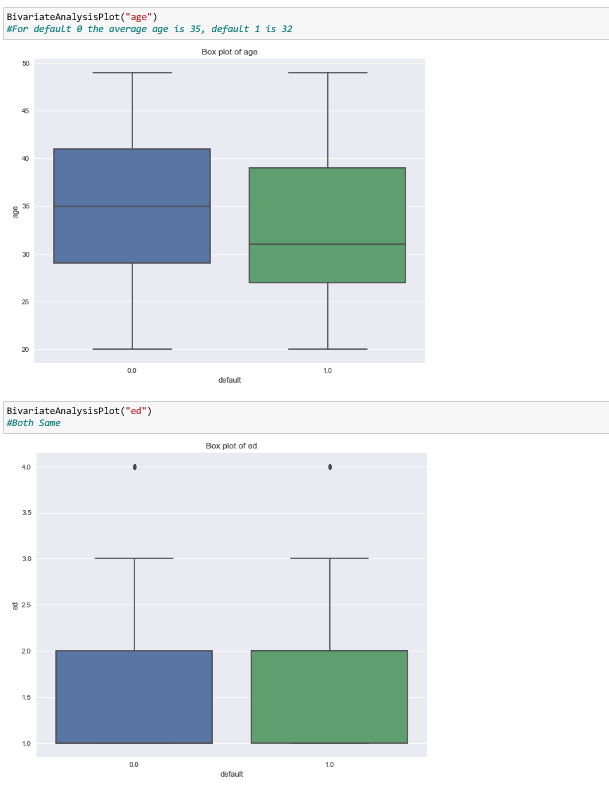
We will do T-Test as we have all Numeric variables.

* The independent samples t-test is a test that compares two groups, on the mean value of a continuous (i.e., interval or ratio), normally distributed variable.
* The model assumes that a difference in the mean score of the dependent variable is found because of the influence of the independent variable that distinguishes the two groups.
* Examples of typical questions that the independent samples t-test answers are as follows:
* Medicine – Has the quality of life improved for patients who took drug A as opposed to patients who took drug B?
* Sociology – Are men more satisfied with their jobs than women?



### 6.2 Bivariate Analysis Visualization.

* It is easier to interpret the findings using visualization.
* So we will visualize all the variables.
* Example :



* + Difference in the mean score of the dependent variable is found because of the influence of the independent variable that distinguishes the two groups.

Chapter 7 : Modelling

After thorough preprocessing and exploratory data analysis phase, the data is ready to enter the model building phase.

In this phase multiple machine learning algorithm will be used and tested to predict the test case i.e the new customers.

Here our target variable i.e. default which is binary (Classification problem) so that here we are using classification models to predict test case.

Different Models we have implemented are :

1.Logistic Regression Model.

2. Random Forest Model.

3. Decision tree model.

Here is the approach used to implement all the below Models.

\*Model Building\*

**1.Fit the model.**

**2.Find coefficient of features.**

**3.Find Intercept of the model.**

\* MODEL PERFORMANCE \*

**4.Predict using the fitted model.**

->Create confusion matrix,Visualize it.

->Find precision,recall,Overall accuracy,support,f1 Score

**Conclusion : Accuracy is good, but many wrong classification.**

**5.Therefore Create ptimum cut-off value.**

\*Find auc score, Draw roc curve

\*Find cutoff value using(Cutoff would be optimum where specificity and sensitivity would be maximum for the given cutoff)

\*Plot it.

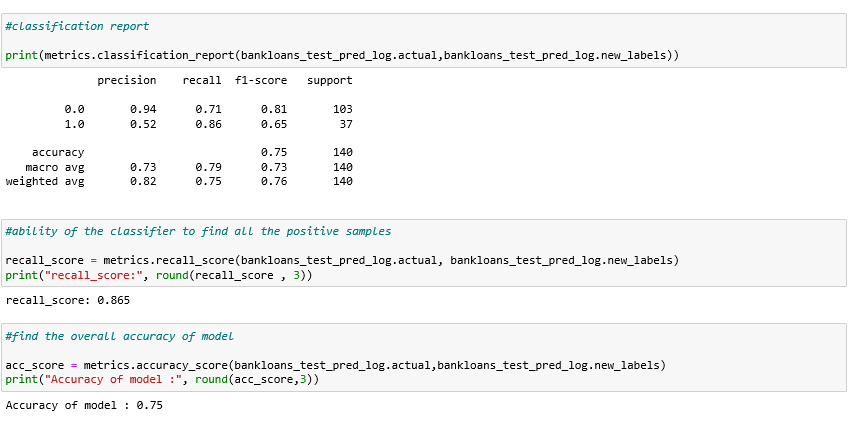
**6.Create confusion matrix,Visualize it.**

**7.Find precision,recall,Overall accuracy,support,f1 Score.**

7.1 Logistic Regression Model.

* [**Logistic regression**](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
* Like all regression analyses, the logistic regression is a predictive analysis.
* Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.





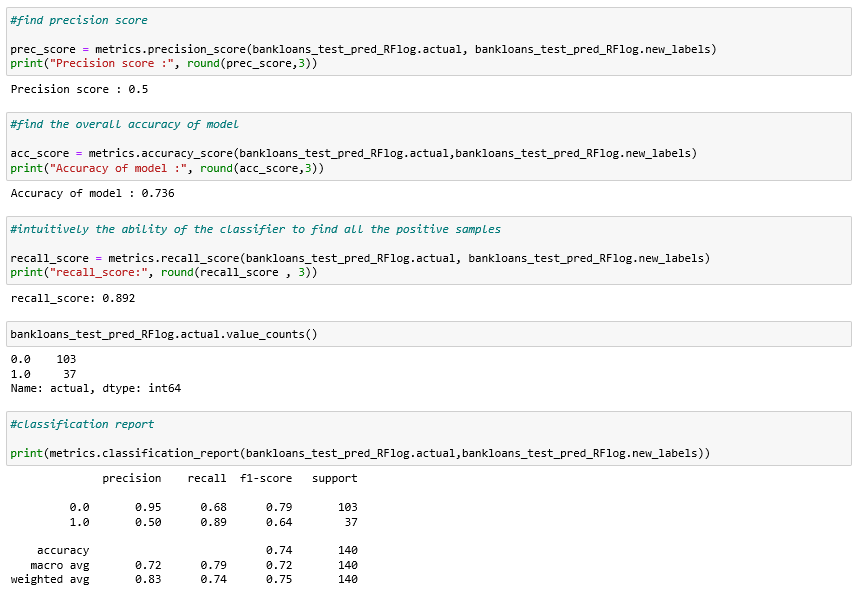
7.2 Random Forest Model.

* Random forests is a supervised learning algorithm. It can be used both for classification and regression.
* It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is.
* It also provides a pretty good indicator of the feature importance.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

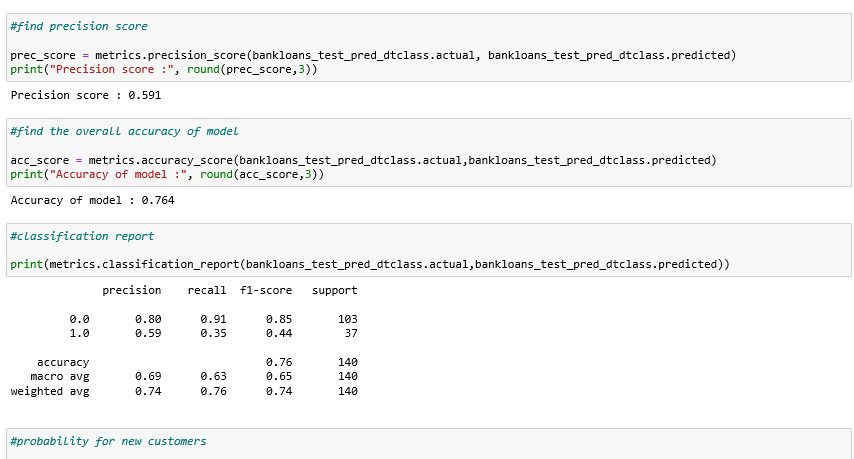
**It works in four steps:**

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.



7.3 Decision Tree Classifier.

* Decision tree can be used for both classification and Regression.
* A **decision tree** is a decision support tool that uses a tree like model of decisions and their possible consequences.
* Decision tree model is a predictive model based on Branching series of Boolean test.
* Each branch connects nodes with “and” and multiple branches are connected by “OR”
* It is Easy to understand for non technical user too.
* **Decision tree learning** is one of the predictive modeling approaches used in [statistics](https://en.wikipedia.org/wiki/Statistics), [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning).
* It uses a [decision tree](https://en.wikipedia.org/wiki/Decision_tree) (as a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling)) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).
* People are able to understand decision tree models after a brief explanation.
* Trees can also be displayed graphically in a way that is easy for non-experts to interpret.



Chapter 8 : Finalize the model.

After implementing different models they need to be studied and one model has to be finalized.

Selection of the final model depends on the type of dataset , different metrics considered.

8.1 Model Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Reg** | **Random Forest** | **Decision Tree** |
| Before optimum cutoff value | | | |
| recall\_score | 0.54 | 0.40 | 0.35 |
| acc\_score | 0.81 | 0.77 | 0.59 |
| After Cutoff | | | |
| recall\_score | 0.86 | 0.86 | 0.86 |
| acc\_score | 0.75 | 0.72 | 0.73 |
| F1 score | 0.66 | 0.63 | 0.63 |

8.2 Model Selection and Business Insights

**- Based on the F1-score (harmonic mean of precision and recall), logistic model with f1 score (for positive labels - default customers) of 0.66 is giving better results than decision tree model with f1 score of 0.44 And Better than Random Forest 0.63. -So we will use the logistic regression model to predict the credit worthiness of the customers**

**-We will Predict the credit risk for remainimg 150 customers using the logistic model with cutoff as 0.224**