Report On loan Prediction by customer behavior

About the data set ➖

The number of rows in the data set is ➖`252000`

The number of columns in the data set is ➖13

```
#   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Id                 252000 non-null  int64
 1   Income             252000 non-null  int64
 2   Age                252000 non-null  int64
 3   Experience         252000 non-null  int64
 4   Married/Single     252000 non-null  object
 5   House_Ownership    252000 non-null  object
 6   Car_Ownership      252000 non-null  object
 7   Profession         252000 non-null  object
 8   CITY               252000 non-null  object
 9   STATE              252000 non-null  object
 10  CURRENT_JOB_YRS    252000 non-null  int64
 11  CURRENT_HOUSE_YRS  252000 non-null  int64
 12  Risk_Flag          252000 non-null  int64
```

```
This is the info on the data set
There are 7 int columns and 6 categorical columns
This data contains not any null value and neither duplicate value
```

Data cleaning on columns ➖

1. we removed the “_” value that contains in the Profession columns
2. We remove the [5] from the State columns.
3. We remove the “_” from the City columns.

Some Important facts about the data set is ➖

There are a total of 13 columns which there are 12 independent columns and 1 dependent column also known as the target variable.

Independent columns are➖

```
'Id', 'Income', 'Age', 'Experience', 'Married/Single',
       'House_Ownership', 'Car_Ownership', 'Profession', 'CITY', 'STATE',
       'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS'
```

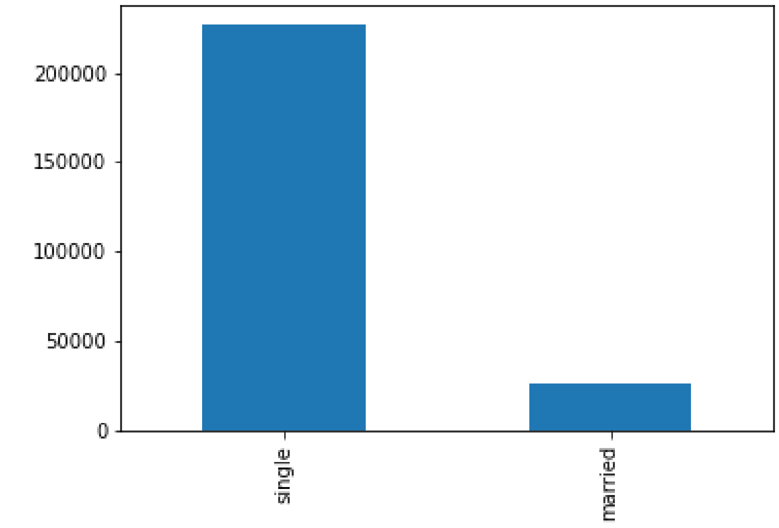
Dependent columns are ➖

```
"Risk_Flag"
```

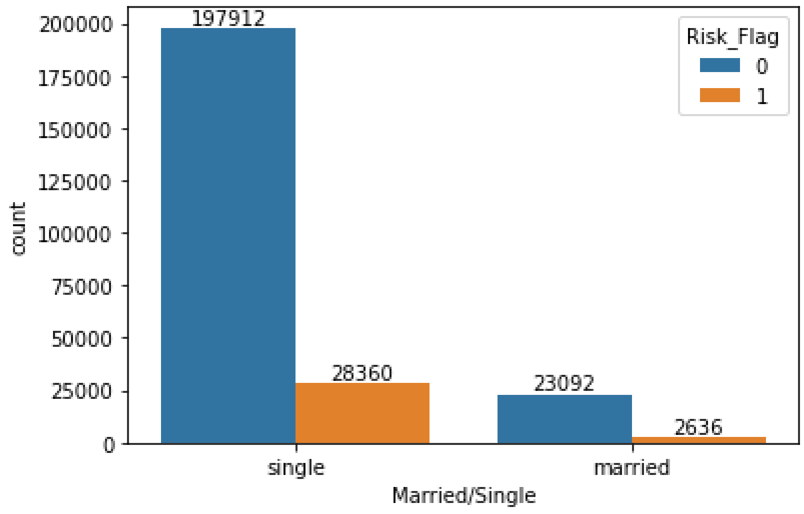
Visualization of the data set ➖

On Married/Single columns➖

```
Count of data present in Married/single columns wrt to their Unique value:-
single     226272
married     25728
```

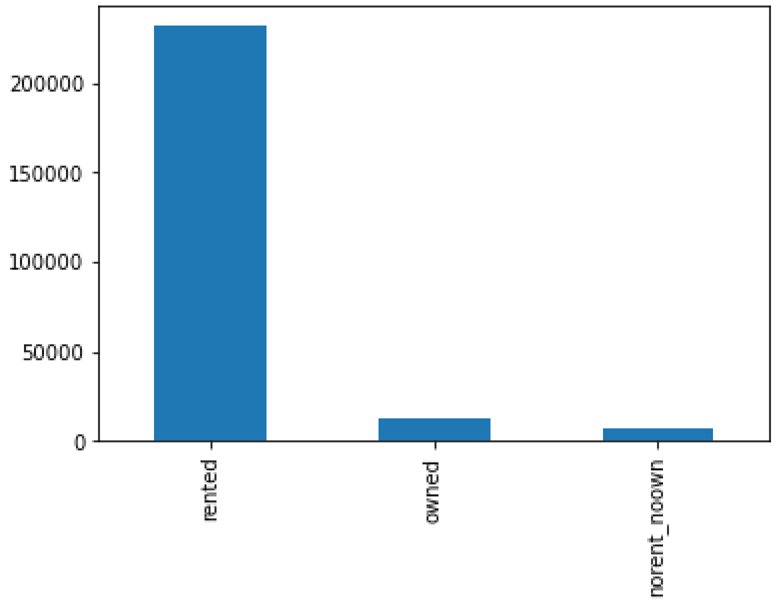


Distribution of the target variable according to Married/Single

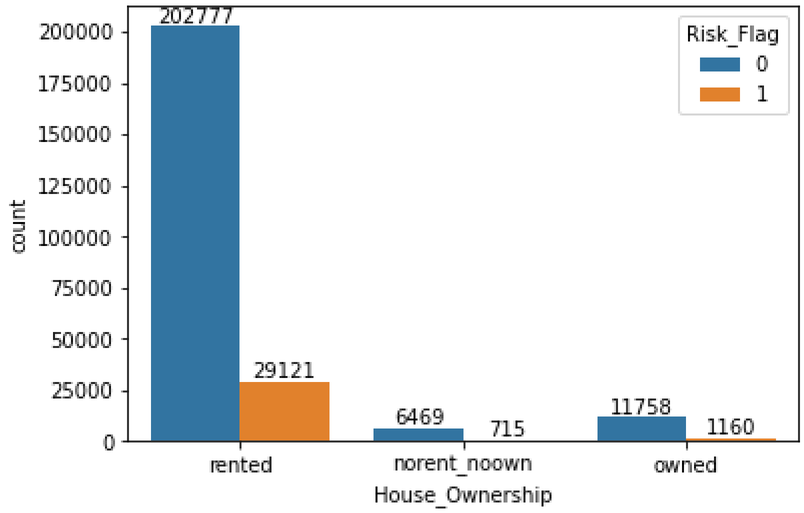


On House_ownership columns➖

```
Count of data present in House_Ownership columns wrt to their Unique value:-
rented          231898
owned            12918
norent_noown      7184
Name: House_Ownership, dtype: int64
```

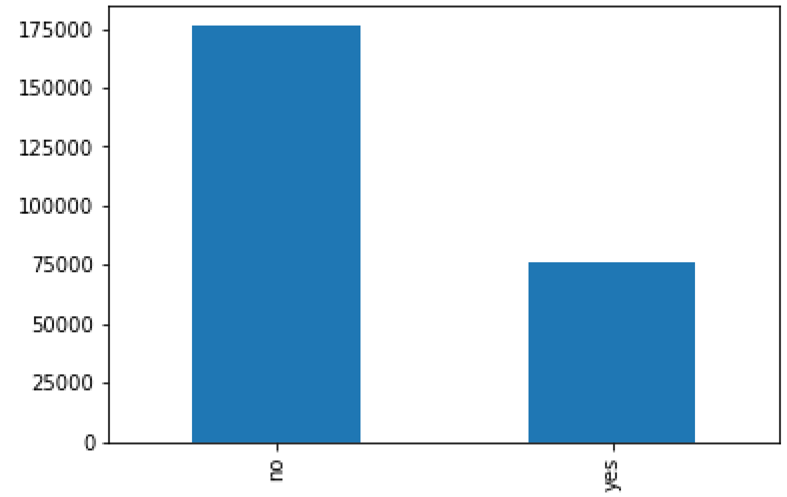


Distribution of the target variable according to House_Ownership

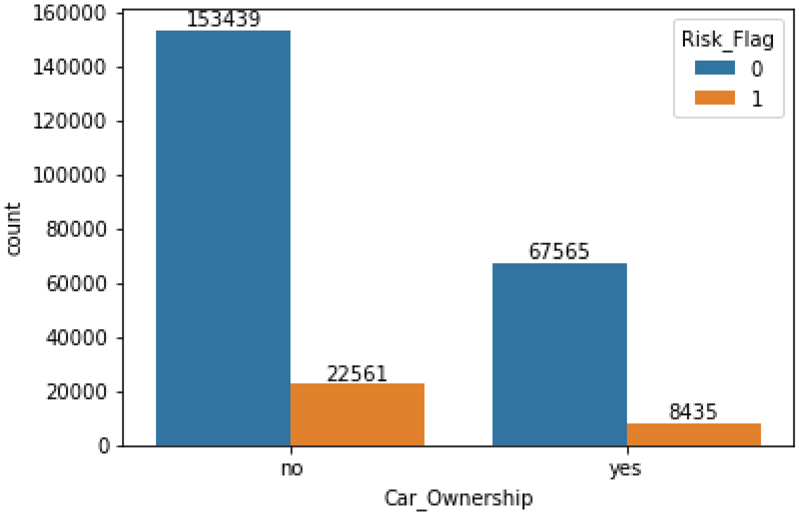


On House_ownership columns➖

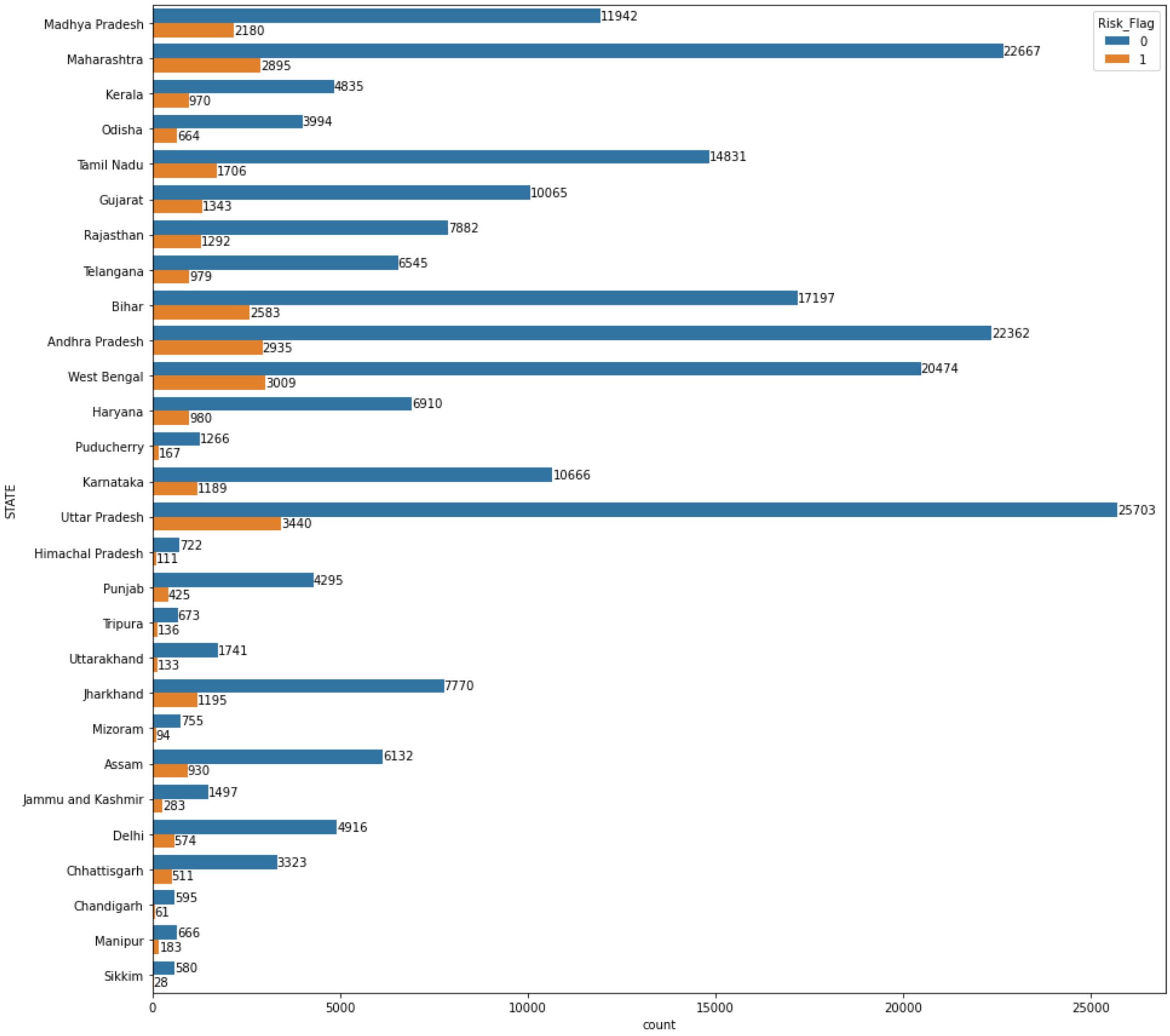
```
Count of data present in House_Ownership columns wrt to their Unique value:-
no      176000
yes      76000
Name: Car_Ownership, dtype: int64
```



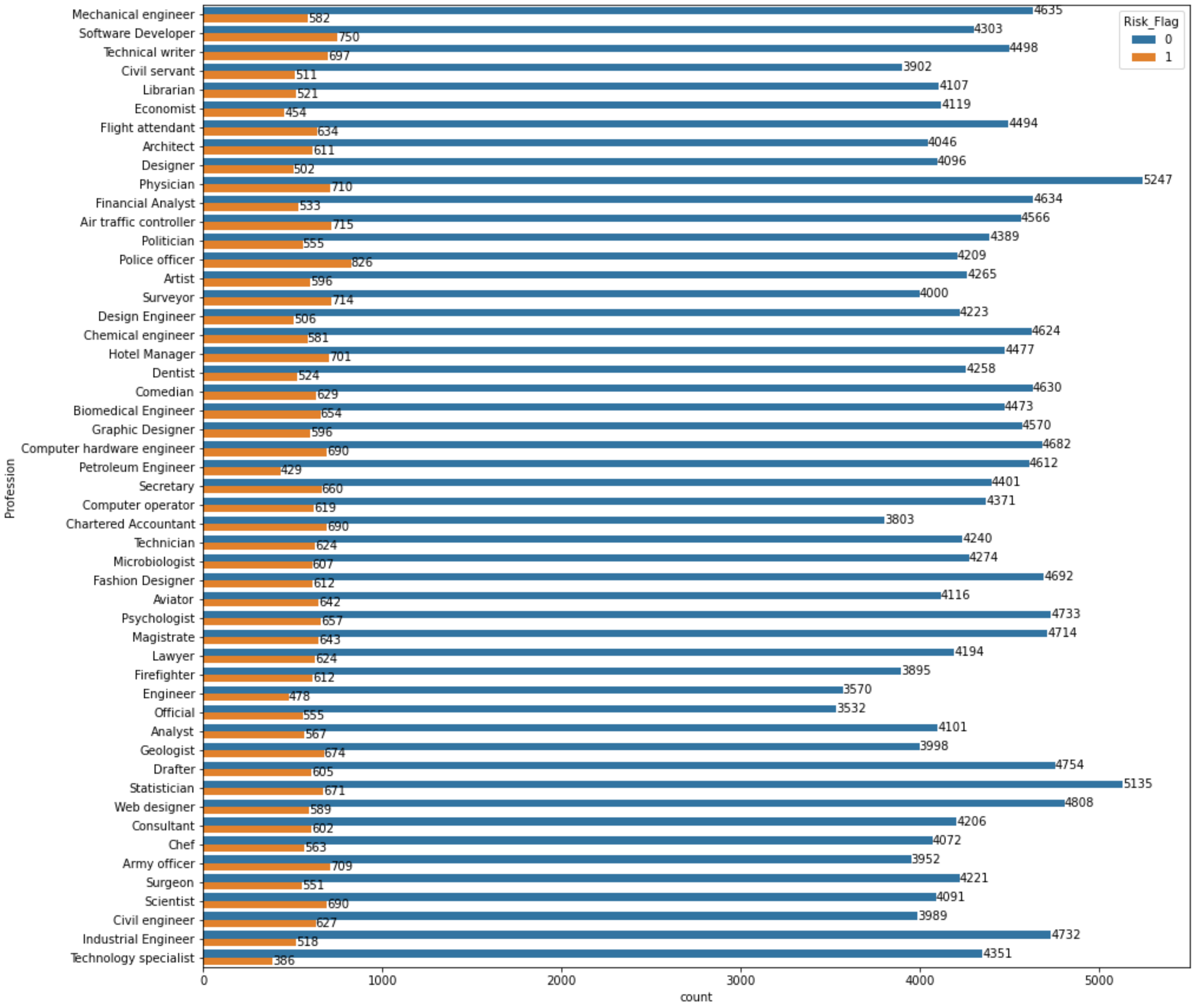
Distribution of the target variable according to Car_Ownership



Distribution of the target variable according to the State column

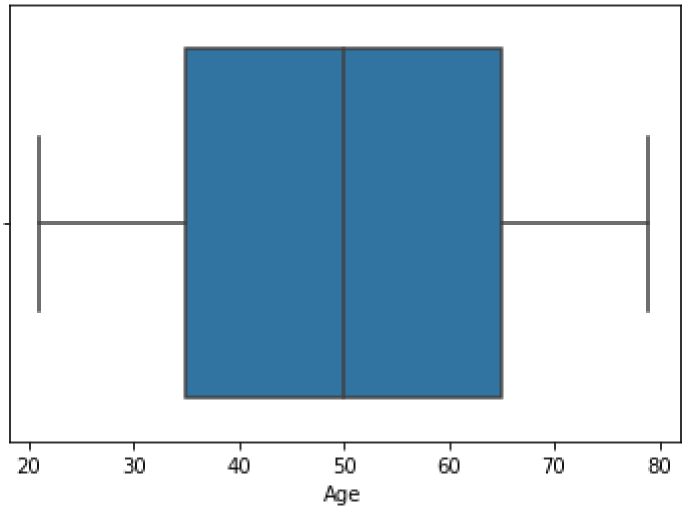


Distribution of the target variable according to the Profession

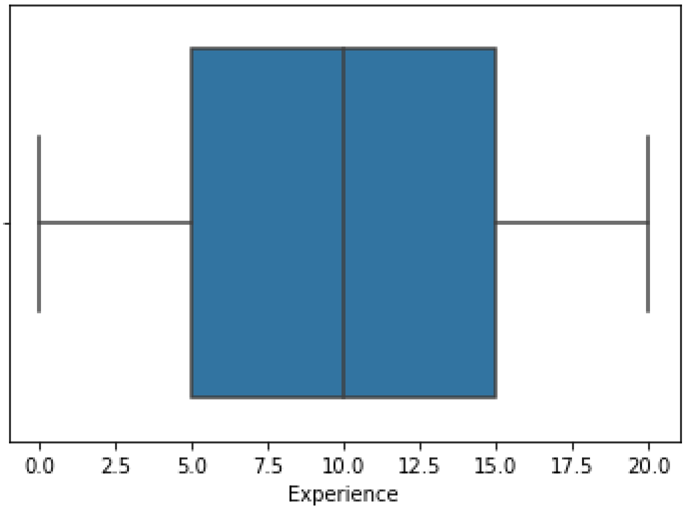


Data visualization on Numerical columns

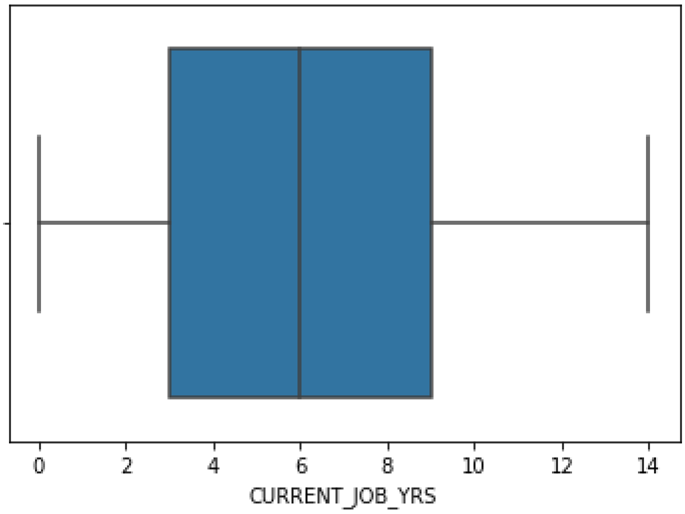
For Age columns



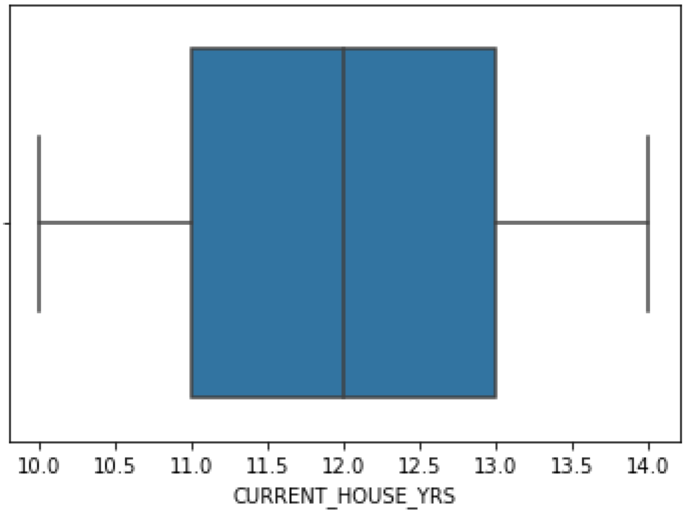
For Experience columns ➖



For CURRENT_JOB_YRS columns ➖



For CURRENT_HOUSE_YRS columns ➖



Conclusion of the Visualization➖

- The single person has more chance to not repay the loan
- The value in the current house year columns lies between 11 to 13
- The value in the current job year lies between 2.4 to 8.7
- The value in the experience lies between 5 to 15
- The software developer has more chances to repay the loan on time
- The physician has more chance to repay the loan
- Uttar Pradesh has more number of customer who no repays their loan on time

Point to be Noted➖

- Before applying the ML algorithm we have to convert the object column to numerical columns
- This is an imbalanced dataset
- The important point is it is a Classification problem so we have to apply classifier model for better accuracy
-

ML algorithm without Sampling➖

- In that process, we remove the State, City, and Profession columns

The Accuracy Score of Logistic regression is :

`87.72668650793651`%

The Accuracy Score of KNeighborsClassifier is :

`86.1984126984127`%

The Accuracy Score of the Random forest Classifier is :

`89.57688492063492`%

The Accuracy Score of the Gradient Booster Classifier is :

`87.73164682539683`%

On applying different ML models we found out that the Random Forest Classifier fits our model very well so we check the prediction on any data by applying Random Forest Model.

By using Under sampling method ➖

In this prediction, we dropped the State Columns for a better prediction method.

The Accuracy Score of the Logistic Regression is :

`50.00403258327284`%

The Accuracy Score of the Decision Tree Classifier is :

`86.73280103234132`%

The Accuracy Score of the Random Forest Classifier is :

`85.41011371884829`%

The Accuracy Score of the Gradient Boosting Classifier is :

`61.03718041777563`%

On applying Under Sampling we found out that the Decision Tree Classifier fits our model very well so we check the prediction on any data by applying the Decision Tree Classifier.

By using the Over_Sampling method➖

In this method, we didn't drop any columns for prediction.

The Accuracy Score of the Logistic Regression is :

`50.1380059274677`%

The Accuracy Score of the Decision Tree Classifier is :

`90.63595846247823`%

The Accuracy Score of the Random Forest Classifier is :

`92.58048460442071`%

On applying Over Sampling we found out that the Random forest Classifier fits our model very well so we check the prediction on any data by applying the Random forest Classifier.