INTRODUCTION

BUSINESS PROBLEM

This dataset is related to micro finance services. A Microfinance Institution (MFI) is an organization that offers financial services to low-income populations.

MFI becomes very useful when targeting especially poor families living in remote areas with not much sources of income. The Microfinance services provided by MFI are group loans, agriculture, Individual business loans. Today, microfinance is widely accepted as a poverty-reduction tool in a remote-areas

But an important step for the Microfinance Institution is that while giving a Loan to the customer in remote area, a lot of attention has to be paid that the Loan is not defaulter

EXERCISE

Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been paid i.e. Non- defaulter, while, Label '0' indicates that the loan has not been paid i.e. defaulter

CONCEPTUAL BACKGROUND OF THE BACKGROUND

The role of Microfinance institution is very important in rural areas or for those people who have low income , because such poor families do not even get Loan from the Bank in this case the Microfinance Institution is very helpful in this process and loan is easily available from these people

The number of Microfinance Institution has increased in few years, therefore now easily Loan available from MFI

Eg. As the number of MFI has increased in India in a few years and due to this , people are getting (Group loan, individuals business loans) easily in rural areas nowadays

ANALYTICAL PROBLEM

DATA SOURCES & THEIR FORMAT

Format of this Dataset is in excel

The Dataset has 37 Variables and 209593 rows

This dataset has no any Null value and most of Variables are (int, float) types , only three variables are object datatype. Dataset has also Outliers & Skewness problem

ANALYTICAL MODELLING OF THE PROBLEM

. We used Describe function on the dataset which got Statistical description and from describe it is known that what is the

min, max, mean & Std of each Attributes and the distribution of minimum & maximum elements of the Attributes is also known

. From Describe function we also know that in which Attributes might have outliers according to this describe function , I saw in this dataset that the distribution of data was very much spread in 30 Attributes , as the distribution of 85 percent values was somewhat similar and 15 percent values of some Variables are 40 or 50 times higher than their another elements of Variables like (cnt_ma_rech90, medianmarechprebal90, medianmarechprebal30, last_rech_date_ma , last_rech_date_da, aon, maxamnt_loans30, fr_da_rech30, cnt_da_rech30,..)

. Analyzing correlation of the dataset , then I saw that

(aon, last_rech_date_ma, last_rech_date_da,fr_ma_rech30, medianmarechprebal30, cnt_da_rech30, fr_da_rech30, cnt_da_rech90, maxamnt_loans30, cnt_loans90) these all variables are such types which are zero percent correlation with all other variables.

. some variables are highly positive correlated with each other

. half of all variables are low positive correlated with Target-Attributes and no any Variables are highly positive correlated with Target

. Plotting the Pairplot after the correlation plot , it also gives the same information as the correlation plot , the only difference is that it shows the relationship between each of the Attributes in a visual way

DATA PREPROCESSING DONE

. First load the dataset on jupyter notebook and print it then check the shape of the dataset

. I started looking at randomly some rows and columns of dataset and tried to understand and after check the Null value in this dataset but in this Dataset there was no any Null value or anything to do like data cleaning , so were moved ahead

. I described the Dataset for Statistical analysis and after that there were some categorical Variables on which value_counts() used to how many types of values are in this variable, then I saw that 'pcircle' variable has only one type value so I will have to drop it.

. Important point of Data-Preprocessing = analyzing outliers , skewness, Label or One-hot encoding, Feature-Selection, Multicollinearity problem, Scaling-dataset and Balanced the dataset using with SMOTE

Now let's do EDA on the Dataset

. I visualize the data using Matplotlib & seaborn

. maximum variables in this dataset have float or integer type , so we we analyzing correlation of the dataset then plotting

scatterplot, lineplot, some countplot

. After all plotting we get information about the Dataset that more Variables are randomly distributed with each other and some

input-variable highly positive correlated with each other

. All input variables are randomly distributed with Target

. when we plotting distribution plot ,we can see that in distribution plot , above 90 percent elements were spread near about zero of maximum Variables

HARDWARE AND SOFTWARE REQUIREMENT TOOLS

. we have done this entire project through (sklearn, pandas numpy, matplotlib, Seaborn, scipy,)

. Load the dataset, and all data cleaning & analyzing part done by Pandas

. we have done Visualization the dataset through matplotlib, & seaborn

. All Machine Learning model have been created in this project through Sklearn , xgboost

MODEL DEVELOPMENT & EVALUATION

IDENTIFICATION OF PROBLEM-SOLVING APPROACHES

. After the analyzing the dataset , it was found that Logistic-Regression and SVM-Classifier will not able to perform well on such data, but still I used all these Algorithms to build the model.

. The values in this Dataset were distributed in such a way that a best accuracy-score could be obtained only from a Ensemble

method like RandomForest, GradientBoosting & xgboost .

TRAING & TESTING OF IDENTIFIED APPROACHES

. First we do Train_test_split on input-data & target-attribute then data split into train & test data , fit the model on train data then test the model accuracy on test data

. we used different types of algorithms like Logistic Regression, KNN Classifier, RandomForest Classifier, xgboost Classifier and GradientBoosting Classifier for making a model

. There is trained the model with these all algorithms for have best accuracy-score

. we got the highest accuracy_score & roc_auc_score on this Dataset from xgboost, RandomForest Model

. I used different types of metrics to find which one model is perform better on this dataset

. confusion_matrix, classification_report, accuracy_score, roc_auc_score we used all these metrics

. we used cross_validation with KFold on all five model , cross_validation_score of RandomForest model is highest compare to all models

. We done Hyperparameter tuning over RandomForest_Model because 'accuracy_score' & cross_validation_score of

RandomForest was highest

. From confusion_matrix , we get about performance of a model in a better way such as False-Positive, False-Negative ,

True-Negative, True-Positive and from classification_report we got precision, recall & f1_score of the model

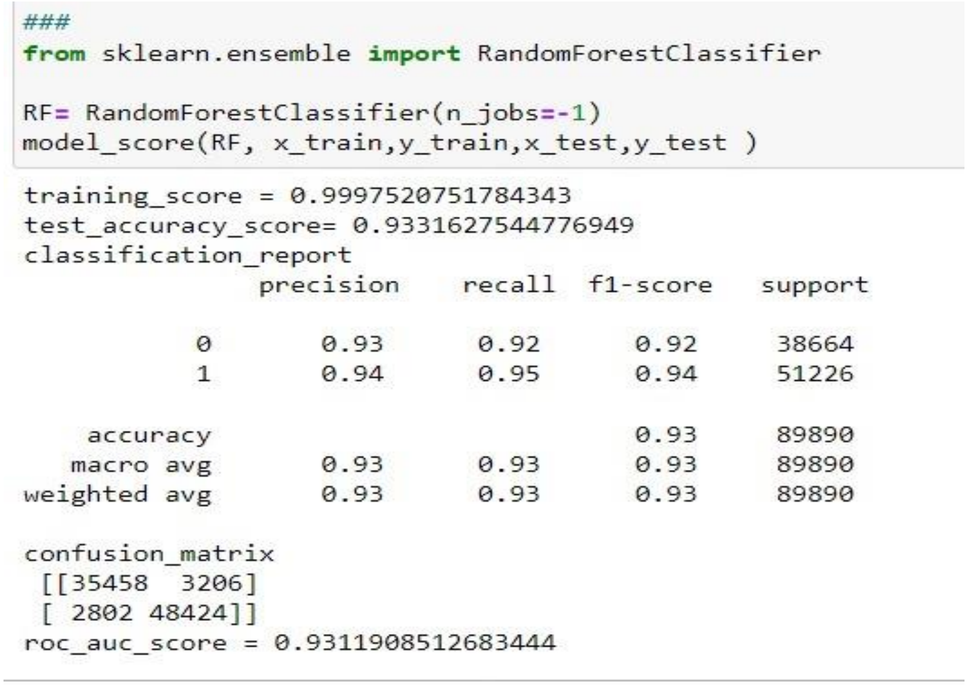
RUN & EVALUATE SELECTED MODELS

RANDOMFOREST MODEL

. Its highest accuracy_score & roc_auc_score on this dataset

compare to another model

. The values of the Variables of this dataset is distributed in such a way that RandomForest or xgboost performs well on such data because RandomForest, XGBoost are outliers robust type algorithm and in Classification problem if input-data is randomly distributed with target then these algorithm perform best compare to Logistic-Regression, SVM



```
###
from sklearn.ensemble import RandomForestClassifier

RF= RandomForestClassifier(n_jobs=-1)
model_score(RF, x_train,y_train,x_test,y_test )
```

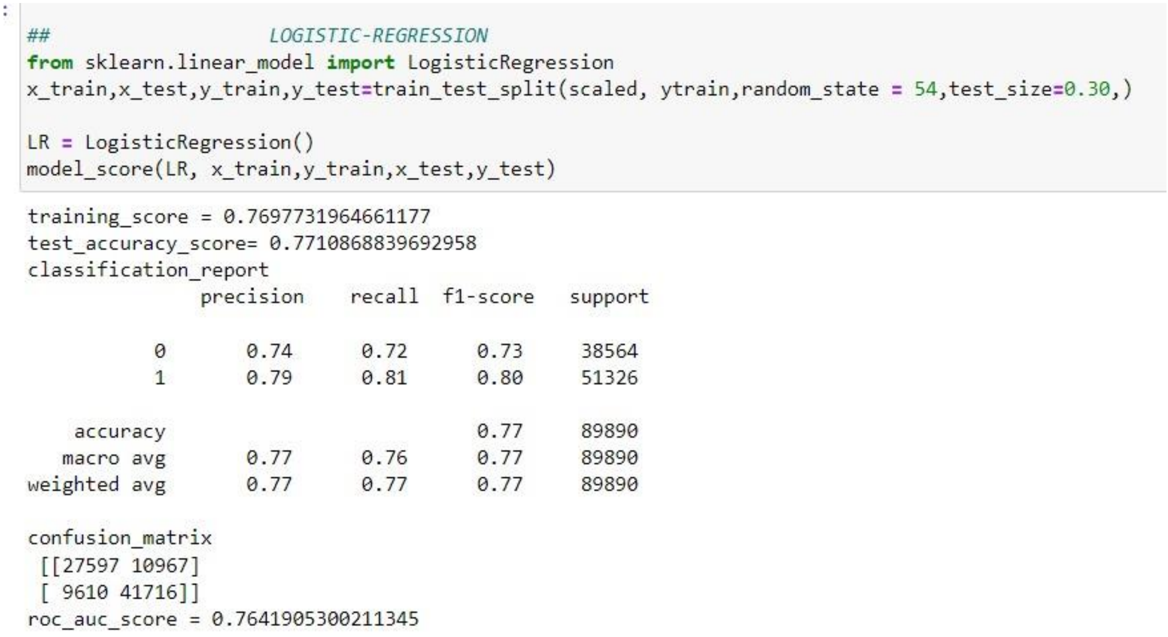
```
training_score = 0.9997520751784343
test_accuracy_score= 0.9331627544776949
classification_report
               precision    recall  f1-score   support

           0       0.93      0.92      0.92     38664
           1       0.94      0.95      0.94     51226

    accuracy                           0.93     89890
   macro avg       0.93      0.93      0.93     89890
weighted avg       0.93      0.93      0.93     89890

confusion_matrix
 [[35458  3206]
 [ 2802 48424]]
roc_auc_score = 0.9311908512683444
```

LOGISTIC REGRESSION MODEL

. It is a Logistic Regression Model which performance is not much more on this dataset.

. we can see that its accuracy_score , roc_auc_score , precision and recall not so good as should be

. I had already said about Logistic Regression that you will not be able to get more accuracy



```
##                  LOGISTIC-REGRESSION
from sklearn.linear_model import LogisticRegression
x_train,x_test,y_train,y_test=train_test_split(scaled, ytrain,random_state = 54,test_size=0.30,)

LR = LogisticRegression()
model_score(LR, x_train,y_train,x_test,y_test)
```

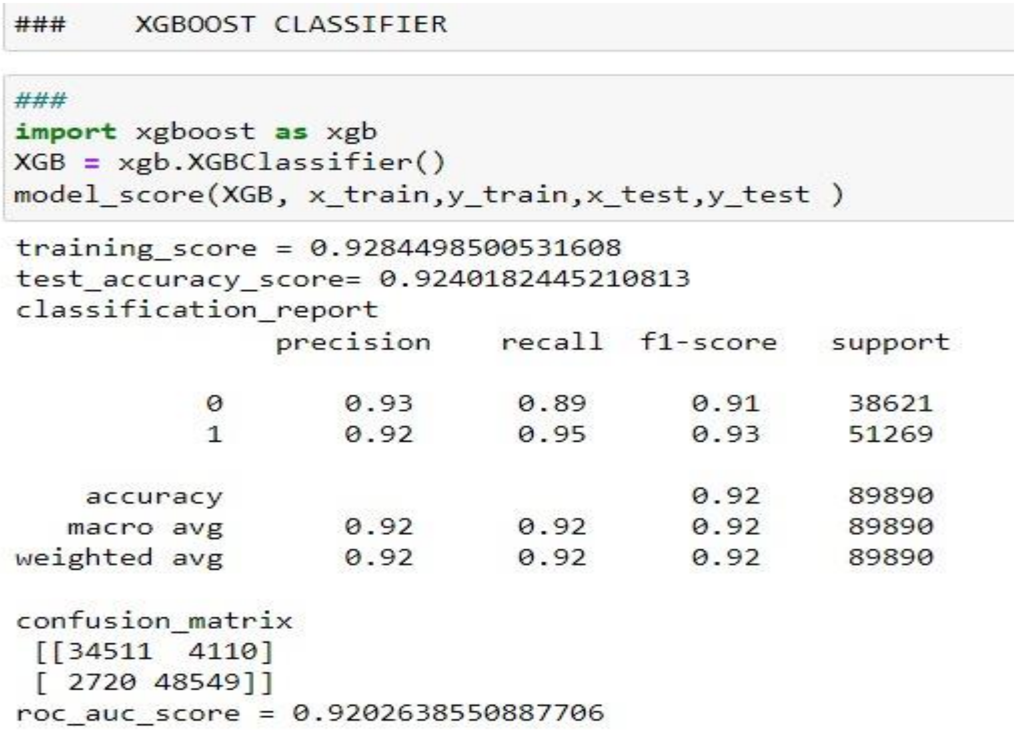
```
training_score = 0.7697731964661177
test_accuracy_score= 0.7710868839692958
classification_report
              precision    recall  f1-score   support

           0       0.74      0.72      0.73     38564
           1       0.79      0.81      0.80     51326

    accuracy                           0.77     89890
   macro avg       0.77      0.76      0.77     89890
weighted avg       0.77      0.77      0.77     89890

confusion_matrix
 [[27597 10967]
 [ 9610 41716]]
roc_auc_score = 0.7641905300211345
```

XGBOOST classifier Model

The performance of this Model is also very good on this Dataset

because its accuracy_score, roc_auc_score, precision and recall is more

compare to another model



```
###    XGBOOST CLASSIFIER
```

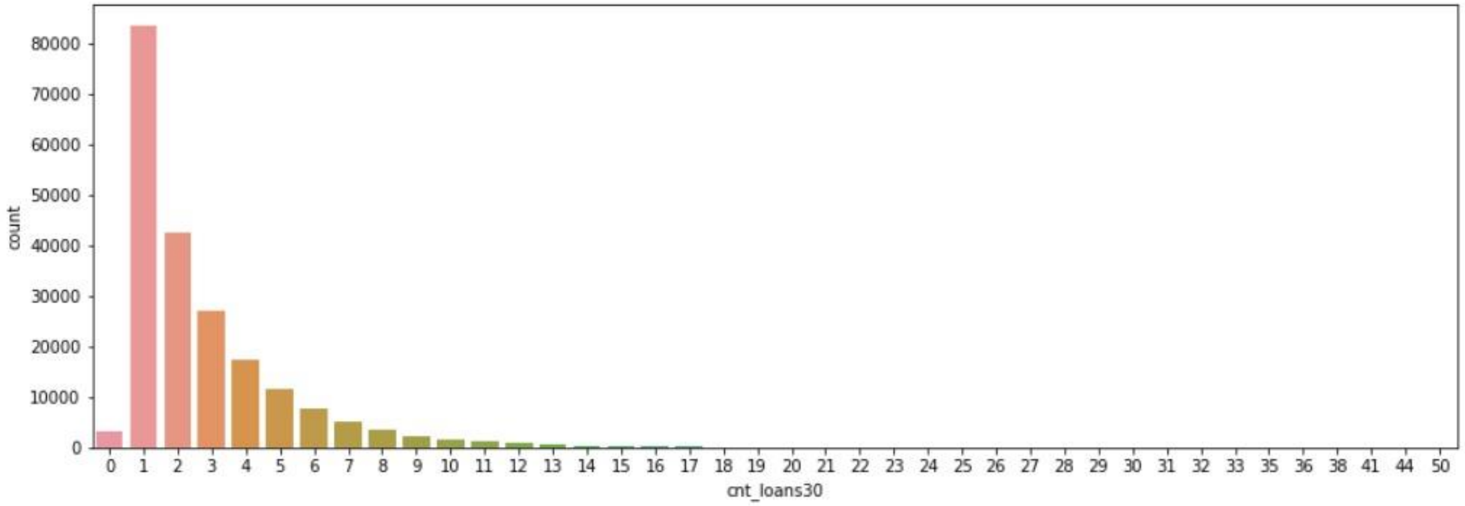
```
###
import xgboost as xgb
XGB = xgb.XGBClassifier()
model_score(XGB, x_train,y_train,x_test,y_test )
```

```
training_score = 0.9284498500531608
test_accuracy_score= 0.9240182445210813
classification_report
               precision    recall  f1-score   support

           0       0.93      0.89      0.91     38621
           1       0.92      0.95      0.93     51269

    accuracy                           0.92     89890
   macro avg       0.92      0.92      0.92     89890
weighted avg       0.92      0.92      0.92     89890

confusion_matrix
 [[34511  4110]
 [ 2720 48549]]
roc_auc_score = 0.9202638550887706
```

VISUALIZATIONS

. In this diagram we can that more than 90 percent value is spread from 0 to 17 and less than 10 percent value is distributed from 18 to 50

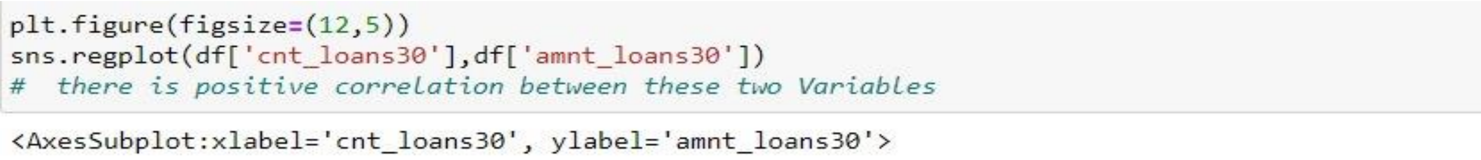
Values that Greater than 30 of ‘cnt_loan30’ variable are equal to 1 percent out of total 209593 rows



....

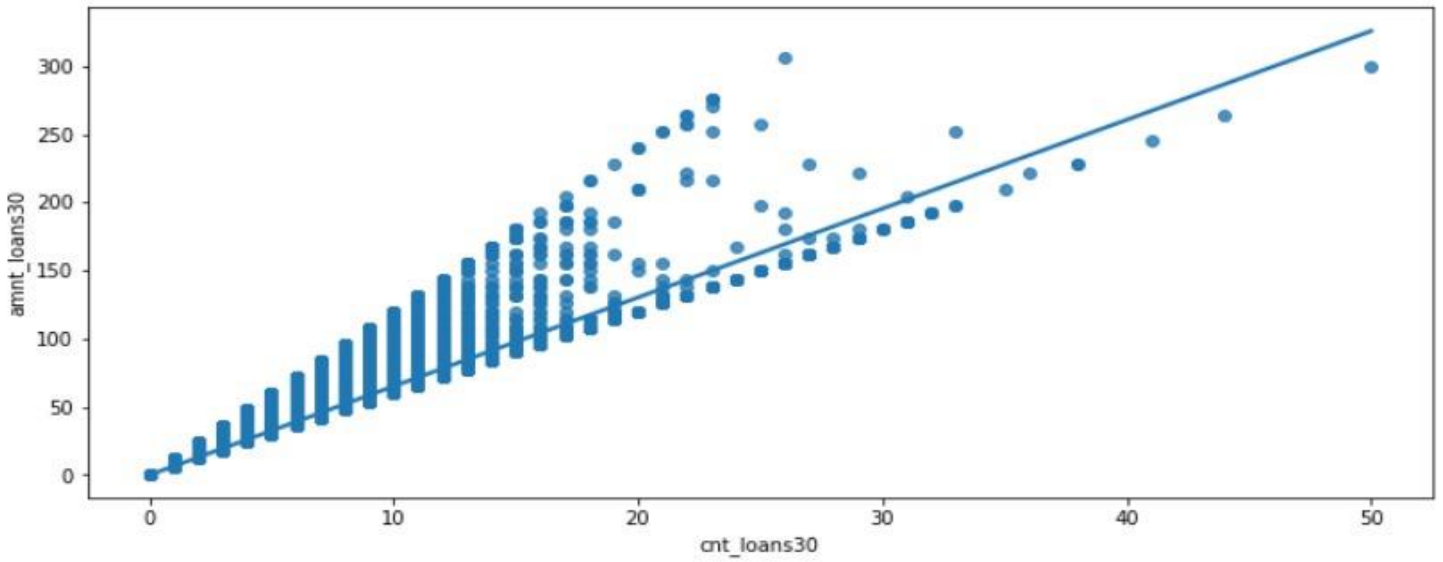
. In this dataset it is example of positive related variable

cnt_loans30 vs amnt_loans30 and some variables are like this



```
plt.figure(figsize=(12,5))
sns.regplot(df['cnt_loans30'],df['amnt_loans30'])
#  there is positive correlation between these two Variables
```

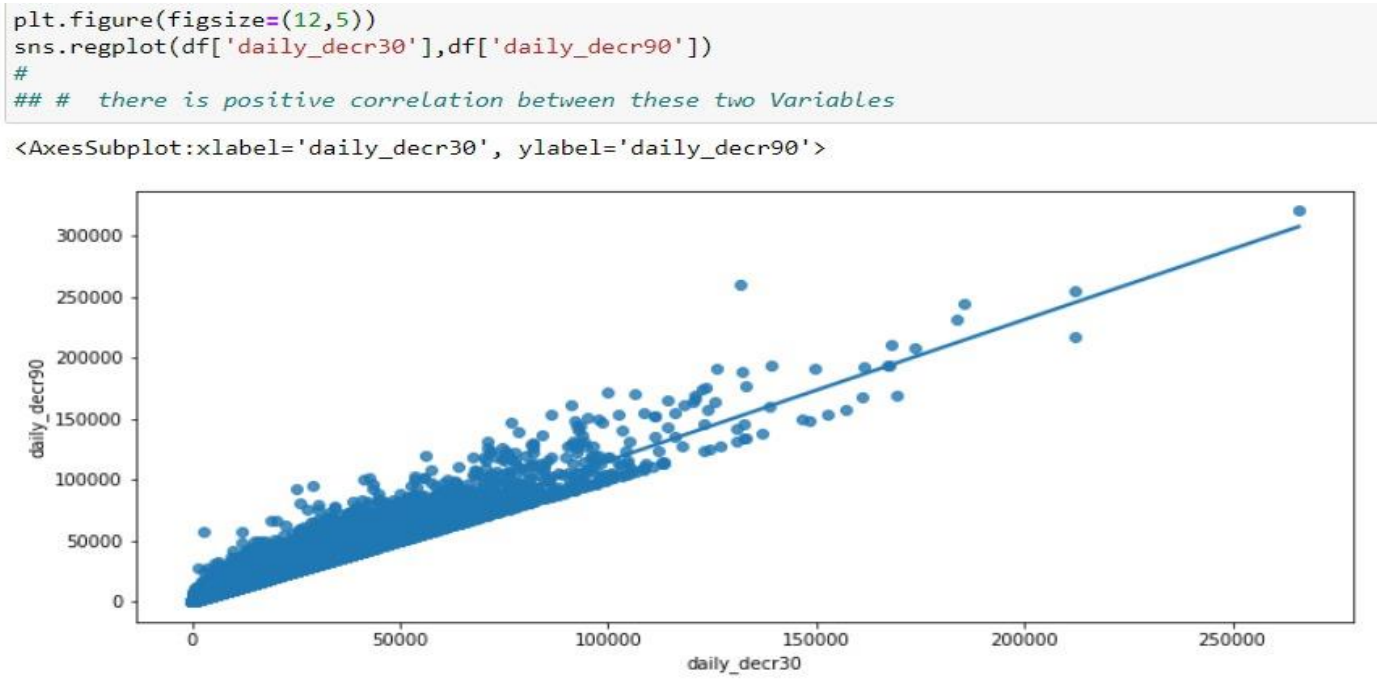
```
<AxesSubplot:xlabel='cnt_loans30', ylabel='amnt_loans30'>
```



....

. It is regression plot between ‘daily_decr30’ and ‘daily_decr90’

Some variables are like this which are highly positive correlated with each other

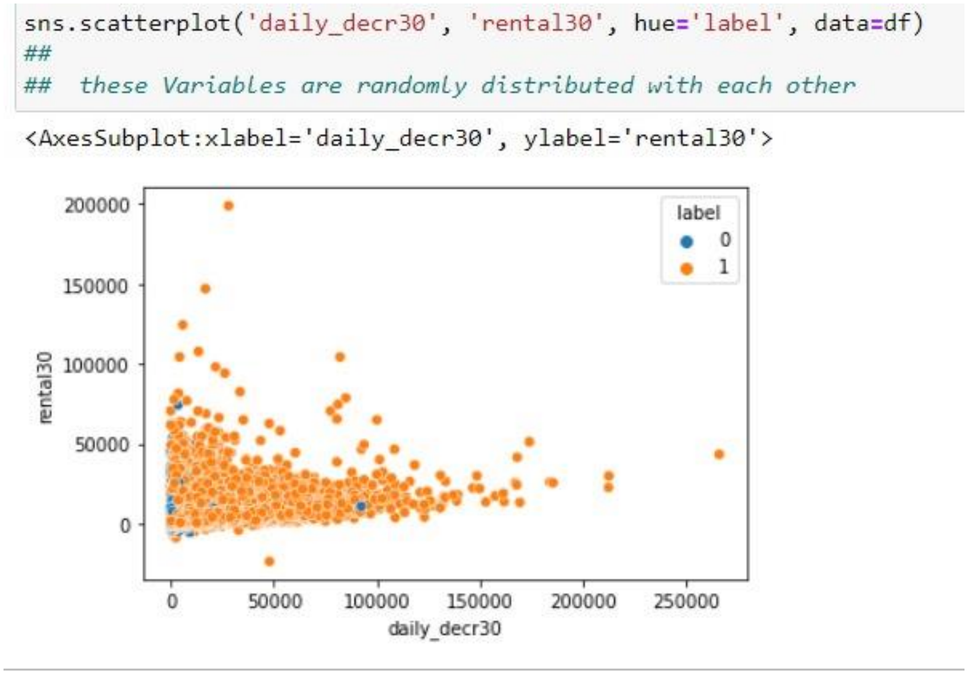


```
plt.figure(figsize=(12,5))
sns.regplot(df['daily_decr30'],df['daily_decr90'])
#
## #  there is positive correlation between these two Variables
```

```
<AxesSubplot:xlabel='daily_decr30', ylabel='daily_decr90'>
```

....

There is no any relationship between daily_decr30 & rental30 or randomly distributed , maximum variables are randomly distributed with each other



```
sns.scatterplot('daily_decr30', 'rental30', hue='label', data=df)
##
##  these Variables are randomly distributed with each other
```

```
<AxesSubplot:xlabel='daily_decr30', ylabel='rental30'>
```

We visualized the different types of diagrams that give information about different types of relation like (highly positive related, randomly distributed) between Variables

. The Important point is that there is no any variable’s distribution which is almost the same as the Normal distribution

CONCLUSION

KEY FINDINGS & CONCLUSION OF THE STUDY

. When I analyzing the dataset then saw that most important point of its is distribution of data because the data was distributed in such a way that about 20 percent values were coming under Outliers , Skewness and more than 8 percent values were not be removed

. I did reduce some skewness and outliers but it didn't go away completely

. The distribution of the dataset matters a lot on the accuracy of the ML model because

. False positive of KNN Classifier is minimum compare to all models means when KNN model predict to customer paid EMI within 5 days and model predicts True then True-Positive , if model predicts wrong then False-Positive

. RandomForest on this dataset performed best because on this data only ensemble method work best or xgboost,

In this data has some skewed or outliers , still RandomForest & XGBoost give good accuracy

. From data Visualization , we also get information about the distribution of dataset and the relationship between Variables

. when we done all Visualization part then we suppose that on this dataset distance-based calculation algorithm will not perform more accurate like Logistic-Regression, SVM classifier and this dataset has maximum numbers of outliers