MACHINE LEARNING WORKFLOW

George

2022-02-28

output: officedown::rdocx\_document: mapstyles: Normal: [‘First Paragraph’]

Table of content

List of figures

List of tables

# Introduction.

The goal of this analysis is to predict with precision the probability of customer churn and consequently help in customer retention and mitigation of the risks related to customer churn.

## Business Value

Below are some the business values that this analysis will create:

1. Reduce churn rate by detecting early clients who are about to churn.
2. Increase customer retention.
3. Understand customer base better based on their activity.

## Dataset

The data set used in this project was sourced from client transaction data recorded since the inception of the financial institution.Based in this data set,informative feature sets were created which were subsequently used in the predictive model.

The clients records included in this study were selected from those who were determined to have done at least one transaction within the banking platform.

# Dataset Overview

## Data Quality

### Dataset Dimensions.

## Warning: The `x` argument of `as\_tibble.matrix()` must have unique column names if `.name\_repair` is omitted as of tibble 2.0.0.  
## Using compatibility `.name\_repair`.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

| Variable | Values |
| --- | --- |
| rows | 12512 |
| columns | 91 |
| discrete\_columns | 8 |
| continuous\_columns | 83 |
| all\_missing\_columns | 0 |
| total\_missing\_values | 120456 |
| complete\_rows | 2820 |
| total\_observations | 1138592 |

The dataset under study had 12512 records and 91 variables inclusive of target variable.

### Variable Explanations.

| Variable | Explanation |
| --- | --- |
| ClientID | Unique Client Identifier |
| CreatedOn | Date when the Account Was created |
| No\_ofProducts | Number of bank products the client is subscribed to |
| Tenure | The length of time a client has stayed with the institution since onboarding in days |
| GenderID | Clients Gender |
| FirstTrxDate | The first time the client made a transaction after onboarding |
| LastTrxDate | The last time the client made a transaction after onboarding |
| TrxTenure | Length of time in days between the first and the last transaction |
| Lag\_Debit | Difference between the current withdrawal and the preceeding withdrawal |
| Lag\_Credit | Difference between the current deposit and the preceeding deposit |
| No\_Trx | Number of transactions the client had done with the bank |
| WeekendTrx | Number of transactions done on weekends |
| EndMonthTrx | Number of transactions done at the end of the month |
| MidMonthTrx | Number of transactions done at the mod of the month |
| TransactedProducts | Number of products the client has actualy transacted after subscription |
| Transactions\_per\_week | Number of transactions a client made on a weeekly basis |
| Transactions\_per\_month | Number of transactions a client made on a monthly basis |
| Transactions\_per\_year | Number of transactions a client made on a yearly basis |
| Transactions\_per\_day | Number of transactions a client made on a daily basis |
| DepositTrx | Number of deposit transactions |
| DepositTrx\_per\_day | Number of deposit transactions made on a daily basis |
| DepositTrx\_per\_week | Number of deposit transactions made on a weekly basis |
| DepositTrx\_per\_month | Number of deposit transactions made on a monthly basis |
| DepositTrx\_per\_year | Number of deposit transactions made on a yearly basis |
| Total\_Deposits | Sum of deposit amount made |
| Total\_Deposits\_per\_day | Sum of deposit amount saved on a daily basis |
| Total\_Deposits\_per\_week | Sum of deposit amount saved on a weekly basis |
| Total\_Deposits\_per\_month | Sum of deposit amount saved on a monthly basis |
| Total\_Deposits\_per\_year | Sum of deposit amount saved on a yearly basis |

### Quality

| variables | types | missing\_count | missing\_percent | unique\_count | unique\_rate |
| --- | --- | --- | --- | --- | --- |
| HasActiveLoans | character | 9667 | 77.26 | 3 | 0.00 |
| Max\_LoansTaken | numeric | 9667 | 77.26 | 30 | 0.00 |
| Max\_LoanTaken | numeric | 9667 | 77.26 | 280 | 0.02 |
| Min\_LoanTaken | numeric | 9667 | 77.26 | 269 | 0.02 |
| Ave\_LoanTaken | numeric | 9667 | 77.26 | 1092 | 0.09 |
| First\_LoanTaken | numeric | 9667 | 77.26 | 245 | 0.02 |
| Last\_LoanTaken | numeric | 9667 | 77.26 | 323 | 0.03 |
| sd\_LoanTaken | numeric | 9667 | 77.26 | 966 | 0.08 |
| var\_LoanTaken | numeric | 9667 | 77.26 | 966 | 0.08 |
| Lag\_Debit | numeric | 1671 | 13.36 | 1765 | 0.14 |
| Lag\_Credit | numeric | 1671 | 13.36 | 979 | 0.08 |
| Var\_Deposits | numeric | 1671 | 13.36 | 7306 | 0.58 |
| sd\_Deposits | numeric | 1671 | 13.36 | 7306 | 0.58 |
| Var\_Withdrawals | numeric | 1671 | 13.36 | 7091 | 0.57 |
| sd\_Withdrawals | numeric | 1671 | 13.36 | 7091 | 0.57 |
| Var\_fixed\_dep | numeric | 1671 | 13.36 | 354 | 0.03 |
| sd\_fixed\_dep | numeric | 1671 | 13.36 | 354 | 0.03 |
| Var\_Deposits\_3 | numeric | 1671 | 13.36 | 2910 | 0.23 |
| sd\_Deposits\_3 | numeric | 1671 | 13.36 | 2910 | 0.23 |
| Var\_Withdrawals\_3 | numeric | 1671 | 13.36 | 2651 | 0.21 |
| sd\_Withdrawals\_3 | numeric | 1671 | 13.36 | 2651 | 0.21 |
| Var\_Deposits\_6 | numeric | 1671 | 13.36 | 3877 | 0.31 |
| sd\_Deposits\_6 | numeric | 1671 | 13.36 | 3876 | 0.31 |
| Var\_Withdrawals\_6 | numeric | 1671 | 13.36 | 3690 | 0.29 |
| sd\_Withdrawals\_6 | numeric | 1671 | 13.36 | 3690 | 0.29 |
| Var\_Deposits\_12 | numeric | 1671 | 13.36 | 5140 | 0.41 |
| sd\_Deposits\_12 | numeric | 1671 | 13.36 | 5140 | 0.41 |
| Var\_Withdrawals\_12 | numeric | 1671 | 13.36 | 4925 | 0.39 |
| sd\_Withdrawals\_12 | numeric | 1671 | 13.36 | 4925 | 0.39 |
| GenderID | character | 33 | 0.26 | 3 | 0.00 |
| ClientID | character | 0 | 0.00 | 12512 | 1.00 |
| CreatedOn | Date | 0 | 0.00 | 731 | 0.06 |
| No\_ofProducts | integer | 0 | 0.00 | 6 | 0.00 |
| Tenure | numeric | 0 | 0.00 | 731 | 0.06 |
| FirstTrxDate | Date | 0 | 0.00 | 773 | 0.06 |
| LastTrxDate | Date | 0 | 0.00 | 761 | 0.06 |
| TrxTenure | numeric | 0 | 0.00 | 742 | 0.06 |
| No\_Trx | integer | 0 | 0.00 | 262 | 0.02 |
| WeekendTrx | integer | 0 | 0.00 | 108 | 0.01 |
| WeekdayTrx | integer | 0 | 0.00 | 215 | 0.02 |
| EndMonthTrxs | integer | 0 | 0.00 | 135 | 0.01 |
| MidMonthTrxs | integer | 0 | 0.00 | 212 | 0.02 |
| TransactedProducts | integer | 0 | 0.00 | 6 | 0.00 |
| Transactions\_per\_day | numeric | 0 | 0.00 | 4462 | 0.36 |
| Transactions\_per\_week | numeric | 0 | 0.00 | 4644 | 0.37 |
| Transactions\_per\_month | numeric | 0 | 0.00 | 4612 | 0.37 |
| Transactions\_per\_year | numeric | 0 | 0.00 | 4660 | 0.37 |
| DepositTrx | numeric | 0 | 0.00 | 95 | 0.01 |
| DepositTrx\_per\_day | numeric | 0 | 0.00 | 1961 | 0.16 |
| DepositTrx\_per\_week | numeric | 0 | 0.00 | 2034 | 0.16 |
| DepositTrx\_per\_month | numeric | 0 | 0.00 | 2011 | 0.16 |
| DepositTrx\_per\_year | numeric | 0 | 0.00 | 2042 | 0.16 |
| Total\_Deposits | numeric | 0 | 0.00 | 5555 | 0.44 |
| Total\_Deposits\_per\_day | numeric | 0 | 0.00 | 7515 | 0.60 |
| Total\_Deposits\_per\_week | numeric | 0 | 0.00 | 7609 | 0.61 |
| Total\_Deposits\_per\_month | numeric | 0 | 0.00 | 7574 | 0.61 |
| Total\_Deposits\_per\_year | numeric | 0 | 0.00 | 7723 | 0.62 |
| Max\_Deposits | numeric | 0 | 0.00 | 1167 | 0.09 |
| Av\_Deposits | numeric | 0 | 0.00 | 6303 | 0.50 |
| Total\_Withdrawals | numeric | 0 | 0.00 | 5011 | 0.40 |
| Max\_Withdrawals | numeric | 0 | 0.00 | 1257 | 0.10 |
| Av\_Withdrawals | numeric | 0 | 0.00 | 5898 | 0.47 |
| total\_fixed\_dep | numeric | 0 | 0.00 | 146 | 0.01 |
| max\_fixed\_dep | numeric | 0 | 0.00 | 115 | 0.01 |
| Av\_fixed\_dep | numeric | 0 | 0.00 | 318 | 0.03 |
| No\_Trx\_3 | numeric | 0 | 0.00 | 93 | 0.01 |
| No\_Trx\_6 | numeric | 0 | 0.00 | 139 | 0.01 |
| No\_Trx\_12 | numeric | 0 | 0.00 | 204 | 0.02 |
| Total\_Deposits\_3 | numeric | 0 | 0.00 | 1670 | 0.13 |
| Max\_Deposits\_3 | numeric | 0 | 0.00 | 638 | 0.05 |
| Av\_Deposits\_3 | numeric | 0 | 0.00 | 2551 | 0.20 |
| Total\_Withdrawals\_3 | numeric | 0 | 0.00 | 1663 | 0.13 |
| Max\_Withdrawals\_3 | numeric | 0 | 0.00 | 625 | 0.05 |
| Av\_Withdrawals\_3 | numeric | 0 | 0.00 | 2373 | 0.19 |
| Total\_Deposits\_6 | numeric | 0 | 0.00 | 2544 | 0.20 |
| Max\_Deposits\_6 | numeric | 0 | 0.00 | 750 | 0.06 |
| Av\_Deposits\_6 | numeric | 0 | 0.00 | 3431 | 0.27 |
| Total\_Withdrawals\_6 | numeric | 0 | 0.00 | 2587 | 0.21 |
| Max\_Withdrawals\_6 | numeric | 0 | 0.00 | 824 | 0.07 |
| Av\_Withdrawals\_6 | numeric | 0 | 0.00 | 3279 | 0.26 |
| Total\_Deposits\_12 | numeric | 0 | 0.00 | 3691 | 0.29 |
| Max\_Deposits\_12 | numeric | 0 | 0.00 | 951 | 0.08 |
| Av\_Deposits\_12 | numeric | 0 | 0.00 | 4546 | 0.36 |
| Total\_Withdrawals\_12 | numeric | 0 | 0.00 | 3520 | 0.28 |
| Max\_Withdrawals\_12 | numeric | 0 | 0.00 | 985 | 0.08 |
| Av\_Withdrawals\_12 | numeric | 0 | 0.00 | 4219 | 0.34 |
| days\_since\_last\_trx | numeric | 0 | 0.00 | 761 | 0.06 |
| HasFixedDeposit | numeric | 0 | 0.00 | 2 | 0.00 |
| LastDepositBalance | numeric | 0 | 0.00 | 1090 | 0.09 |
| ChurnStatus | character | 0 | 0.00 | 2 | 0.00 |
| LoanHistory | character | 0 | 0.00 | 2 | 0.00 |

**Observations and Inferences**

From the above ,it was noticed that a number of variables had missing values.The missing values occurrence where however systematic showing an indication that the client had never used the service/product before.These will later require imputation.

# Exploratory Data Analysis.

The main interest under exploratory analysis is to understand the underlying distrubutions on every variable and their interactions. ## Univariate Analysis. ### Target Variable

| ChurnStatus | n | percent |
| --- | --- | --- |
| Churn | 4,424 | 0.35 |
| Not\_Churn | 8,088 | 0.64 |
| Total | 12,512 | 1.0 |

**Observations and inferences**

As we can observe from above,the institution had a 35% churn rate.

### Explanatory Variables

#### Categorical Variables

## Note: Using an external vector in selections is ambiguous.  
## i Use `all\_of(discretevars)` instead of `discretevars` to silence this message.  
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.  
## This message is displayed once per session.

## Warning: attributes are not identical across measure variables; they will be  
## dropped

TRUE

**Observations and inference**

1. Majority of customers have subscribed for 2 products.
2. Most customers apear to be males interms of gender
3. Of the rigistered customers,majority have transacted in more than 2 products.
4. Majority of the customers have no fixed deposit with the bank.
5. Most clients who have transacted didnt have a running loan at the time pf the analysis
6. A bigger percentage of customers have never taken loans with the bank.

#### Numerical Variables

##### Deposits transactions count Related

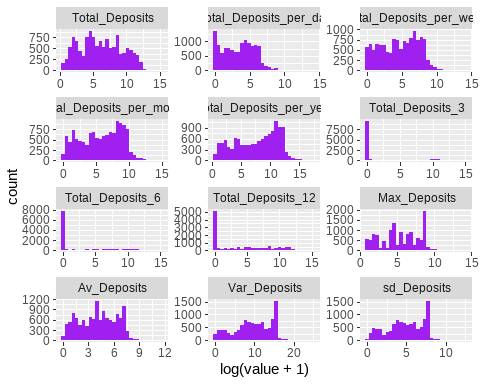
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

TRUE

##### Deposit transaction amount Related

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3342 rows containing non-finite values (stat\_bin).



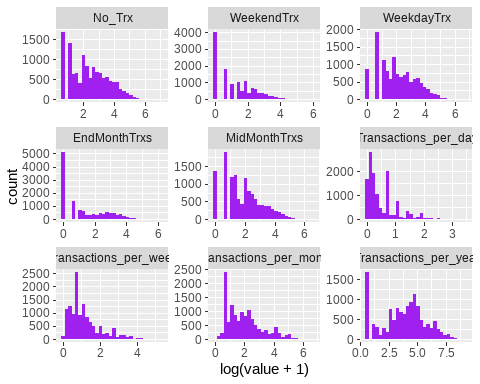
TRUE

##### Transaction Count Related

###### Descriptive Statistics

###### Visualization

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 **Observations and Inferences**

* Number of transactions appear to be right skewed at Mean of 22.41 and sd = 41, transactions.Most clients have done 1-3 transactions only.
* Yearly transactions were observed to be evenly distributed across the year.
* Most customers make at most 3 transactions per month and less than 2 per week.

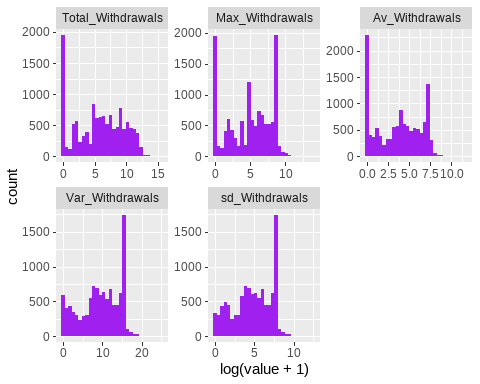
##### Withdrawals Related

###### Descriptive Statistics table

###### Visualization

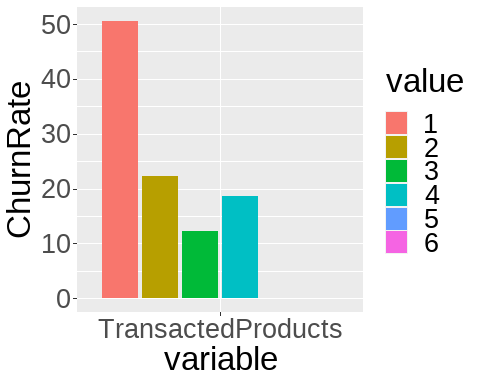
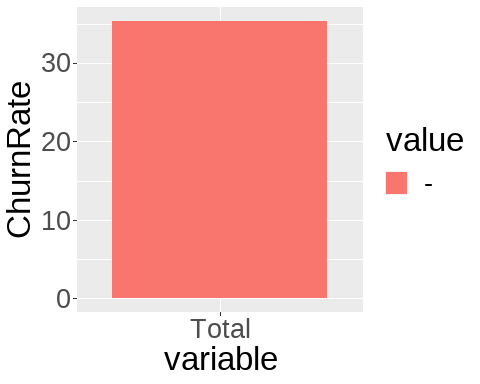
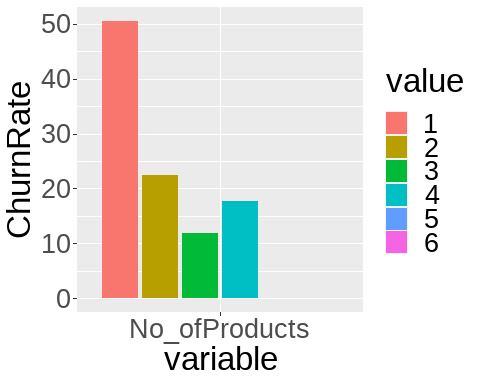
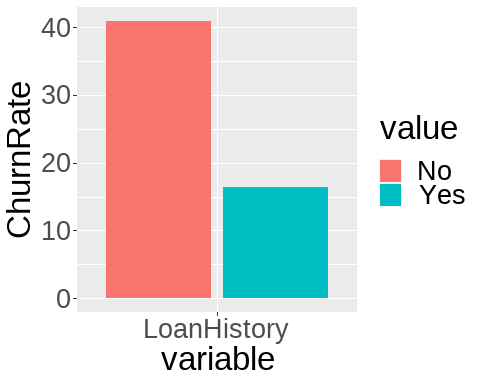
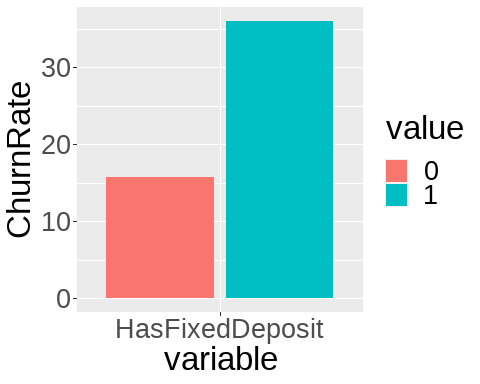
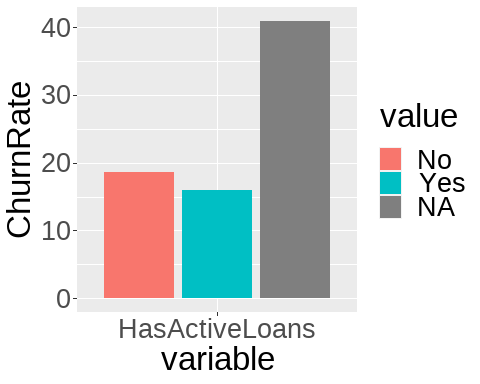
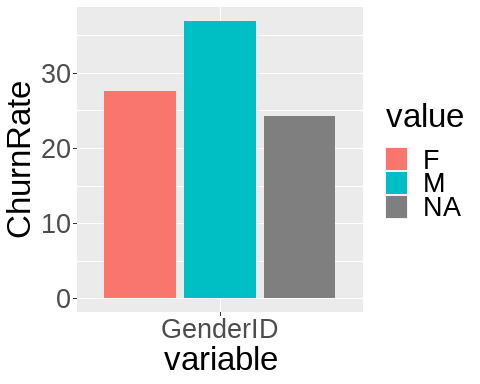
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3342 rows containing non-finite values (stat\_bin).

 ## Bivariate Analaysis ### Target Variable Vs Categorical Variables

#### Descriptive Statistics

| variable | value | Churn | Not\_Churn | Total | ChurnRate |
| --- | --- | --- | --- | --- | --- |
| No\_ofProducts | 1 | 2999 | 2934 | 5933 | 50.55 % |
| No\_ofProducts | 2 | 1360 | 4689 | 6049 | 22.48 % |
| No\_ofProducts | 3 | 57 | 421 | 478 | 11.92 % |
| No\_ofProducts | 4 | 8 | 37 | 45 | 17.78 % |
| No\_ofProducts | 5 | 0 | 6 | 6 | 0 % |
| No\_ofProducts | 6 | 0 | 1 | 1 | 0 % |
| Total | - | 4424 | 8088 | 12512 | 35.36 % |
| GenderID | F | 565 | 1488 | 2053 | 27.52 % |
| GenderID | M | 3851 | 6575 | 10426 | 36.94 % |
| GenderID | NA | 8 | 25 | 33 | 24.24 % |
| Total | - | 4424 | 8088 | 12512 | 35.36 % |
| TransactedProducts | 1 | 3005 | 2936 | 5941 | 50.58 % |
| TransactedProducts | 2 | 1354 | 4699 | 6053 | 22.37 % |
| TransactedProducts | 3 | 57 | 411 | 468 | 12.18 % |
| TransactedProducts | 4 | 8 | 35 | 43 | 18.6 % |
| TransactedProducts | 5 | 0 | 6 | 6 | 0 % |
| TransactedProducts | 6 | 0 | 1 | 1 | 0 % |
| Total | - | 4424 | 8088 | 12512 | 35.36 % |
| HasFixedDeposit | 0 | 68 | 365 | 433 | 15.7 % |
| HasFixedDeposit | 1 | 4356 | 7723 | 12079 | 36.06 % |
| Total | - | 4424 | 8088 | 12512 | 35.36 % |
| HasActiveLoans | No | 91 | 399 | 490 | 18.57 % |
| HasActiveLoans | Yes | 377 | 1978 | 2355 | 16.01 % |
| HasActiveLoans | NA | 3956 | 5711 | 9667 | 40.92 % |
| Total | - | 4424 | 8088 | 12512 | 35.36 % |
| LoanHistory | No | 3956 | 5711 | 9667 | 40.92 % |
| LoanHistory | Yes | 468 | 2377 | 2845 | 16.45 % |
| Total | - | 4424 | 8088 | 12512 | 35.36 % |



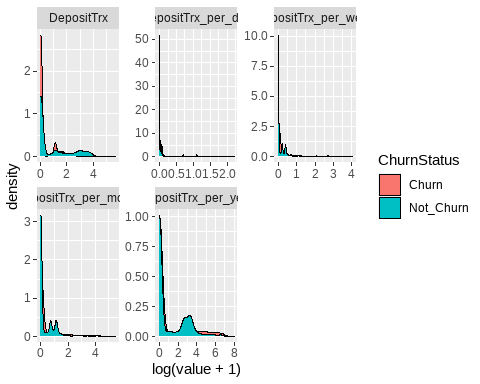
### Target Variable Vs Numerical Variables

##### Deposits transactions count Related

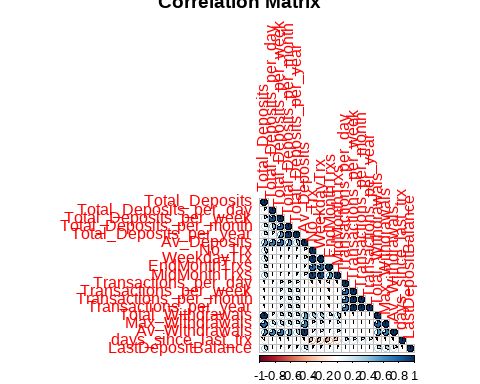
###### Descriptive Statistics

| variable | ChurnStatus | n | na | mean | sd | se\_mean | IQR | skewness | kurtosis | p00 | p01 | p05 | p10 | p20 | p25 | p30 | p40 | p50 | p60 | p70 | p75 | p80 | p90 | p95 | p99 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DepositTrx | Churn | 4424 | 0 | 0.94 | 3.13 | 0.05 | 0.00 | 9.44 | 150.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 2.00 | 3.00 | 5.00 | 13.0 | 79 |
| DepositTrx | Not\_Churn | 8088 | 0 | 6.09 | 13.68 | 0.15 | 4.00 | 4.81 | 51.68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.00 | 4.00 | 9.00 | 23.00 | 35.00 | 52.0 | 301 |
| DepositTrx\_per\_day | Churn | 4424 | 0 | 0.16 | 0.53 | 0.01 | 0.00 | 5.23 | 39.88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.03 | 0.40 | 1.00 | 2.5 | 8 |
| DepositTrx\_per\_day | Not\_Churn | 8088 | 0 | 0.09 | 0.35 | 0.00 | 0.04 | 6.41 | 52.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.04 | 0.06 | 0.09 | 0.46 | 2.0 | 5 |
| DepositTrx\_per\_month | Churn | 4424 | 0 | 3.35 | 12.59 | 0.19 | 0.00 | 6.41 | 60.46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.83 | 6.00 | 20.00 | 60.0 | 240 |
| DepositTrx\_per\_month | Not\_Churn | 8088 | 0 | 1.96 | 8.55 | 0.10 | 1.07 | 8.68 | 97.09 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.89 | 1.07 | 1.52 | 2.37 | 6.00 | 60.0 | 150 |
| DepositTrx\_per\_week | Churn | 4424 | 0 | 0.82 | 2.95 | 0.04 | 0.00 | 6.32 | 59.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.19 | 2.00 | 4.67 | 14.0 | 56 |
| DepositTrx\_per\_week | Not\_Churn | 8088 | 0 | 0.48 | 2.00 | 0.02 | 0.26 | 8.57 | 95.56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.21 | 0.26 | 0.40 | 0.65 | 1.92 | 14.0 | 35 |
| DepositTrx\_per\_year | Churn | 4424 | 0 | 40.26 | 153.23 | 2.30 | 0.00 | 6.41 | 60.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 2.00 | 73.00 | 243.33 | 730.0 | 2920 |
| DepositTrx\_per\_year | Not\_Churn | 8088 | 0 | 23.60 | 104.08 | 1.16 | 12.88 | 8.68 | 97.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7.56 | 12.88 | 16.61 | 28.79 | 73.00 | 730.0 | 1825 |

###### Visualization

 ### Correlation Analysis.

In order to plot correlation,features that exhibited some distributions were selected.

 Observations and Inferences

* Some variables are highly correlated with others. eg Total withdrawals vs Total deposits.
* This correlation will be dealt with at feature engineering phase.

**Variable Selection for the next stage of analysis**