Heri Purwadi

Should This Loan be Approved or Denied?

Business Understanding

- Small businesses have been a primary source of job creation in the United States
- Fostering small business formation and growth has social benefits by creating job opportunities and reducing unemployment.
- They need loan for business improvement and expansion
- They could be successes or defaulted on their SBA-guaranteed loans

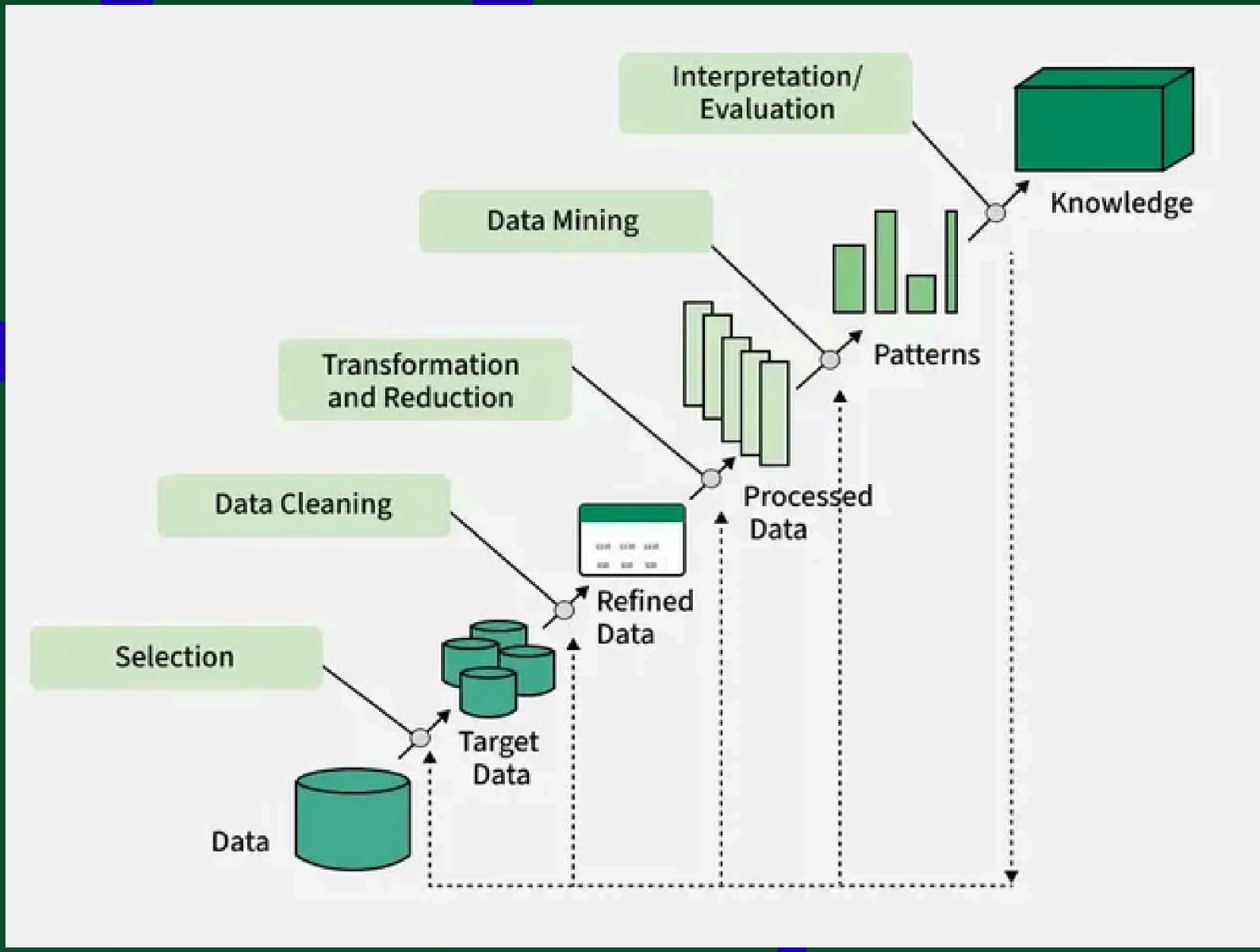
What solution could be offered?

Developing machine learning model to indicate which businesses would be defaulted on their loans.

KDD Framework

Solve the problem using the Knowledge Discovery in Databases (KDD) process:

1. Data Selection
2. Data Cleaning and Preprocessing
3. Data Transformation and Reduction
4. Data Mining
5. Evaluation and Interpretation of Results



picture from https://www.geeksforgeeks.org/dbms/kdd-process-in-data-mining/

Data Selection

- Downloaded from https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied
- 27 columns
- 670,119 rows

Variables

(NOT Used)

- LoanNr_ChkDgt
- Name
- City
- State
- Zip
- Bank
- BankState
- NAICS
- ApprovalDate
- ApprovalFY
- ChgOffDate
- DisbursementDate
- BalanceGross
- ChgOffPrinGr
- GrAppv
- SBA_Appv

Covariates

- Term
- NoEmp
- NewExist
- CreateJob
- RetainedJob
- FranchiseCode
- UrbanRural
- RevLineCr
- LowDoc
- DisbursementGross

Response

- MIS_Status:
  - P I F (0)
  - CHGOFF (1)

Data Cleaning and Preprocessing

- Handling Missing Data:
  - Using Mode for NewExist, RevLineCr, Low Doc.
  - Using values in ChgOffPrinGr For MIS_Status
- Handling Outliers:
  - Keep Outliers based on their values
- Handling Duplicate:
  - No duplicate

	0
Term	0
NoEmp	0
NewExist	136
CreateJob	0
RetainedJob	0
FranchiseCode	0
UrbanRural	0
RevLineCr	4528
LowDoc	2582
DisbursementGross	0
BalanceGross	0
MIS_Status	1997
ChgOffPrinGr	0

Data Transformation and Reduction

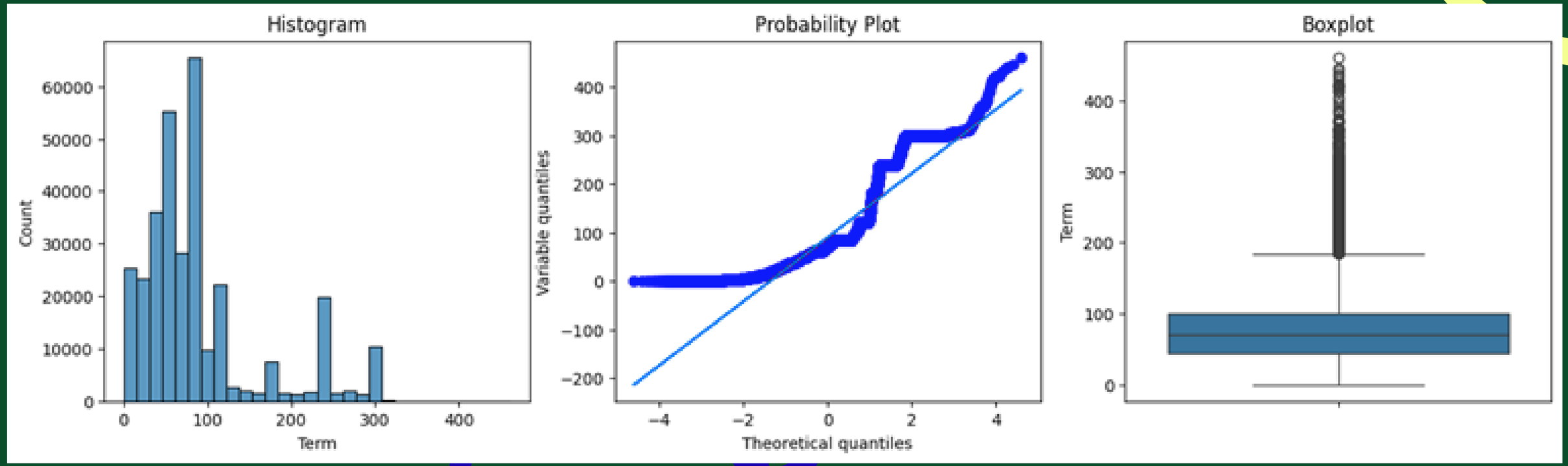
- Imbalance Handling:
  - Undersampling, considering the number of samples
  - Resulting 159,555 for each status
- Label Encoding:
  - NewExist
  - FranchiseCode
  - RevLineCr
  - LowDoc
  - MIS_Status
- One Hot Encoding:
  - UrbanRural
- Feature Engineeing:
  - Standard Scaler for:
    - Term
    - NoEmp
    - CreateJob
    - DisbursementGross

Data Mining

- Exploratory Data Analysis:
  - Histogram
  - QQ/ Probability Plot
  - Boxplot
  - Heatmap Correlation
- Data Modeling
  - Logistic Regression
  - Bernoulli Naive Bayes
  - K-Nearest Neighbors (KNN)
  - Decision Tree
  - Random Forest

Exploratory Data Analysis

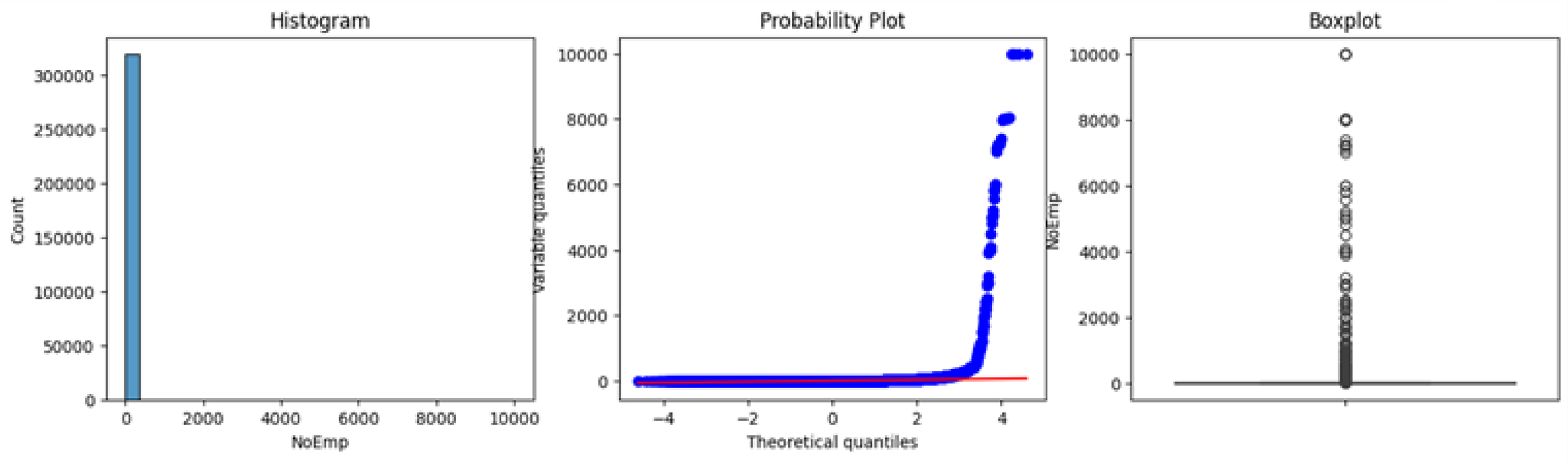
Term



Data is not normal (skew to the left), with so many outliers. However, the outliers is kept because of its relevant values.

Exploratory Data Analysis

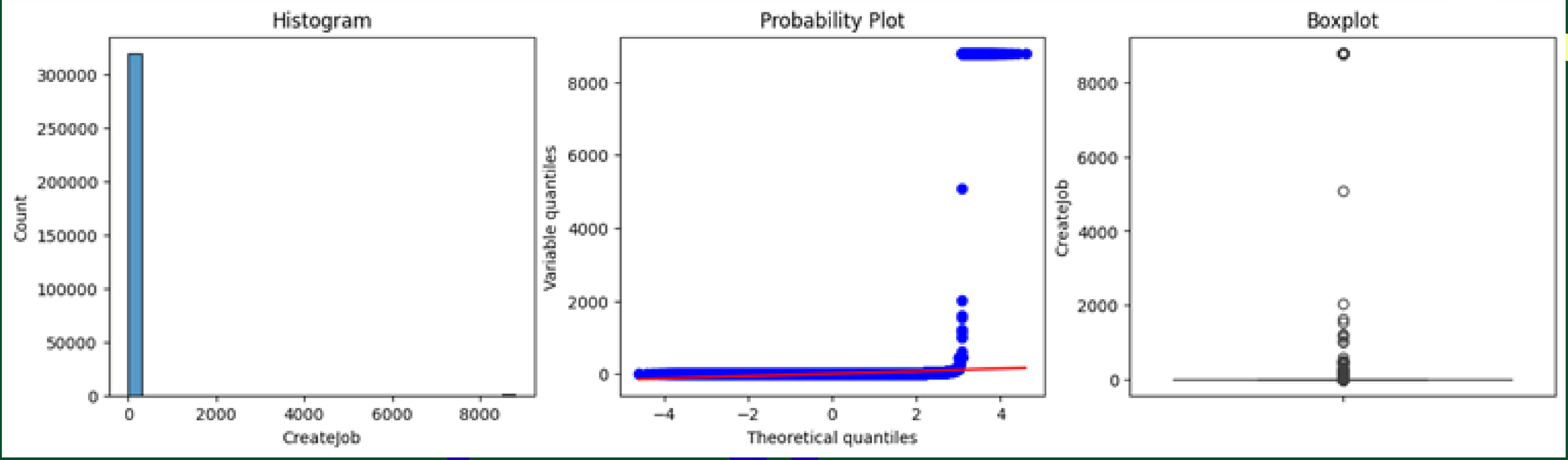
NoEmp



Data is not normal (look like one bar chart only) and the boxplot is truncated with so many outliers. We still keep the outliers here as the value is still make sense.

Exploratory Data Analysis

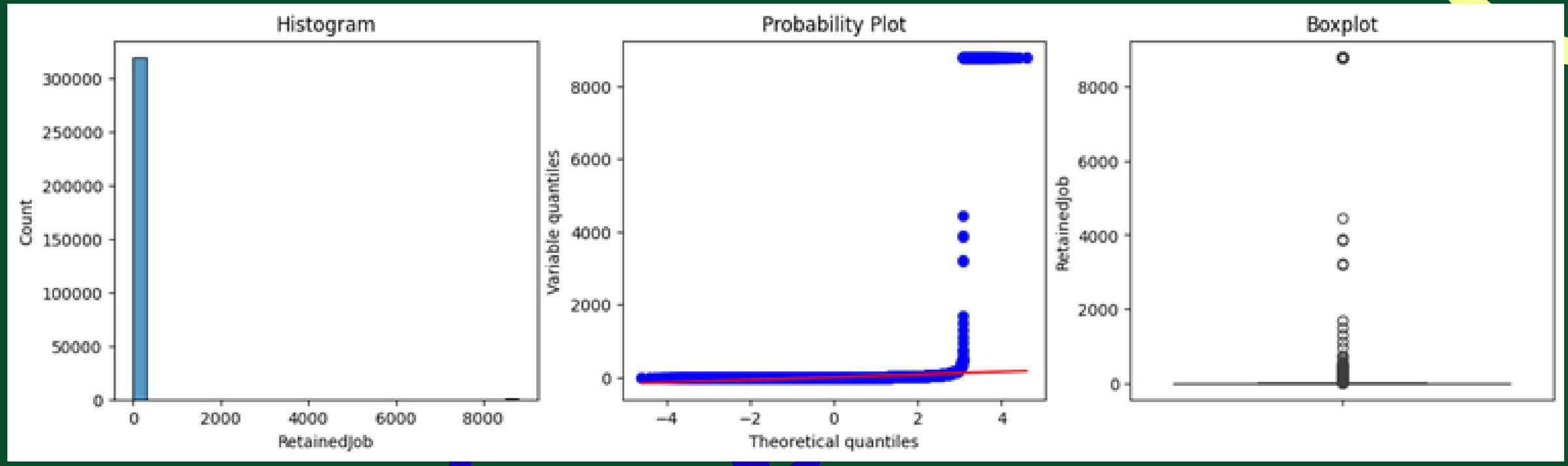
CreateJob



Data is also not normal with truncated boxplot. For the outliers, we keep them.

Exploratory Data Analysis

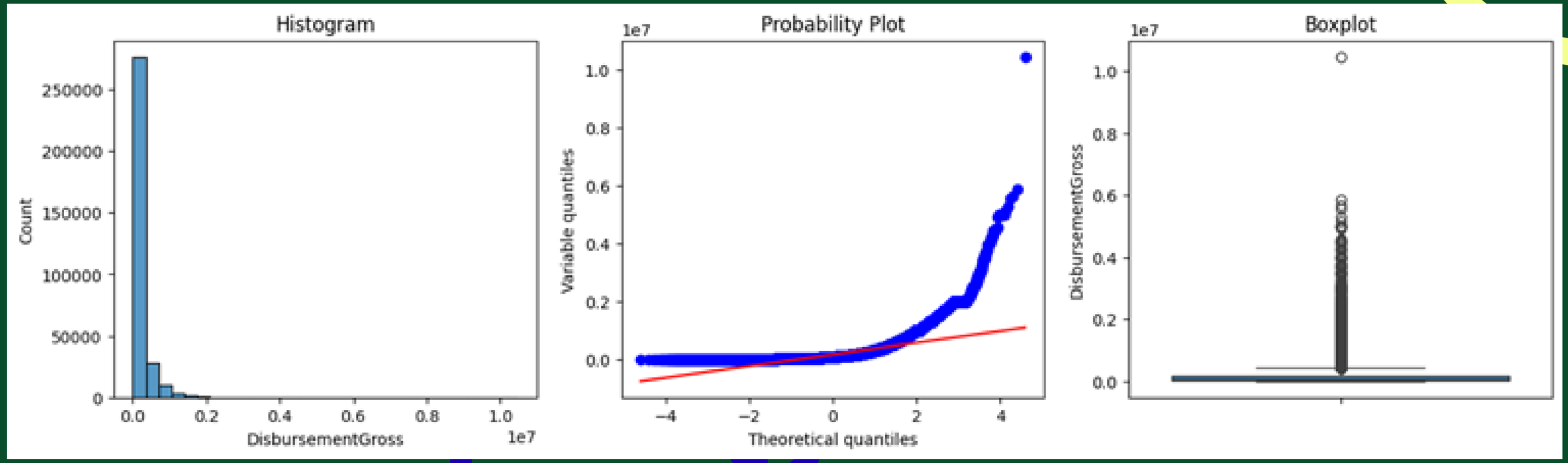
RetainedJob



Almost identical figures with the previous page.

Exploratory Data Analysis

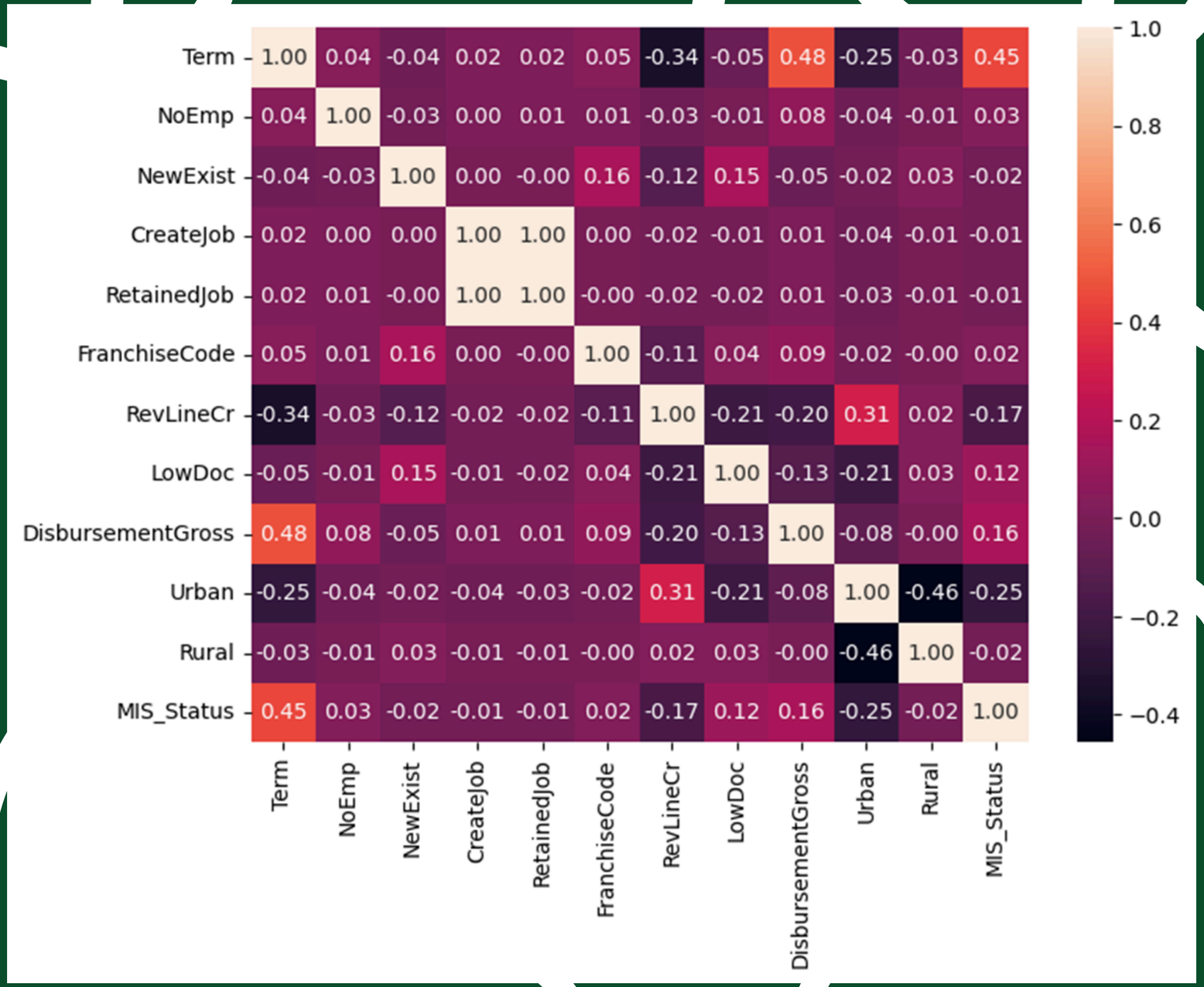
DisbursementGross



Also not normal but we still need the outlier values.

Heatmap

- No Significant Correlation, except for CreatedJob and RetainedJob
- Drop RetainedJob



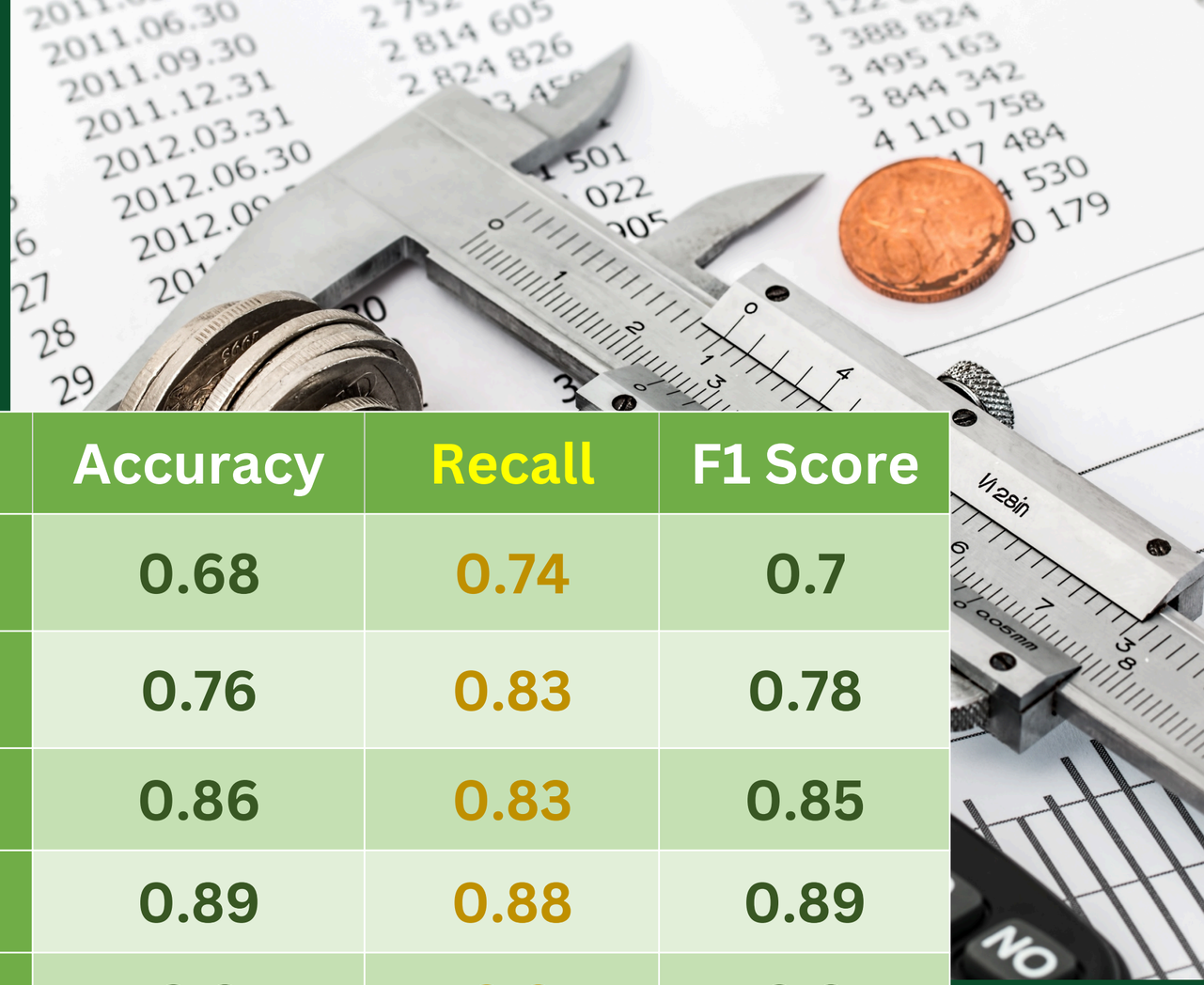
Modeling

- Logistic Regression
  - A statistical approach
- Bernoulli Naive Bayes
  - Based on Bayes theorm
- K-Nearest Neighbours
  - Considering closest neighbours
- Decision Tree
  - Tree-Like model of decisions
- Random Forest
  - Multiple Decision trees

Preferred Metric: Recall

- Recall is the ratio of true positive values divided by the sum of true positive and false negative values from the prediction.
- As we need to minimize the number of defaulted business, we prefer recall than other metrics.
- Accuracy and F1 Score is not use for choosing the best model but we show it for comparison.

Results



Model	Accuracy	Recall	F1 Score
Bernoulli Naïve Bayes	0.68	0.74	0.7
Logistic Regression	0.76	0.83	0.78
KNN	0.86	0.83	0.85
Random Forest	0.89	0.88	0.89
Decision Tree	0.9	0.9	0.9
Decision Tree (Model Checking)	0.9	0.91	0.9

The best model (Decision Tree) is checked using its train data and obtain slightly better result in recall. This show that the model is not overfitting.

Evaluation

- Most data are not normal, so using parametric statistical approach would not be useful
- Outliers cannot be deleted as it values are still reasonable
- Correlations of existed variables are not significant
- Decision Tree is the best model in this simulation possibly because 6 of 10 covariates are binary

Interpretation of Results

- Use the Decision Tree model for the new data to minimize the risk of defaulted loans
- Deploy the model for its application
- Compare the model using oversampling technique or by adding covariates
- For future modeling, consider spatial statistics approach by including state in the variables
- Remember, all model are wrong but some are useful. So, improving the model accuracy is the key.

Thank You