DATA WAREHOUSING & DATA MINING

Introduction

In the present world we are generating huge amount of data through which we can make some business decisions and to predict future outcomes.

This project is totally based on the concepts we learnt in big data.

We have used the following resources:

Platform/Software: Jupiter Notebook

Language Used: Python

Concepts: Python, Data science, Big data.

As our project is house loan approval prediction, it is a classification model. So, we make logistic regression, decision tree classifier, random forest classifier and extra trees classifier.

Logistic regression:

Logistic regression predicts the output of a categorical dependent variable. So the outcome must be in the format of Yes or No, 0 or 1 etc, true or False, etc. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Random Forest Classifier:

Random forest is a commonly-used machine learning algorithm, It can be used for both Classification and Regression problems in ML. It combines the output of multiple decision trees to reach a single result.

Decision Tree Classifier:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Extra Tree Classifier:

It is also called Extremely Randomized Trees Classifier. Extra Trees Classifier is an ensemble learning method fundamentally based on decision trees. Extra Trees Classifier, like Random Forest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting

Out Comes:

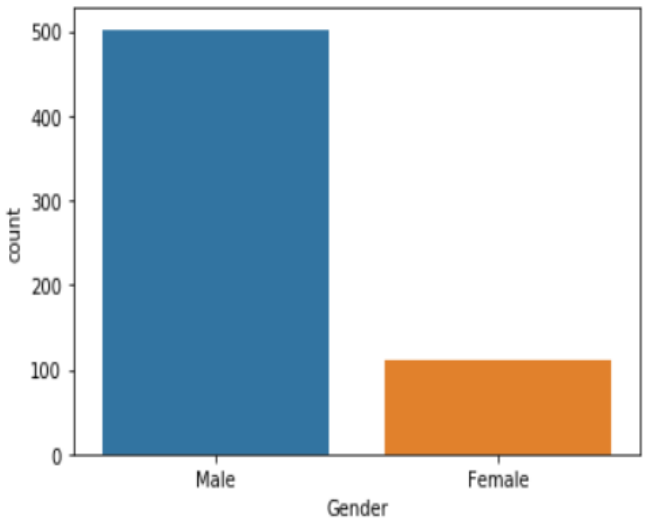
- After this project we are able to find the identify the difference between classification and regression models.
- We understood how the null values effect our model results.
- We understood how to train the data and test the data.
- We increased problem solving skills and programming capability.

Graphs:

Gender:

```
Male      502
Female    112
Name: Gender, dtype: int64
```

```
Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0x1fa489224a8>
```

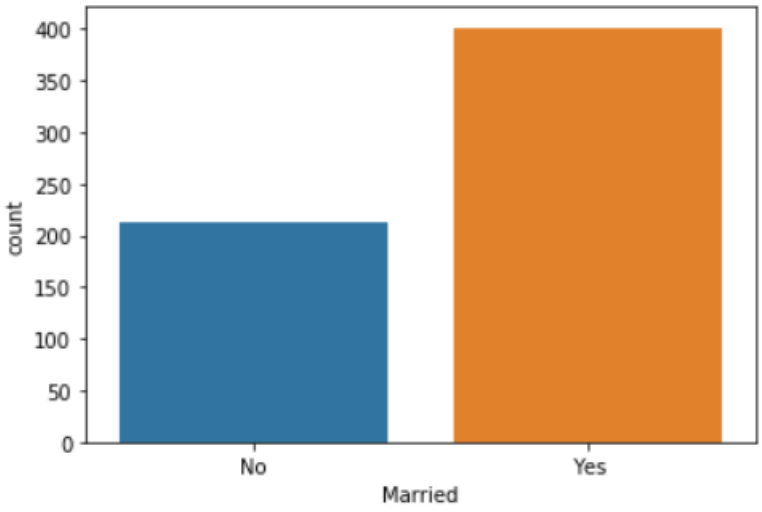


By looking the graph we can analyse that The number of male present are approx. 500

And the female present are approx.-120

Marrital status:-

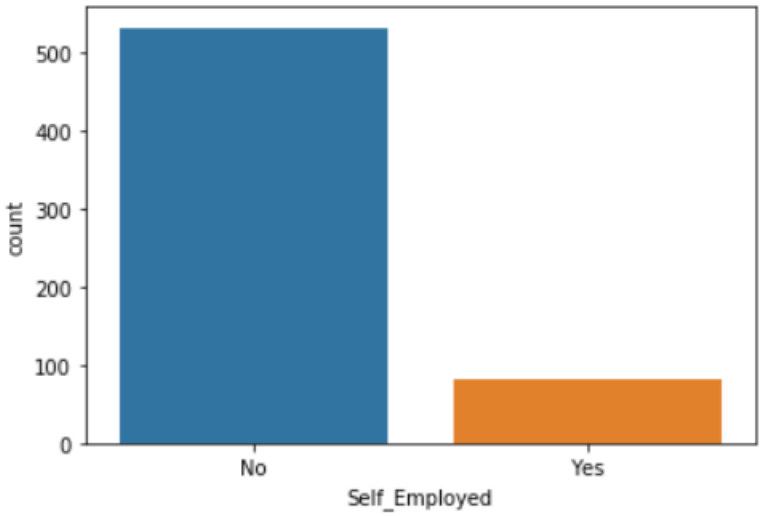
```
Yes    401
No     213
Name: Married, dtype: int64
```



Around 400 are married and 215 are unmarried

Employment:-

```
No     532
Yes     82
Name: Self_Employed, dtype: int64
```

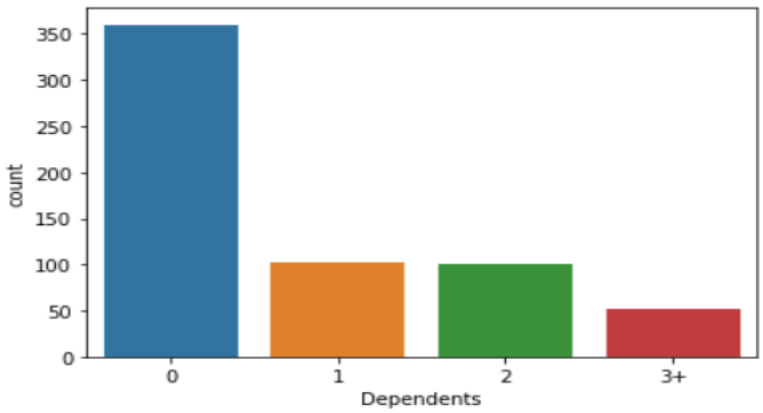


We can see majority of them are not self employed they are nearly:532

And 82 people are employed

Dependents:-

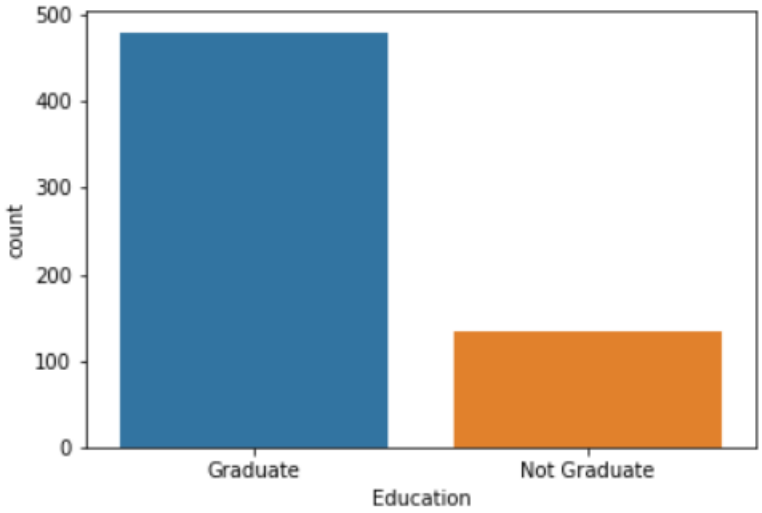
```
0      360
1      102
2      101
3+      51
Name: Dependents, dtype: int64
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1fa48755da0>
```



Education:-

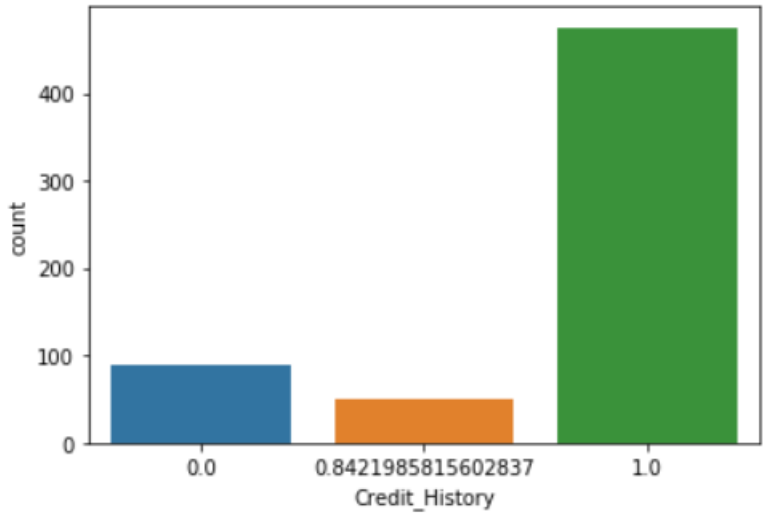
```
Graduate        480
Not Graduate    134
Name: Education, dtype: int64
```

```
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x1fa4880c9e8>
```



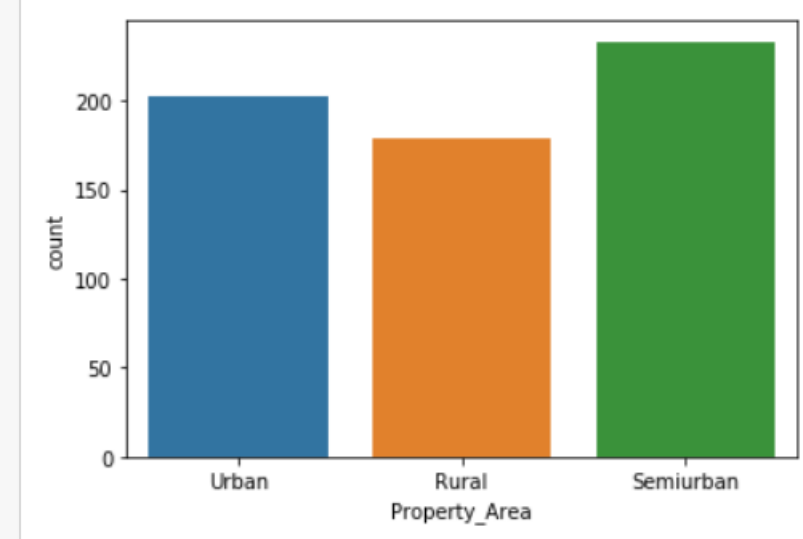
Credit history:-

```
1.000000    475
0.000000     89
0.842199     50
Name: Credit_History, dtype: int64
```



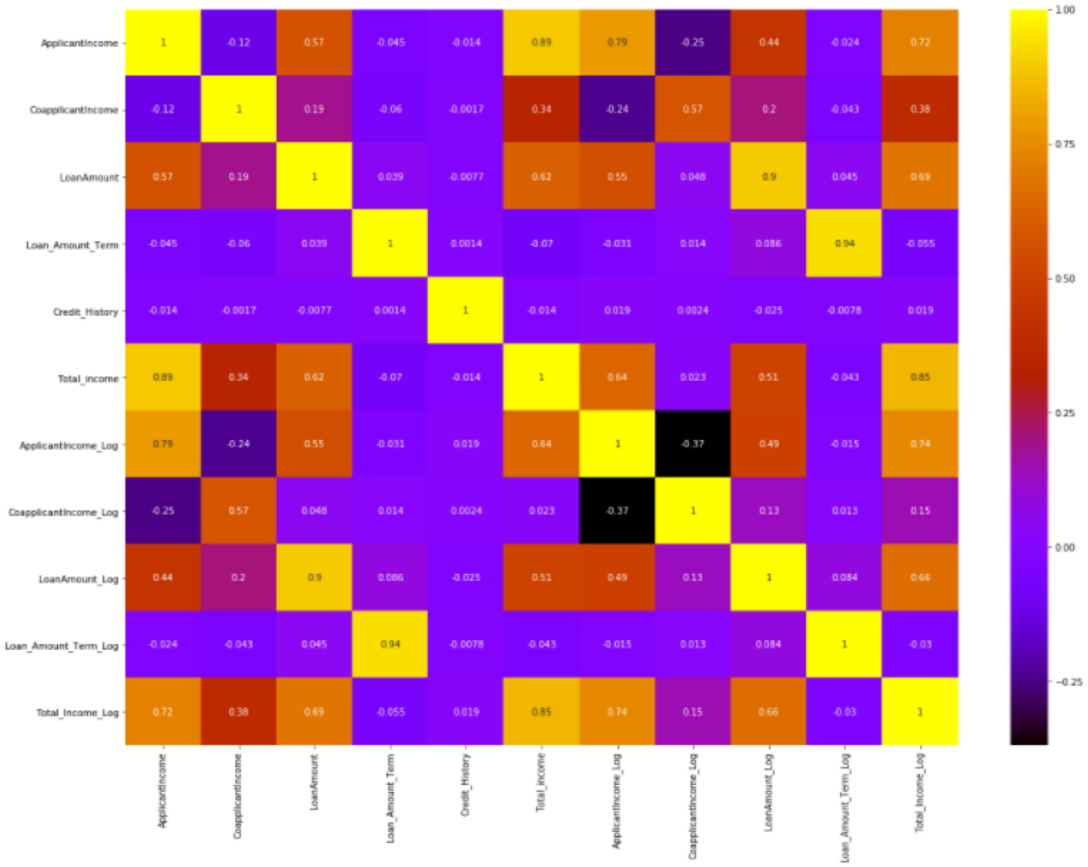
Property area:-

```
Semiurban    233
Urban        202
Rural        179
Name: Property_Area, dtype: int64
```



Correlation Matrix:-

This correlation graph tells the relation between two variables.



Result:

As the data set is classification model we used logistic regression, Decision Tree classifier, Random-forest classifier, Extra Tress classifier by doing all these classifications we got best accuracy for logistic regression.

By doing this project we understood how this classification is working. We can predict weather the person is eligible for getting loan or not by giving some inputs like education, graduation, marital status etc this will tell us to approve the loan or not.