MICRO CREDIT PROJECT

Submitted by:

Akriti Kakkar

ACKNOWLEDGMENT

I have referred to datatrained course material, fliprobo use case documentation and packages documentation for this project.

INTRODUCTION

- ## Business Problem Framing

    i. 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.
    ii. Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.
    iii. This case study is intended to distinguish between honest and dishonest loan seekers.

- ## Conceptual Background of the Domain Problem

- 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 donorsare 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 MFShas 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 arecollaborating 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 be6(in Indonesian

Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).

- ## Review of Literature

  This is a comprehensive and introductive summary of research done and background of the problem.

  Please follow the link to google dashboard:

  https://docs.google.com/spreadsheets/d/1wEgp5xKLCykpivGxlDrialuBTrEhKa_AEeKH01wnXMY/edit?usp=sharing

- ## Motivation for the Problem Undertaken

  I have financial background and such social concepts, like, MFI, have always touched my heart. It has been a privilege to work on this project write this exhaustive research documentation. To every pro there is an aligned con. Behind this noble cause, there is always a need to detect who is honest loan seeker and who is dishonest loan seeker. With empathy to the client, I found this case study really interesting and could detect many anomalies and trends that I have shared in various dashboards. Along with that based on the trend, I could also come up with an improvisation scheme for customer retention and more recovery of the loans disbursed.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

1. Data Extraction: Using read_ csv function of pandas library to read the data in tabulated format and analyze it.

2. Data Cleaning for missing values detection and its handiling.

3. Feature Engineering for encoding object format data and deriving more features.

4. EDA For data visualization and biasness detection:

· HEAD VIEW OF DATA

· TAIL VIEW OF DATA

· SAMPLE VIEW OF DATA

· GROUPBY EXPLORATION

· DESCRIPTIVE STATISTICS

· SCATTER PLOTS

· CORRELATION ANALYSIS

· BOX PLOTS EXPLORATION

· DESCRIPTIVE STATISTICS

· DISTRIBUTION PLOTS

5. VIF Test for multicollinearity reduction.

6. Power Transformation for standard scaling and outliers transformation.

7. Data PreProcessing for data transformation, scaling and vectorization.

8. Feature Selection (Ensemble Methods): ANOVA Test, p value, ftest, constant threshold filter to classify features based on relevance and biasness and select the most relevant features.

9. Model Development, Evaluation And Selection (Ensemble Methods and Grid Search CV) to do best hyper parameter tuning and develop low bias and low variance with right fit and minimal difference between test metrics and train metrics.

10. Resampling: To make 0 and 1 equal in number to remove biasness in learning of the models.

- ## Data Sources and their formats

  Data Source: Client Database.

  Data Transfer Tool: CSV File.

- ## Data Preprocessing Done

- · Data Pre Processing
- · POWER TRANSFORM
- · MIN MAX SCALING
- · VECTORIZATION
- · TRAIN TEST SPLIT
- The data was further used for feature selection.
- Assumption made:
- • Acceptable Skewness Range Is +/-0.65
- • Acceptable VIF Score Is Than 6

- • Acceptable P Value Is Less Than 0.05
- • Variance Threshold Is 0.01
- Data Inputs- Logic- Output Relationships
- Data Inputs Include:

```
1 p_values
```

	Features	P Values
fr_ma_rech30	fr_ma_rech30	0.000000e+00
fr_ma_rech90	fr_ma_rech90	0.000000e+00
medianamnt_loans30	medianamnt_loans30	1.283229e-252
medianamnt_loans90	medianamnt_loans90	4.513084e-146
cnt_da_rech90	cnt_da_rech90	1.557350e-37
fr_da_rech90	fr_da_rech90	7.178386e-11
cnt_da_rech30	cnt_da_rech30	7.197254e-10
fr_da_rech30	fr_da_rech30	3.731212e-04

-
- Data Input Type: float; Min Max Scaling in the range of 0 to 1.
- Impact On Output: fr_da_rech90 and fr_da_rech30 are negatively correlated with label and others are positively correlated with label.

fr_ma_rech30	0.000867
fr_ma_rech90	0.083645
fr_da_rech90	-0.00716
cnt_da_rech30	0.005345
cnt_da_rech90	0.00111
medianamnt_loans30	0.043686
medianamnt_loans90	0.034773
fr_da_rech30	-0.00285

- State the set of assumptions (if any) related to the problem under consideration
- · Acceptable Skewness Range Is +/-0.65

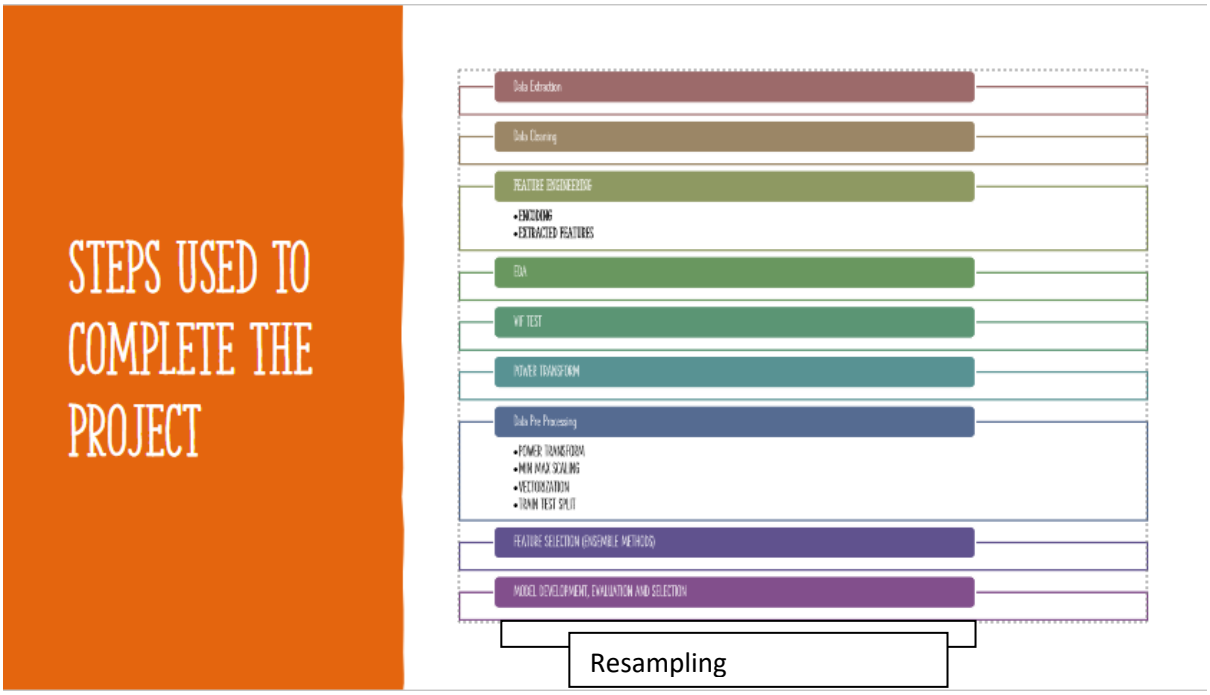
- · Acceptable VIF Score Is Than 6
- · Acceptable P Value Is Less Than 0.05
- · Variance Threshold Is 0.01

- Hardware and Software Requirements and Tools Used

- Installation Of Anaconda Community.
- Required Installations:
- · Pandas (Within environment)
- · Numpy (Within environment)
- · Seaborn (Within environment)
- · Matplotlib (Within environment)
- · Cufflinks
- · Plotly Express
- · Sklearn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)



1. Data Extraction: Using read_ csv function of pandas library to read the data in tabulated format and analyze it.

2. Data Cleaning for missing values detection and its handiling.

3. Feature Engineering for encoding object format data and deriving more features.

4. EDA For data visualization and biasness detection:

· HEAD VIEW OF DATA

· TAIL VIEW OF DATA

· SAMPLE VIEW OF DATA

· GROUPBY EXPLORATION

· DESCRIPTIVE STATISTICS

· SCATTER PLOTS

· CORRELATION ANALYSIS

· BOX PLOTS EXPLORATION

· DESCRIPTIVE STATISTICS

· DISTRIBUTION PLOTS

5. VIF Test for multicollinearity reduction.

6. Power Transformation for standard scaling and outliers transformation.

7. Data PreProcessing for data transformation, scaling and vectorization.

8. Feature Selection (Ensemble Methods): ANOVA Test, p value, ftest, constant threshold filter to classify features based on relevance and biasness and select the most relevant features.

9. Model Development, Evaluation And Selection (Ensemble Methods and Grid Search CV) to do best hyper parameter tuning and develop low bias and low variance with right fit and minimal difference between test metrics and train metrics.

10. Resampling: To make 0 and 1 equal in number to remove biasness in learning of the models.

- Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing (Total Model=7):

 # Model 1: Random Forest Classifier With Intuitional Hyper Parameter Tuning

 # Model 2: Random Forest Classifier With Default Hyper Parameter Tuning

 # Model 3: RFC With Grid Search CV

 # Model 4: Bagging Classifier With Grid Search CV Hyper Parameter Tuning

 # Model 5: Decision Tree Regressor With Default Hyper Parameter Tuning

 # Model 6: RFC On Resampled Data With Intuitional Hyper Parameter Tuning

 # Model 7: RFC On Resampled Data With Default Hyper Parameter Tuning

- Run and Evaluate selected models

  Follow the link to dashboard:

  https://docs.google.com/spreadsheets/d/1TuNyvHS9SMSplSkHTN0_OEhV4isocrpgNdzR_IpV1xs/edit?usp=sharing

- Key Metrics for success in solving problem under consideration

- 1. Power Transform: To remove outliers from extremely spread out data.
- 2. VIF Scores: To reduce multicollinearity from a highly biased dataset.
- 3. Ensemble Methods: To remove over fitting in a complex dataset and finding maximum explanatory power .

- 4. Resampling: To remove biasness and optimize models by improving the way of learning patterns.
-
- Visualizations
  1. Log Transformed and Cumulative Visualizations:

     Follow the link to the dashboard:

     https://docs.google.com/spreadsheets/d/1wEgp5xKLCykpivGxlDrialuBTrEhKa_AEeKH01wnXMY/edit?usp=sharing

  2. Head, Sample, and Tail View Of Data

     Follow the link to the dashboard:

     https://docs.google.com/spreadsheets/d/1d1ixblgxC86MrFE2zjB3BepgPMtgIJuCtrhRenFrWQw/edit?usp=sharing

  3. Groupby

     Follow the link to the dashboard:

     https://docs.google.com/spreadsheets/d/1OwcCvO9vY5b5UTBR8_r0-rsChG4K9pLL2Hxo1aev-4M/edit?usp=sharing

  4. Scatter Plots

     Follow the link to the dashboard:

     https://docs.google.com/spreadsheets/d/1p_B4bUvKvDGGjXBbwzjADuIIphEWqnYApzoEfekEtj8/edit?usp=sharing

  5. Histograms

     Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1v0qh6j2l5-ZO5i7Ya-T6QMnx_CF31fprDkTW4GvTAT8/edit?usp=sharing

6. Correlation

   Follow the link to the dashboard:

   https://docs.google.com/spreadsheets/d/1QnC77IXeSpTk__Rfz_WJlXwNaDyPHbWJhwFVpm4Q4HY/edit?usp=sharing

7. Box Plots

   Follow the link to the dashboard:

   https://docs.google.com/spreadsheets/d/186CvJpOSw521aU54LZn4TYK0m-2o96ub98xZ_MTFjBw/edit?usp=sharing

8. Descriptive Statistics

   Follow the link to the dashboard:

   https://docs.google.com/spreadsheets/d/1_qmUX1jXKiEGZDP3symS1VlZKxOg_7nlhq8Kl7ZrXRI/edit?usp=sharing

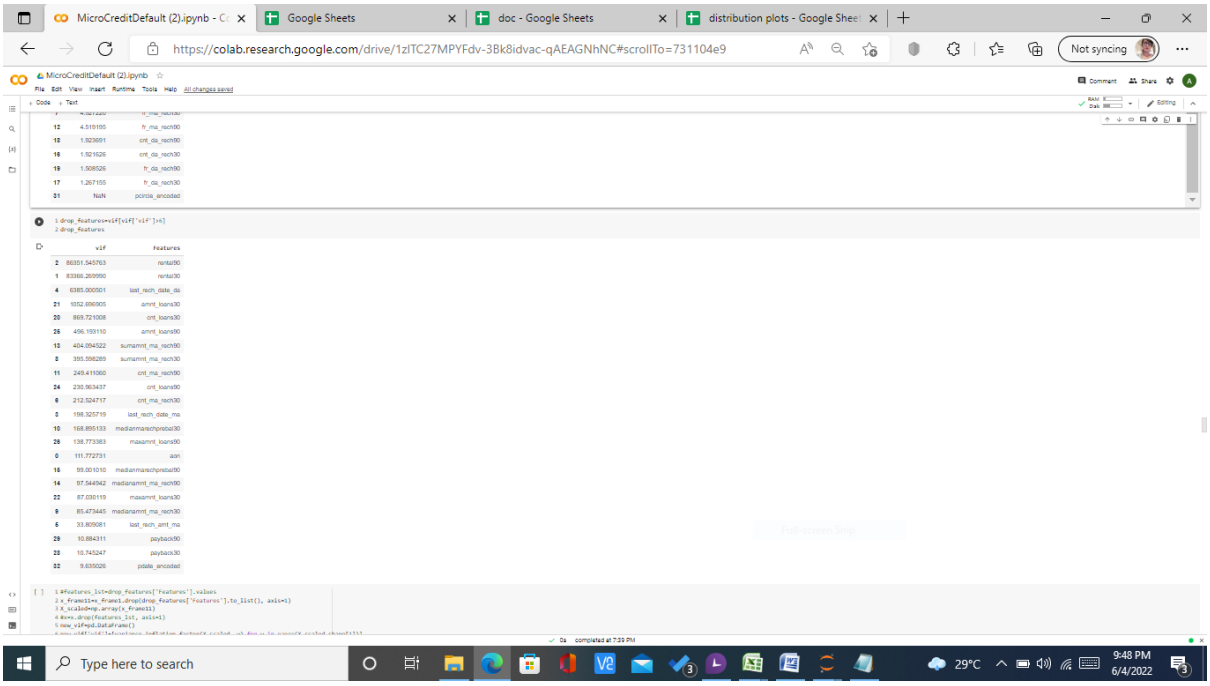
9. Distribution Plots

   Follow the link to the dashboard:

   https://docs.google.com/spreadsheets/d/171asS_59gZxz8uhHiNHUUtlg4gj0dvLSZWllNb95u14/edit?usp=sharing

- Interpretation of the Results

- # Based on above analysis, these columns have significant outliers, hence, are platykurtic, hence, I am removing these columns.
-
- i. Daily Decr 30
-
- ii. Daily Decr 90
-
- I will do further vif test to remove multicollinearity from the data.



- a. Based on VIF, I have dropped these features. I have done VIF after outliers transformation and feature scaling.
- b. Final features passed for anova test include:

```
#features_lst=drop_features['Features'].values
x_frame11=x_frame1.drop(drop_features['Features'].to_list(), axis=1)
X_scaled=np.array(x_frame11)
#x=x.drop(features_lst, axis=1)
new_vif=pd.DataFrame()
new_vif['vif']=[variance_inflation_factor(X_scaled, w) for w in range(X_scaled.shape[1])]
new_vif['Features']=x_frame11.columns
new_vif=new_vif.sort_values(ascending=False, by='vif')
new_vif
```

	vif	Features
8	4.894064	medianamnt_loans30
7	4.842441	medianamnt_loans90
1	3.631734	fr_ma_rech90
0	3.453946	fr_ma_rech30
8	2.131133	msisdn_encoded
2	1.912742	cnt_da_rech30
4	1.893472	cnt_da_rech90
5	1.503855	fr_da_rech90
3	1.266128	fr_da_rech30
9	NaN	pcircle_encoded

  - i.
- c. After 2 level anova test, final features, used in the best model are:

```
1 p_values
```

	Features	P Values
fr_ma_rech30	fr_ma_rech30	0.000000e+00
fr_ma_rech90	fr_ma_rech90	0.000000e+00
medianamnt_loans30	medianamnt_loans30	1.283229e-252
medianamnt_loans90	medianamnt_loans90	4.513084e-146
cnt_da_rech90	cnt_da_rech90	1.557350e-37
fr_da_rech90	fr_da_rech90	7.178386e-11
cnt_da_rech30	cnt_da_rech30	7.197254e-10
fr_da_rech30	fr_da_rech30	3.731212e-04

i.

- For visualizations based conclusion follow the below links to individual dashboards:
- Visualizations
-

10. Log Transformed and Cumulative Visualizations:

    Follow the link to the dashboard:

    https://docs.google.com/spreadsheets/d/1wEgp5xKLCykpivGxlDrialuBTrEhKa_AEeKH01wnXMY/edit?usp=sharing

11. Head, Sample, and Tail View Of Data

    Follow the link to the dashboard:

    https://docs.google.com/spreadsheets/d/1d1ixblgxC86MrFE2zjB3BepgPMtgIJuCtrhRenFrWQw/edit?usp=sharing

12. Groupby

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1OwcCvO9vY5b5UTBR8_r0-rsChG4K9pLL2Hxo1aev-4M/edit?usp=sharing

13. Scatter Plots

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1p_B4bUvKvDGGjXBbwzjADuIIphEWqnYApzoEfekEtj8/edit?usp=sharing

14. Histograms

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1v0qh6j2l5-ZO5i7Ya-T6QMnx_CF31fprDkTW4GvTAT8/edit?usp=sharing

15. Correlation

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1QnC77IXeSpTk__Rfz_WJlXwNaDyPHbWJhwFVpm4Q4HY/edit?usp=sharing

16. Box Plots

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/186CvJpOSw521aU54LZn4TYK0m-2o96ub98xZ_MTFjBw/edit?usp=sharing

17. Descriptive Statistics

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1_qmUX1jXKiEGZDP3symS1VlZKxOg_7nlhq8Kl7ZrXRI/edit?usp=sharing

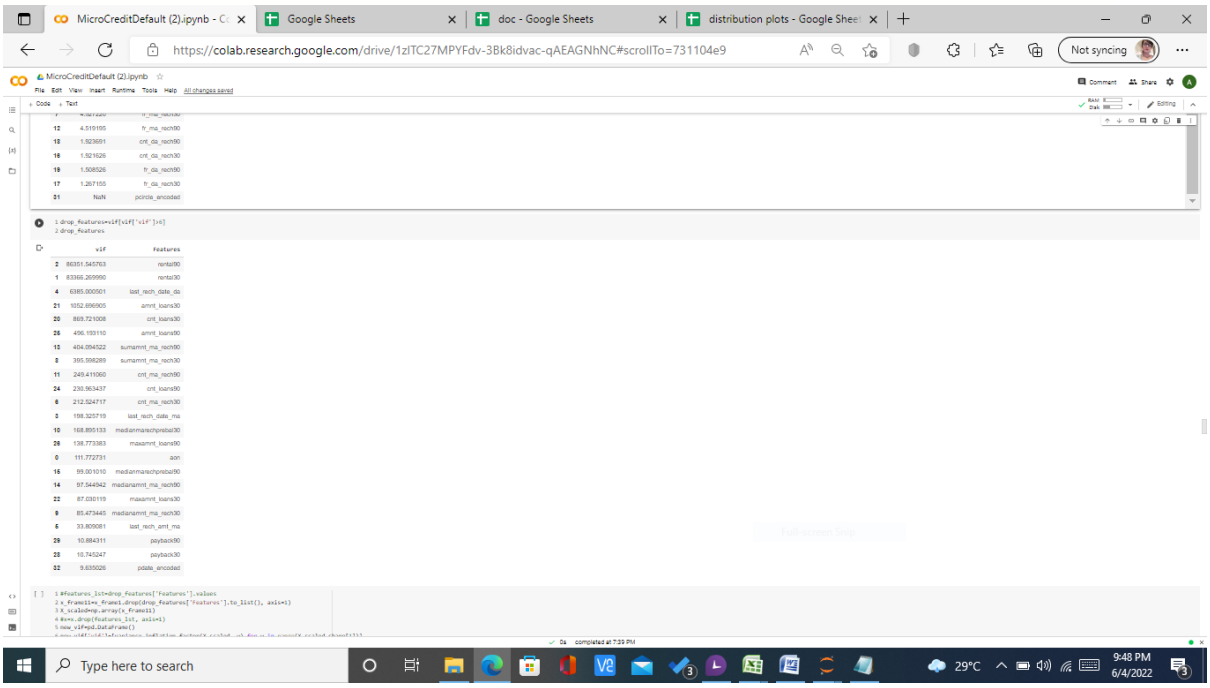
18. Distribution Plots

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/171asS_59gZxz8uhHiNHUUtlg4gj0dvLSZWllNb95u14/edit?usp=sharing

CONCLUSION

- Key Findings and Conclusions of the Study
- # Based on above analysis, these columns have significant outliers, hence, are platykurtic, hence, I am removing these columns.
-
- i. Daily Decr 30
-
- ii. Daily Decr 90
-
- I will do further vif test to remove multicollinearity from the data.
-



a. Based on VIF, I have dropped these features. I have done VIF after outliers transformation and feature scaling.

b. Final features passed for anova test include:

```
[ ]
#features_lst=drop_features['Features'].values
x_frame11=x_frame1.drop(drop_features['Features'].to_list(), axis=1)
X_scaled=np.array(x_frame11)
#x=x.drop(features_lst, axis=1)
new_vif=pd.DataFrame()
new_vif['vif']=[variance_inflation_factor(X_scaled, w) for w in range(X_scaled.shape[1])]
new_vif['Features']=x_frame11.columns
new_vif=new_vif.sort_values(ascending=False, by='vif')
new_vif
```

	vif	Features
8	4.894064	medianamnt_loans30
7	4.842441	medianamnt_loans90
1	3.631734	fr_ma_rech90
0	3.453946	fr_ma_rech30
8	2.131133	msisdn_encoded
2	1.912742	cnt_da_rech30
4	1.893472	cnt_da_rech90
6	1.503855	fr_da_rech90
3	1.266128	fr_da_rech30
9	NaN	pcircle_encoded

i.

c. After 2 level anova test, final features, used in the best model are:

```
1 p_values
```

	Features	P Values
fr_ma_rech30	fr_ma_rech30	0.000000e+00
fr_ma_rech90	fr_ma_rech90	0.000000e+00
medianamnt_loans30	medianamnt_loans30	1.283229e-252
medianamnt_loans90	medianamnt_loans90	4.513084e-146
cnt_da_rech90	cnt_da_rech90	1.557350e-37
fr_da_rech90	fr_da_rech90	7.178386e-11
cnt_da_rech30	cnt_da_rech30	7.197254e-10
fr_da_rech30	fr_da_rech30	3.731212e-04

i.

- For visualizations based conclusion follow the below links to individual dashboards:
- Visualizations
-

19. Log Transformed and Cumulative Visualizations:

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1wEgp5xKLCykpivGxlDrialuBTrEhKa_AEeKH01wnXMY/edit?usp=sharing

20. Head, Sample, and Tail View Of Data

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1d1ixblgxC86MrFE2zjB3BepgPMtgIJuCtrhRenFrWQw/edit?usp=sharing

21. Groupby

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1OwcCvO9vY5b5UTBR8_r0-rsChG4K9pLL2Hxo1aev-4M/edit?usp=sharing

22. Scatter Plots

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1p_B4bUvKvDGGjXBbwzjADuIIphEWqnYApzoEfekEtj8/edit?usp=sharing

23. Histograms

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1v0qh6j2l5-ZO5i7Ya-T6QMnx_CF31fprDkTW4GvTAT8/edit?usp=sharing

24. Correlation

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1QnC77IXeSpTk__Rfz_WJlXwNaDyPHbWJhwFVpm4Q4HY/edit?usp=sharing

25. Box Plots

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/186CvJpOSw521aU54LZn4TYK0m-2o96ub98xZ_MTFjBw/edit?usp=sharing

26. Descriptive Statistics

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/1_qmUX1jXKiEGZDP3symS1VlZKxOg_7nlhq8Kl7ZrXRI/edit?usp=sharing

27. Distribution Plots

Follow the link to the dashboard:

https://docs.google.com/spreadsheets/d/171asS_59gZxz8uhHiNHUUtlg4gj0dvLSZWllNb95u14/edit?usp=sharing

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

- Visualizations and data cleaning convert a whole complex and messy dataset into insightful and interesting representation, which make it easier to reach the core of the problem and solve it.
-
- The best model is RFC With Default Hyper Parameter tuning on resampled data, the most challenging part in models development process was to reduce overfitting and biasness in the data, that is why I have applied ensemble methods on base estimators... RFC provided the best framework to reduce overfitting. By running it on resampled data, I achieved an accuracy score of 0.74.

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

Further optimatization can be obtained by applying deep learning solutions. Since, it requires very high RAM capacity, it could not be displayed in jupyter notebook... I would like to update Google Colab Notebook for future projects, if acceptable... That can help me to submit a completely optimized model.