

MICRO-CREDIT RISK ANALYSIS

Submitted by:

Harshvardhan Palawat

**ACKNOWLEDGMENT**

I want to extend my sincere regards to the below mentioned sources and references who helped me a lot in completion of my Project:

Team FlipRobo

Team DataTrained

scikit-learn official documentation

<https://scikit-learn.org/stable/>

geeksforgeeks

https://www.geeksforgeeks.org/

programiz

[https://www.programiz.com](https://www.programiz.com/)

Machine Learning Mastery

<https://machinelearningmastery.com/>

Medium

[https://www.medium.com](https://www.programiz.com/)

**INTRODUCTION**

* Business Problem Framing

We need to build a machine learning model that can classify a given loan transaction whether the customer will return the amount within 5 days or not based on a probability score.

There are a lot of Micro Financial institutions which deals with customer loans and credits at large scale. The similar model can help bring efficiency in their operations by indicating fraudsters based on certain characteristic features.

* Conceptual Background of the Domain Problem

The problem is related to credit assessment for Micro Finance institutions. Credit history of every customer along with his repayment pattern and other crucial factors are recorded. Now, we can convert this enormous data into a structured Machine Learning model which can predict the Repayment likelihood for a given customer even before the loan is credited.

* Review of Literature

The idea of Mobile Financial services is being supported by a lot of of Micro Finance institutions as they feel FMS are convenient and efficient than traditional lending models. Right now, MFI industry is primarily focusingon low income families which makes the implementation of these services even more challenging.

* Motivation for the Problem Undertaken
* India is a country where common man looks up to financial institutions for loans and credits in order to fulfill his dreams whether it be a home or a car or starting a new business. This project helps build a modern approach towards credit risk assessment at a very small scale and industry-oriented manner. The broader concepts and techniques used here can be extended to the mainstream finance industry which benefits both financial institutions and common man by increasing efficiency of operation and mitigating the risk.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This is a supervised classification problem where we need to predict whether the customer will repay the loan within a period of 5 days or not. Thus, we have used linear models like Logistic Regression, tree based models like Decision Tree, Ensemble models like Random Forest for the modelling of this task.

The models were tested on various performance metrices such as accuracy, precision and recall to ensure generalization on future data.

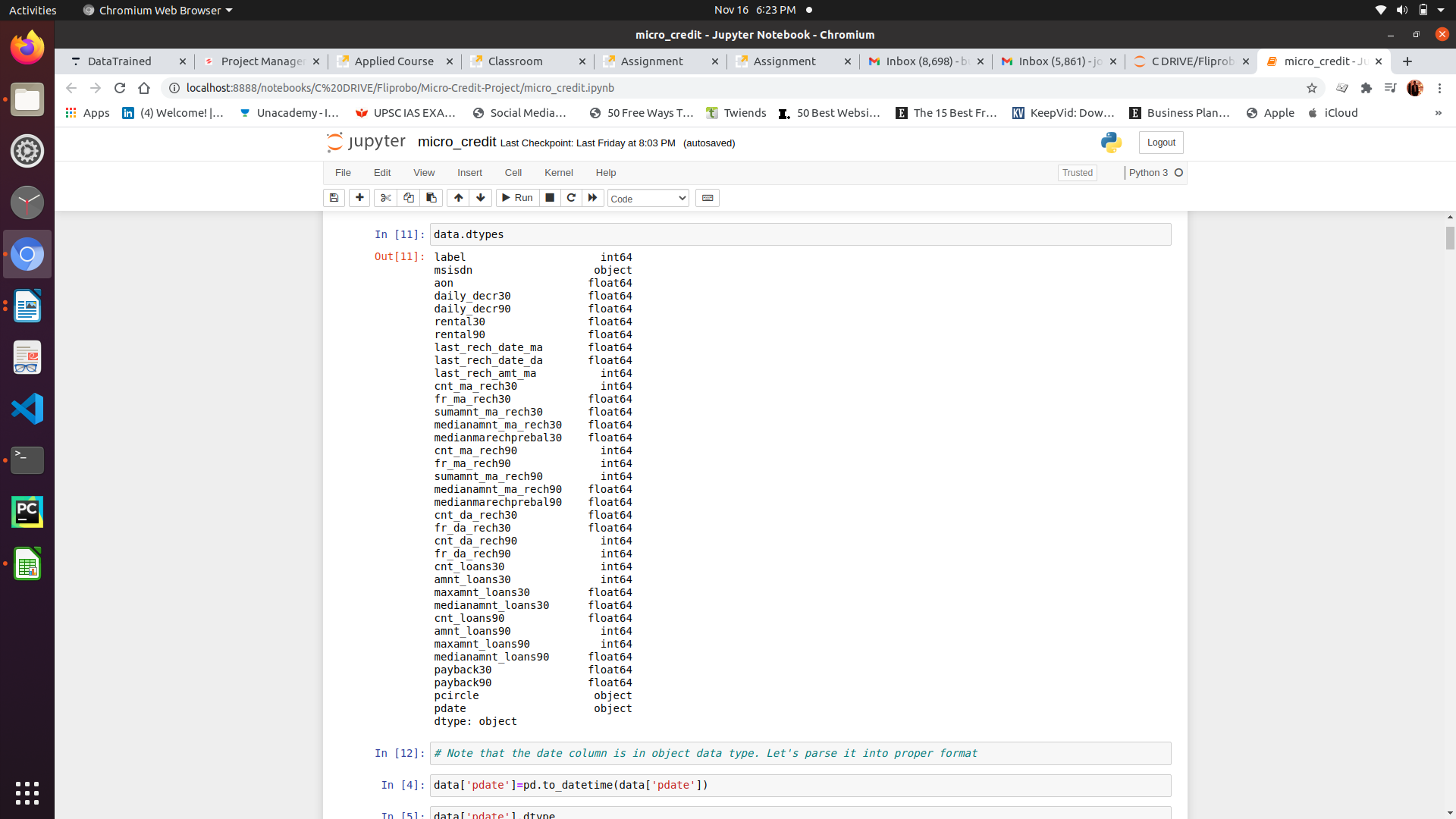
* Data Sources and their formats

The data was provided to us by our client who is in Telecom industry. The data provided was in csv (comma seperated value) format. There are 35 independent features and one target named 'label'.

Most of the features belong to float64 data-type with a few columns with int64 and object data types.

The data is imbalanced with around 85% observations belonging to the majority class.

Below is a snapshot describing the format of all the features along with the target.



* Data Preprocessing Done

The data has zero missing values but the data has a lot of outliers. It is highly skewed and it has a few impossibly large observations in certain features.

Also, there are fractional and negative values in features containing count, which is illogical.

In pre-processing, we have applied several techniques to find what works best. For example, all the outliers were first replaced with the threshold value as per the Inter Quartile Range. Then, we tried removing them as well.

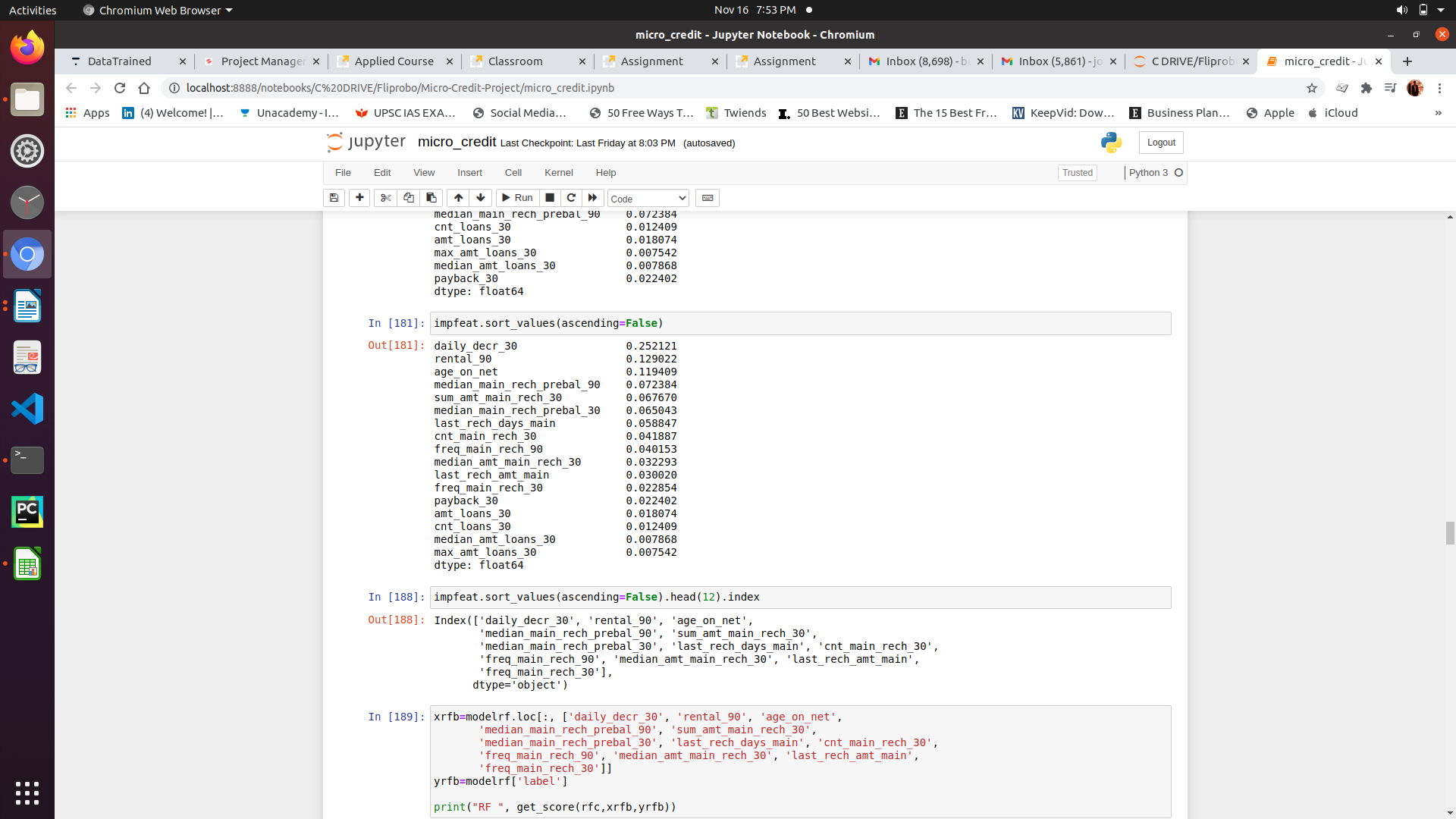
The skewness of the data was removed using various transformation strategies such as log and power transformations of the features.

All the illogical fractional and negative values were either removed or treated with appropriate replacement.

* Data Inputs- Logic- Output Relationships

The data is high-dimensional and can not be visualized directly, but the fact that logistic regression is also performing well on the data indicates that data is almost linearly seperable in higher dimensional space.

We have also used feature importance of Random Forest to have an insight of which features are the most important ones for the classification. Because the data is multi-collinear, we can not use feature weights by Logistic Regression to estimate a relationship between features and target. But, the pearson correlation gives an idea about how input and output are correlated.



* State the set of assumptions (if any) related to the problem under consideration
* We have assumed that any values above and below 1.5 times the IQR is an outlier and will either be treated (replaced) or removed.We have also assumed the features are gaussian distributed and skewness is removed using proper transformation.

* Hardware and Software Requirements and Tools Used

The size of data is very small, therefore any system running on Windows 7 or higher, Mac or Linux based operating systems with 4 GB of RAM is more than sufficient for the given task. We can use any Python IDE or Jupyter notebooks or Google Colab for modelling.

Below is the list of tools used for the task:

sklearn for model building,

pandas for reading and manipulation of data,

numpy for numerical operations,

matplotlib and seaborn for data visualization

imblearn for upsampling the data

scipy for scientific operations and outlier detection

joblib for saving the model

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Since the task is classification based on supervised learning, we can use Logistic Regression, Tree based classification algorithms, Ensemble models and Nearest Neighbors approach etc.

Since there are a lot of outliers, we should use algorithms which are robust to outliers. Also, the size of data is in lakhs, so we should use algorithms with lower time complexity otherwise training time will be huge. Still, it is not that important factor.

We may also use artificial neural networks but the size of data is not enough, so it has a high risk of over-fitting.

* Testing of Identified Approaches (Algorithms)
* We have used following algorithms for training and testing:

Logistic Regression

K-Nearest Neighbors

Decision Tree Classifier

GaussianNB

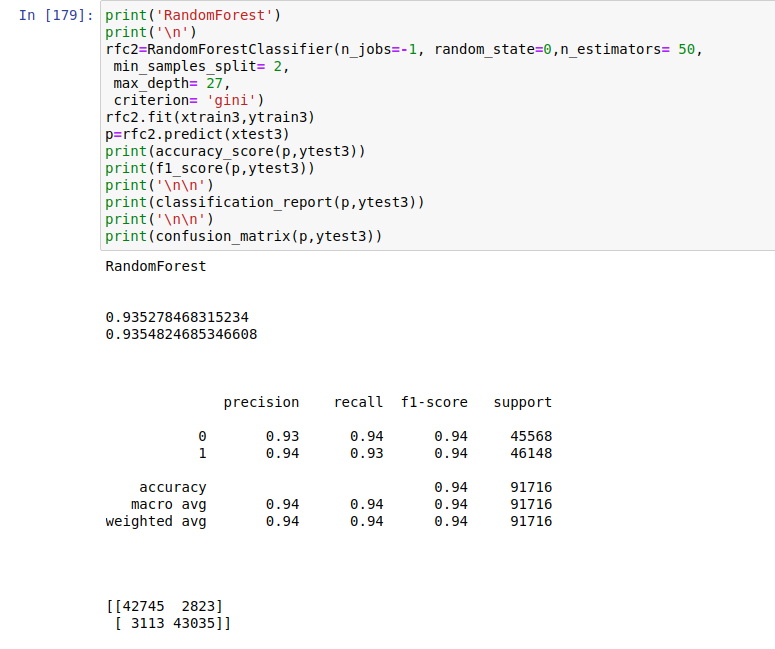
Gradient Boosting Classifier

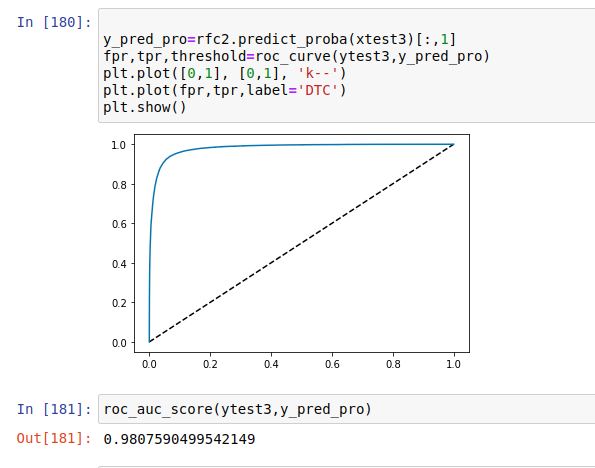
Random Forest Classifier

XGBoost Classifier

* Run and Evaluate selected models
* Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

As listed above, we have used several classification algorithms for training and evaluating the model. Random forest classifier came out as the best performing algorithm of all. Below are the snapshots of the code and results over various evaluation metrices:





* Key Metrics for success in solving problem under consideration

As the data is imbalanced, accuracy would not be a correct measure of performance evaluation. Though we may use accuracy as our performance metric once the data is treated for imbalance using upsampling.

Here we focus more on correctly classifying all the defaulters. We give less emphasis on False Positive for default than False negative as it would cost our client if a defaulter is credited with the loan. Therefore, important metrices here are Recall and Precision. Thus we can use F1 score as our evaluation metrics for a trade-off between both Precision and Recall.

* Visualizations
* A lot of plots were made as part of data visualisation. We used libraries like matplotlib, pandas and seaborn for data visualisation.

The key findings are:

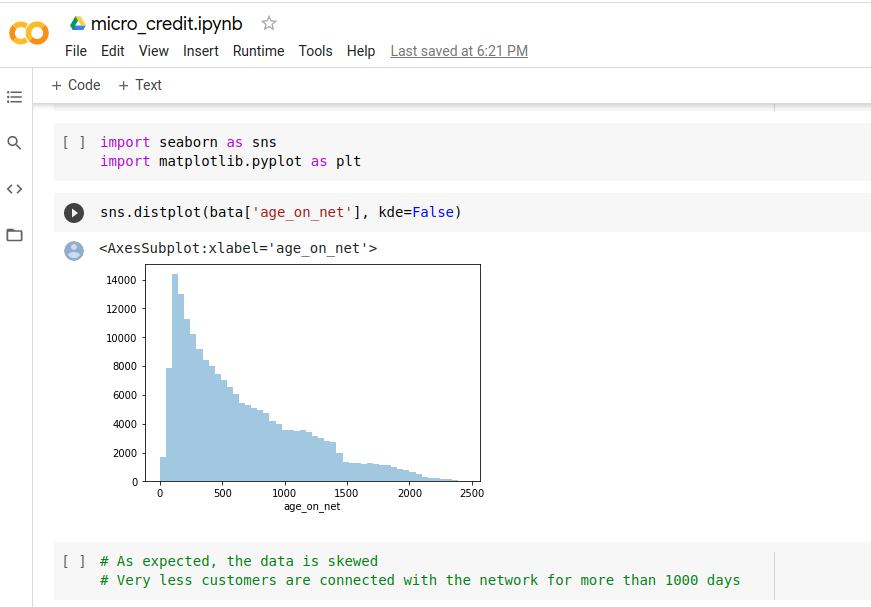
There are negative values in few columns.

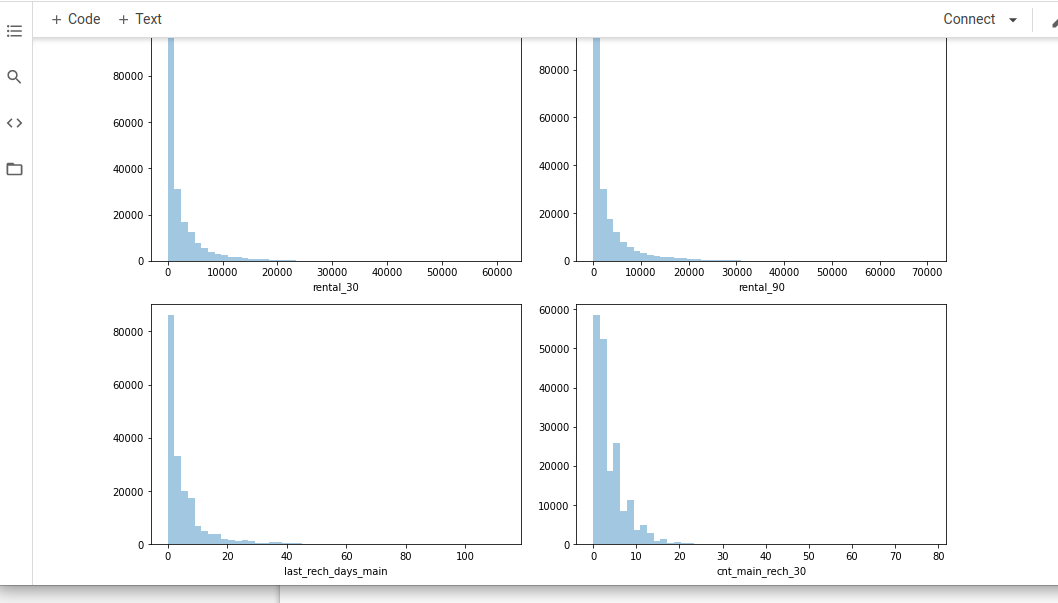
There are exceptionally high values in many columns.

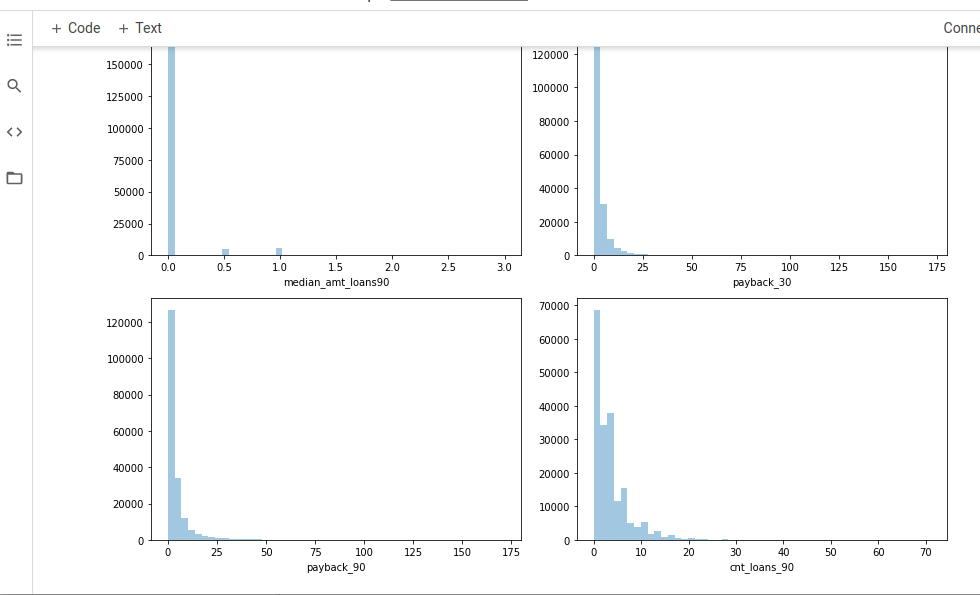
There are few columns which have fractional values.

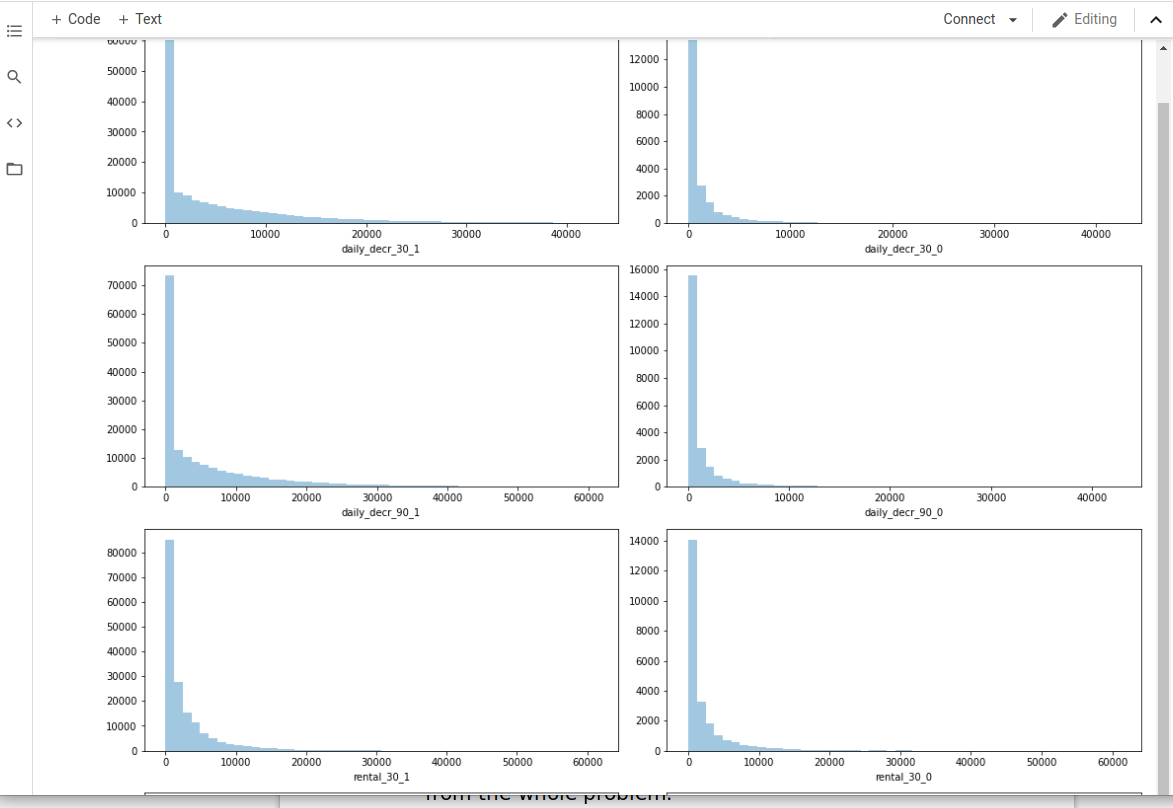
The data is highly skewed.

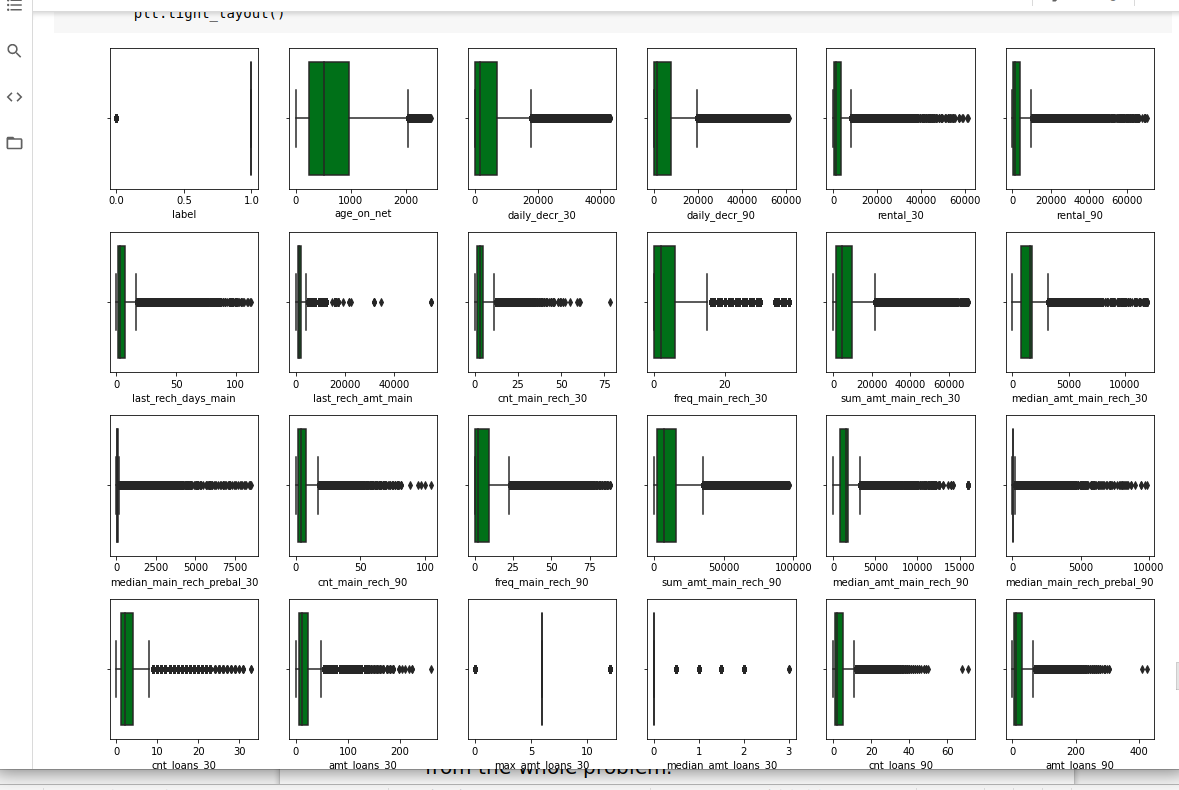
Few of the columns are highly correlated.

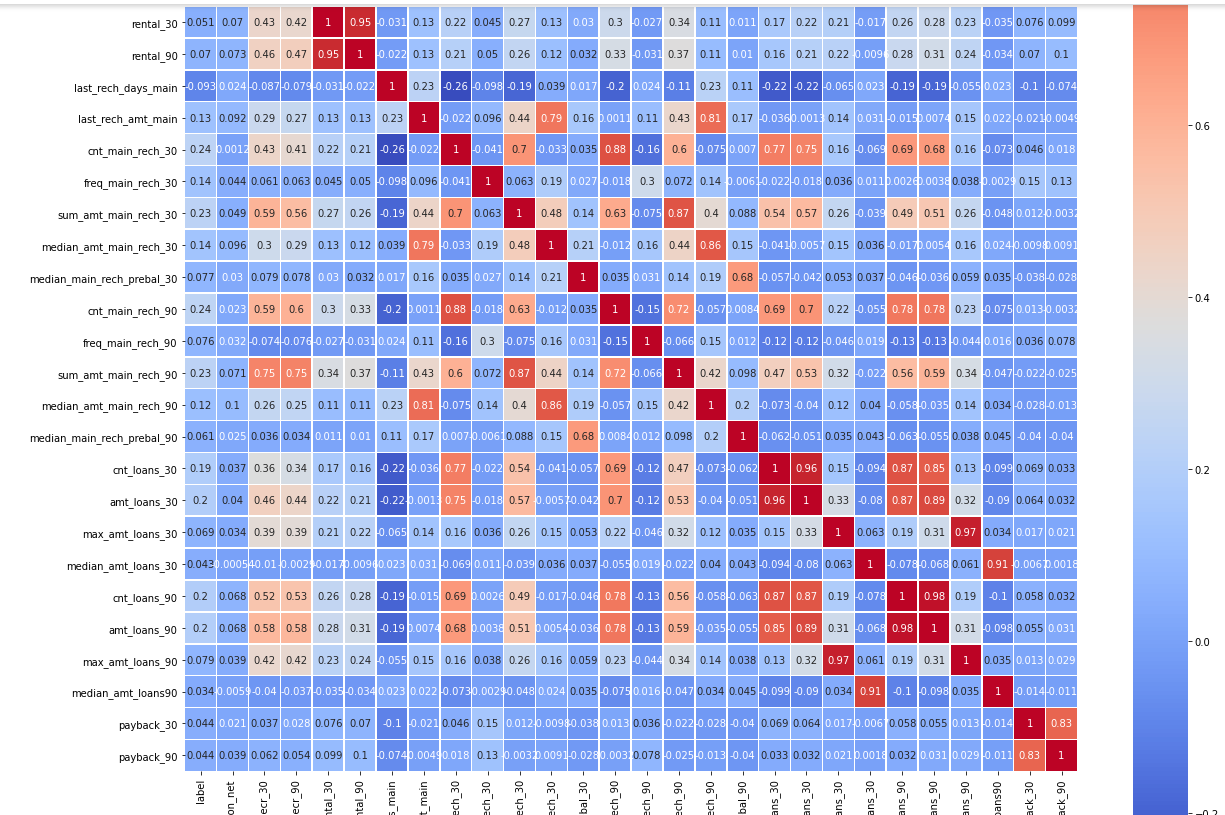












* Interpretation of the Results
* The process for outlier removal, outlier treatment and skewness treatment were designed as per the interpretations from data visualization steps. Statistical methods were used to treat or remove outliers. Also, we used various transformations in our pre-processing steps to remove the skewness. A lot of discrepencies were caught while visualising the data using methods like percentiles, value counts and plotting. Proper measures were adopted to clean the data and get rid of all the discrepencies without losing much data.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

# It is a binary classification problem.

# First of all, there are no missing values

# Looking at the results of describe, we can say that something is fishy in:

# age\_on\_net (min is -ve and max is too large, also days can't be in fraction)

# daily\_dec\_30, daily\_dec\_90, rental\_30, rental\_90 (all has -ve min and way too large max)

# last\_rech\_days\_main, last\_rech\_days\_data (unusual min and max values, also days can't be in fraction)

# last\_rech\_amt\_main (how the value of recharge can be 0, also very high maxm)

# cnt\_main\_rech\_30, freq\_main\_rech\_30 (unusually high values)

# sum\_amt\_main\_rech\_30, median\_amt\_main\_rech\_30 (unusually high values)

# median\_main\_rech\_prebal\_30 (-ve min and too high max)

# cnt\_main\_rech\_90, freq\_main\_rech\_90 (unusually high values)

# sum\_amt\_main\_rech\_90, median\_amt\_main\_rech\_90 (unusually high values)

# median\_main\_rech\_prebal\_90 (-ve min and too high max)

# cnt\_data\_rech\_30, freq\_data\_rech\_30 (most of the values are 0 but still max is very very high, count in fraction)

# cnt\_data\_rech\_90, freq\_data\_rech\_90 (most of the values are 0 but still max is high)

# cnt\_loans\_30, amt\_loans\_30 (max values are little high)

# max\_amt\_loans\_30 (max is tooo high)

# a lot of people didn't even take any loan

# cnt\_loans\_90 (max is very high, also count can't be fraction), amt\_loans\_90 (max is a little high)

# payback\_30, payback\_90 (max is too high)

# Also check whether there are fraction values in the count columns

* Learning Outcomes of the Study in respect of Data Science

Data visualization is the utmost important step for any Machine Learning project as it paves the foundation for data cleaning by giving us a detailed insight about the data. It also gives us an idea about which algorithms might work well for the given data. All the insights we received during data cleaning process are listed above. We planned our data cleaning in accordance with those insights.

The main challenge was to clean the data without loss of it. The power of data was utilised in order to formulate proper cleaning strategies. Also, the imbalance of data was a big challenge. Upsampling technique was used to over come that. Formulation of proper metrics was crucial for the problem.

Selection of algorithm can be done according to the task in hand. For example, Logistic Regression works very well for Linearly Seperable data. Also, there are other factors like Latency, complexity, Interpretability that help us choose the model that can be used.

In our case, Random Forest classifier was giving the best results. The model is a little complex, but we don't have a very low latency requirement as per our use case. It is highly interpretable (which is very necessary in our case) as we can check the feature importances.

* Limitations of this work and Scope for Future Work
* The provided solution can be made better using more complex models like Artificial Neural Networks. Also, improvements can be made in data gathering and cleaning pipeline, as present data has a lot of discrepencies. We can also construct new features by consulting a domain expert.