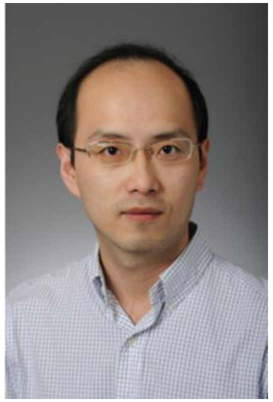
TEAM VERSATILE — VEHICLE LOAN DEFAULT PREDICTION

GEORGE MASON UNIVERSITY | Data Analytics Engineering

SYST/OR 568 - DL
Applied Predictive Analytics
Data Analytics Engineering Program
Spring 2023

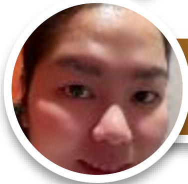
May 4, 2023



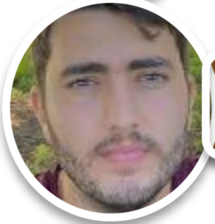
Dr. Jie Xu, Associate Professor
Systems Engineering & Operations Research
VOLGENAU SCHOOL OF ENGINEERING,
COLLEGE OF ENGINEERING & COMPUTING
GEORGE MASON UNIVERSITY – FAIRFAX CAMPUS
4400 UNIVERSITY DR MSN 4A7 FAIRFAX, VA 22030

Ph. 703-993-4620
Email: jxu13@gmu.edu

Hill, Faith



Le, Huong Thi Thu



Jarrouje, Tarex

Sharma, Indu Priya

Teh, B. Bloti

AGENDA

- Raw Data Set & Source
  - About The Data Set
- High-level Summary (Raw Data Sets)
  - Predictors & Responses
- Data Preprocessing
  - Check for Missing Values
  - Removal of Near-Zero Variance Predictors
  - Box-Cox Transformation, Centering and Scaling
- High-level Summary (Cleaned-up Data Sets)
  - Predictors & Responses
- Predictive Models (Already Tried or Plan To Use)
  - Classification Tree
  - Logistic Regression
  - Random Forest
- Roadblocks
- Conclusion

About The Data Set/Project:

- **Importance/Essence of problem:** Financial organizations suffer substantial losses from defaults on car loans, resulting in stricter vehicle loan underwriting standards and higher rejection rates. Consequently, financial institutions are calling for a more effective credit risk scoring model to assess the likelihood of vehicle loan defaults. This necessitates an investigation to identify the factors that contribute to the default of vehicle loans. Doing so will ensure that consumers capable of repayment are not rejected and important determinants can be identified which can be further used for minimizing the default rates.

Prediction

- **Likelihood of default or nondefault loaner:** based on the information collected from the borrower, we can predict the likelihood of that a loan can be returned or not.

TEAM VERSATILE – RAW DATA SET & SOURCE

Project Dataset

Dataset Name: Vehicle Loan

- Dataset Owners: Kaggle (Avik Paul)
- Dataset Type: Open Source
- Dataset Size: 233,154 Obs. & 41 Variables

Below are pieces of information regarding the loan and loanee in the data set:

1. Loanee Information (Demographic data like age, income, Identity proof etc.)
2. Loan Information (Disbursal details, amount, EMI, loan to value ratio etc.)
3. Bureau data & history (Bureau score, number of active accounts, the status of other loans, credit history etc.)

Extract of Vehicle Loan Dataset

UNIQUEID	DISBURSED_AMOUNT	ASSET_COST	LTV	BRANCH_ID	SUPPLIER_ID	MANUFACTURER_ID	CURRENT_PINCODE_ID	DATE_OF_BIRTH
420825	50578	58400	89.55	67	22807	45	1441	1/1/1984
537409	47145	65550	73.23	67	22807	45	1502	31-07-1985
417566	53278	61360	89.63	67	22807	45	1497	24-08-1985
624493	57513	66113	88.48	67	22807	45	1501	30-12-1993
539055	52378	60300	88.39	67	22807	45	1495	9/12/1977
518279	54513	61900	89.66	67	22807	45	1501	8/9/1990
529269	46349	61500	76.42	67	22807	45	1502	1/6/1988
510278	43894	61900	71.89	67	22807	45	1501	4/10/1989
490213	53713	61973	89.56	67	22807	45	1497	15-11-1991
510980	52603	61300	86.95	67	22807	45	1492	1/6/1968
548567	53278	61230	89.83	67	22807	45	1493	1/1/1979
486821	64769	74190	89.23	67	22807	45	1446	7/9/1984
478647	53278	61330	89.68	67	22807	45	1497	1/6/1974
479533	49478	57010	89.46	67	22807	45	1497	16-08-1984
483869	49278	57080	89.35	67	22807	45	1495	18-02-1973
600655	47549	61400	79.8	67	22807	45	1440	5/7/1994
513916	57713	65750	89.28	67	22807	45	1440	1/6/1976
522020	53503	62100	87.28	67	22807	45	1498	27-02-1983
492995	70017	86760	82.99	67	22807	45	1479	10/8/1988
568857	58259	68500	86.13	67	22807	45	1468	16-04-1980
590630	58013	69650	84.71	67	22807	45	1497	1/11/1978
467015	31184	57110	56.91	67	22807	45	1498	29-02-1984
563215	43594	78256	57.5	67	22744	86	1499	14-07-1994
513139	54513	61900	89.66	67	22807	45	1468	31-05-1979
498082	73123	92900	79.66	67	22807	45	1480	2/1/1989
586411	55213	68600	83.09	67	22807	45	1494	1/1/1986

Vehicle Loan Data Dictionary

Variable Name	Description
UniqueID	Identifier for customers
loan_default	Payment default in the first EMI on due date
disbursed_amount	Amount of Loan disbursed
ltv	Loan to Value of the asset
branch_id	Branch where the loan was disbursed
supplier_id	Vehicle Dealer where the loan was disbursed
manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)
Current_pincode	Current pin code of the customer
Date_of_Birth	Date of birth of the customer
Employment_Type	Employment Type of the customer (Salaried/Self Employed)
Disbursal_Date	Date of birth of the customer
State_ID	State of disbursement

Vehicle Loan Data Dictionary

Variable Name	Description
Employee_code_ID	Employee of the organization who logged the disbursement
MobileNo_Avl_Flag	if Mobile no. was shared by the customer, then flagged as 1
Aadhar_flag	if aadhar was shared by the customer then flagged as 1
PAN_flag	if voter was shared by the customer, then flagged as 1
VoterID_flag	if voter was shared by the customer, then flagged as 1
Driving_flag	if DL was shared by the customer, then flagged as 1
Passport_flag	if passport was shared by the customer, then flagged as 1
PERFORM_CNS_SCORE	Bureau Score
PERFORM_CNS_SCORE_DESCRIPTION	Bureau score description
PRI_NO_OF_ACCTS	count of total loans taken by the customer at the time of disbursement
DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS	Loans defaulted in the last 6 months

Variable Name	Description
PRI_ACTIVE_ACCTS	count of active loans taken by the customer at the time of disbursement
PRI_OVERDUE_ACCTS	count of default accounts at the time of disbursement
PRI_CURRENT_BALANCE	total Principal outstanding amount of the active loans at the time of disbursement
PRI_SANCTIONED_AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement
PRI_DISBURSED_AMOUNT	total amount that was disbursed for all the loans at the time of disbursement
SEC_NO_OF_ACCTS	count of total loans taken by the customer at the time of disbursement
SEC_ACTIVE_ACCTS	count of active loans taken by the customer at the time of disbursement
SEC_OVERDUE_ACCTS	count of default accounts at the time of disbursement
AVERAGE_ACCT_AGE	Average loan tenure

TEAM VERSATILE – RAW DATA SET & SOURCE

41 Tentative Predictors:

```
$ UNIQUEID
$ DISBURSED_AMOUNT
$ ASSET_COST
$ LTV
$ BRANCH_ID
$ SUPPLIER_ID
$ MANUFACTURER_ID
$ CURRENT_PINCODE_ID
$ DATE_OF_BIRTH
$ EMPLOYMENT_TYPE
$ DISBURSAL_DATE
$ STATE_ID
$ EMPLOYEE_CODE_ID
$ MOBILENO_AVL_FLAG
$ AADHAR_FLAG
$ PAN_FLAG
$ VOTERID_FLAG
$ DRIVING_FLAG
$ PASSPORT_FLAG
$ PERFORM_CNS_SCORE
$ PERFORM_CNS_SCORE_DESCRIPTION
$ PRI_NO_OF_ACCTS
$ PRI_ACTIVE_ACCTS
$ PRI_OVERDUE_ACCTS
$ PRI_CURRENT_BALANCE
$ PRI_SANCTIONED_AMOUNT
$ PRI_DISBURSED_AMOUNT
$ SEC_NO_OF_ACCTS
$ SEC_ACTIVE_ACCTS
$ SEC_OVERDUE_ACCTS
$ SEC_CURRENT_BALANCE
$ SEC_SANCTIONED_AMOUNT
$ SEC_DISBURSED_AMOUNT
$ PRIMARY_INSTAL_AMT
$ SEC_INSTAL_AMT
$ NEW_ACCTS_IN_LAST_SIX_MONTHS
$ DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS
$ AVERAGE_ACCT_AGE
$ CREDIT_HISTORY_LENGTH
$ NO_OF_INQUIRIES
$ LOAN_DEFAULT
```

Response/Outcome: Default or Not Default

Vehicle Loan Data Dictionary

Variable Name	Description
SEC_CURRENT_BALANCE	total Principal outstanding amount of the active loans at the time of disbursement
CREDIT_HISTORY_LENGTH	Time since first loan
NO_OF_INQUIRIES	Enquires done by the customer for loans
SEC_SANCTIONED_AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement
SEC.DISBURSED_AMOUNT	total amount that was disbursed for all the loans at the time of disbursement
PRIMARY_INSTA_AMT	EMI Amount of the primary loan
SEC_INSTAL_AMT	EMI Amount of the secondary loan
NEW_ACCTS_IN_LAST_SIX_MONTHS	New loans taken by the customer in last 6 months before the disbursement.

Primary (PRI) accounts are those which the customer has taken for his personal use.

Secondary (SEC) accounts are those which the customer act as a co-applicant or gaurantor.

Vehicle Loan Data Set (RAW DATA) Internal Structure/Statistics

```
> str(df) # Displays vehicle loans data set internal structure
'data.frame':	233154 obs. of  41 variables:
 $ UNIQUEID                           : int  420825 537409 417566 624493 539055 518279 529269 510278 490213 510980 ...
 $ DISBURSED_AMOUNT                   : int  50578 47145 53278 57513 52378 54513 46349 43894 53713 52603 ...
 $ ASSET_COST                         : int  58400 65550 61360 66113 60300 61900 61500 61900 61973 61300 ...
 $ LTV                                : num  89.5 73.2 89.6 88.5 88.4 ...
 $ BRANCH_ID                          : int  67 67 67 67 67 67 67 67 67 67 ...
 $ SUPPLIER_ID                        : int  22807 22807 22807 22807 22807 22807 22807 22807 22807 22807 ...
 $ MANUFACTURER_ID                    : int  45 45 45 45 45 45 45 45 45 45 ...
 $ CURRENT_PINCODE_ID                 : int  1441 1502 1497 1501 1495 1501 1502 1501 1497 1492 ...
 $ DATE_OF_BIRTH                      : chr  "1/1/1984" "31-07-1985" "24-08-1985" "30-12-1993" ...
 $ EMPLOYMENT_TYPE                    : chr  "Salaried" "Self employed" "Self employed" "Self employed" ...
 $ DISBURSAL_DATE                     : chr  "3/8/2018" "26-09-2018" "1/8/2018" "26-10-2018" ...
 $ STATE_ID                           : int  6 6 6 6 6 6 6 6 6 6 ...
 $ EMPLOYEE_CODE_ID                   : int  1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
 $ MOBILENO_AVL_FLAG                  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ AADHAR_FLAG                        : int  1 1 1 1 1 1 1 1 1 0 ...
 $ PAN_FLAG                           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ VOTERID_FLAG                       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ DRIVING_FLAG                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PASSPORT_FLAG                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PERFORM_CNS_SCORE                  : int  0 598 0 305 0 825 0 17 718 818 ...
 $ PERFORM_CNS_SCORE_DESCRIPTION      : chr  "No Bureau History Available" "I-Medium Risk" "No Bureau History Available" "L-Very High Risk" ...
 $ PRI_NO_OF_ACCTS                    : int  0 1 0 3 0 2 0 1 1 1 ...
 $ PRI_ACTIVE_ACCTS                   : int  0 1 0 0 0 0 0 1 1 0 ...
 $ PRI_OVERDUE_ACCTS                  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ PRI_CURRENT_BALANCE                : int  0 27600 0 0 0 0 0 72879 -41 0 ...
 $ PRI_SANCTIONED_AMOUNT              : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ PRI_DISBURSED_AMOUNT               : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ SEC_NO_OF_ACCTS                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_ACTIVE_ACCTS                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_OVERDUE_ACCTS                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_CURRENT_BALANCE                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_SANCTIONED_AMOUNT              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_DISBURSED_AMOUNT               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PRIMARY_INSTAL_AMT                 : int  0 1991 0 31 0 1347 0 0 0 2608 ...
 $ SEC_INSTAL_AMT                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NEW_ACCTS_IN_LAST_SIX_MONTHS       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS: int  0 1 0 0 0 0 0 0 0 0 ...
 $ AVERAGE_ACCT_AGE                   : chr  "0yrs 0mon" "1yrs 11mon" "0yrs 0mon" "0yrs 8mon" ...
 $ CREDIT_HISTORY_LENGTH              : chr  "0yrs 0mon" "1yrs 11mon" "0yrs 0mon" "1yrs 3mon" ...
 $ NO_OF_INQUIRIES                    : int  0 0 0 1 1 0 0 0 1 0 ...
 $ LOAN_DEFAULT                       : int  0 1 0 1 1 0 0 0 0 0 ...
```

```
# Display of "DATE_OF_BIRTH" and # #
"DISBURSAL_DATE" data type using # the
class fn.
> class(df$DATE_OF_BIRTH)
[1] "character"
> class(df$DISBURSAL_DATE)
[1] "character"
```

Vehicle Loan Data Set (RAW DATA) Internal Structure/Statistics

```
> str(dates_df)
```

```
> str(dates_df_modified)
'data.frame':	233154 obs. of  43 variables:
 $ UNIQUEID                           : int  420825 537409 417566 624493 539055 518279 529269 510278 490213 510980 ...
 $ DISBURSED_AMOUNT                   : int  50578 47145 53278 57513 52378 54513 46349 43894 53713 52603 ...
 $ ASSET_COST                         : int  58400 65550 61360 66113 60300 61900 61500 61900 61973 61300 ...
 $ LTV                                : num  89.5 73.2 89.6 88.5 88.4 ...
 $ BRANCH_ID                          : int  67 67 67 67 67 67 67 67 67 67 ...
 $ SUPPLIER_ID                        : int  22807 22807 22807 22807 22807 22807 22807 22807 22807 22807 ...
 $ MANUFACTURER_ID                    : int  45 45 45 45 45 45 45 45 45 45 ...
 $ CURRENT_PINCODE_ID                 : int  1441 1502 1497 1501 1495 1501 1502 1501 1497 1492 ...
 $ DATE_OF_BIRTH                      : chr  "1/1/1984" "31-07-1985" "24-08-1985" "30-12-1993" ...
 $ EMPLOYMENT_TYPE                    : chr  "Salaried" "Self employed" "Self employed" "Self employed" ...
 $ DISBURSAL_DATE                     : chr  "3/8/2018" "26-09-2018" "1/8/2018" "26-10-2018" ...
 $ STATE_ID                           : int  6 6 6 6 6 6 6 6 6 6 ...
 $ EMPLOYEE_CODE_ID                   : int  1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
 $ MOBILENO_AVL_FLAG                  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ AADHAR_FLAG                        : int  1 1 1 1 1 1 1 1 1 0 ...
 $ PAN_FLAG                           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ VOTERID_FLAG                       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ DRIVING_FLAG                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PASSPORT_FLAG                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PERFORM_CNS_SCORE                  : int  0 598 0 305 0 825 0 17 718 818 ...
 $ PERFORM_CNS_SCORE_DESCRIPTION      : chr  "No Bureau History Available" "I-Medium Risk" "No Bureau History Available" "L-Very High Risk" ...
 $ PRI_NO_OF_ACCTS                    : int  0 1 0 3 0 2 0 1 1 1 ...
 $ PRI_ACTIVE_ACCTS                   : int  0 1 0 0 0 0 0 1 1 0 ...
 $ PRI_OVERDUE_ACCTS                  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ PRI_CURRENT_BALANCE                : int  0 27600 0 0 0 0 0 72879 -41 0 ...
 $ PRI_SANCTIONED_AMOUNT              : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ PRI_DISBURSED_AMOUNT               : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ SEC_NO_OF_ACCTS                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_ACTIVE_ACCTS                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_OVERDUE_ACCTS                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_CURRENT_BALANCE                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_SANCTIONED_AMOUNT              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_DISBURSED_AMOUNT               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PRIMARY_INSTAL_AMT                 : int  0 1991 0 31 0 1347 0 0 0 2608 ...
 $ SEC_INSTAL_AMT                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NEW_ACCTS_IN_LAST_SIX_MONTHS       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS: int  0 1 0 0 0 0 0 0 0 0 ...
 $ AVERAGE_ACCT_AGE                   : chr  "0yrs 0mon" "1yrs 11mon" "0yrs 0mon" "0yrs 8mon" ...
 $ CREDIT_HISTORY_LENGTH              : chr  "0yrs 0mon" "1yrs 11mon" "0yrs 0mon" "1yrs 3mon" ...
 $ NO_OF_INQUIRIES                    : int  0 0 0 1 1 0 0 0 1 0 ...
 $ LOAN_DEFAULT                       : int  0 1 0 1 1 0 0 0 0 0 ...
 $ formatted_DATE_OF_BIRTH            : Date, format: "1984-01-01" "1985-07-31" "1985-08-24" "1993-12-30" ...
 $ formatted_DISBURSAL_DATE           : Date, format: "2018-03-08" "2018-09-26" "2018-01-08" "2018-10-26" ...
> 
```

The **lubridate** package provides a convenient and user-friendly way to work with dates and times in R.

```
#---------------------------------------------------------------------------------------------#
# Convert the character columns to date columns using ymd() or mdy() or dmy()
# depending on the format for both formatted_DATE_OF_BIRTH & formatted_DISBURSAL_DATE

# Convert the character columns to date columns using ymd() or mdy() or dmy()
# depending on the format
dates_df$formatted_DATE_OF_BIRTH <- mdy(dates_df$formatted_DATE_OF_BIRTH) # month-day-year format
dates_df$formatted_DISBURSAL_DATE <- mdy(dates_df$formatted_DISBURSAL_DATE) # month-day-year format

# Displays internal structure of the data frame with formatted date columns
str(dates_df)
```

New formatted date columns

Vehicle Loan Data Set (RAW DATA) Internal Structure/Statistics

Added a new column, AGE at the bottom of to the data frame

```
> str(age_df)
'data.frame':	233154 obs. of  44 variables:
 $ UNIQUEID                           : int  420825 537409 417566 624493 539055 518279 529269 510278 490213 510980 ...
 $ DISBURSED_AMOUNT                   : int  50578 47145 53278 57513 52378 54513 46349 43894 53713 52603 ...
 $ ASSET_COST                         : int  58400 65550 61360 66113 60300 61900 61500 61900 61973 61300 ...
 $ LTV                                : num  89.5 73.2 89.6 88.5 88.4 ...
 $ BRANCH_ID                          : int  67 67 67 67 67 67 67 67 67 67 ...
 $ SUPPLIER_ID                        : int  22807 22807 22807 22807 22807 22807 22807 22807 22807 22807 ...
 $ MANUFACTURER_ID                    : int  45 45 45 45 45 45 45 45 45 45 ...
 $ CURRENT_PINCODE_ID                 : int  1441 1502 1497 1501 1495 1501 1502 1501 1497 1492 ...
 $ DATE_OF_BIRTH                      : chr  "1/1/1984" "31-07-1985" "24-08-1985" "30-12-1993" ...
 $ EMPLOYMENT_TYPE                    : chr  "Salaried" "Self employed" "Self employed" "Self employed" ...
 $ DISBURSAL_DATE                     : chr  "3/8/2018" "26-09-2018" "1/8/2018" "26-10-2018" ...
 $ STATE_ID                           : int  6 6 6 6 6 6 6 6 6 6 ...
 $ EMPLOYEE_CODE_ID                   : int  1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
 $ MOBILENO_AVL_FLAG                  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ AADHAR_FLAG                        : int  1 1 1 1 1 1 1 1 1 0 ...
 $ PAN_FLAG                           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ VOTERID_FLAG                       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ DRIVING_FLAG                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PASSPORT_FLAG                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PERFORM_CNS_SCORE                  : int  0 598 0 305 0 825 0 17 718 818 ...
 $ PERFORM_CNS_SCORE_DESCRIPTION      : chr  "No Bureau History Available" "I-Medium Risk" "No Bureau History Available" "L-Very High Risk" ...
 $ PRI_NO_OF_ACCTS                    : int  0 1 0 3 0 2 0 1 1 1 ...
 $ PRI_ACTIVE_ACCTS                   : int  0 1 0 0 0 0 0 1 1 0 ...
 $ PRI_OVERDUE_ACCTS                  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ PRI_CURRENT_BALANCE                : int  0 27600 0 0 0 0 0 72879 -41 0 ...
 $ PRI_SANCTIONED_AMOUNT              : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ PRI_DISBURSED_AMOUNT               : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ SEC_NO_OF_ACCTS                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_ACTIVE_ACCTS                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_OVERDUE_ACCTS                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_CURRENT_BALANCE                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_SANCTIONED_AMOUNT              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SEC_DISBURSED_AMOUNT               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PRIMARY_INSTAL_AMT                 : int  0 1991 0 31 0 1347 0 0 0 2608 ...
 $ SEC_INSTAL_AMT                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NEW_ACCTS_IN_LAST_SIX_MONTHS       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS: int  0 1 0 0 0 0 0 0 0 0 ...
 $ AVERAGE_ACCT_AGE                   : chr  "0yrs 0mon" "1yrs 11mon" "0yrs 0mon" "0yrs 8mon" ...
 $ CREDIT_HISTORY_LENGTH              : chr  "0yrs 0mon" "1yrs 11mon" "0yrs 0mon" "1yrs 3mon" ...
 $ NO_OF_INQUIRIES                    : int  0 0 0 1 1 0 0 0 1 0 ...
 $ LOAN_DEFAULT                       : int  0 1 0 1 1 0 0 0 0 0 ...
 $ formatted_DATE_OF_BIRTH            : Date, format: "1984-01-01" "1985-07-31" "1985-08-24" "1993-12-30" ...
 $ formatted_DISBURSAL_DATE           : Date, format: "2018-03-08" "2018-09-26" "2018-01-08" "2018-10-26" ...
 $ AGE                                : num  39 38 38 30 46 33 35 34 32 55 ...
```

Removal of Near-Zero Variance Predictors

	freqRatio	percentUnique	zeroVar	nzv
UNIQUEID	1.000000	1.000000e+02	FALSE	FALSE
DISBURSED_AMOUNT	1.007059	1.053595e+01	FALSE	FALSE
ASSET_COST	1.142617	1.983753e+01	FALSE	FALSE
LTV	4.300971	2.821740e+00	FALSE	FALSE
BRANCH_ID	1.159781	3.516989e-02	FALSE	FALSE
SUPPLIER_ID	1.101538	1.266545e+00	FALSE	FALSE
MANUFACTURER_ID	1.934341	4.717912e-03	FALSE	FALSE
CURRENT_PINCODE_ID	1.086077	2.872779e+00	FALSE	FALSE
EMPLOYMENT_TYPE	1.304288	1.286703e-03	FALSE	FALSE
STATE_ID	1.316685	9.435824e-03	FALSE	FALSE
EMPLOYEE_CODE_ID	1.250996	1.402506e+00	FALSE	FALSE
AADHAR_FLAG	5.262530	8.578021e-04	FALSE	FALSE
PAN_FLAG	12.231599	8.578021e-04	FALSE	FALSE
VOTERID_FLAG	5.899272	8.578021e-04	FALSE	FALSE
PERFORM_CNS_SCORE	13.326117	2.457603e-01	FALSE	FALSE
PERFORM_CNS_SCORE_DESCRIPTION	7.288875	8.578021e-03	FALSE	FALSE
PRI_NO_OF_ACCTS	3.343530	4.632132e-02	FALSE	FALSE
PRI_ACTIVE_ACCTS	3.258019	1.715604e-02	FALSE	FALSE
PRI_OVERDUE_ACCTS	10.359489	9.435824e-03	FALSE	FALSE
PRI_CURRENT_BALANCE	1171.041322	3.059823e+01	FALSE	FALSE
PRI_SANCTIONED_AMOUNT	91.880240	1.903892e+01	FALSE	FALSE
PRI_DISBURSED_AMOUNT	98.858369	2.054822e+01	FALSE	FALSE
PRIMARY_INSTAL_AMT	546.291096	1.203797e+01	FALSE	FALSE
NEW_ACCTS_IN_LAST_SIX_MONTHS	5.654195	1.115143e-02	FALSE	FALSE
DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS	14.387190	6.004615e-03	FALSE	FALSE
NO_OF_INQUIRIES	9.062643	1.072253e-02	FALSE	FALSE
LOAN_DEFAULT	3.606785	8.578021e-04	FALSE	FALSE
formatted_DATE_OF_BIRTH	1.001382	6.619230e+00	FALSE	FALSE
formatted_DISBURSAL_DATE	1.317117	3.602769e-02	FALSE	FALSE
AGE	1.005128	2.058725e-02	FALSE	FALSE
DRIVING_FLAG	42.025281	8.578021e-04	FALSE	TRUE
PASSPORT_FLAG	469.068548	8.578021e-04	FALSE	TRUE
SEC_NO_OF_ACCTS	65.576746	1.586934e-02	FALSE	TRUE
SEC_ACTIVE_ACCTS	85.445976	9.864725e-03	FALSE	TRUE
SEC_OVERDUE_ACCTS	205.329495	3.860110e-03	FALSE	TRUE
SEC_CURRENT_BALANCE	22979.000000	1.392213e+00	FALSE	TRUE
SEC_SANCTIONED_AMOUNT	2764.072289	9.534471e-01	FALSE	TRUE
SEC_DISBURSED_AMOUNT	3888.983051	1.094984e+00	FALSE	TRUE
SEC_INSTAL_AMT	32991.000000	8.226323e-01	FALSE	TRUE
AVERAGE_ACCT_AGE	19.803086	8.234901e-02	FALSE	TRUE
CREDIT_HISTORY_LENGTH	25.021424	1.260969e-01	FALSE	TRUE
MOBILENO_AVL_FLAG	0.000000	4.289011e-04	TRUE	TRUE

To Be Removed!
12 Near-Zero Variance Variables Identified

Removal of Near-zero Variance Predictors:

- **12 data points or non-informative predictors** identified during the variable selection process.
- **30 variables** with high predictive capabilities were identified as final PREDICTORS for modeling.

Vehicle Loan Data Set (RAW DATA): Missing Values

```
> # Check for missing values in numeric data type
>  sum(is.na(df_nzv))
[1] 0
>
> # Check for missing values specifically for EMPLOYMENT_TYPE column
> # Filter the data set to show only rows with blank spaces in the
> # EMPLOYMENT_TYPE column
> missing_employment <- df %>% filter(EMPLOYMENT_TYPE == "")
>
> # Count the number of rows with blank spaces in the EMPLOYMENT_TYPE column
> nrow(missing_employment)
[1] 7661
> # - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - #
>
```

UNIQUEID	DISBURSED_AMOUNT	ASSET_COST	LTV	BRANCH_ID	SUPPLIER_ID	MANUFACTURER_ID	CURRENT_PINCODE_ID	EMPLOYMENT_TYPE
525234	52428	67405	81.60	78	17014	45	2099	
637252	51653	63896	86.08	78	17014	45	2079	
584433	49488	63306	83.72	78	17014	45	2069	
515149	40884	59313	70.81	78	17014	45	2099	
547112	49683	62577	83.10	78	17014	45	2099	
497986	17850	97311	19.53	11	22976	51	5969	
535877	49303	68885	74.04	11	15893	86	5969	
562770	56013	80906	71.69	11	24654	49	5940	
623921	51003	65606	79.26	20	23502	45	6188	
635397	45549	73104	63.61	20	14158	45	6207	
505026	70123	95213	74.57	20	23569	48	6207	
432582	49078	55957	89.89	20	14158	45	6207	
629054	46952	63872	75.15	20	14158	45	6207	
609242	48045	69367	71.94	63	16309	45	7093	
485779	32484	61346	55.42	63	16309	45	7085	
438496	53078	61346	88.84	63	16309	45	7093	

EMPLOYMENT_TYPE:

- 7,661 missing values identified

Imputation of Missing Values

Dynamic imputation of missing values with mode depending on the data type

```
> # Validating imputation of missing values for the categorical variable,
> # EMPLOYMENT_TYPE.
> # Filter the data set to show only rows with blank spaces in the
> # EMPLOYMENT_TYPE column
> missing_employment <- df_Imp %>% filter(EMPLOYMENT_TYPE == "")
>
> # Count the number of rows with blank spaces in the EMPLOYMENT_TYPE column
> nrow(missing_employment)
[1] 0
```

EMPLOYMENT_TYPE:

- **Zero missing values identified AFTER IMPUTATION**

Categorical Data: Impute missing values with mode.

UNIQUEID	DISBURSED_AMOUNT	ASSET_COST	LTV	BRANCH_ID	SUPPLIER_ID	MANUFACTURER_ID	CURRENT_PINCODE_ID	EMPLOYMENT_TYPE
637252	51653	63896	86.08	78	17014	45	2079	Self employed
584433	49488	63306	83.72	78	17014	45	2069	Self employed
515149	40884	59313	70.81	78	17014	45	2099	Self employed
547112	49683	62577	83.10	78	17014	45	2099	Self employed
497986	17850	97311	19.53	11	22976	51	5969	Self employed
535877	49303	68885	74.04	11	15893	86	5969	Self employed
562770	56013	80906	71.69	11	24654	49	5940	Self employed
623921	51003	65606	79.26	20	23502	45	6188	Self employed
635397	45549	73104	63.61	20	14158	45	6207	Self employed
505026	70123	95213	74.57	20	23569	48	6207	Self employed
432582	49078	55957	89.89	20	14158	45	6207	Self employed
629054	46952	63872	75.15	20	14158	45	6207	Self employed
609242	48045	69367	71.94	63	16309	45	7093	Self employed
485779	32484	61346	55.42	63	16309	45	7085	Self employed
438496	53078	61346	88.84	63	16309	45	7093	Self employed

DATA PRE-PROCESSING: IMPUTATION (RAW DATA SET

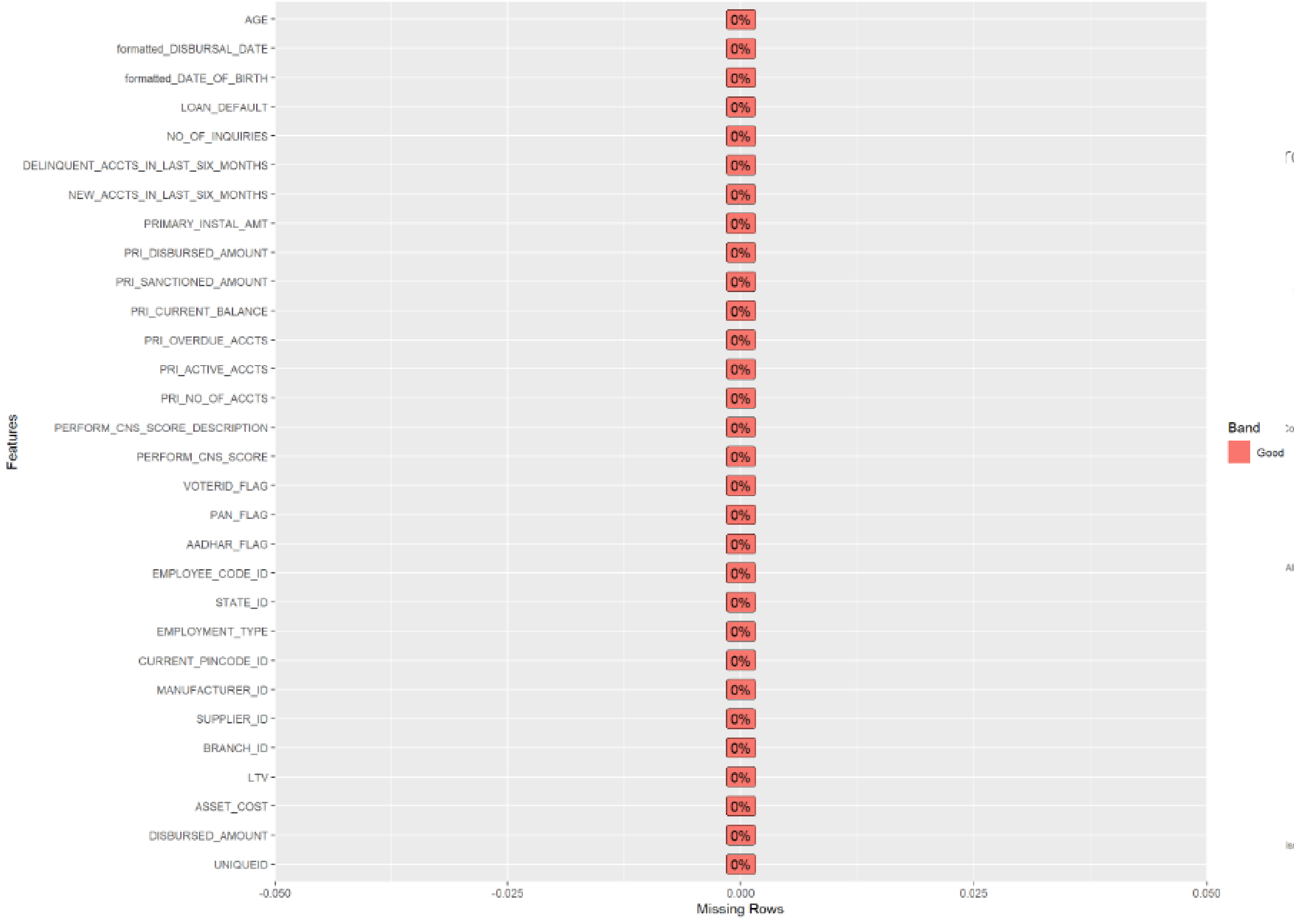
Post-Imputation Results

Data Explorer Package displaying missing data profile and basic statistics

Basic Statistics
Raw Counts

Name	Value
Rows	233,154
Columns	30
Discrete columns	4
Continuous columns	26
All missing columns	0
Missing observations	0
Complete Rows	233,154
Total observations	6,994,620
Memory allocation	53.4 Mb

Missing Data Profile



Vehicle Loan Data Set (RAW DATA) Internal Structure/Statistics

30 Possible Predictors

Response/Outcome: Default or Not Default

Data Structure

root (Classes 'data.table' and 'data.frame': 233154 obs. of 30 variables:)

- UNIQUEID (num)
- DISBURSED_AMOUNT (num)
- ASSET_COST (num)
- LTV (num)
- BRANCH_ID (num)
- SUPPLIER_ID (num)
- MANUFACTURER_ID (num)
- CURRENT_PINCODE_ID (num)
- EMPLOYMENT_TYPE (chr)
- STATE_ID (num)
- EMPLOYEE_CODE_ID (num)
- AADHAR_FLAG (num)
- PAN_FLAG (num)
- VOTERID_FLAG (num)
- PERFORM_CNS_SCORE (num)
- PERFORM_CNS_SCORE_DESCRIPTION (chr)
- PRI_NO_OF_ACCTS (num)
- PRI_ACTIVE_ACCTS (num)
- PRI_OVERDUE_ACCTS (num)
- PRI_CURRENT_BALANCE (num)
- PRI_SANCTIONED_AMOUNT (num)
- PRI_DISBURSED_AMOUNT (num)
- PRIMARY_INSTAL_AMT (num)
- NEW_ACCTS_IN_LAST_SIX_MONTHS (num)
- DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS (num)
- NO_OF_INQUIRIES (num)
- LOAN_DEFAULT (num)
- formatted_DATE_OF_BIRTH (Date, format)
- formatted_DISBURSAL_DATE (Date, format)
- AGE (num)

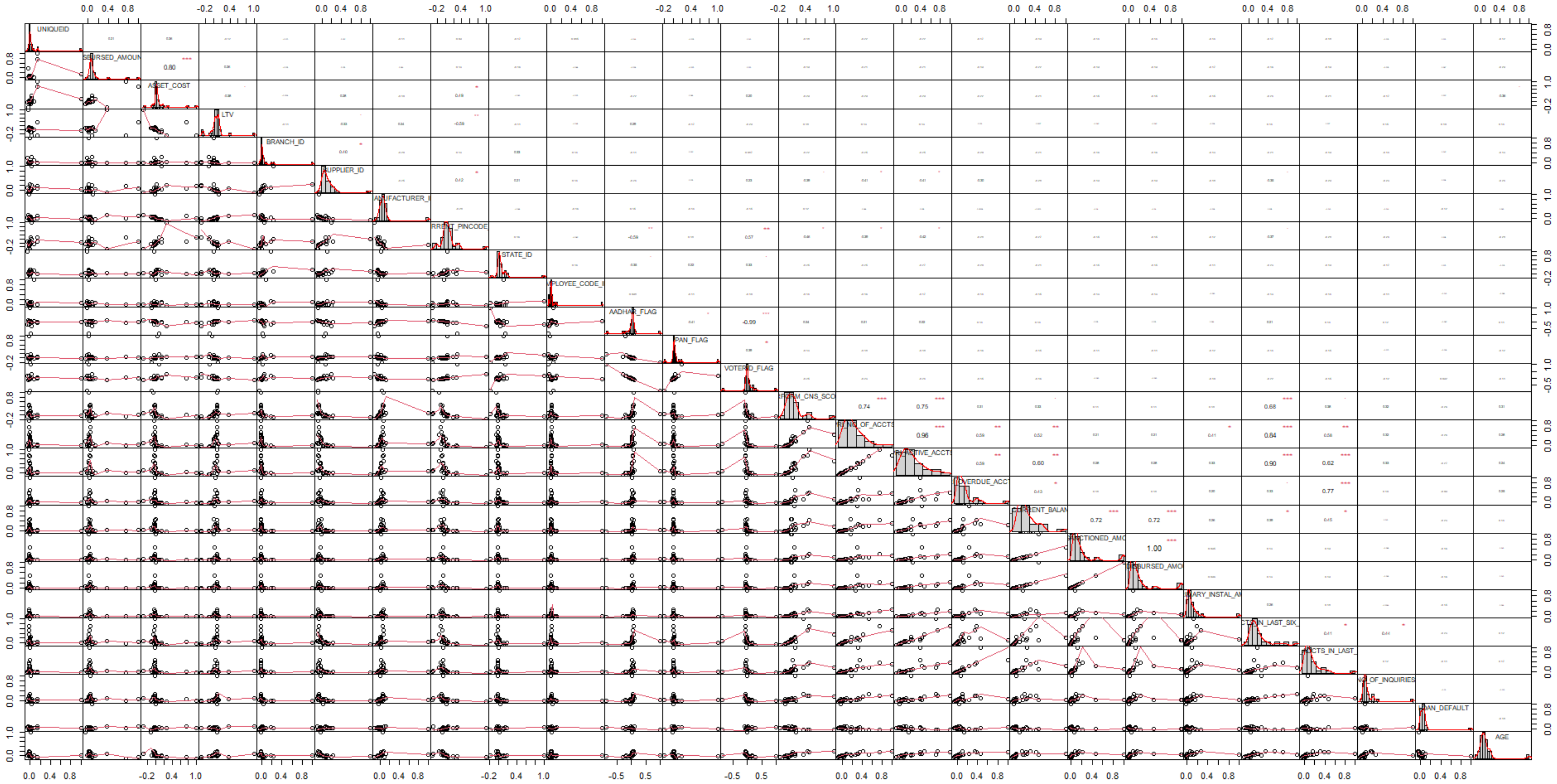
Vehicle Loan Data Set (RAW DATA) Internal Structure/Statistics

30 Possible Predictors:

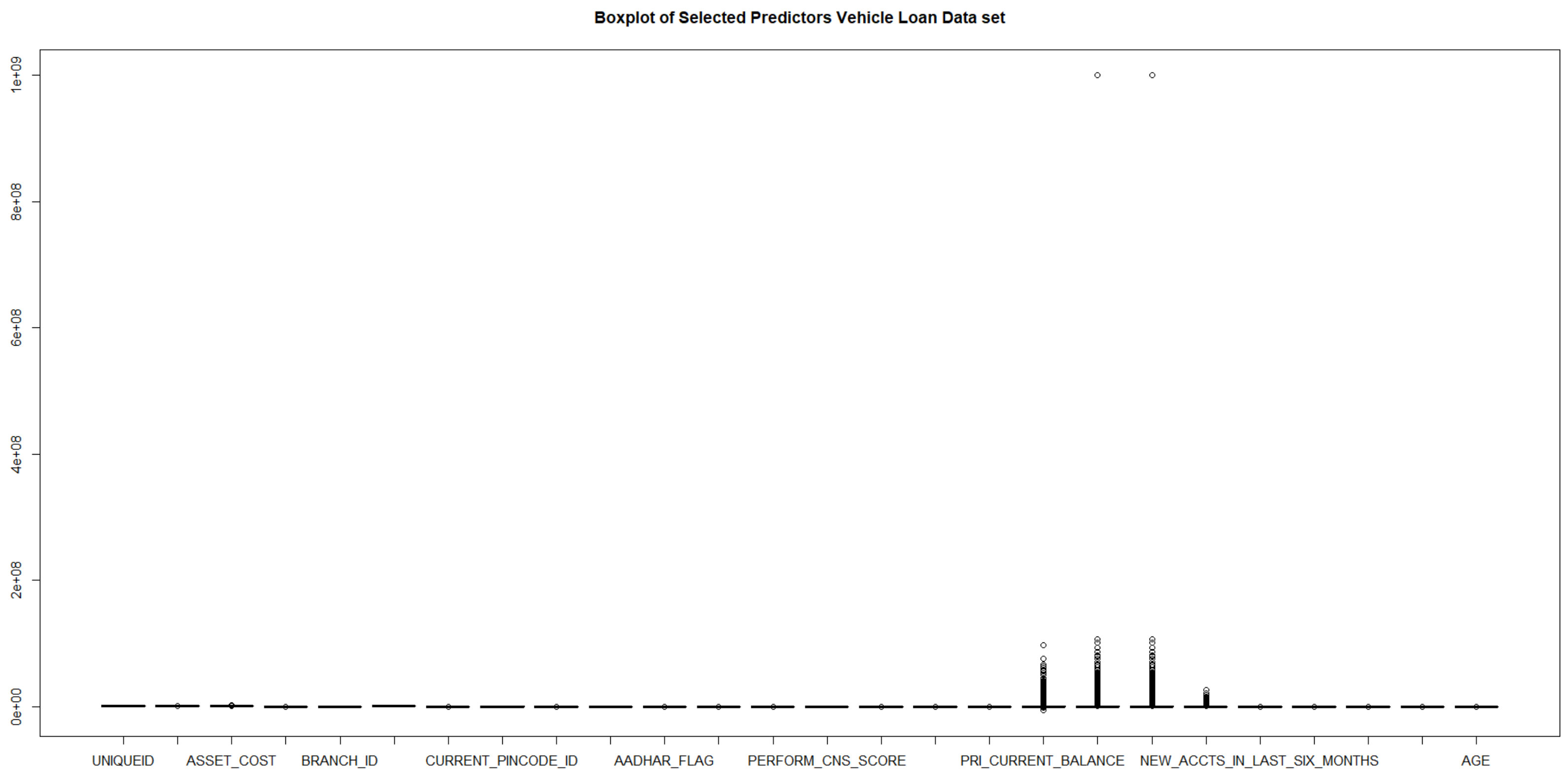
Response/Outcome: Default or Not Default

```
> str(df_nzv) # Displays new_df_nzv_VehicleLoan.csv data set internal structure
'data.frame':	233154 obs. of  30 variables:
 $ UNIQUEID                           : int  420825 537409 417566 624493 539055 518279 529269 510278 490213 510980 ...
 $ DISBURSED_AMOUNT                   : int  50578 47145 53278 57513 52378 54513 46349 43894 53713 52603 ...
 $ ASSET_COST                         : int  58400 65550 61360 66113 60300 61900 61500 61900 61973 61300 ...
 $ LTV                                : num  89.5 73.2 89.6 88.5 88.4 ...
 $ BRANCH_ID                          : int  67 67 67 67 67 67 67 67 67 67 ...
 $ SUPPLIER_ID                        : int  22807 22807 22807 22807 22807 22807 22807 22807 22807 22807 ...
 $ MANUFACTURER_ID                    : int  45 45 45 45 45 45 45 45 45 45 ...
 $ CURRENT_PINCODE_ID                 : int  1441 1502 1497 1501 1495 1501 1502 1501 1497 1492 ...
 $ EMPLOYMENT_TYPE                    : chr  "Salaried" "Self employed" "Self employed" "Self employed" ...
 $ STATE_ID                           : int  6 6 6 6 6 6 6 6 6 6 ...
 $ EMPLOYEE_CODE_ID                   : int  1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
 $ AADHAR_FLAG                        : int  1 1 1 1 1 1 1 1 1 0 ...
 $ PAN_FLAG                           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ VOTERID_FLAG                       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ PERFORM_CNS_SCORE                  : int  0 598 0 305 0 825 0 17 718 818 ...
 $ PERFORM_CNS_SCORE_DESCRIPTION      : chr  "No Bureau History Available" "I-Medium Risk" "No Bureau History Available" "L-Very High Risk" ...
 $ PRI_NO_OF_ACCTS                    : int  0 1 0 3 0 2 0 1 1 1 ...
 $ PRI_ACTIVE_ACCTS                   : int  0 1 0 0 0 0 0 1 1 0 ...
 $ PRI_OVERDUE_ACCTS                  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ PRI_CURRENT_BALANCE                : int  0 27600 0 0 0 0 0 72879 -41 0 ...
 $ PRI_SANCTIONED_AMOUNT              : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ PRI_DISBURSED_AMOUNT               : int  0 50200 0 0 0 0 0 74500 365384 0 ...
 $ PRIMARY_INSTAL_AMT                 : int  0 1991 0 31 0 1347 0 0 0 2608 ...
 $ NEW_ACCTS_IN_LAST_SIX_MONTHS       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS: int  0 1 0 0 0 0 0 0 0 0 ...
 $ NO_OF_INQUIRIES                    : int  0 0 0 1 1 0 0 0 1 0 ...
 $ LOAN_DEFAULT                       : int  0 1 0 1 1 0 0 0 0 0 ...
 $ formatted_DATE_OF_BIRTH            : Date, format: "1984-01-01" "1985-07-31" "1985-08-24" "1993-12-30" ...
 $ formatted_DISBURSAL_DATE           : Date, format: "2018-03-08" "2018-09-26" "2018-01-08" "2018-10-26" ...
 $ AGE                                : num  39 38 38 30 46 33 35 34 32 55 ...
```

Correlation Plot: If no stars, the variable is NOT statistically significant, while one, two and three stars mean that the corresponding variable is significant at 10%, 5% and 1% levels, respectively)



Vehicle Loan Data Set (RAW DATA): Identifying Outliers Using Boxplot

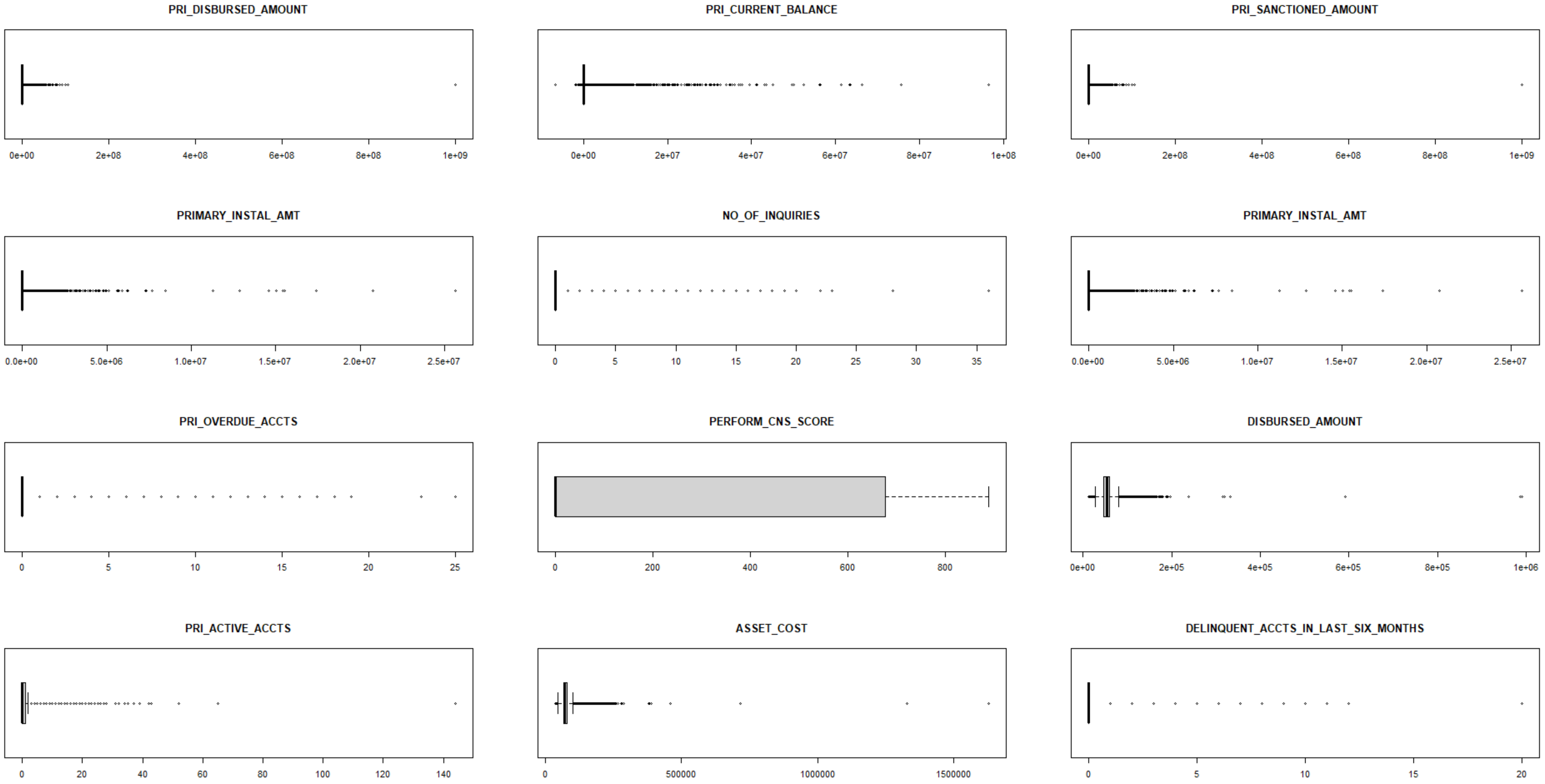


Why is it important to identify Outliers?

It is important to identify outliers in a dataset in R (or any other statistical software) because they can have a significant impact on the statistical analysis and modeling results. Outliers are data points that are significantly different from other observations in the dataset, and they can be caused by various factors such as measurement errors, data entry errors, or rare events.

The boxplot above does show example of variables with outliers: PRI_OVERDUE_ACCTS, PRIMARY_INSTAL_AMT, as well as other variables (i.e., NO_OF_INQUIRIES, DISBURSED_AMOUNT, ASSET_COST, etc.)

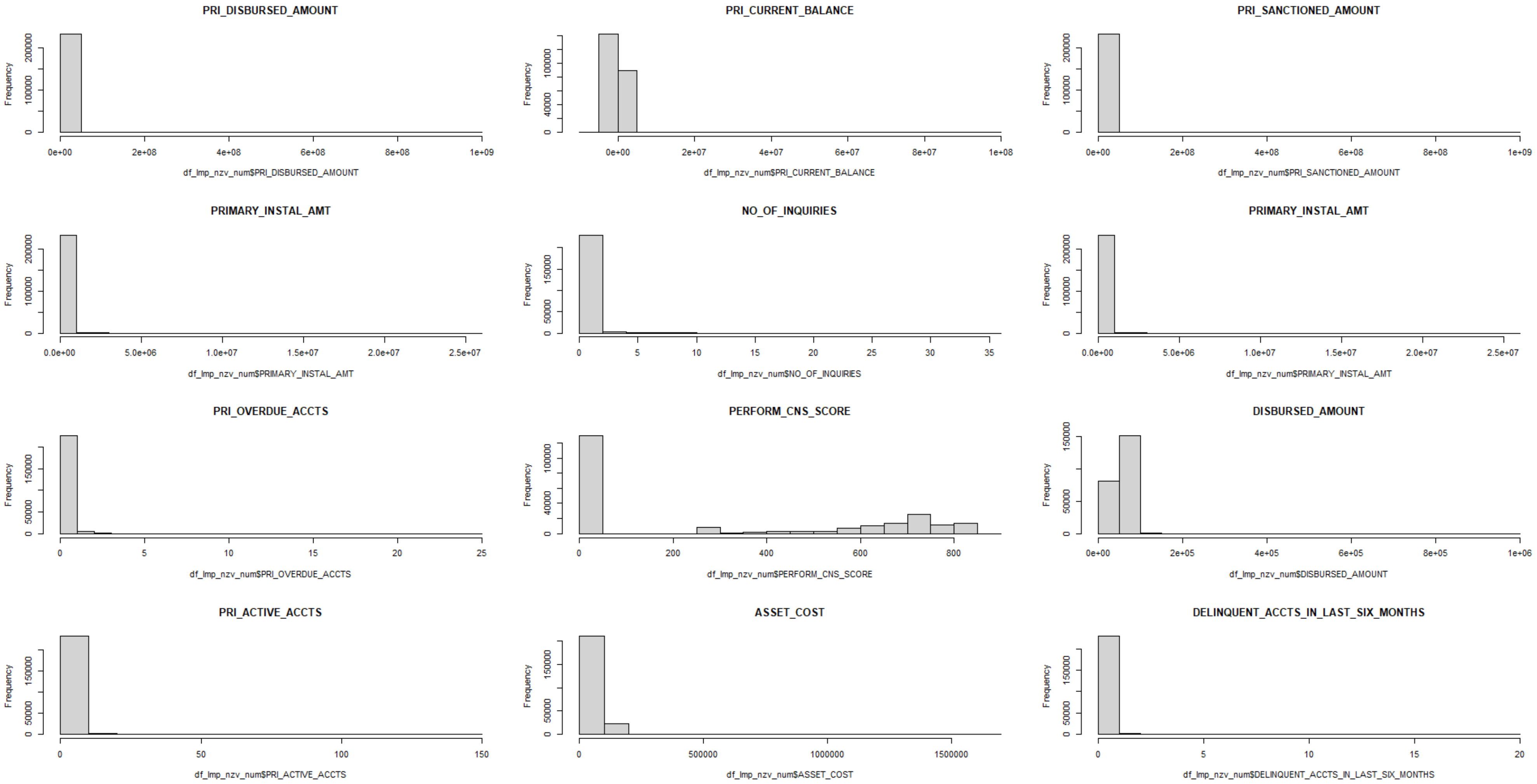
Boxplot of Selected Predictor Variables



There appears to be outliers in the data set, and some are skewed. Clearly, the dots above/below the ends of the "whiskers" in the boxplots are indications of outliers.

HIGH-LEVEL SUMMARY (CLEANED-UP DATA SET)

Histograms of Selected Predictor Variables

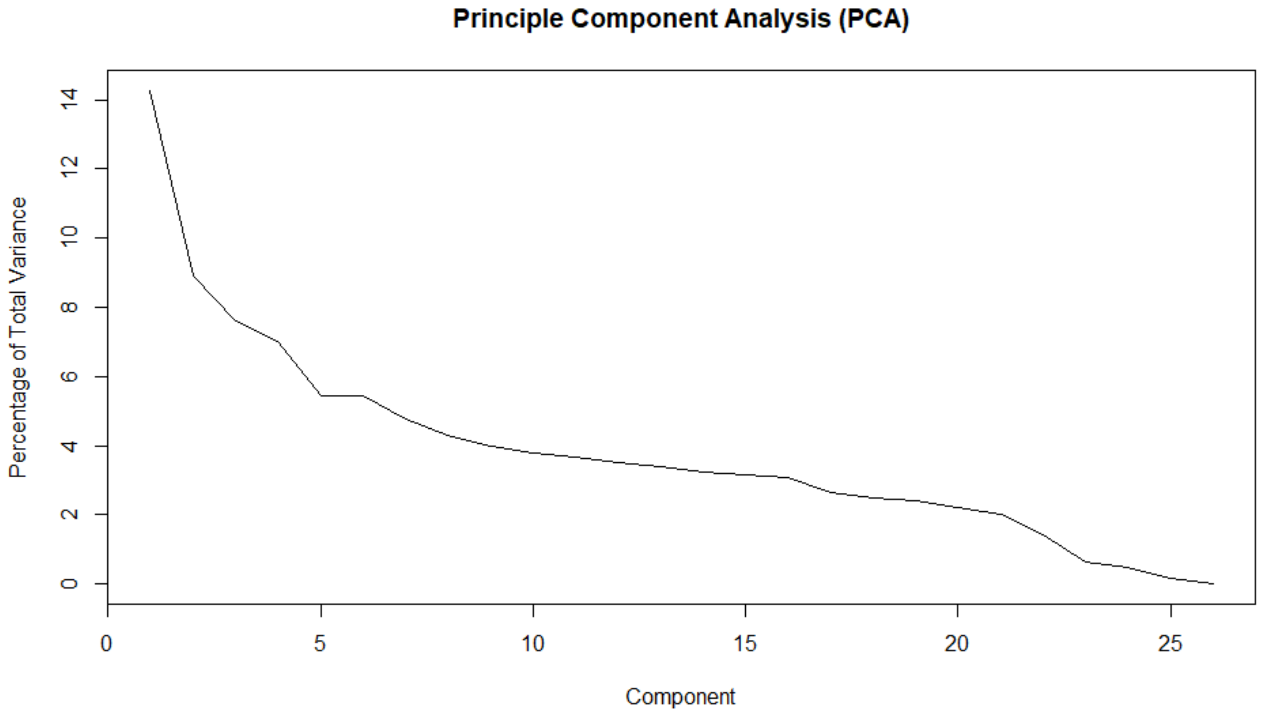


There appears to be outliers in the data set, and most of the data points are skewed to the right, which are indicators of outliers. The PERFORM_CNS_SCORE appears to be bi-modal.

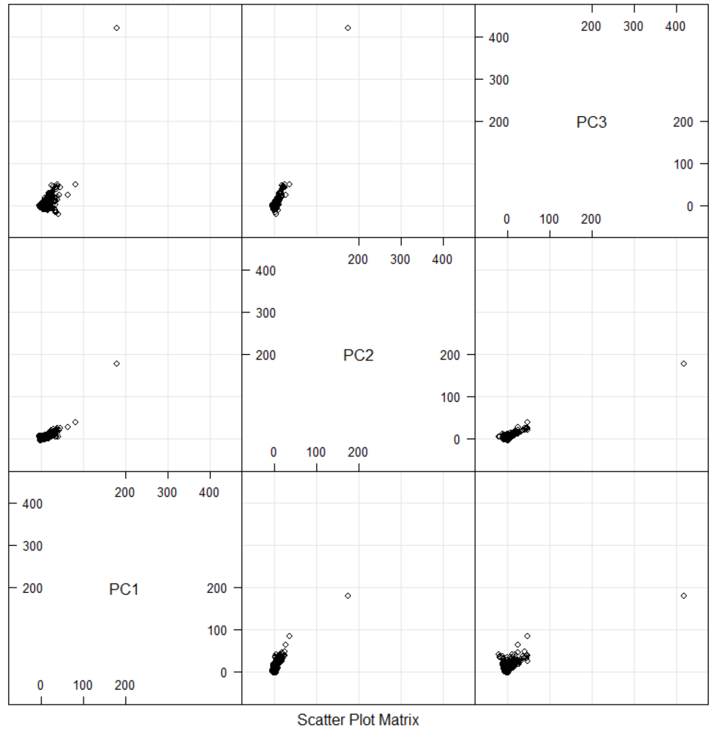
DATA PRE-PROCESSING: BOX–COX TRANSFORM, CENTER, AND SCALE

```
> # Administration of a series of transformation (i.e. trans) to the data set
>
> ## Use caret's preProcess function to transform for skewness
> # preProcess estimates the transformation (centering, scaling etc.)
> # function from the training data and can be applied to any data set with
> # the same variables.
> df_Imp_nzv_PP <- preProcess(df_ImpTrain, c("BoxCox", "center", "scale"))
> df_Imp_nzv_PP
Created from 186475 samples and 30 variables

Pre-processing:
  - Box-Cox transformation (11)
  - centered (26)
  - ignored (4)
  - scaled (26)

Lambda estimates for Box-Cox transformation:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.3000  0.0000  0.3000  0.3727  0.7000  2.0000

>
```

DATA PRE-PROCESSING: PCA



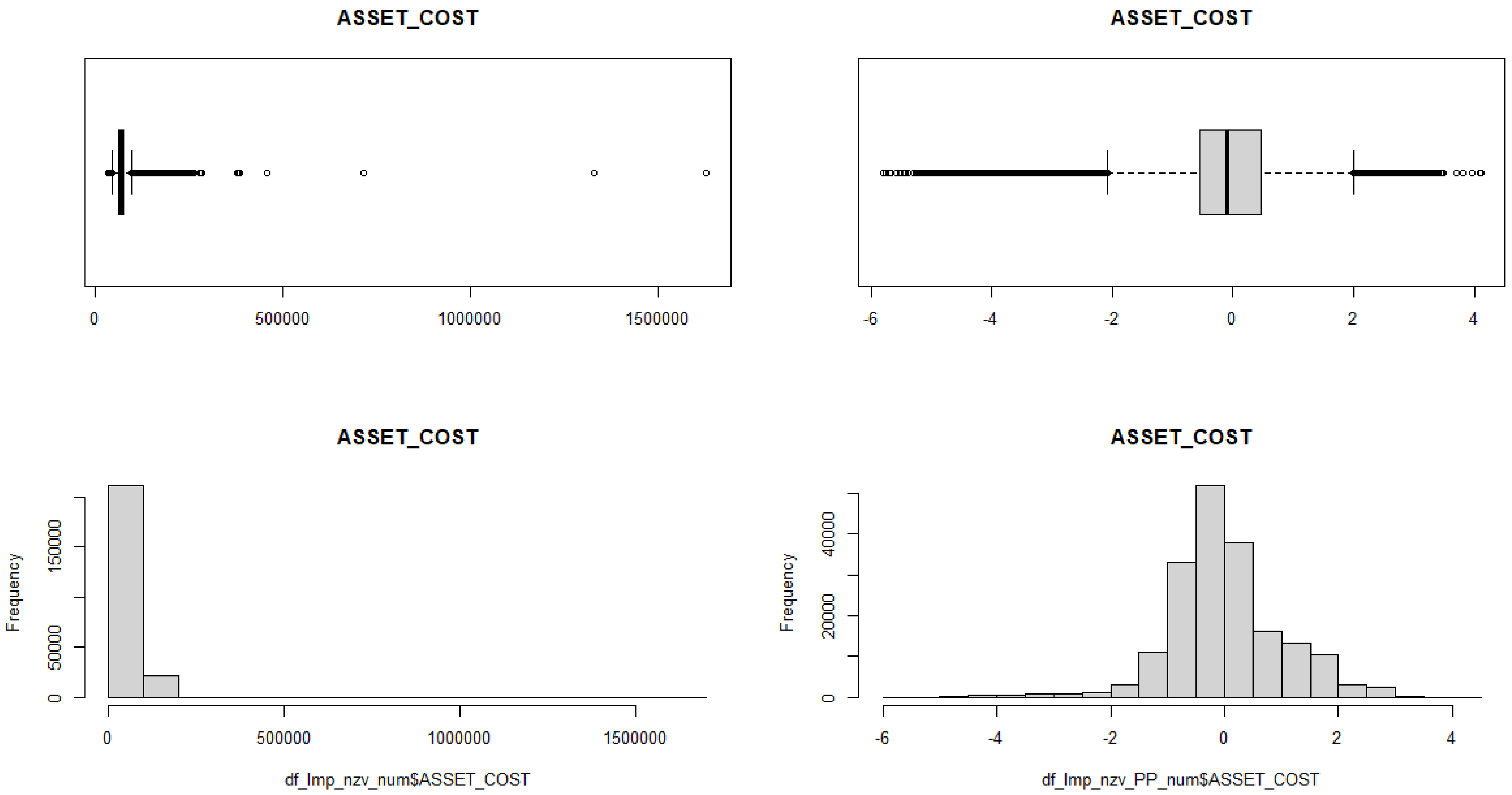
Scatter Plot Matrix of the First Three Components




```
> ## compute the percentage of variance for each component
> percentVariancePCA = df_Imp_nzv_PCA$sd^2/sum(df_Imp_nzv_PCA$sd^2)*100
>
> percentVariancePCA = df_Imp_nzv_PCA$sd^2/sum(df_Imp_nzv_PCA$sd^2)*100
>
> percentVariancePCA[1:3]   # first 3 components account for 31% of variance
[1] 14.273330  8.888238  7.595638
> plot(percentVariancePCA, xlab="Component", ylab="Percentage of Total Variance", type="l",
+      main="Principle Component Analysis (PCA)")
>
```

```
> ## show the transformed values
> head(df_Imp_nzv_PCA$x[,1:3])
          PC1        PC2         PC3
1 -0.71410465 -1.4982547  0.90352717
2  0.90944704 -0.4822288  0.02066826
3 -0.74952060 -1.3416952  0.80006689
4 -0.35645181 -0.7659506 -0.11543563
5 -0.54788798 -1.1733791  0.37532743
6 -0.01628318 -1.2710953  0.19795584
>
```

Comparison of Box plots & Histograms BEFORE & AFTER Clean-Up



❑ Predictive Models: Linear Classification Models

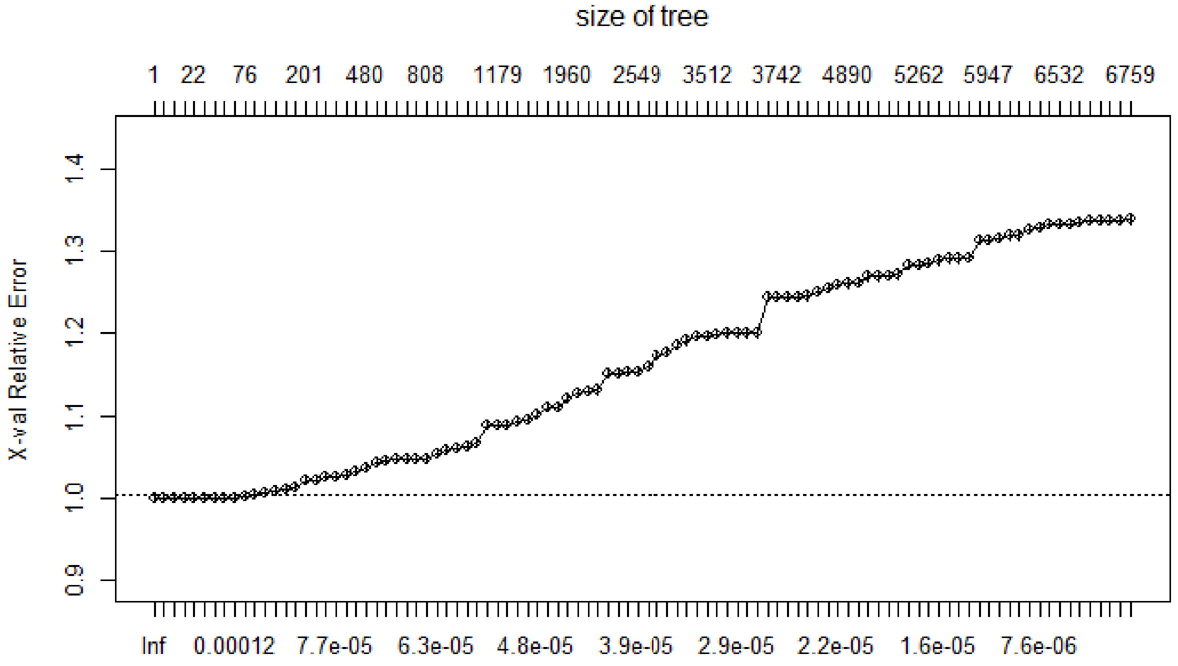
- Classification Tree
- Logistic Regression
- Random Forest

Predictive Models: Linear Classification Models

Classification And Regression Tree (CART) – **Decision Tree**

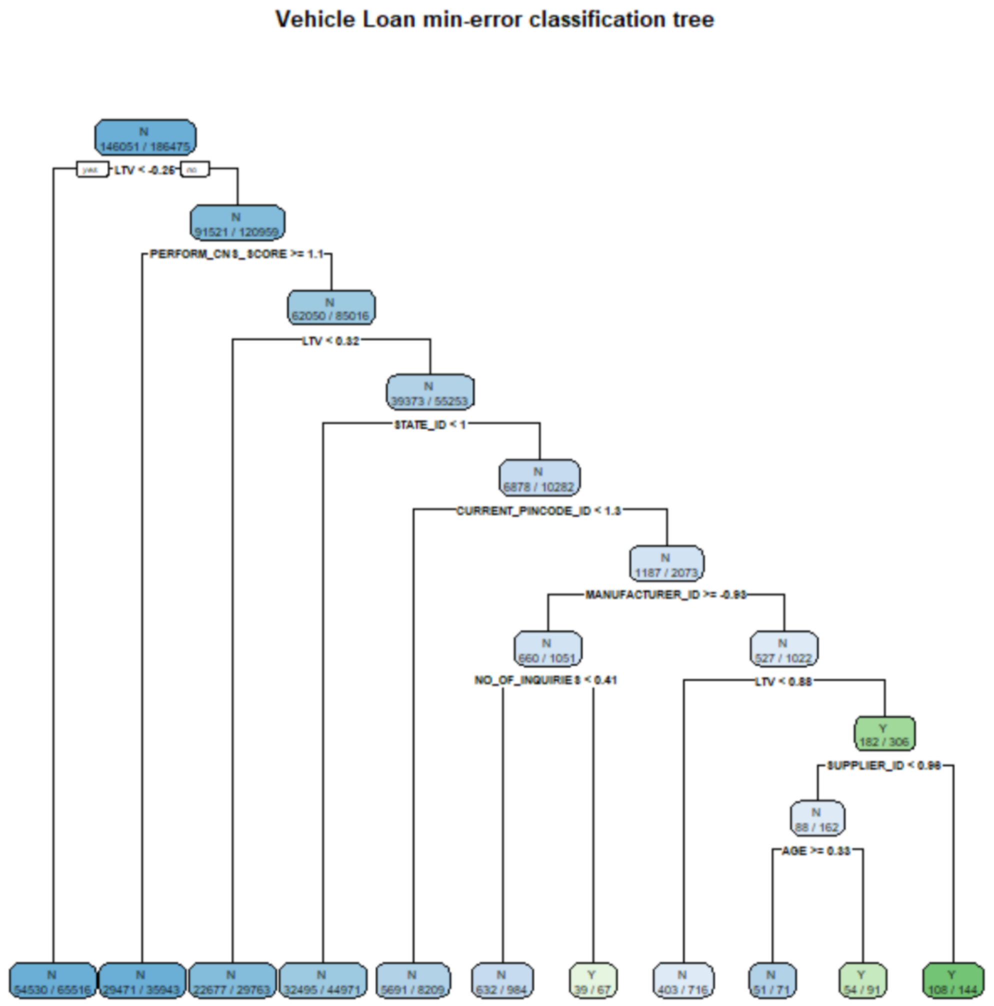
Fitting Classification Tree Using rpart

```
Classification tree:
rpart(formula = DEFAULT ~ . - LOAN_DEFAULT, data = vl_Traindf,
    cp = 1e-13)

Variables actually used in tree construction:
 [1] AADHAR_FLAG                        AGE
 [3] ASSET_COST                         BRANCH_ID
 [5] CURRENT_PINCODE_ID                 DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS
 [7] DISBURSED_AMOUNT                   LTV
 [9] MANUFACTURER_ID                    NEW_ACCTS_IN_LAST_SIX_MONTHS
[11] NO_OF_INQUIRIES                    PAN_FLAG
[13] PERFORM_CNS_SCORE                  PRI_ACTIVE_ACCTS
[15] PRI_DISBURSED_AMOUNT               PRI_NO_OF_ACCTS
[17] PRI_SANCTIONED_AMOUNT              PRIMARY_INSTAL_AMT
[19] STATE_ID                           SUPPLIER_ID
[21] UNIQUEID                           VOTERID_FLAG

Root node error: 40424/186475 = 0.21678

n= 186475

           CP nsplit rel error  xerror      xstd
1  2.0497e-04      0   1.00000 1.00000 0.0044017
2  1.7316e-04     10   0.99753 0.99973 0.0044013
3  1.5255e-04     11   0.99735 0.99970 0.0044012
4  1.4843e-04     20   0.99594 0.99963 0.0044011
5  1.4018e-04     21   0.99579 0.99926 0.0044005
6  1.2864e-04     25   0.99518 0.99983 0.0044014
7  1.2369e-04     30   0.99453 0.99983 0.0044014
8  1.1819e-04     41   0.99263 1.00017 0.0044020
9  1.1309e-04     68   0.98909 1.00025 0.0044021
10 1.1132e-04     75   0.98830 1.00079 0.0044030
11 9.8951e-05     85   0.98716 1.00272 0.0044060
12 9.4004e-05    110   0.98444 1.00562 0.0044106
13 9.0705e-05    116   0.98380 1.00804 0.0044145
14 8.6582e-05    122   0.98325 1.01044 0.0044182
15 8.2459e-05    197   0.97558 1.01247 0.0044214
16 8.0398e-05    200   0.97534 1.02165 0.0044358
17 7.8336e-05    204   0.97501 1.02165 0.0044358
18 7.6962e-05    214   0.97403 1.02476 0.0044406
```



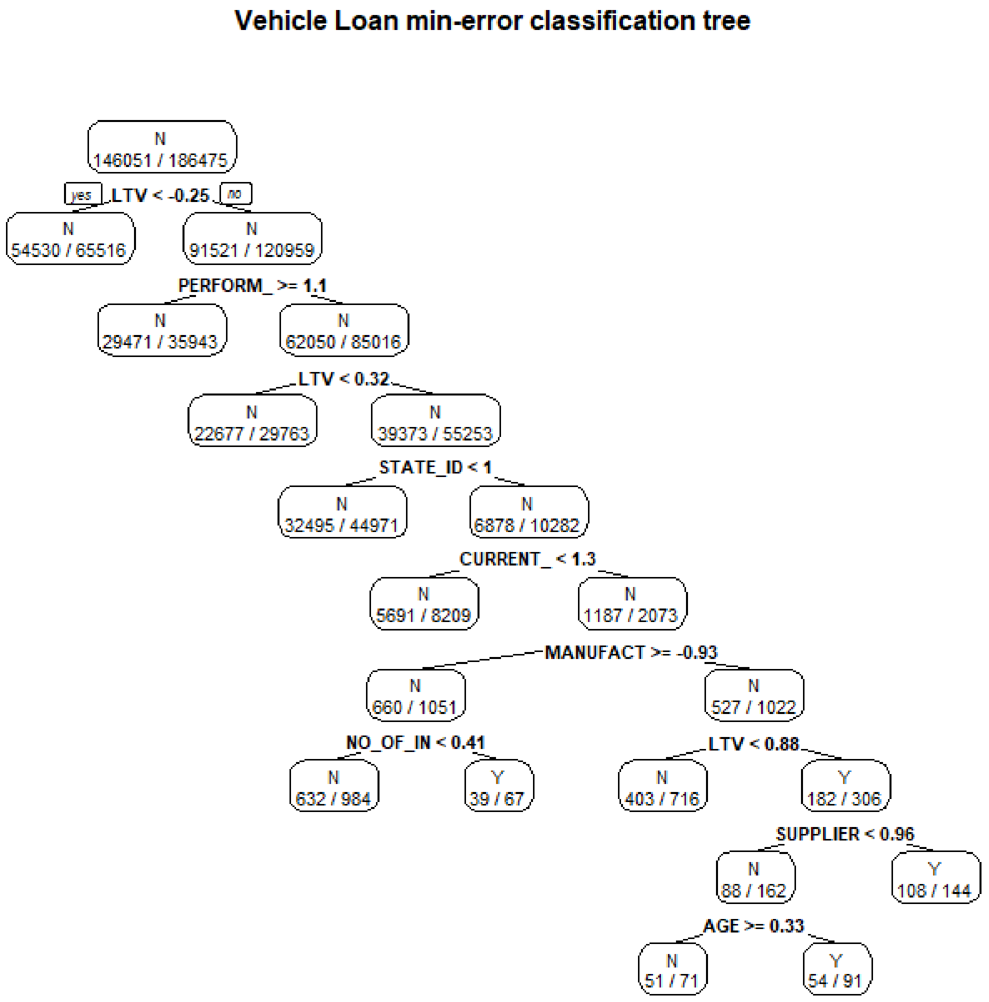
- CART – **Decision Tree**
  - **Fitting Classification Tree Using rpart**

```
> rpart.vl_Traindf <- prune(rpart.vl_Traindf, cp=0.0002)
> printcp(rpart.vl_Traindf)

Classification tree:
rpart(formula = DEFAULT ~ . - LOAN_DEFAULT, data = vl_Traindf,
    cp = 1e-13)

Variables actually used in tree construction:
[1] AGE                CURRENT_PINCODE_ID LTV                MANUFACTURER_ID
[5] NO_OF_INQUIRIES    PERFORM_CNS_SCORE  STATE_ID           SUPPLIER_ID

Root node error: 40424/186475 = 0.21678

n= 186475

          CP nsplit rel error  xerror      xstd
1 0.00020497      0   1.00000 1.00000 0.0044017
2 0.00020000     10   0.99753 0.99973 0.0044013
```



- CART – **Decision Tree**
  - **Fitting CART tree model prp() function**

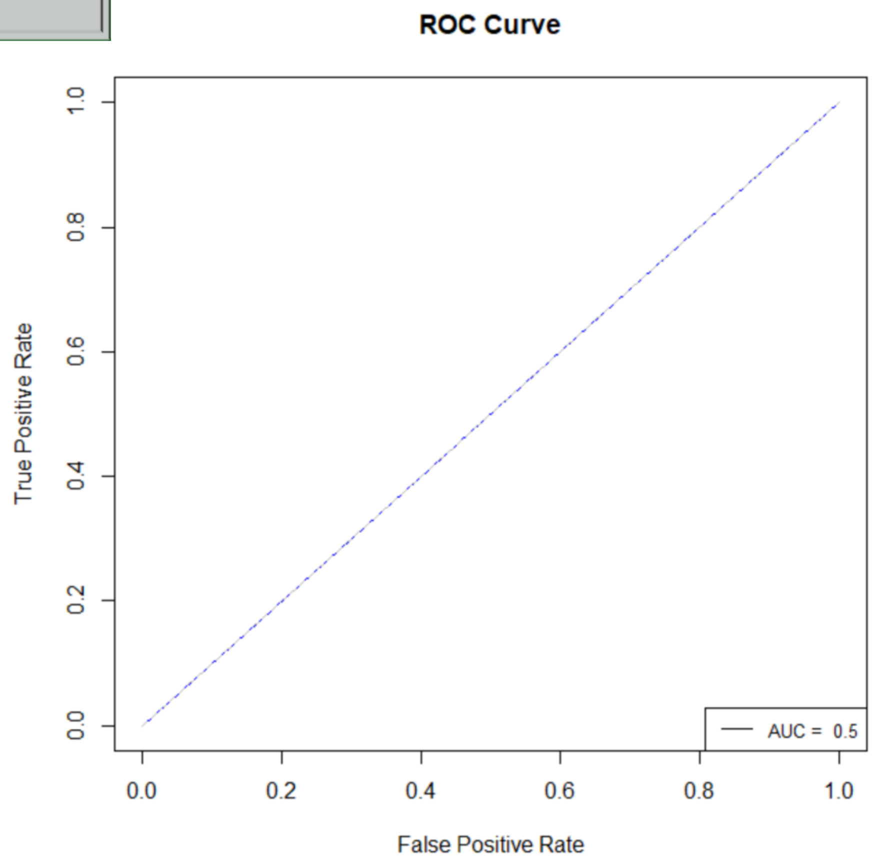
```
> ## ii) CART Tree Model (The use of prp() function)
> # Display in the plot window the pruned tree (that is, the
> # minimum-error tree).
> prp(rpart.vl_Traindf, type =2, extra=2,
+     main="Vehicle Loan min-error classification tree")
> rpart.vl_Traindf
n= 186475

node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 186475 40424 N (0.7832203 0.2167797)
     2) LTV< -0.2507237 65516 10986 N (0.8323158 0.1676842) *
     3) LTV≥-0.2507237 120959 29438 N (0.7566283 0.2433717)
       6) PERFORM_CNS_SCORE≥1.051031 35943  6472 N (0.8199371 0.1800629) *
       7) PERFORM_CNS_SCORE< 1.051031 85016 22966 N (0.7298626 0.2701374)
        14) LTV< 0.3237075 29763  7086 N (0.7619192 0.2380808) *
        15) LTV≥0.3237075 55253 15880 N (0.7125948 0.2874052)
          30) STATE_ID< 1.00897 44971 12476 N (0.7225768 0.2774232) *
          31) STATE_ID≥1.00897 10282  3404 N (0.6689360 0.3310640)
            62) CURRENT_PINCODE_ID< 1.289521 8209  2518 N (0.6932635 0.3067365) *
            63) CURRENT_PINCODE_ID≥1.289521 2073   886 N (0.5726001 0.4273999)
             126) MANUFACTURER_ID≥-0.9315015 1051   391 N (0.6279734 0.3720266)
               252) NO_OF_INQUIRIES< 0.4143625 984   352 N (0.6422764 0.3577236) *
               253) NO_OF_INQUIRIES≥0.4143625 67    28 Y (0.4179104 0.5820896) *
             127) MANUFACTURER_ID< -0.9315015 1022   495 N (0.5156556 0.4843444)
               254) LTV< 0.8814643 716   313 N (0.5628492 0.4371508) *
               255) LTV≥0.8814643 306   124 Y (0.4052288 0.5947712)
                 510) SUPPLIER_ID< 0.9613553 162    74 N (0.5432099 0.4567901)
                  1020) AGE≥0.3256264 71    20 N (0.7183099 0.2816901) *
                  1021) AGE< 0.3256264 91    37 Y (0.4065934 0.5934066) *
                 511) SUPPLIER_ID≥0.9613553 144    36 Y (0.2500000 0.7500000) *
>
```



- CART – **Decision Tree**
  - **Confusion Matrix, ROC Curve & AUC**

Accuracy	CI	P-Value	Mcnemar's Test P-Value	Sensitivity
0.78	95%	0.5	<2e-16	1.0

```
> # 9.9.2
> # Evaluate the model's accuracy using a confusion matrix
> conf_mat <- confusionMatrix(my_data3[my_data3$type == "prediction",1],
+                             my_data3[my_data3$type == "real",1],
+                             dnn = c("Prediction", "Reference"))
> conf_mat
Confusion Matrix and Statistics

          Reference
Prediction     N     Y
         N 36492 10187
         Y     0     0

               Accuracy : 0.7818
                 95% CI : (0.778, 0.7855)
    No Information Rate : 0.7818
    P-Value [Acc > NIR] : 0.5027

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.7818
         Neg Pred Value :    NaN
             Prevalence : 0.7818
         Detection Rate : 0.7818
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : N
```



- CART – **Decision Tree**
  - **Model's Performance Statistics**

```
> ##
> # Get the model's performance statistics
> perf <- list(
+   Accuracy = conf_mat$overall['Accuracy'],
+   Precision = conf_mat$byClass['Precision'],
+   Recall = conf_mat$byClass['Recall'],
+   F1 = conf_mat$byClass['F1']
+ )
>
> # Print the performance statistics
> print(perf)
$Accuracy
 Accuracy
0.7817648

$Precision
Precision
0.7817648

$Recall
Recall
     1

$F1
       F1
0.8775174

> # Calculate AUC
> auc_val <- performance(roc_obj, "auc")@y.values[[1]]
> auc_val
[1] 0.5
```

Accuracy	Precision	Recall	F1	AUC
0.78	0.78	1.00	0.88	0.5

Predictive Models: Linear Classification Models

- Logistic Regression

Accuracy	Precision	Recall	RMSE	R-squared
0.7817	0.7818	0.9984	0.4044	0.0415

Confusion Matrix		
Prediction	0	1
0	36434	10128
1	58	59

```
> print(confusion_matrixLog$table)
          Reference
Prediction     0     1
         0 36434 10128
         1    58    59
> #
> # calculate RMSE and R-squared
> rmseLog <- sqrt(mean((predictionsLog - df_ImpTest$LOAN_DEFAULT)^2))
> rsquaredLog <- cor(predictionsLog, df_ImpTest$LOAN_DEFAULT)^2
>
> # print results
> print(paste0("Accuracy: ", round(accuracyLog, 4)))
[1] "Accuracy: 0.7818"
> print(paste0("Precision: ", round(precisionLog, 4)))
[1] "Precision: 0.7825"
> print(paste0("Recall: ", round(recallLog, 4)))
[1] "Recall: 0.9984"
> print(paste0("RMSE: ", round(rmseLog, 4)))
[1] "RMSE: 0.4044"
> print(paste0("R-squared: ", round(rsquaredLog, 4)))
[1] "R-squared: 0.0415"
```

Predictive Models: Linear Classification Models

- Random Forest with n=5

Accuracy	Precision	Recall	RMSE	R-squared
0.74	0.79	0.91	0.51	0.01

Confusion Matrix		
Prediction	0	1
0	33129	8628
1	3363	1559

```
> # print confusion matrix
> print(confusion_matrixrf$table)
          Reference
Prediction     0     1
         0 33129  8628
         1  3363  1559
> #
> # calculate RMSE and R-squared
> rmserf <- sqrt(mean((predicted_classesrf - df_ImpTest$LOAN_DEFAULT)^2))
> rsquaredrf <- cor(predicted_classesrf, df_ImpTest$LOAN_DEFAULT)^2
>
> # print evaluation metrics
> cat("Accuracy: ", round(accuracyrf, 2), "\n")
Accuracy:  0.74
> cat("Precision: ", round(precisionrf, 2), "\n")
Precision:  0.79
> cat("Recall: ", round(recallrf, 2), "\n")
Recall:  0.91
> cat("RMSE: ", round(rmserf, 2), "\n")
RMSE:  0.51
> cat("R-squared: ", round(rsquaredrf, 2), "\n")
R-squared:  0.01
```

The Random Forest model runs well, as it has predicted more accurately than the other n values.

Predictive Models: Linear Classification Models

- Random Forest continue…

N value	Accuracy	Precision	Recall	RMSE	R-squared
10	0.76	0.79	0.95	0.49	0.01
4	0.75	0.79	0.92	0.5	0
6	0.76	0.79	0.93	0.49	0.01

Reference N=10		
Prediction	0	1
0	34645	9218
1	1847	969

Reference N=4		
Prediction	0	1
0	33740	8945
1	2752	1242

Reference N=6		
Prediction	0	1
0	34114	9058
1	2378	1129

Predictive Models: Models Comparison

- Preprocessing: Box-Cox Transformation (11); Centering (26); Ignored (4); Scaled (26)
- Models: Classification Tree, Logistic Regression, Random Forest

Model	N-Value	Accuracy	Precision	Recall	F1	AUC	RMSE	R-Squared
LS		0.78	0.78	0.99	-	-	0.40	0.04
RF	5	0.74	0.79	0.91	-	-	0.51	0.01
	10	0.76	0.79	0.95	-	-	0.49	0.01
	4	0.75	0.79	0.92	-	-	0.5	0
	6	0.76	0.79	0.93	-	-	0.49	0.01
CART	186475	0.78	0.78	1.00	0.88	0.5	-	-

ROADBLOCKS

Risk Name	Description	Probability	Impact	Mitigation
NonZeroVariance	**data[, -nearZeroVar(data)]** & **nearZeroVar(data, saveMetrics = TRUE)**	Low	Low	Run in sequence
Confounding Variables	The Simpson's Paradox: the trend disappears with different combination groups.	Low	Low	Look for correlation not causation.
Modeling	Some predictive models prefer predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the model's numerical stability.	Low	Medium	PCA preprocessing creates new predictors with desirable characteristics; Meanwhile it delivers new predictors with desirable characteristics, it must be used with understanding and care.

ROADBLOCKS

Risk Name	Description	Probability	Impact	Mitigation
Data set	Initial data set – Organ Transplant Prediction for a Fe/Male as donor recipient was not a viable data set for the anticipated prediction. In order to predict for particular procedure, we need to have a patient level data set – which is not available. With limited aggregated data set – we can only predict aggregated information.	Medium	High	Instead, we could determine the trend of surgery over the span of years on the percentage of Fe/Male for a particular organ using a Time Series Model. Changed our data set.
Order of Operation/ Sequence of Coding	Coding for NonZeroVar threw several errors as we tried to understand the output. After several trials and observations – we came to a conclusion that: **data[, -nearZeroVar(data)]** function/method is performed on the entire dataset whilst nearZeroVar **(data, saveMetrics = TRUE)** performs nearZeroVar on Columns(predictors) and output corresponding Statistic Metrics.	Low	Medium	Separate the functions/methods and run.

Note: The initial risk as identified to be associated with the Organ Transplant dataset (OTD) has been mitigated with the replacement of the entire dataset with the Vehicle Loan data set. **Thus, there is zero impact to the project since the OTD no longer exist.**

CONCLUSION

The risks and mitigations have been thoroughly dealt with up until halfway through the project. As we proceeded to the viable predictive model, we paid close attention to confounding variables as we may locate the **Simpson's Paradox** as explained in the table.

From the entire output performance of each of the given methods in this project with the chosen data, it is apparent that results from preprocessing with **PCA** has a considerably low performance. On the contrary, the output results could be worse with PCA preprocessing.

Comparing the results in our previous slide (33), it is indicative that the Logistic Regression outperforms the Random Forest showing a larger R-Squared value of 0.04.

The CART model, the lower the cp (complex parameter) value the better the performance. The **Mcnemar's Test P-value** determines if there are differences on a dichotomous dependent variable between two related groups. A dichotomous variable is a categorical variables with two categories only – which is the loan default column for either a Yes (1) or No (0). Also, with a P-value of **<2e-16** which is less than **0.05** it is evident that there a significant difference between the categorical variables.

Based on the model comparison or in terms of model accuracy, CART and Logistics Regression models are the best.

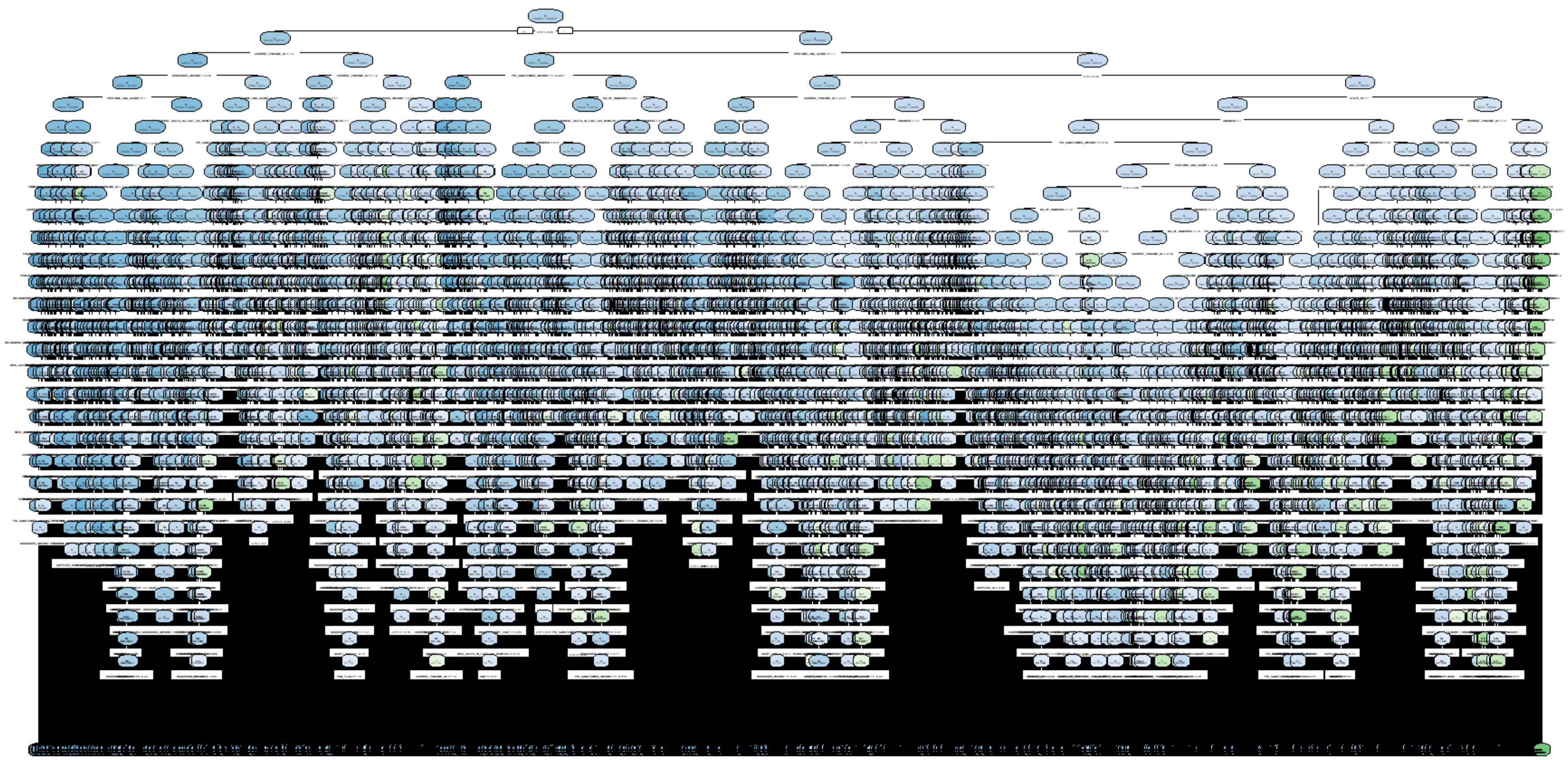
Back-up

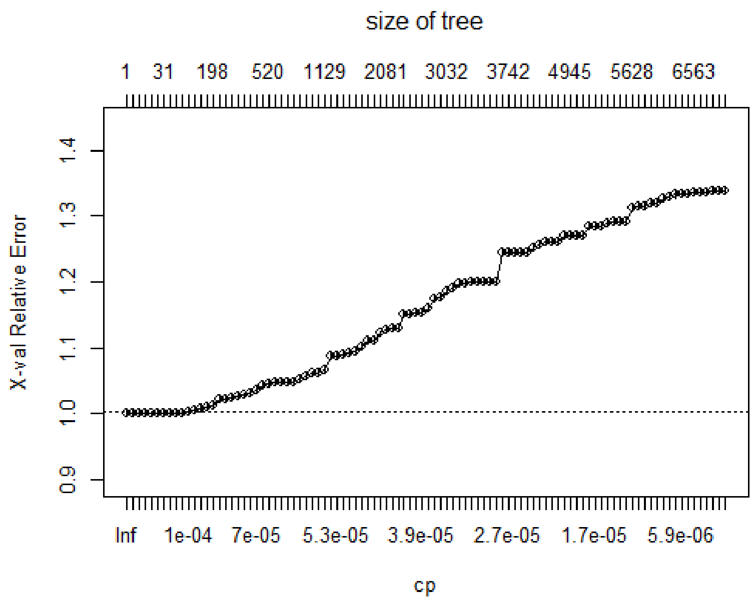
Source: Dr. Gang (DAEN 690 Project Report Template)

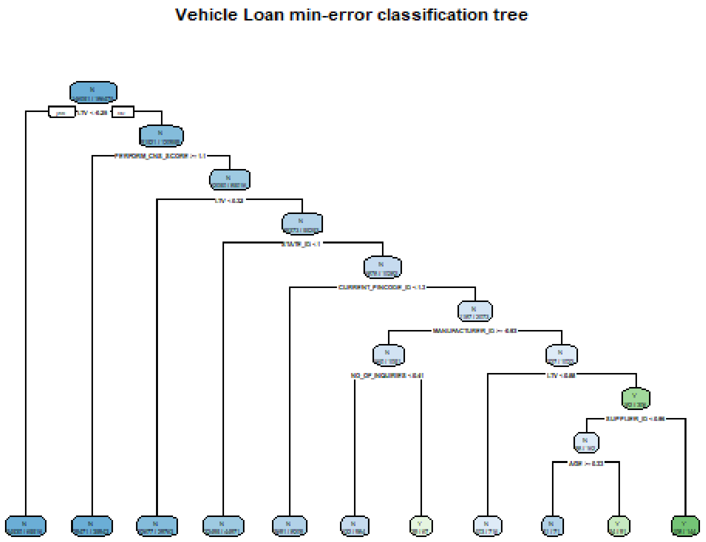
Predictive Models: Linear Classification Models

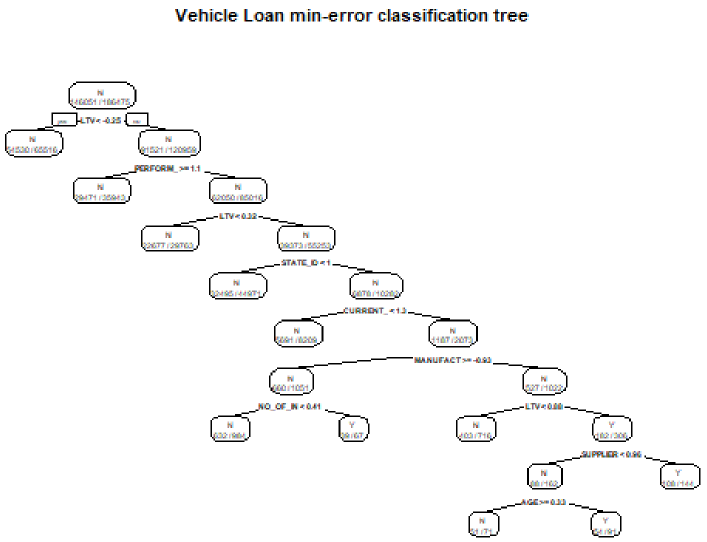
- A smaller cp value (example: cp=0.0000000000001) will result in a larger tree with more splits, which lead to overfitting.

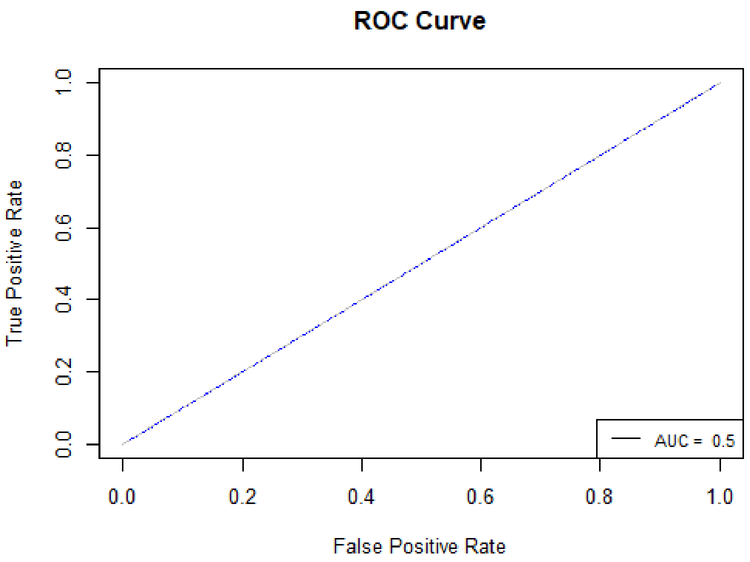
Vehicle Loan min-error classification tree











Source: Dr. Gang (DAEN 690 Project Report Template)