FLIP ROBO

# Micro-Credit Defaulter Model


Submitted by:

Kumar Gourabh

kumargourabh94@gmail.com

Contact: 8130803199

# ACKNOWLEDGMENT

Data was provided to me by the Flip Robo Team as part of a practice Assignment.

I'd like to thank my mentor – Nitin Mishra for giving me such a nice use case to work upon. This has increased my skills as well as confidence level as a Data Science enthusiast.

Below mentioned are some of the websites I took help from when stuck or when I came across any error:

- geeksforgeeks.org
- medium.com
- stackoverflow.com

# INTRODUCTION

- ## Business Problem:

  The goal is to predict whether a customer will be paying back the loaned amount within 5 days of insurance of loan be a defaulter.

- ## Conceptual Background of the Domain Problem

- The data belongs to a client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

- They understand the importance of communication and how it affects a person's life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah). The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

- ## Review of Literature

  A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on. Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes. Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

- ## Motivation for the Problem Undertaken

  As discussed, In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers

# Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

After trying various machine learning algorithms, the XGBoost Classifier Technique was used to build the model for predicting credit defaulters. As, the label was imbalanced i.e., Label '1' (Non-Defaulters) has approximately 87.5% records, while label '0' (Defaulters) has approximately 12.5% records, I've used the SMOTE function while separating the dataset into Target and Test Data

- ## Data Sources and their formats

- The data belongs to a client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

- They understand the importance of communication and how it affects a person's life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah). The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

- Here's a snap of the actual data: df.head()

| label | msisdn | aon | daily_deci | daily_deci | rental30 | rental90 | last_rech_ | last_rech_ | last_rech_ | cnt_ma_re | fr_ma_rec | sumamnt_ | mediananı | medianma |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 21408I707 | 272 | 3055.05 | 3065.15 | 220.13 | 260.13 | 2 | 0 | 1539 | 2 | 21 | 3078 | 1539 | 7.5 |
| 1 | 76462I703 | 712 | 12122 | 12124.75 | 3691.26 | 3691.26 | 20 | 0 | 5787 | 1 | 0 | 5787 | 5787 | 61.04 |
| 1 | 17943I703 | 535 | 1398 | 1398 | 900.13 | 900.13 | 3 | 0 | 1539 | 1 | 0 | 1539 | 1539 | 66.32 |
| 1 | 55773I707 | 241 | 21.228 | 21.228 | 159.42 | 159.42 | 41 | 0 | 947 | 0 | 0 | 0 | 0 | 0 |
| 1 | 03813I827 | 947 | 150.6193 | 150.6193 | 1098.9 | 1098.9 | 4 | 0 | 2309 | 7 | 2 | 20029 | 2309 | 29 |

| cnt_ma_re | fr_ma_rec | sumamnt_ | mediananı | medianma | cnt_da_re | fr_da_rech | cnt_da_re | fr_da_rech | cnt_loans | amnt_loar | maxamnt_ | mediananı | cnt_loans | amnt_loar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 21 | 3078 | 1539 | 7.5 | 0 | 0 | 0 | 0 | 2 | 12 | 6 | 0 | 2 | 12 |
| 1 | 0 | 5787 | 5787 | 61.04 | 0 | 0 | 0 | 0 | 1 | 12 | 12 | 0 | 1 | 12 |
| 1 | 0 | 1539 | 1539 | 66.32 | 0 | 0 | 0 | 0 | 1 | 6 | 6 | 0 | 1 | 6 |
| 1 | 0 | 947 | 947 | 2.5 | 0 | 0 | 0 | 0 | 2 | 12 | 6 | 0 | 2 | 12 |
| 8 | 2 | 23496 | 2888 | 35 | 0 | 0 | 0 | 0 | 7 | 42 | 6 | 0 | 7 | 42 |

| maxamnt_ | mediananı | payback3( | payback9( | pcircle | pdate |
|---|---|---|---|---|---|
| 6 | 0 | 29 | 29 | UPW | 7/20/2016 |
| 12 | 0 | 0 | 0 | UPW | 8/10/2016 |
| 6 | 0 | 0 | 0 | UPW | 8/19/2016 |
| 6 | 0 | 0 | 0 | UPW | 6/6/2016 |
| 6 | 0 | 2.333333 | 2.333333 | UPW | 6/22/2016 |

Fig: The above figures are snapshots of the first five rows of the given dataset

## • Data Pre-processing

- Dropped irrelevant columns – ‘Unnamed’: ’Contains S.No.’; ‘pcircle’: ‘Only 1 value throughout’; ‘pdate’: ‘Not Clear’; ‘msisdn’: ‘Simply the mobile number of the individual’

- Deleted rows with ‘maxamnt_loans30’ having values other than 0,6 and 12.

- Dropped daily_decr90 as it has very high correlation with daily_decr30
- Dropped rental90 as it has very high correlation with rental30
- Dropped amnt_loans30 as it has very high correlation with cnt_loans30
- Dropped maxamnt_loans90 as it has very high correlation with maxamnt_loans30
- Dropped medianamnt_loans90 as it has very high correlation with medianamnt_loans30

- Treated for outliers upto z-score of six by deleting the rows. Also applied cube root on continuous features to make the data normalized.

- Performed Standard Scaling of the data.

- Did PCA to reduce curse of dimensionality, with n_components =15, after looking at the explained variance

## • Data Inputs- Logic- Output Relationships

The input data is the data that I preprocessed in the above-mentioned step. The ‘label’ column was separated and stored in y, while the feature columns were saved as x.

Used Classification algorithms and selected the one with the best roc-auc scores. Futher did hyper-parameter tuning on the selected algorithm to further improve the obtained score.

## • State the set of assumptions (if any) related to the problem under consideration

The phone number and date features are not directly linked to defaulting.

## • Hardware and Software Requirements and Tools Used

Hardware: Simple System with basic configurations
Software & Tools: Anacondas, Jupyter Notebook
Libraries: Numpy, Pandas, Seaborn, matplotlib, imblearn, sklearn, joblib etc.

# Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

This is a classification problem. I will be checking various Algorithms and compare the roc-auc scores. I'll further be tuning the best performing model using randomized search CV and iterating over 42-100 r_state values to find the best score.

- ## Testing of Identified Approaches (Algorithms)

Algorithms Tested:

1. Random Forest Classifier
2. KNN Classifier
3. DecisionTree Classifier
4. Logistic Regression Classifier
5. Gaussian NB Classifier
6. Gradient Boost Classifier
7. Adaboost Classifier
8. XGBoostClassifier

- ## Run and Evaluate selected models

Below is the snapshot of cross validations for above models with scoring= 'roc_ auc' and cv = 5

```
Random Forest Classifier


Mean ROC_AUC score for classifier:  0.8691888881541587
standard deviation in ROC_AUC score for classifier:  0.0019453576794263567
[0.86812682 0.86639525 0.87206196 0.87047151 0.8688889 ]


KNN Classifier


Mean ROC_AUC score for classifier:  0.8009490710658633
standard deviation in ROC_AUC score for classifier:  0.0037918211885474883
[0.79823678 0.79618058 0.80374006 0.80668571 0.79990224]
```

```
DecisionTree Classifier


Mean ROC_AUC score for classifier:  0.6792487917762691
standard deviation in ROC_AUC score for classifier:  0.003029115795545029
[0.67756418 0.67457708 0.67923225 0.68310903 0.68176141]


Logistic Regression Classifier


Mean ROC_AUC score for classifier:  0.8331919164684749
standard deviation in ROC_AUC score for classifier:  0.0016801064814120116
[0.83496282 0.83010498 0.83324385 0.83325525 0.83439268]


Gaussian NB Classifier


Mean ROC_AUC score for classifier:  0.7925785133876686
standard deviation in ROC_AUC score for classifier:  0.0025209568249097217
[0.79443502 0.78817069 0.79136182 0.79406169 0.79486334]


Gradient Boost


Mean ROC_AUC score for classifier:  0.857854772731781
standard deviation in ROC_AUC score for classifier:  0.0016922297393971413
[0.85802712 0.85482854 0.85872353 0.85994399 0.85775068]


Adaboost Classifier


Mean ROC_AUC score for classifier:  0.8401743466666746
standard deviation in ROC_AUC score for classifier:  0.0017865000885765013
[0.83929956 0.83705221 0.84177275 0.84129106 0.84145614]


XGBoostClassifier


Mean ROC_AUC score for classifier:  0.8745928672997124
standard deviation in ROC_AUC score for classifier:  0.0018031761625965895
[0.87397969 0.87216134 0.87645135 0.87689059 0.87348137]
```

- ## Key Metrics for success in solving problem under consideration

  As mentioned before, I've used auc_roc score to determine the best performing model. As shown below, after hyper parameter tuning, the score obtained was = 79.28

```
from sklearn.model_selection import RandomizedSearchCV
xg=XGBClassifier()
parameters = {"learning_rate"    : [0.05, 0.10, 0.15, 0.20, 0.25, 0.30 ] ,
 "max_depth"        : [ 3, 4, 5, 6, 8, 10, 12, 15],
 "min_child_weight" : [ 1, 3, 5, 7 ],
 "gamma"            : [ 0.0, 0.1, 0.2 , 0.3, 0.4 ],
 "colsample_bytree" : [ 0.3, 0.4, 0.5 , 0.7 ] }
clf = RandomizedSearchCV(xg, parameters, cv=5,scoring="roc_auc")
clf.fit(x,y)
clf.best_params_
```

```
{'min_child_weight': 7,
 'max_depth': 12,
 'learning_rate': 0.05,
 'gamma': 0.4,
 'colsample_bytree': 0.5}
```

```
Confusion matrix
 [[ 3578  1300]
 [ 5102 28822]]
classification report
               precision    recall  f1-score   support

           0       0.41      0.73      0.53      4878
           1       0.96      0.85      0.90     33924

    accuracy                           0.84     38802
   macro avg       0.68      0.79      0.71     38802
weighted avg       0.89      0.84      0.85     38802

AUC_Score: 0.7915511671918983
[0 1 0 ... 1 1 1]
```

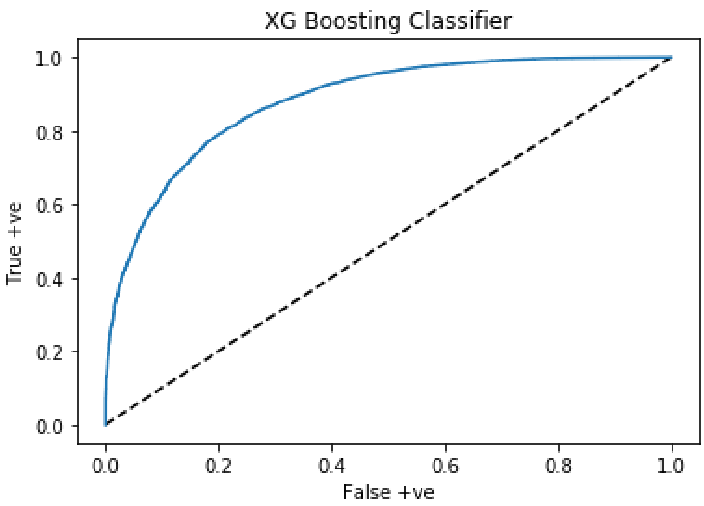


Fig: AUC Curve

- # Visualizations

Below attached are some of the visualizations done to understand some of the features and correlations.

```
Proportion of Defaulted to Not Defaulted = 14.26258375083819 %
1    0.875177
0    0.124823
Name: label, dtype: float64
```

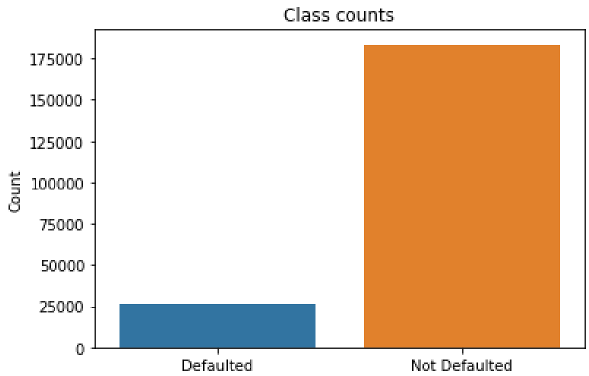


Fig: Imbalance in the Label

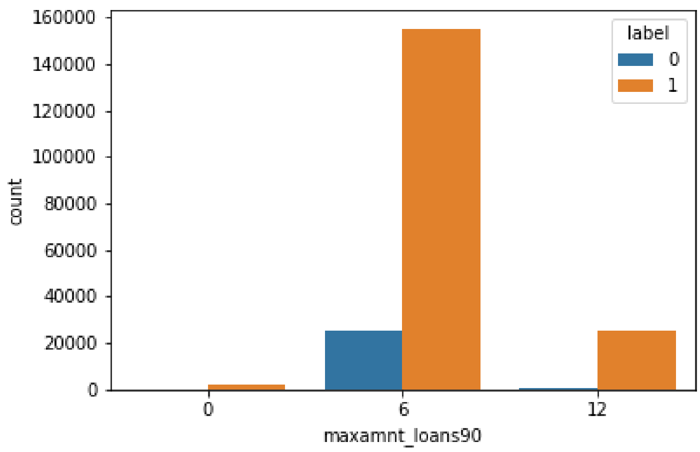


Fig: Max amount(90 days) v count, hued with label

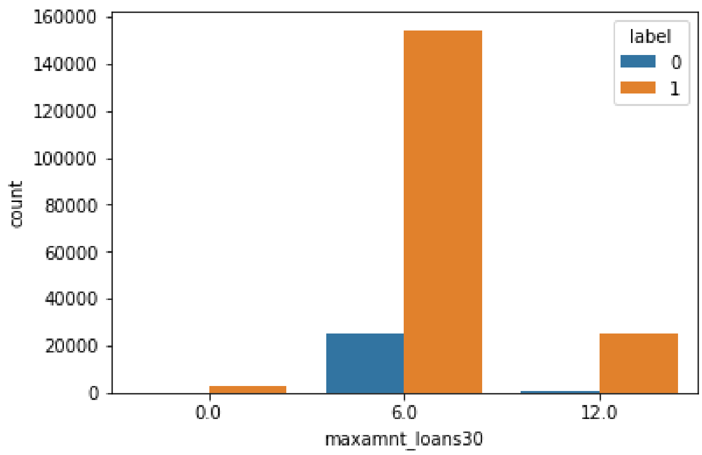


Fig: Max amount(30 days) v count, hued with label

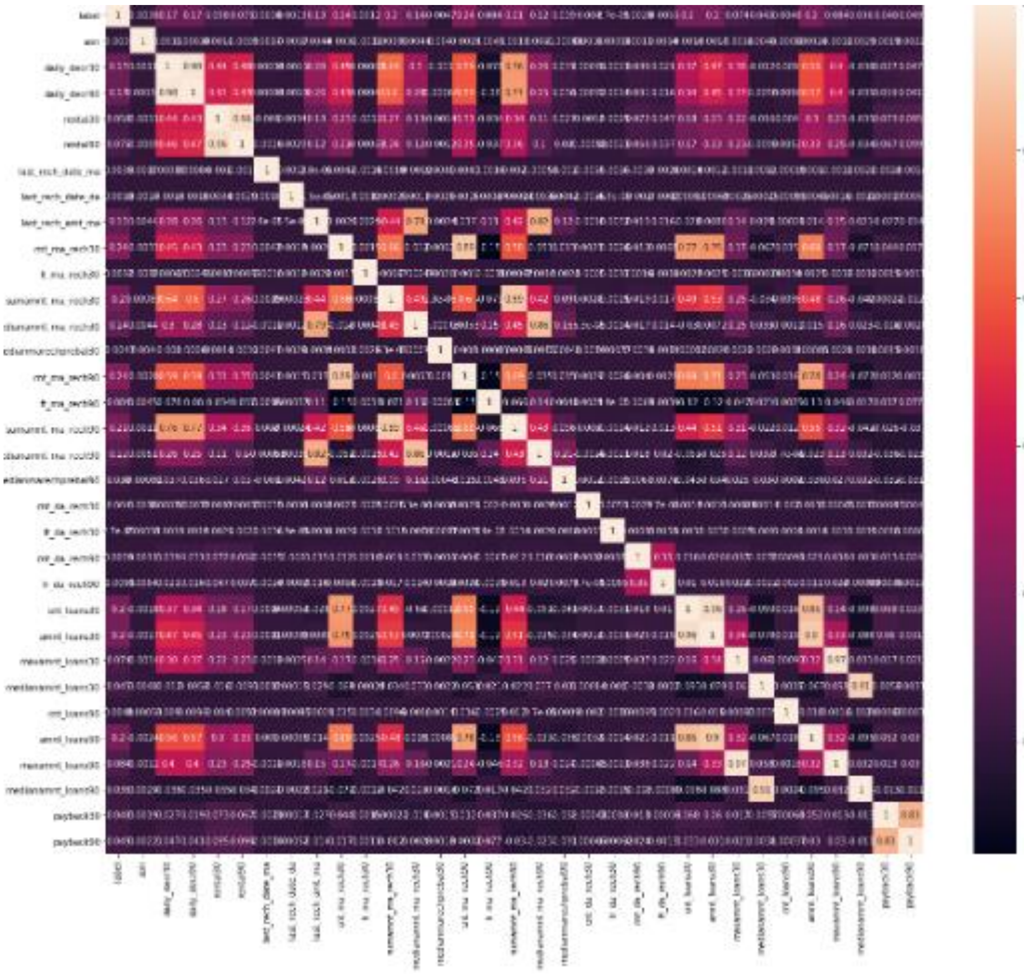

Fig: Correlation Heatmap

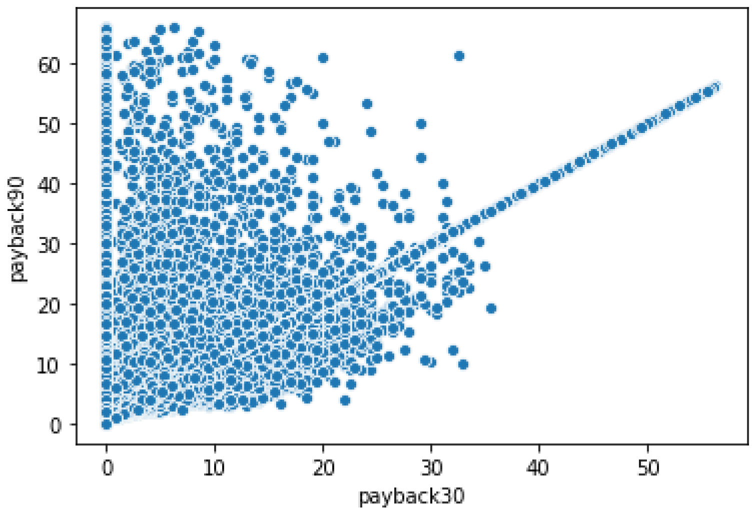


Fig: Payback 30 v 90

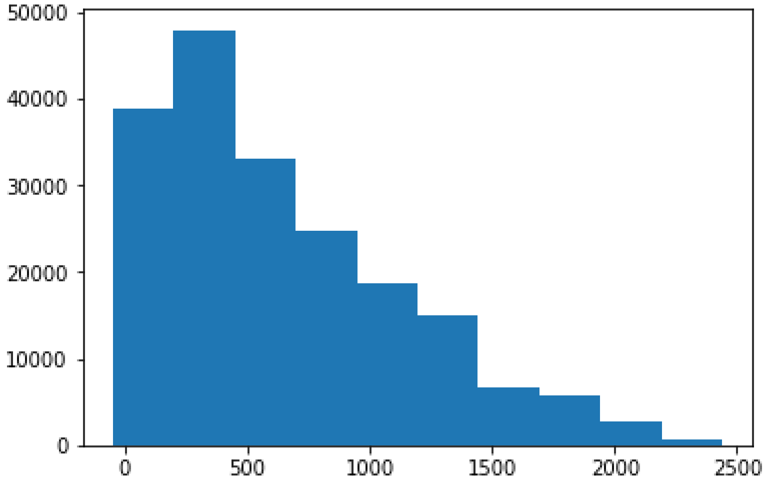


Fig: Aon histplot

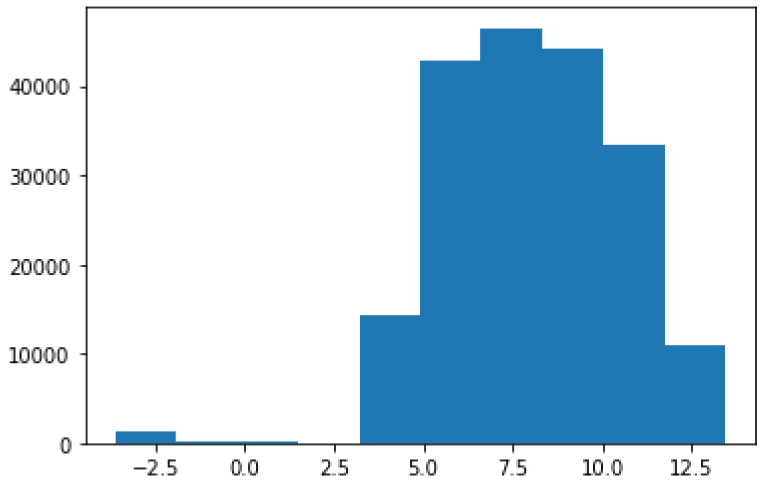


Fig: Aon histplot after cube root

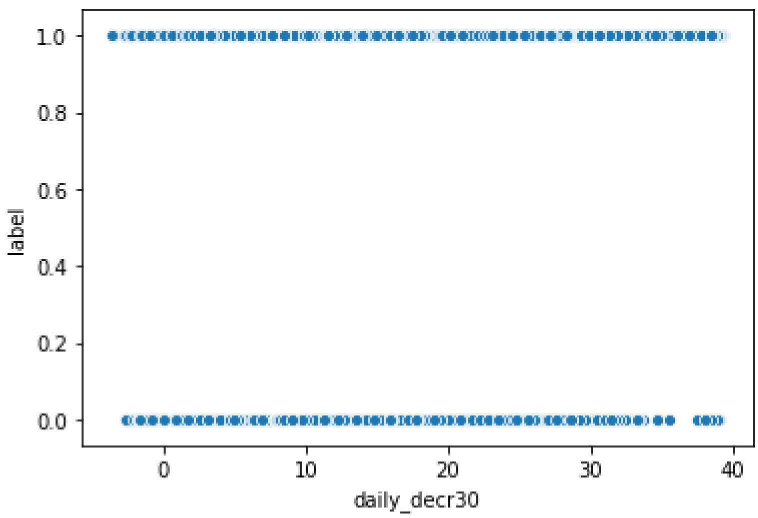


Fig: Scatterplot of daily_decr30 vs label

- ## Interpretation of the Results
- There was a lot of imbalance in the dataset
- There seem to be negative values for columns like aon, i.e. age on network and others as well, which seems unlikely.
- Also, some values are very large.
- Columns like maxamnt_loans30 etc should have values 0 or 6 or 12 bur we see other values as well.
- Max amt for both 30 and 90 days are more for 6, i.e 5
- We can see that there is strong positive as well as negative correlation within features, and I've used PCA also to eliminate the curse of dimensionality.
- Between Payback 30 and 90, A lot of data points are correlated strongly, but there are also a few anomalies.
- Most of the data is not normalized, so I've done a cube root transform on the continuous features.

# CONCLUSION

- ## Key Findings and Conclusions of the Study

Following are the results of the prediction.

```
Confusion matrix
 [[ 3578  1300]
 [ 5102 28822]]
classification report
               precision    recall  f1-score   support

           0       0.41      0.73      0.53      4878
           1       0.96      0.85      0.90     33924

    accuracy                           0.84     38802
   macro avg       0.68      0.79      0.71     38802
weighted avg       0.89      0.84      0.85     38802

AUC_Score: 0.7915511671918983
[0 1 0 ... 1 1 1]
```

An auc_roc score of almost 80% is quite good to predict the results. Overall, the total cases of defaulters are 12-13% which is still a high ratio and there is a need to predict the chances of a new case being defaulter or not, so that action can be taken and the ratio can be brought down.

- ## Learning Outcomes of the Study in respect of Data Science

As the dataset was comparatively large, the testing was taking a lot of time. All the ML Algorithms took long time to execute, as a result I could tune only 1 best performing model.

Also, Grid Search CV seemed to never stop running, so I tried Randomized Search CV for this task, and it gave decent results.

- ## Limitations of this work and Scope for Future Work

This model does not factor in the effect of Date/Month. For future analysis, I would like to explore more on that and try to see if there are any seasonal relationship or trends among defaulters.