FLIP ROBO

Micro Credit Use Case

Submitted by:

Simran Kumari

ACKNOWLEDGMENT

I would like to thank every one who helped me during the making of the projects.
The help was provided by:
1. Data Trained faculty
2. Shubham Yadav

References include
1. Scikit-learn.org
2. Kaggle.com
3. Github.com
4. Stack-Overflow
5. Learning.datatrained.com

INTRODUCTION

• Business Problem Framing

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.

We are working with one such 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.

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 payed i.e. Non-defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter.

- Conceptual Background of the Domain Problem
- Review of Literature

  The main purpose of this part of the study is to understand the application of credit scoring in the financial services sector. Based on previously published research, it is possible to identify commonly used datasets, relevant features and the types of models constructed. This valuable information then provides a basis for developing a credit scoring system for mobile airtime lending. While the literature review does not offer any papers specifically addressing airtime lending, a number of papers are identified that undertake research in related fields, in particular microfinance in developing countries, and these offer useful insights. The following paragraphs summarize the different model structures that have been deployed and describe their performance and the variables commonly used. The studies use several models such as linear and quadratic discriminant analysis (LDA, QDA) (Blanco et al. 2013), logistic regression (Baklouti 2013), binary logit model (Van Gool et al. 2009), classification trees, multilayer perceptron (MLP), support vector machine (SVM), random forest and boosting (Cubiles-de-la-Vega et al. 2013). The best performing models were selected and described in Table 1 in terms of the amount of data and number of variables. Evaluation criteria presented in the papers included area under the receiver operating curve (AUC) (Blanco et al. 2013), Kolmogorov–Smirnov Statistic (KS) (Van Gool et al. 2009) and expected misclassification cost (EMC) (Cubiles-de-la-Vega et al. 2013). Table 1. Summary of best performing models. Papers #Variables Model Defaulters Non-Defaulters ACC% N Cubiles-de-la-Vega et al. (2013) 39 MLP 2673 2778 88.33 5451 Van Gool et al. (2009) 16 Binary Logit Model 1661 5061 76.8 6722 Blanco et al. (2013) 39 MLP 2673 2778 93.22 5451 Baklouti (2013) 10 Logistic Regression 1994 3028 – 5022 None of the papers above explicitly address the issue that a given customer may have taken multiple loans. The behaviour of a customer is often summarised by defining a variable that relates to their historical repayments and the outcome of different loans. It is therefore necessary to design an evaluation scheme that can effectively deal with the temporal dynamics concerning different loans. This temporal dimension has important implications for the type of cross-validation technique that is appropriate for this particular application. Another difference is the fact that loan disbursement is relatively frequent in airtime lending unlike bank loans and microfinance loans. The amount being borrowed is also generally quite small compared to the loans approved by financial institutions. The meta analysis selection criteria endeavors to find statistically robust studies whereby the number of customers and loans are large enough to obtain statistically significant results. The criteria also focused on analysing papers that perform credit scoring research in developing countries where ComzAfrica operates. The objective is to find research relating to countries facing similar challenges in terms of data availability on prospective applicants and countries with a similar economic and development context. Variables used in previous studies were also reviewed in order to identify which variables are most likely to offer predictive information for the present case study. The similarity test involves checking how the variables describe a prospective loan applicant and its relevance to the model measured via statistical significance. Finally, these variables employed in the models developed in the papers that are statistically significant and related to the mobile industry are listed in Table 2. The majority of the variables listed above in Table 2 are customer details. The case study company, ComzAfrica, is a third party company that does not have access to extensive customer details such as demographic information (for example, age and gender). There were no papers that describe the application of building a credit score model without substantial customer details. This paper seeks to demonstrate that it is still possible to create accurate credit scoring models for airtime lending without demographic details of customers.

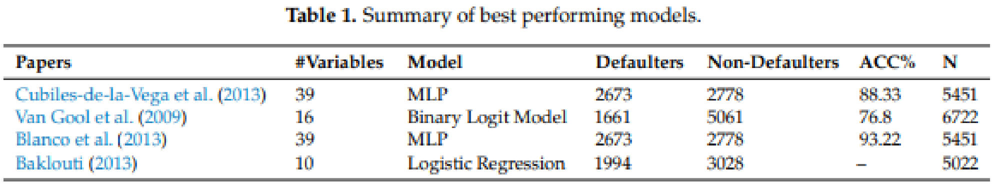


Table 1. Summary of best performing models.

Papers	#Variables	Model	Defaulters	Non-Defaulters	ACC%	N
Cubiles-de-la-Vega et al. (2013)	39	MLP	2673	2778	88.33	5451
Van Gool et al. (2009)	16	Binary Logit Model	1661	5061	76.8	6722
Blanco et al. (2013)	39	MLP	2673	2778	93.22	5451
Baklouti (2013)	10	Logistic Regression	1994	3028	–	5022

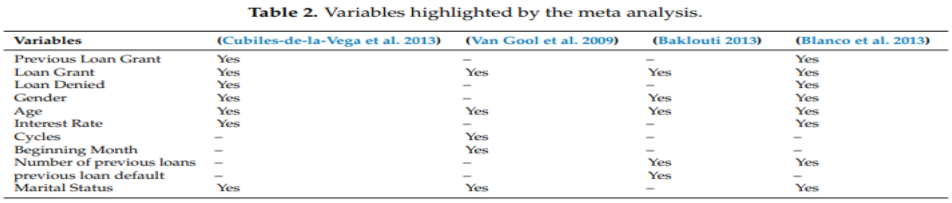


Table 2. Variables highlighted by the meta analysis.

Variables	(Cubiles-de-la-Vega et al. 2013)	(Van Gool et al. 2009)	(Baklouti 2013)	(Blanco et al. 2013)
Previous Loan Grant	Yes	–	–	Yes
Loan Grant	Yes	Yes	Yes	Yes
Loan Denied	Yes	–	–	Yes
Gender	Yes	–	Yes	Yes
Age	Yes	Yes	Yes	Yes
Interest Rate	Yes	–	–	Yes
Cycles	–	Yes	–	–
Beginning Month	–	Yes	–	–
Number of previous loans	–	–	Yes	Yes
previous loan default	–	–	Yes	–
Marital Status	Yes	Yes	–	Yes

- Motivation for the Problem Undertaken

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

In the financial services sector, it is more important to predict those who will default than those who will repay. This is because the financial risk associated with defaulters is high. A confusion matrix is used to evaluate the classification models with positive (negative) outcomes denoting repayment (default), respectively. This 2x2 matrix measures the number of predicted/actual cases that are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). From this matrix, it is possible to calculate the classification accuracy. The accuracy formula is given by:

$$ACC = (TP + TN)/(TP + TN + FN + FP)$$

The confusion matrix for a predictive classifier and describes the impacts associated with the four potential outcomes. The matrix gives the cost or benefit in terms of profit for each prediction and actual outcome where the objective is to identify customers that will repay. Lending is a cost-sensitive business, where the cost of one non-performing loan from a customer defaulting is much greater than the benefit of a customer repaying, a factor of ten in this case. The loss from rejecting a customer that will repay is much less than that suffered from approving a loan for a customer that will default. Accuracy is a useful summary statistic but is not the most relevant performance metric for this particular business application. The greatest threat to financial sustainability arises when the classifier predicts that a customer will repay a loan and they actually default (FP). Therefore, it is most important to correctly predict the customers who will not repay. For applications that require highly effective detection ability for only one class, it is recommended to consider an alternative metric to accuracy (Tang et al. 2009). The loan approval application is best assessed using the classification metric known as specificity defined as:

$$Specificity = TN/(TN + FP)$$

Specificity measures the probability that a classifier correctly predicts default when considering all those that actually default. The priority for profitable lending is to avoid customers that are likely to default, which is achieved by maximizing specificity.

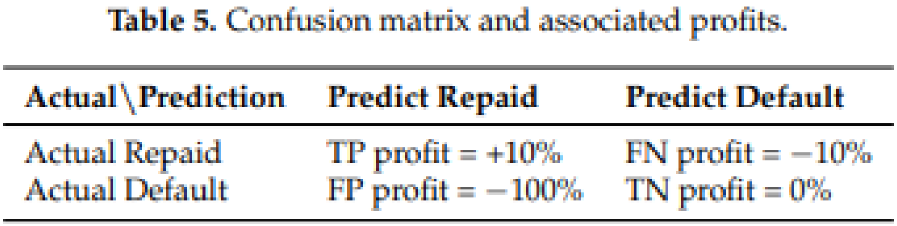


Table 5. Confusion matrix and associated profits.

Actual\Prediction	Predict Repaid	Predict Default
Actual Repaid	TP profit = +10%	FN profit = −10%
Actual Default	FP profit = −100%	TN profit = 0%

- Data Sources and their formats

The data was provided by the flip and robo team.

Below is the data description

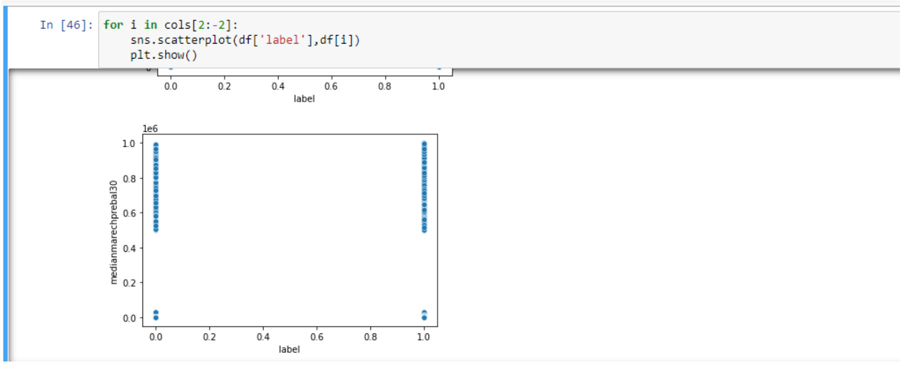
Variable	Definition
label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the 0:failure}
msisdn	mobile number of user
aon	age on cellular network in days
daily_decr30	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah
daily_decr90	Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah
rental30	Average main account balance over last 30 days
rental90	Average main account balance over last 90 days
last_rech_date_ma	Number of days till last recharge of main account
last_rech_date_da	Number of days till last recharge of data account
last_rech_amt_ma	Amount of last recharge of main account (in Indonesian Rupiah)
cnt_ma_rech30	Number of times main account got recharged in last 30 days
fr_ma_rech30	Frequency of main account recharged in last 30 days
sumamnt_ma_rech30	Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
medianamnt_ma_rech30	Median of amount of recharges done in main account over last 30 days at user level (in I
medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level (in Indo
cnt_ma_rech90	Number of times main account got recharged in last 90 days
fr_ma_rech90	Frequency of main account recharged in last 90 days
sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)
medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user level (in I
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indo
cnt_da_rech30	Number of times data account got recharged in last 30 days
fr_da_rech30	Frequency of data account recharged in last 30 days
cnt_da_rech90	Number of times data account got recharged in last 90 days
fr_da_rech90	Frequency of data account recharged in last 90 days
cnt_loans30	Number of loans taken by user in last 30 days
amnt_loans30	Total amount of loans taken by user in last 30 days
maxamnt_loans30	maximum amount of loan taken by the user in last 30 days
medianamnt_loans30	Median of amounts of loan taken by the user in last 30 days
cnt_loans90	Number of loans taken by user in last 90 days

amnt_loans90	Total amount of loans taken by user in last 90 days
maxamnt_loans90	maximum amount of loan taken by the user in last 90 days
medianamnt_loans90	Median of amounts of loan taken by the user in last 90 days
payback30	Average payback time in days over last 30 days
payback90	Average payback time in days over last 90 days
pcircle	telecom circle
pdate	date

Label, pdate, pcircle were of object data type and rest all were in the numerical data type

- Data Preprocessing Done
- Regularised the data using MinMax Scaler.- The next step is to bring the data to a common scale, since there are certain columns with very small values and some columns with high values. This process is important as values on a similar scale allow the model to learn better. We used MinMax scaler for this process
- Removed the skewness using power transform-yeo-Johnson- The Yeo–Johnson transformation allows also for zero and negative values in the dataset.
- Removed the outliers – Outliers are some miss interpreted data which could be far from the range that should be accepted. Removed outliers using the scipy.stats.zscore and removed 9% of the total data.
- The data was imbalanced – Since we had a lot of data used nearMiss (undersampling to balance the output data)

- Data Inputs- Logic- Output Relationships
- We used scatterplot to get the relationship between input and output variables.

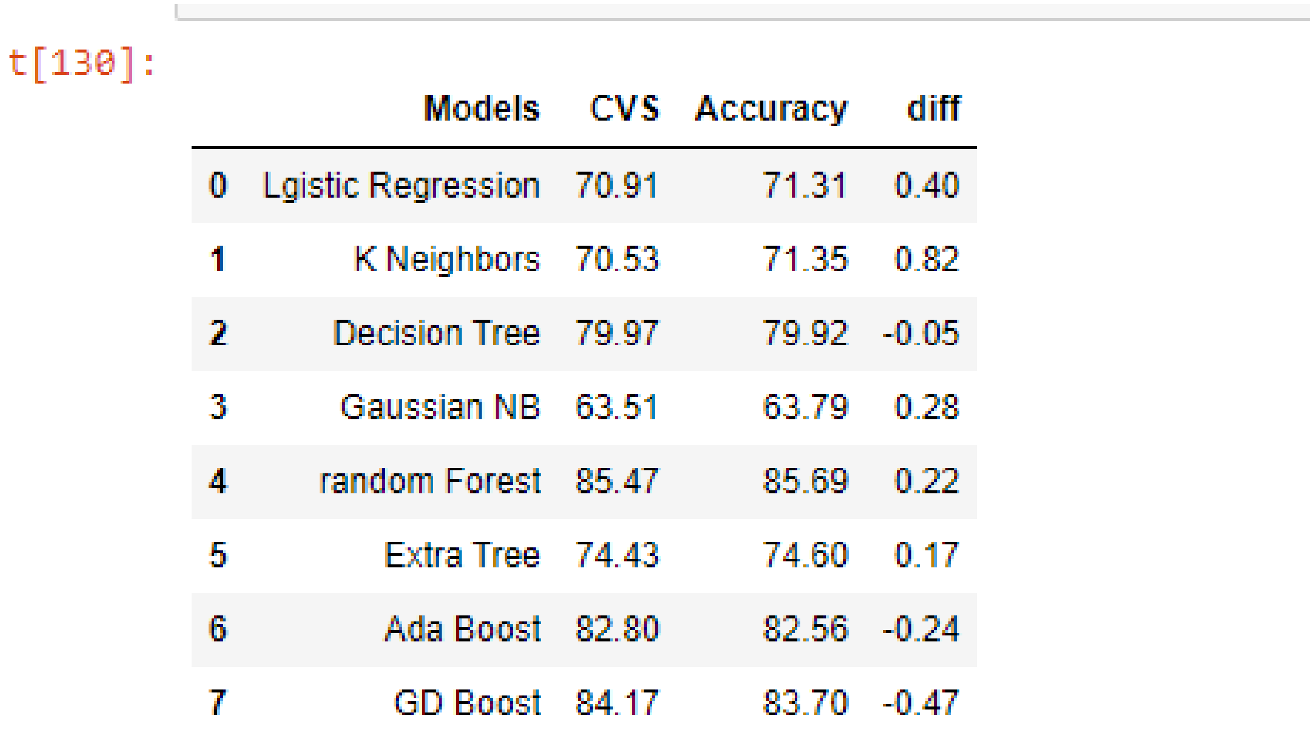


```
In [46]: for i in cols[2:-2]:
             sns.scatterplot(df['label'],df[i])
             plt.show()
```

- State the set of assumptions (if any) related to the problem under consideration
  1. The data was mostly left skewed.
  2. There were a lot of variables that were collinear.
  3. The data was highly spread
  4. Data was imbalanced.
  5. The data had a lot of outliers.
- Hardware and Software Requirements and Tools Used
- Listing down the hardware and software requirements along with the tools, libraries and packages used.
- 1. Jupyter Notebook.- Used python to perform the machine Learning task
- 2. Laptop with 8 GB RAM

Model/s Development and Evaluation

Different models I tried:



t[130]:

	Models	CVS	Accuracy	diff
0	Lgistic Regression	70.91	71.31	0.40
1	K Neighbors	70.53	71.35	0.82
2	Decision Tree	79.97	79.92	-0.05
3	Gaussian NB	63.51	63.79	0.28
4	random Forest	85.47	85.69	0.22
5	Extra Tree	74.43	74.60	0.17
6	Ada Boost	82.80	82.56	-0.24
7	GD Boost	84.17	83.70	-0.47

#From the above analysis Random Forest Classifier has least difference between r2 and cvs

Using hyper parameter tuning on Random Forest Classifier further increased the accuracy.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

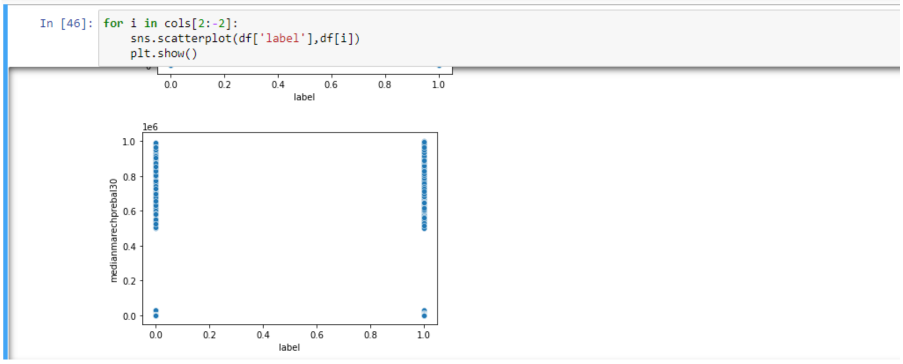
A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. **The reason for this wonderful effect is that the trees protect each other from their individual errors** (as long as they don’t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

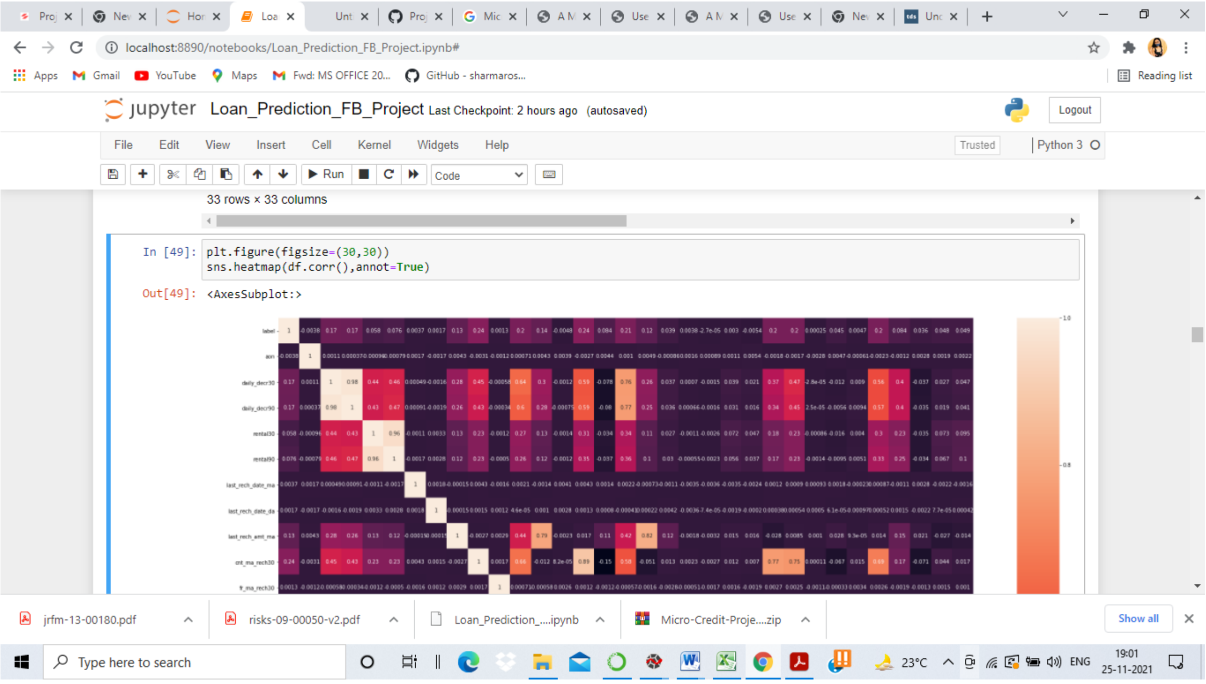
Visualization

Scatter plot



```
In [46]: for i in cols[2:-2]:
             sns.scatterplot(df['label'],df[i])
             plt.show()
```

Violin plot

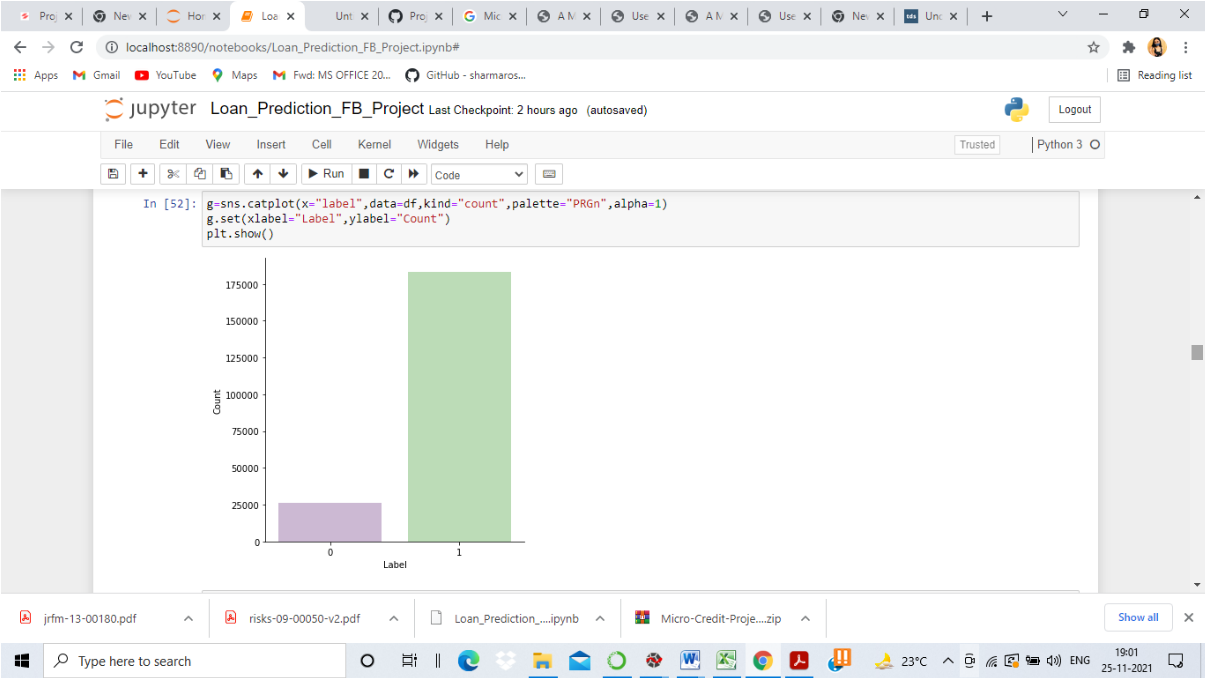


33 rows × 33 columns

```
In [49]: plt.figure(figsize=(30,30))
         sns.heatmap(df.corr(),annot=True)

Out[49]: <AxesSubplot:>
```

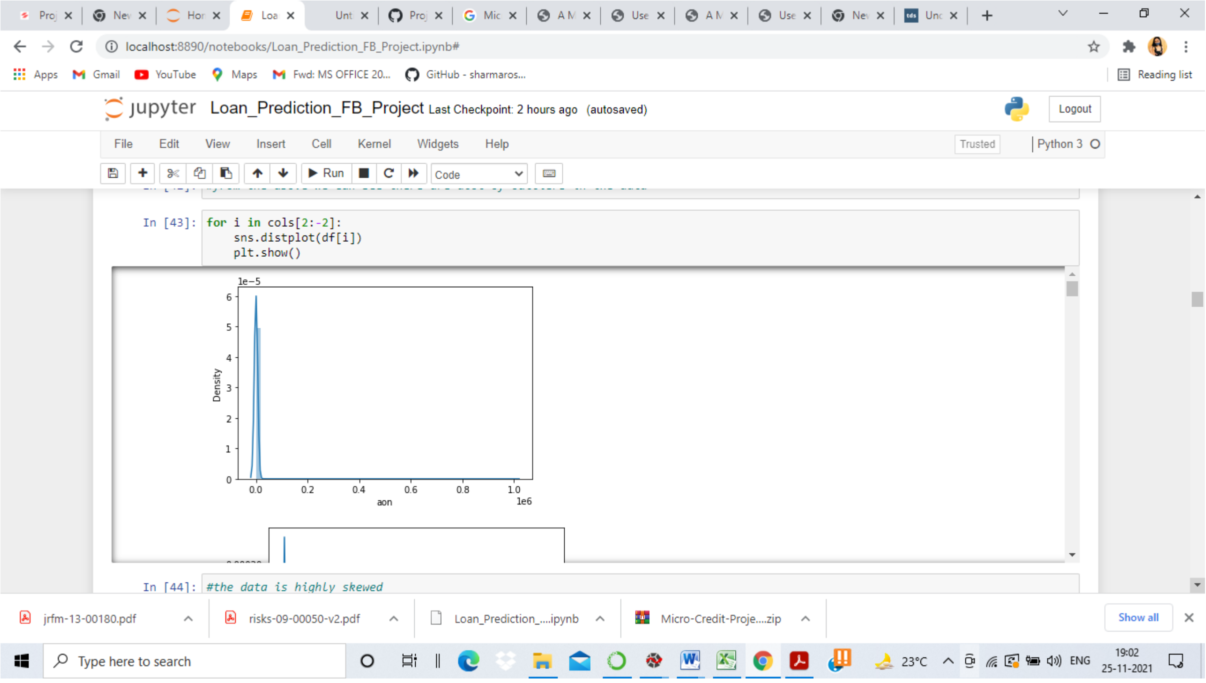
Cat plot



localhost:8890/notebooks/Loan_Prediction_FB_Project.ipynb#
jupyter Loan_Prediction_FB_Project Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
175000
150000
125000
100000
75000
50000
25000
0
Count
0
1
Label
jrfm-13-00180.pdf
risks-09-00050-v2.pdf
Loan_Prediction_....ipynb
Micro-Credit-Proje....zip
Show all
Type here to search
23°C
ENG
19:01
25-11-2021

```
In [52]: g=sns.catplot(x="label",data=df,kind="count",palette="PRGn",alpha=1)
         g.set(xlabel="Label",ylabel="Count")
         plt.show()
```

Distplot

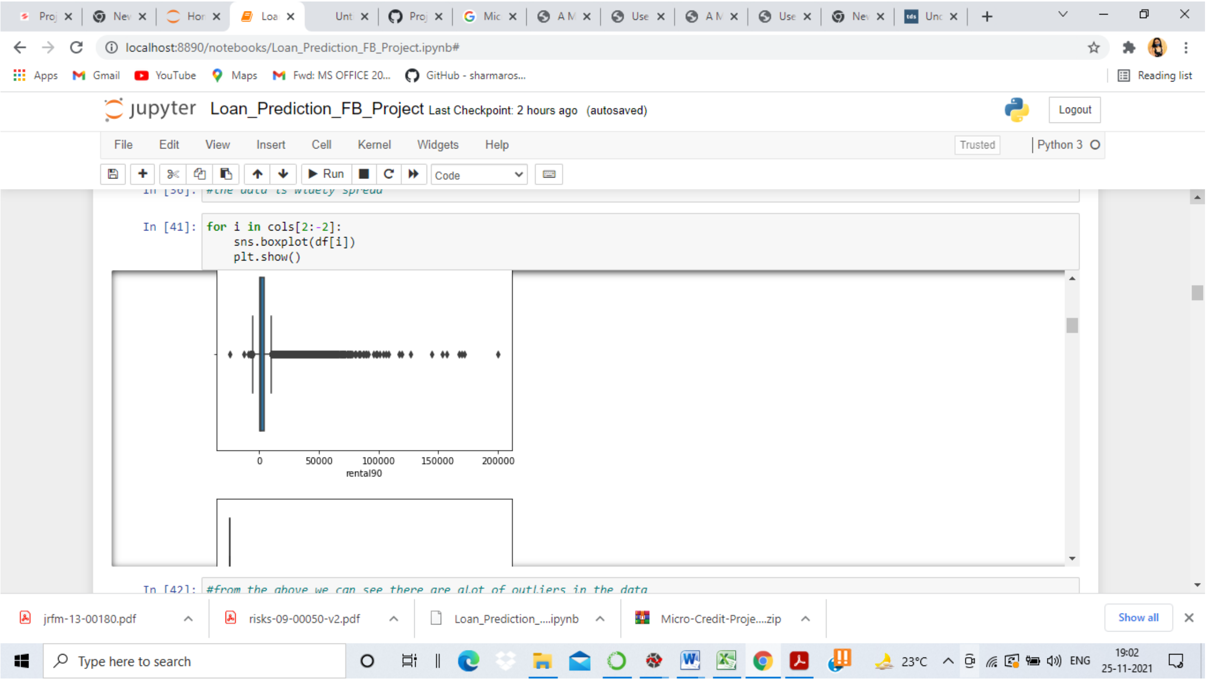


localhost:8890/notebooks/Loan_Prediction_FB_Project.ipynb#
jupyter Loan_Prediction_FB_Project Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
1e−5
6
5
4
3
2
1
0
Density
0.0
0.2
0.4
0.6
0.8
1.0
1e6
aon
jrfm-13-00180.pdf
risks-09-00050-v2.pdf
Loan_Prediction_....ipynb
Micro-Credit-Proje....zip
Show all
Type here to search
23°C
ENG
19:02
25-11-2021

```
In [43]: for i in cols[2:-2]:
             sns.distplot(df[i])
             plt.show()
```

```
In [44]: #the data is highly skewed
```

Boxplot



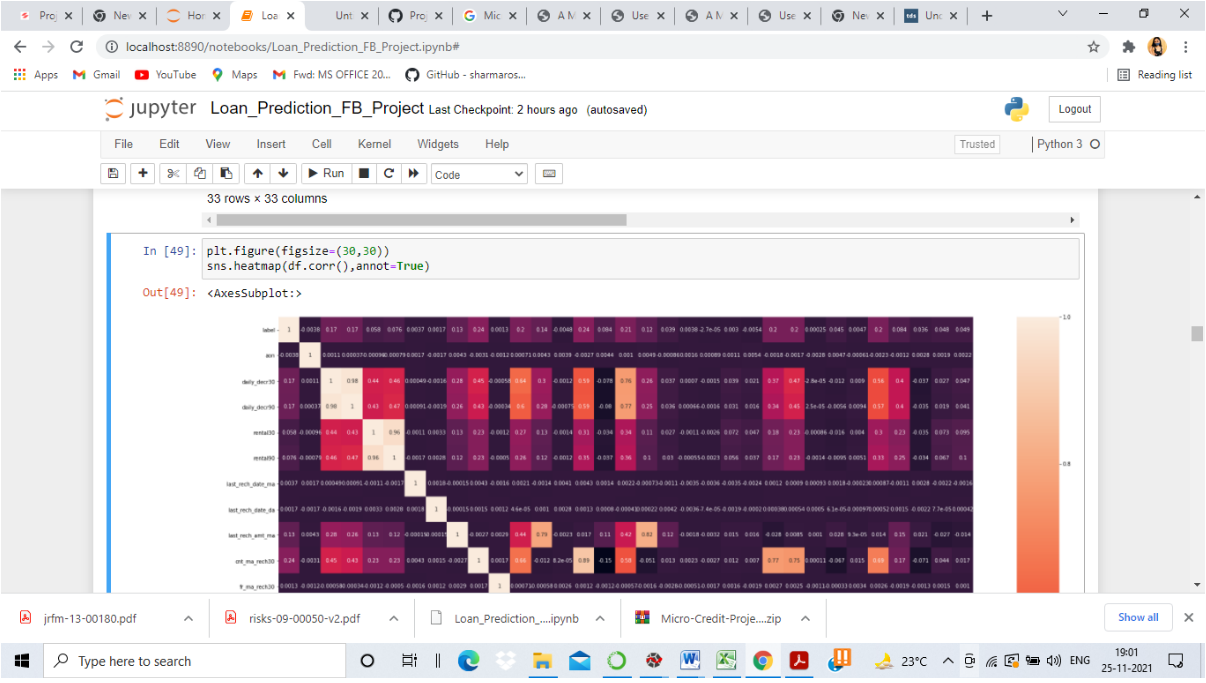
```
In [36]: #the data is widely spread
```

```
In [41]: for i in cols[2:-2]:
             sns.boxplot(df[i])
             plt.show()
```

jupyter Loan_Prediction_FB_Project Last Checkpoint: 2 hours ago (autosaved)
localhost:8890/notebooks/Loan_Prediction_FB_Project.ipynb#
0
50000
100000
150000
200000
rental90

```
In [42]: #from the above we can see there are alot of outliers in the data
```

Testing Corelation



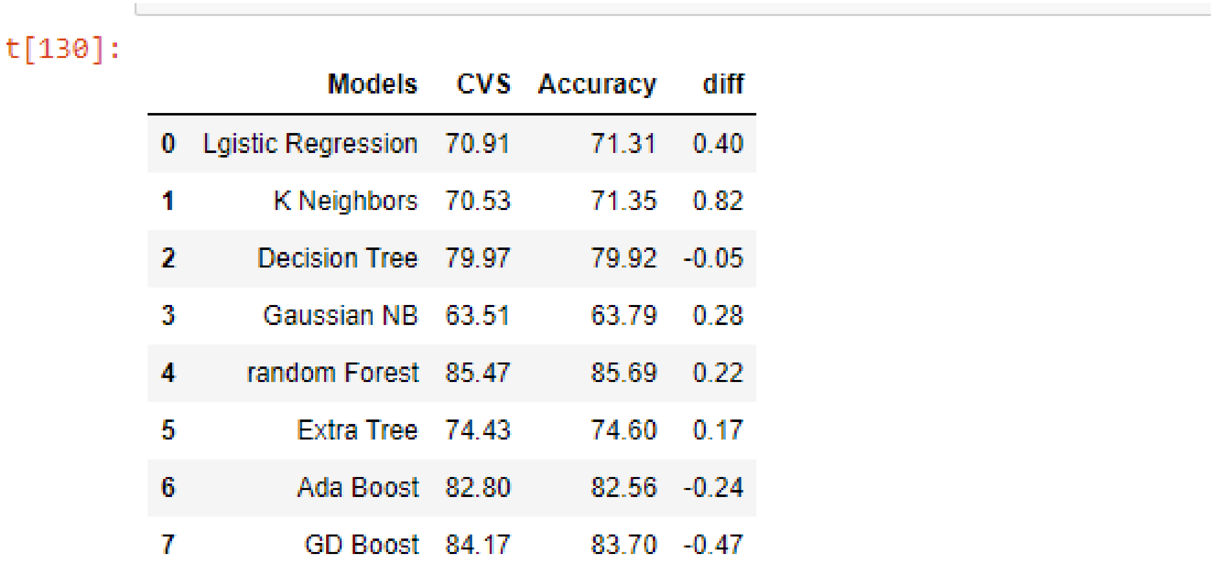
33 rows × 33 columns

```
In [49]: plt.figure(figsize=(30,30))
         sns.heatmap(df.corr(),annot=True)
Out[49]: <AxesSubplot:>
```

jupyter Loan_Prediction_FB_Project Last Checkpoint: 2 hours ago (autosaved)
localhost:8890/notebooks/Loan_Prediction_FB_Project.ipynb#
label
aon
daily_decr30
daily_decr90
rental30
rental90
last_rech_date_ma
last_rech_date_da
last_rech_amt_ma
cnt_ma_rech30
fr_ma_rech30

CONCLUSION

The model that fits best is random forest



t[130]:

	Models	CVS	Accuracy	diff
0	Lgistic Regression	70.91	71.31	0.40
1	K Neighbors	70.53	71.35	0.82
2	Decision Tree	79.97	79.92	-0.05
3	Gaussian NB	63.51	63.79	0.28
4	random Forest	85.47	85.69	0.22
5	Extra Tree	74.43	74.60	0.17
6	Ada Boost	82.80	82.56	-0.24
7	GD Boost	84.17	83.70	-0.47