

Predictive model to classify profitable borrowers

Identify customers who will either fully pay their loans or their loans needs to be charged off

Prepared By: Souvik Ganguly,

Date: 20th May 2024



Index

1. Index
2. Document Management Control
3. Executive summary
4. Project roadmap
5. Data Quality Analysis and Data Quality report
6. Exploratory Data Analysis
 1. Part-1 Columns with Missing Values
 2. Part-2: Columns with no missing values
 3. Part-3: New Feature Creation
7. Segmentation
 1. Part-1 Data Preparation
 2. Part-2 Model Iterations
 3. Part-3 Cluster analysis formed with optimal cluster number=2, 3 and 4
 4. Part-4 Segmentation Conclusion
8. Next Steps to be continued..
9. Appendix

Document Management Control

ID	Classify Profitable Borrowers		
Document Title	Predictive model to classify profitable borrowers_v0.2.pptx		
Document Status	Draft <input type="checkbox"/>	Proposed <input type="checkbox"/>	Approved <input type="checkbox"/>

Version number	Issue Date	Prepared By	Modifications
V0.1	15 th May 2024	Souvik Ganguly	First Draft- DQ reports generated
V0.2	20 th May 2024	Souvik Ganguly	Implement customer segmentation and review comments

Reviewed By	Review Date
Model Owner	TBD
Model Sponsor	TBD
Business Owner	TBD

GitHub Repository: Click [here](#)

Executive Summary

- **Objective:**

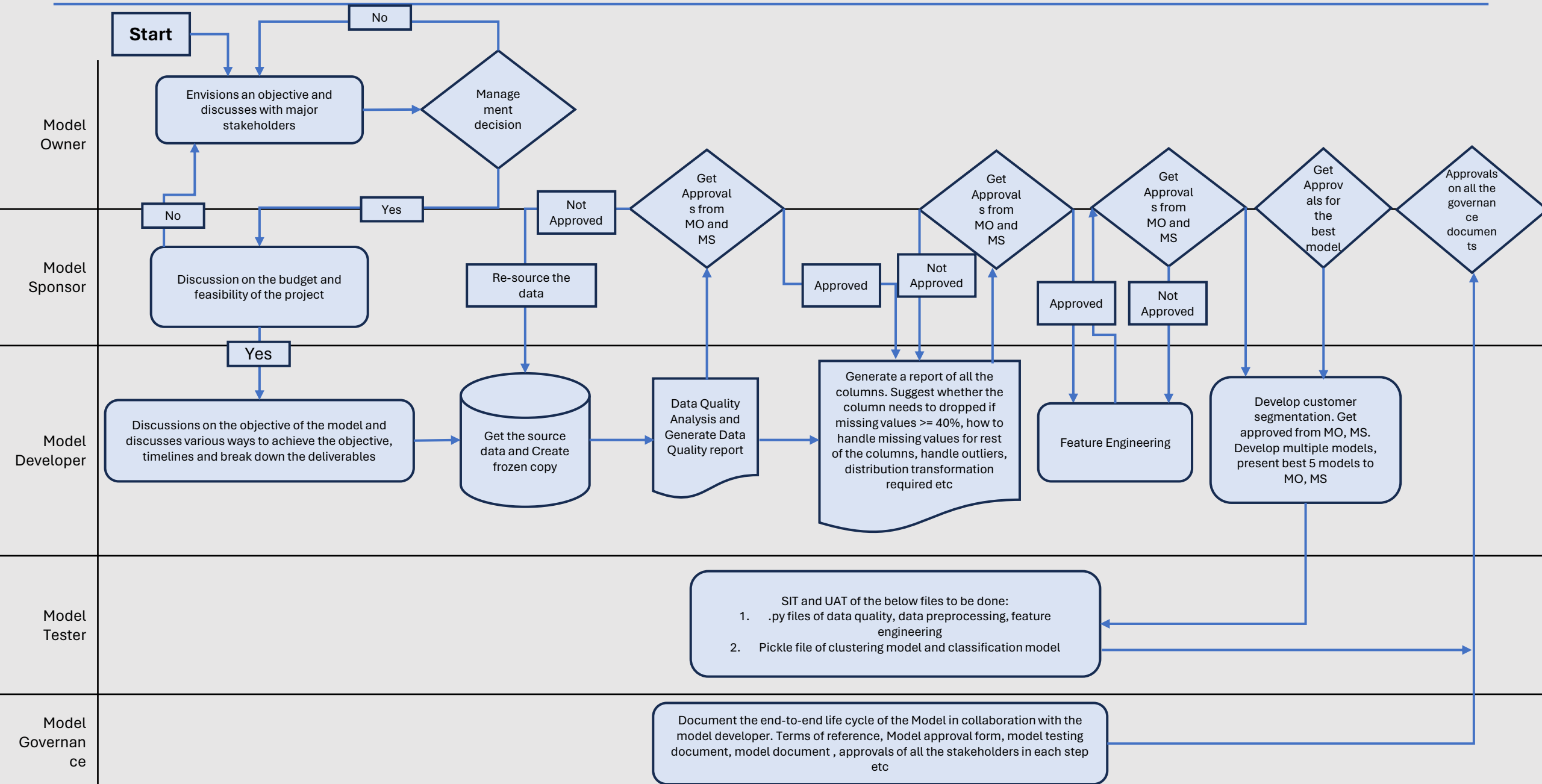
- The bank is facing significant financial losses due to loan defaults, which directly impacts profitability and risk management. To address this issue, the bank aims to develop a predictive model that can accurately identify borrowers who are likely to fully repay their loans versus those who are at risk of defaulting (Charged Off).
- By leveraging comprehensive data on loan applications, borrower demographics, and historical payment behavior, this project seeks to enhance the bank's loan approval process, minimize default rates, and optimize lending decisions. The goal is to improve the bank's overall financial health and customer satisfaction through data-driven insights and predictive analytics.
- Build a robust model that will predict whether the borrower will be profitable for the bank. We have a dataset that consists of:
 - Borrower's identification data
 - Borrower's loan data
 - Borrower's demographic data
 - Borrower's verification data
 - Data/ time data
 - Borrower's transaction data
 - Borrower's hardship data
 - Borrower's debt settlement data
 - Borrower's application data

- **Key Steps to achieve the objective:**

- Data Quality Analysis
- Exploratory data analysis: Handling Missing values and Outliers
- Feature engineering
- Customer Segmentation
- Model development for classification

- **Expected Outcome:** Improved accuracy in predicting loan defaults thus enabling better decision making for loan approvals. [Click here to visit the github repository of this project](#)

Project Roadmap



Data Quality Analysis and Data Quality report

- Data sourced from an open-source platform: [Hugging Face Datasets](#)
- Data Quality script executed on the sourced data.(Click [here](#) to find the Data quality scripts. Section: 2.0 Generate an exhaustive Data Quality Report)
- The script generates a Data Quality report with below data points:
 1. Missing value %
 2. Number of Unique values
 3. Data Types
 4. Descriptive statistical measures
 5. Type of distribution



- For the complete data quality report, click here [data_quality_report.xlsx](#)

- **Summary:**

Summary of Data Quality Report	
No. of numerical variables	113
No. of categorical variables	38
No. of variables with null values > 40%	58
No. of variables with null values > 10% and <=40%	1
No. of variables with null values > 0% and <=10%	46
No. of variables with no null values	46

- **Since there’s lack of stakeholder discussion here, assuming the stakeholder agreed on deleting the columns with > 40% missing values due to lack of data and lack of relevance of the columns. 151 columns reduced to 93 columns**

Exploratory Data Analysis and Feature Engineering

Part-1 Columns with Missing Values

- An initial report is extracted using the dataset which has:
 1. Descriptive statistical data for Numerical Variables: count, mean, median, standard deviation, percentiles(0, 25, 50, 75, 100), min, max and missing values %.
 2. Value data for categorical variables: count of each category, % of fully paid and charged off loans for each categories etc.

Click [here](#) to find the script generating the above output.
- Based on the above report, analysis of each feature with missing values in the dataset is conducted. Click [here](#) to find the script of Exploratory data analysis.
- A summary report is created to address:
 - Columns to be dropped
 - Missing values treatment techniques
 - Outlier treatment techniques
 - Distribution transformation techniques
 - Possible new variables using existing variables



For the complete analysis report on columns with missing values, click [here](#)

Data Cleaning
feature engineerin

- **Summary:**

Top-view Summary of Data Quality Report	
No. of variables under scope	47
No. of categorical variables	12
No. of numerical variables	35
No. of columns to be dropped. <i>(Variable list to be dropped in appendix)</i>	34
No. of variables where outlier imputation is required	0
No. of variables where box_cox transformation is required	11
No of new variables to be created	5

Part-2 Columns with no Missing Values

- An initial report is extracted using the dataset which has:
 1. Descriptive statistical data for Numerical Variables: count, mean, median, standard deviation, percentiles(0, 25, 50, 75, 100), min, max and missing values %.
 2. Value data for categorical variables: count of each category, % of fully paid and charged off loans for each categories etc.

Click [here](#) to find the script generating the above output.
- Based on the above report, analysis of each feature with no missing values in the dataset is conducted. Click [here](#) to find the script of Exploratory data analysis.
- A summary report is created to address:
 - Columns to be dropped
 - Missing values treatment techniques
 - Outlier treatment techniques
 - Distribution treatment techniques
 - Possible new variables using existing variables

For the complete analysis report on columns with no missing values, click [here](#)



Data Cleaning
feature engineerin

- **Summary:**

Top-view Summary of Data Quality Report	
No. of variables under scope	46
No. of categorical variables	18
No. of numerical variables	28
No. of columns to be dropped. <i>(Variable list to be dropped in appendix)</i>	29
No. of variables where outlier imputation is required	0
No. of variables where box_cox transformation is required	8
No of new variables to be created	9

Part-3 New Feature creation

- **Summary:** 15 new features have been created from existing variables. Click [here](#) to find the script of Data Preprocessing and Feature Engineering.

Sr No	New Feature	Old feature	Feature creation technique	Old feature dropped?
1	int_rate_category	int_rate	Binning with a gap of 5%	No
2	installment_category	installment	% based categorization	No
3	annual_inc_category	annual_inc	% based categorization	No
4	delinq_2yrs_category	delinq_2yrs	'Y' if value>=1 else 'N'	Yes
5	revol_bal_category	revol_bal	% based categorization	No
6	perc_loan_paid	total_rec_prncp	100* (total_rec_prncp/ loan_amount)	Yes
7	recoveries_category	Recoveries	'Y' if value is not Null else 'N'	Yes
8	last_fico_range_high_category	last_fico_range_high	Industry standard categorization	No
9	last_fico_range_low_category	last_fico_range_low	Industry standard categorization	No
10	inq_last_6mths_binary	inq_last_6mths	'Y' if value is not Null else 'N'	Yes
11	tot_cur_bal_category	tot_cur_bal	% based categorization	No
12	avg_cur_bal_category	avg_cur_bal	% based categorization	No
13	total_bal_ex_mort_category	total_bal_ex_mort	% based categorization	No
14	total_bc_limit_category	total_bc_limit	% based categorization	No
15	total_il_high_credit_limit_category	total_il_high_credit_limit	% based categorization	No

- **Post EDA and Feature Engineering, remaining 93 columns has reduced to 44 columns.**

Segmentation

Part-1 Data Preparation

- Dataset of 44 features including both numerical and categorical variables
- Data Preprocessing applied to these features:
 1. Creating dummy variables for categorical features
 2. Scaling the numerical variables using StandardScalar
 3. Feature elimination using variance threshold method
 4. Feature elimination using correlation matrix

Data Preprocessing increased 44 features to 64

- **Hopkins Statistical test:** It calculates the Hopkins statistic to evaluate the cluster tendency of the dataset. It prints the Hopkins statistic value.
 1. Null Hypothesis: Dataset is uniformly distributed. Hence no meaningful clusters. Average $H \leq 0.85$
 2. Alternate Hypothesis: Dataset is not uniformly distributed. Hence It contains meaningful clusters. Average $H > 0.85$

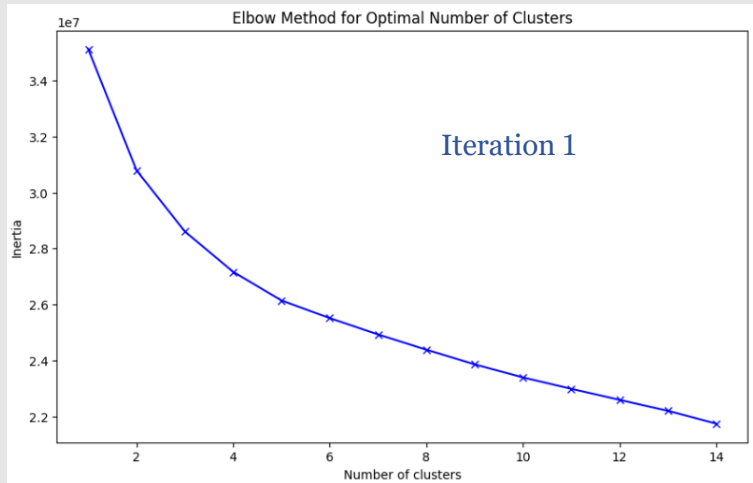
Summary:

Test #	Hopkins statistics
1	0.9648238978692573
2	0.9667271706377367
3	0.9651389560040534
4	0.965542978391139
5	0.9646249880857238

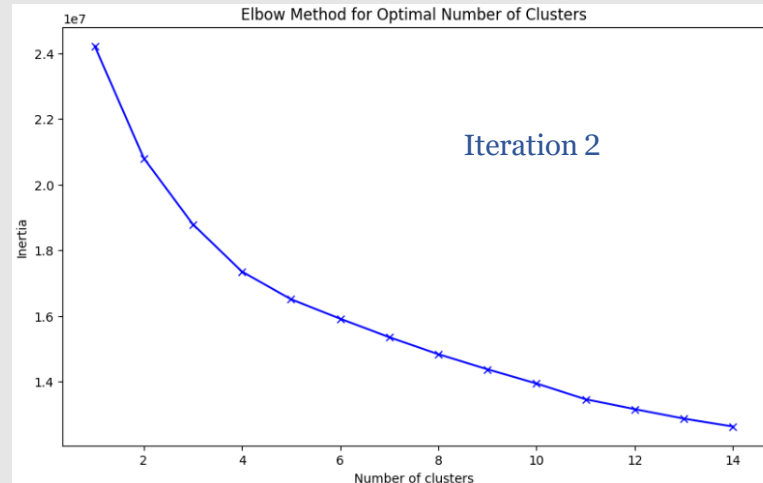
- In this case Alternate hypothesis is true meaning data has meaningful clusters. $H > 0.85$
- Click [here](#) to find the script of Data Preparation and performing Hopkins Statistical test (*Section 2 of the notebook*)

Part-2 Model Iterations

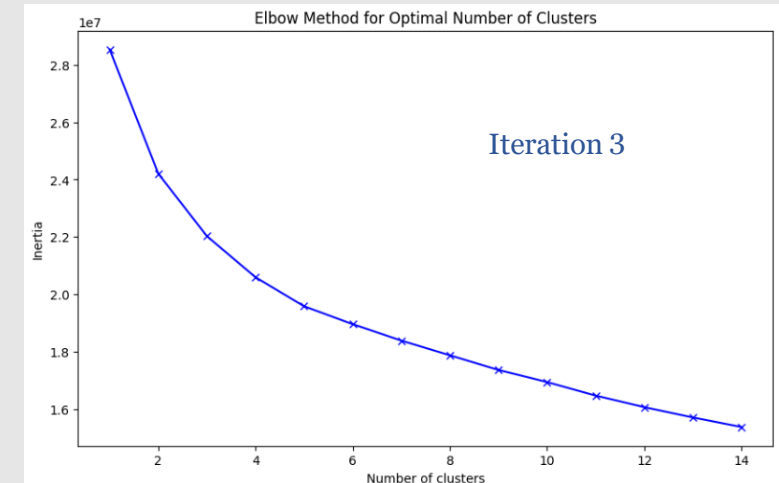
- Model is iterated amongst 3 different combinations of data
 - Preprocessed data in the last step with 64 features
 - Preprocessed numerical data with 18 variables: All the dummy variables, categorical variables are dropped.
 - Principal Component Analysis: Reduced features with cumulative_explained_variance ≥ 0.8 which resulted in 16 features
- Elbow curve analysis of all 3 iterations resulted in optimal number of clusters=3. Click [here](#) to find the script of Elbow curve analysis and Silhouette Analysis for all 3 iterations (Section 3 of the notebook)



# clusters	Silhouette Score
2	0.122
3	0.112
4	0.081



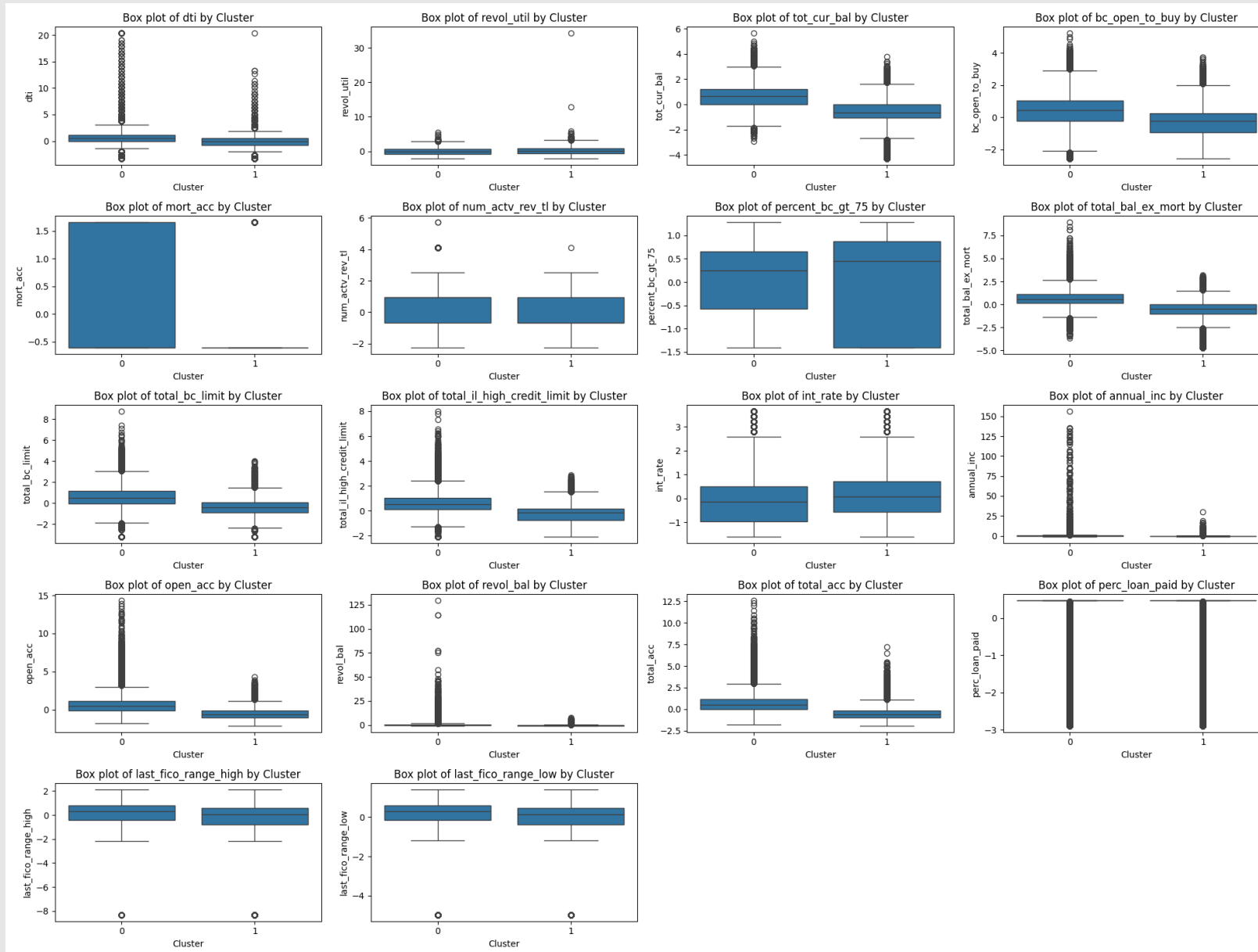
# clusters	Silhouette Score
2	0.140
3	0.140
4	0.116



# clusters	Silhouette Score
2	0.150
3	0.141
4	0.109

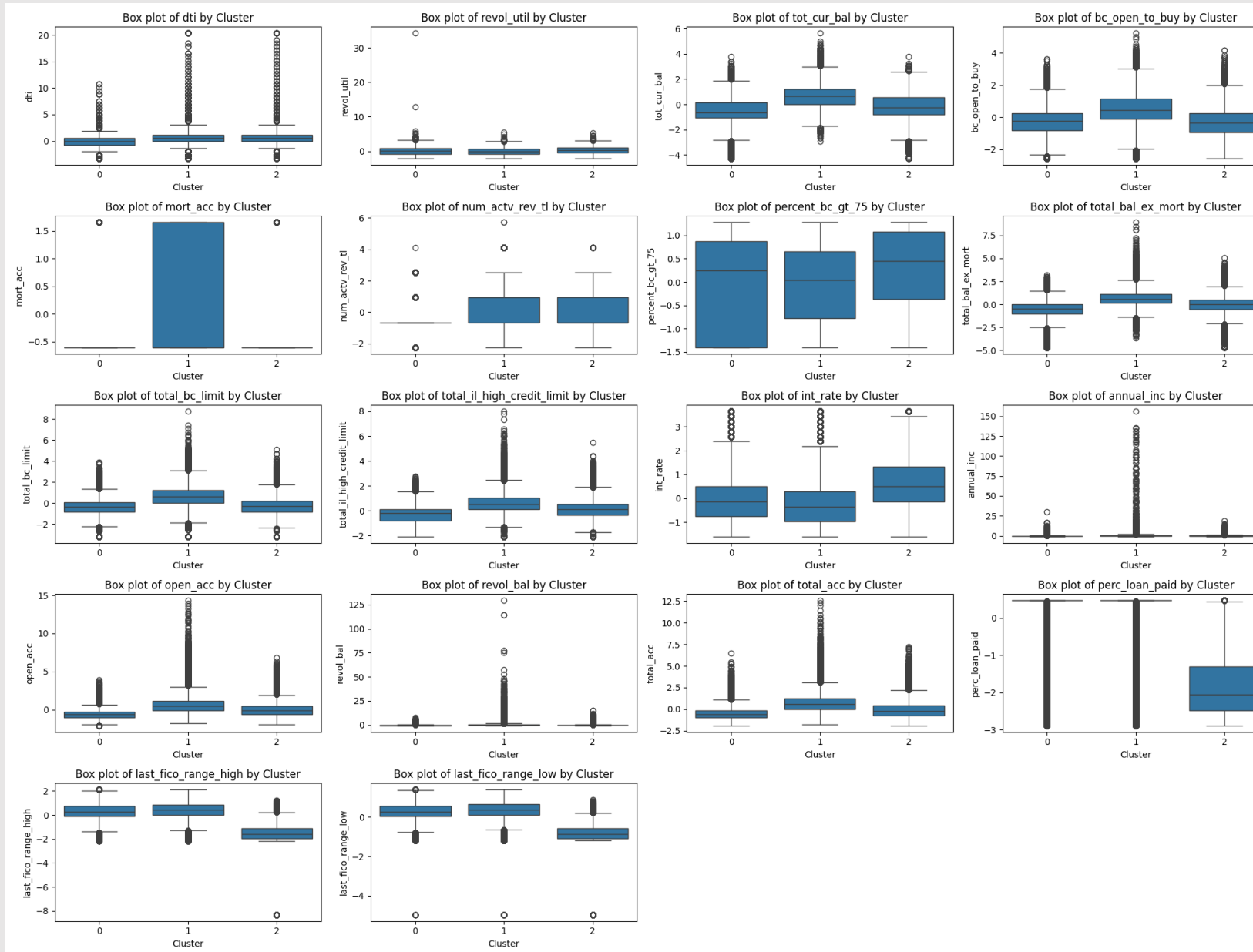
- Based on the statistical results, clustering with reduced dimensions will give the best result. But assuming that the regulators will resist the PCA transformation since there's no interpretability of the clusters w.r.t the variables, I'm going ahead with the clustering approach using 2nd iteration(Preprocessed numerical data with 18 variables)

Part-3 Cluster analysis formed with optimal cluster number=2



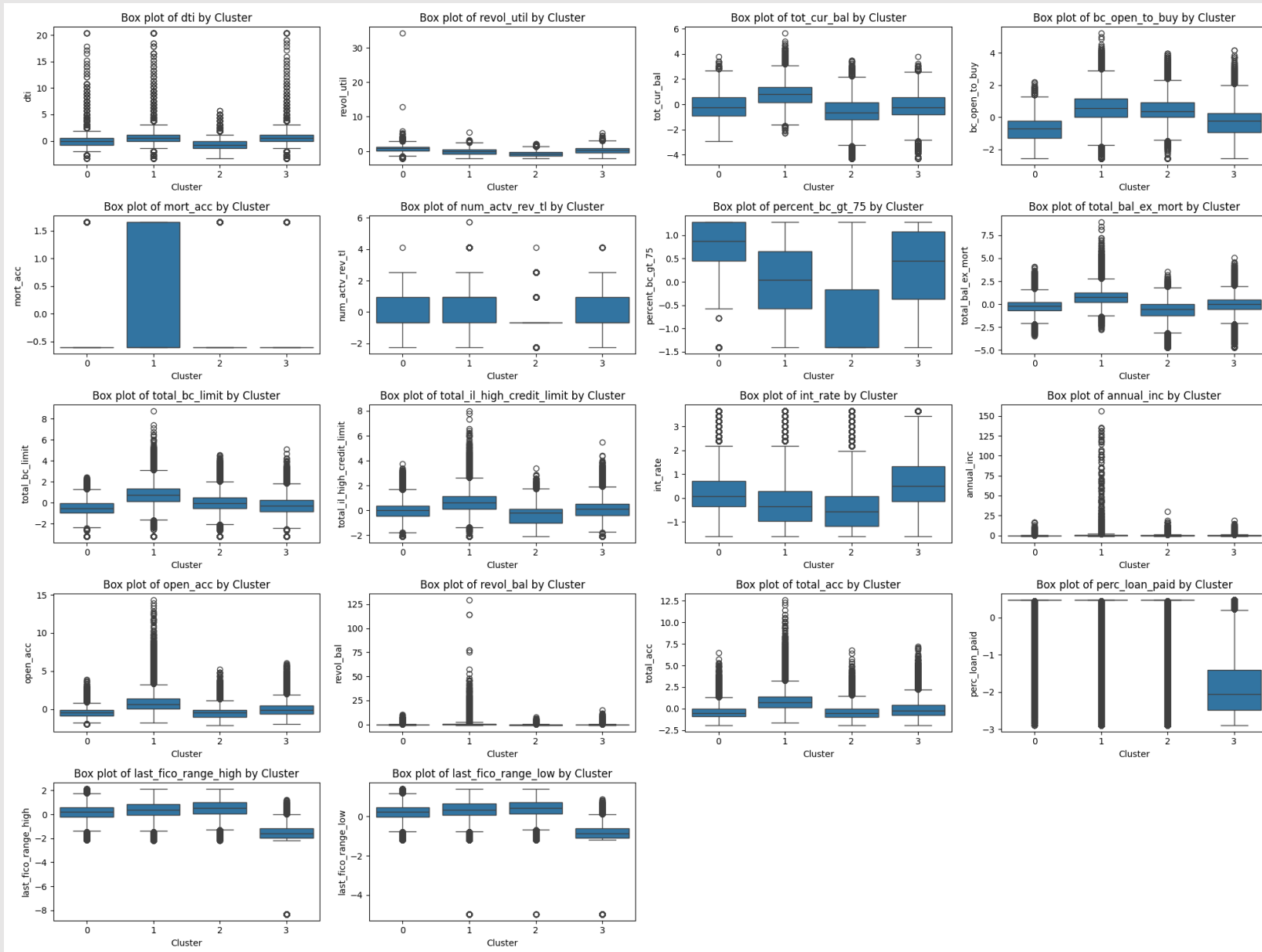
Click [here](#) to find the script of cluster analysis for all k=2, 3 and 4 for the 2nd iteration (Section 3.5 of the notebook)

Part-3 Cluster analysis formed with optimal cluster number=3



Click [here](#) to find the script of cluster analysis for all k=2, 3 and 4 for the 2nd iteration (Section 3.5 of the notebook)

Part-3 Cluster analysis formed with optimal cluster number=4



Click [here](#) to find the script of cluster analysis for all k=2, 3 and 4 for the 2nd iteration (Section 3.5 of the notebook)

Part-4 Segmentation Conclusion

- The Silhouette score for iteration 2 with optimal number of clusters= 2 and 3 are the best
- However, the intra-cluster heterogeneity(*based on last 3 slides*) between the clusters is missing with optimal number of clusters = 2. The variable distributions for most of the features are similar.
- There is required intra-cluster heterogeneity between the clusters with optimal number of clusters = 3 and 4 (*More in 4 than 3*)
- Loan status (*target variable*) distribution within clusters

Clusters	Fully Paid	Charged Off
0	578685	17665
1	473018	28860
2	25048	222034

Clusters	Fully Paid	Charged Off
0	374005	6888
1	352265	27746
2	331463	13293
3	19018	220632

- **Given that, optimal number of clusters=3 has the best Silhouette score and relatively good intra-cluster heterogeneity, our suggestion is to segment the borrowers into 3 clusters.**
- **Review Pending.**

Next Steps to be continued..

1. Classification based on features decided
2. Model Monitoring framework

THANK YOU


Contact me at

[Linkedin](#)

E-mail: Souvik.ganguly.ds@gmail.com

Phone no: (+91) 8141786094

Appendix

Sr No	Description	Attachment
1	List of columns dropped during feature engineering	 List_of_columns_dropped.xlsx