Credit EDA Case Study

PGC Data Science

Submitted By: Ashish Sharma (CDS2010025)

Date : 22nd July 2020

Introduction & Problem Statement

TASK 1

Apply EDA

Use EDA in a real business scenario of risk analytics in banking

TASK 2

Business Understanding

Applicants capable of repaying the loan should not get rejected.

TASK 3

Business Objectives

Identify patterns which indicate if a client has difficulty in repayment

TASK 4

Output

Report the driving factors (Variables) and their impact on loan default

Conclusion :: Default Driving Factors :: Strong Indicators

Service Tenure
Var-DAYS_EMPLOYED
Negative Correlation 0.075

6

1

Region
Var- REGION_RATING_CLIENT
Positive Correlation 0.059

Age
Var-DAYS_BIRTH
Negative Correlation 0.078

5

Correlation with TARGET Variable
TARGET = 0 (Non Default)
TARGET = 1(Default)

2

City
Var-REGION_RATING_CLIENT_W_CITY
Positive Correlation 0.061

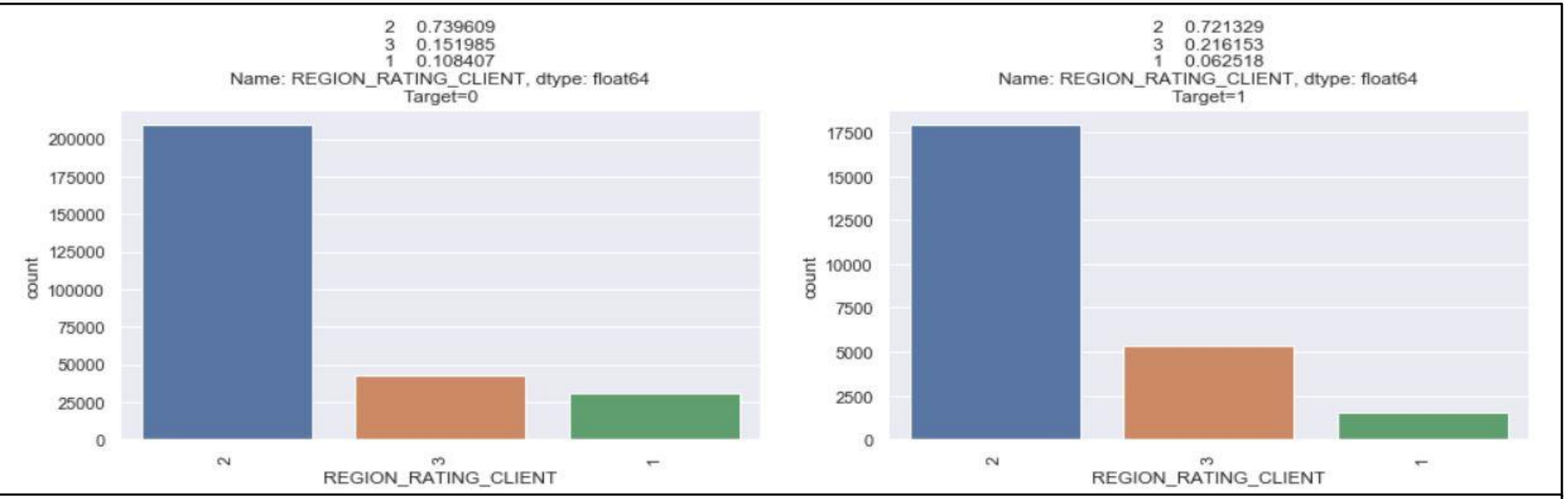
External Agency Score
Var-EXT_SOURCE_3
Negative Correlation 0.179

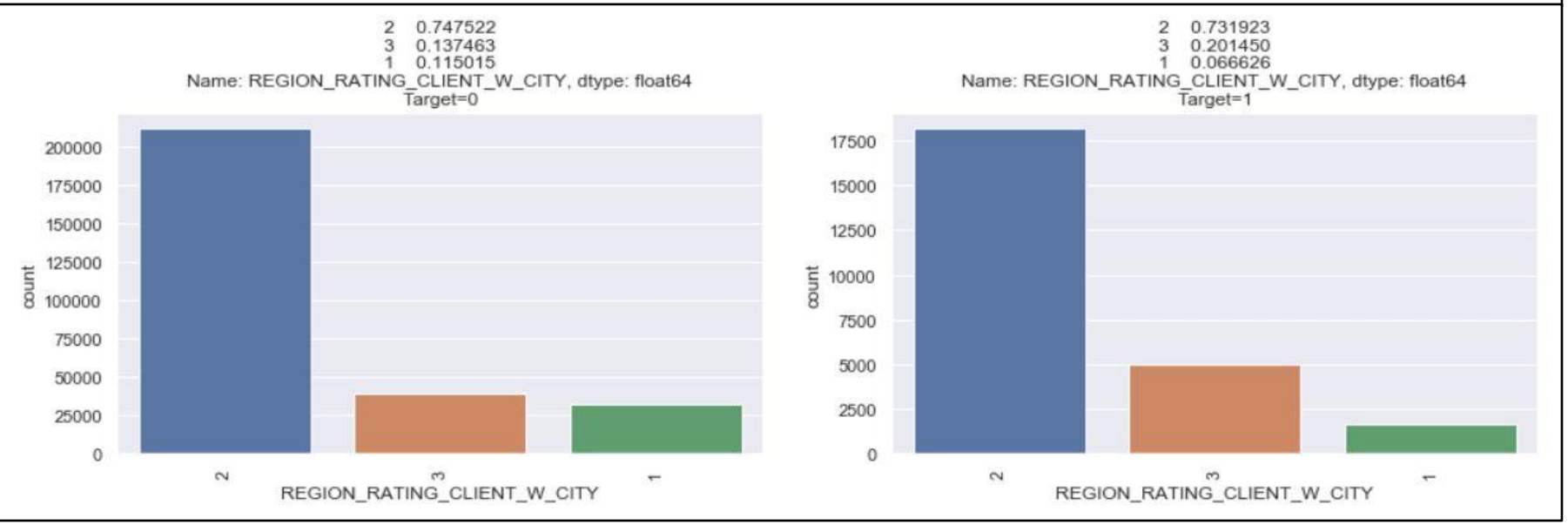
4

3

External Agency Score
Var-EXT_SOURCE_2
Negative Correlation 0.160

Default Driving Factor :: Strong Indicators





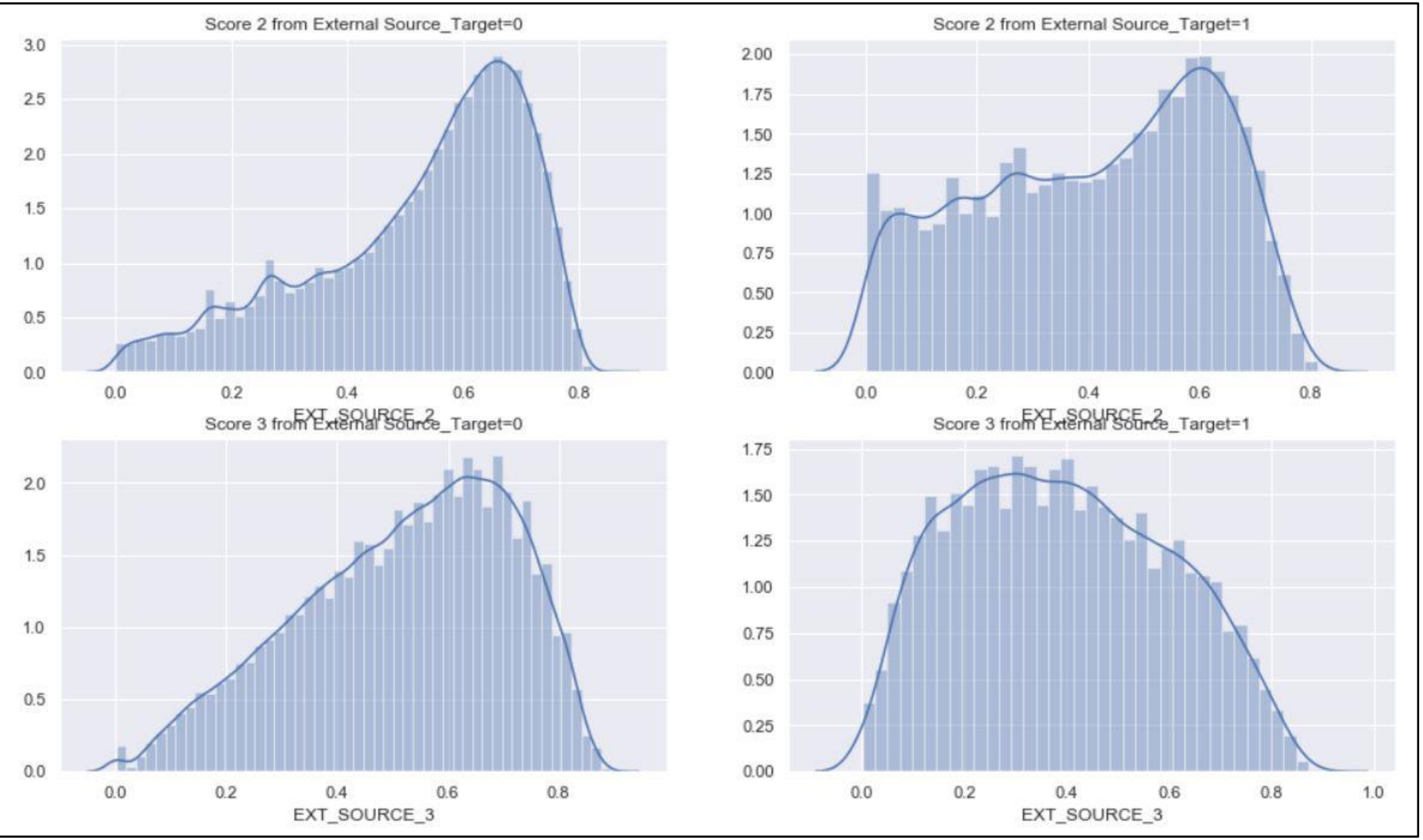
1. Region

As it is clearly indicated in the bar chart that for loan default the % increase in applicant residing in region with rating 3 is ~6.5%. This indicates that applicants residing in region with rating 3 are more likely to default on loan repayment.

2. City

As it is clearly indicated in the bar chart that for loan default the % increase in applicant residing in city with rating 3 is ~6.3%. This indicates that applicants residing in city with rating 3 are more likely to default on loan repayment.

Default Driving Factor :: Strong Indicators



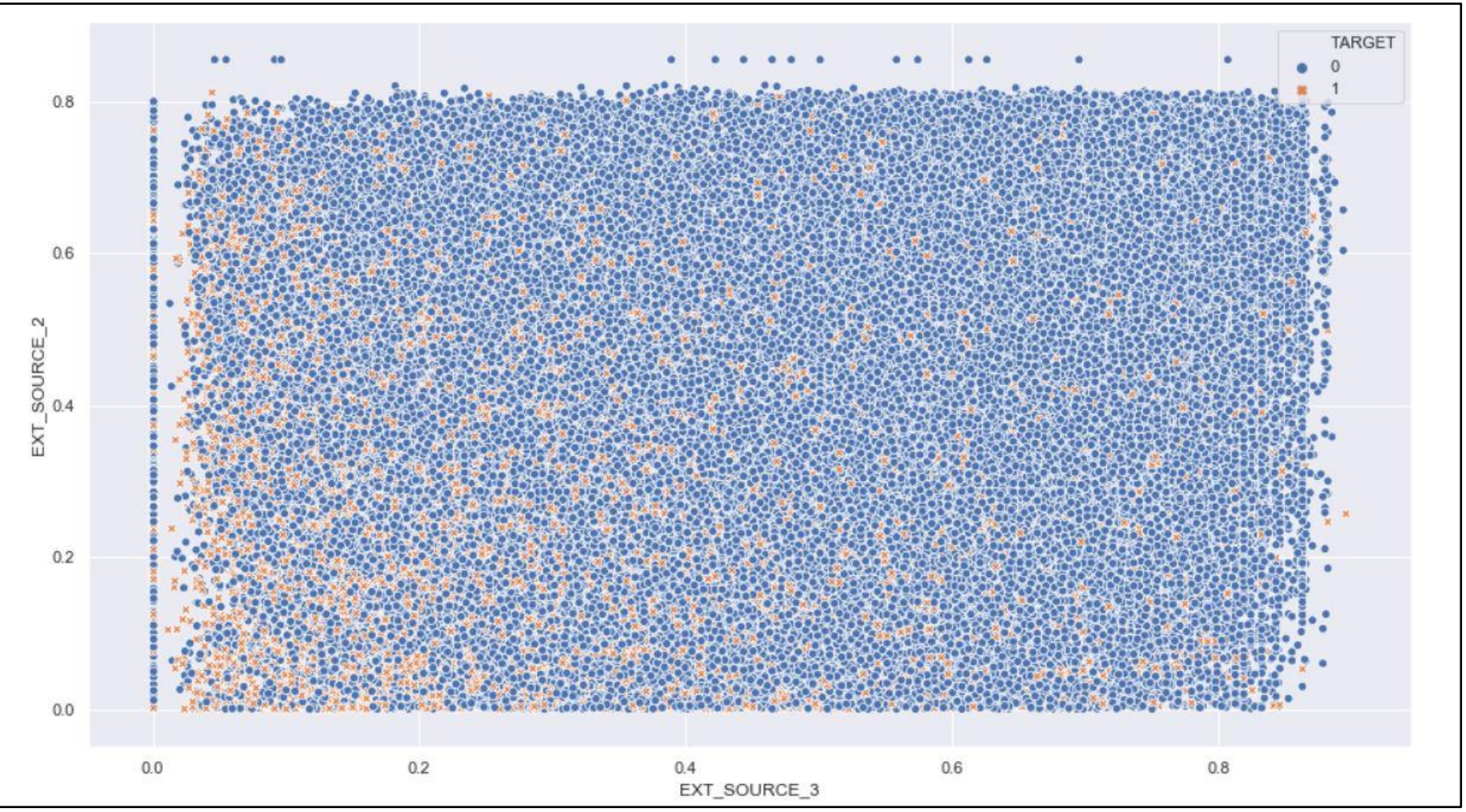
3. External Score2

While comparing the plots, it is indicated that for loan default cases there is an increase in the population density below external agency score 2 of 0.4. So, the applicants with low external agency score more likely to default.

4. External Score3

While comparing the plots, it is indicated that for loan default cases there is an increase in the population density below external agency score 3 of 0.4. So, the applicants with low external agency score more likely to default.

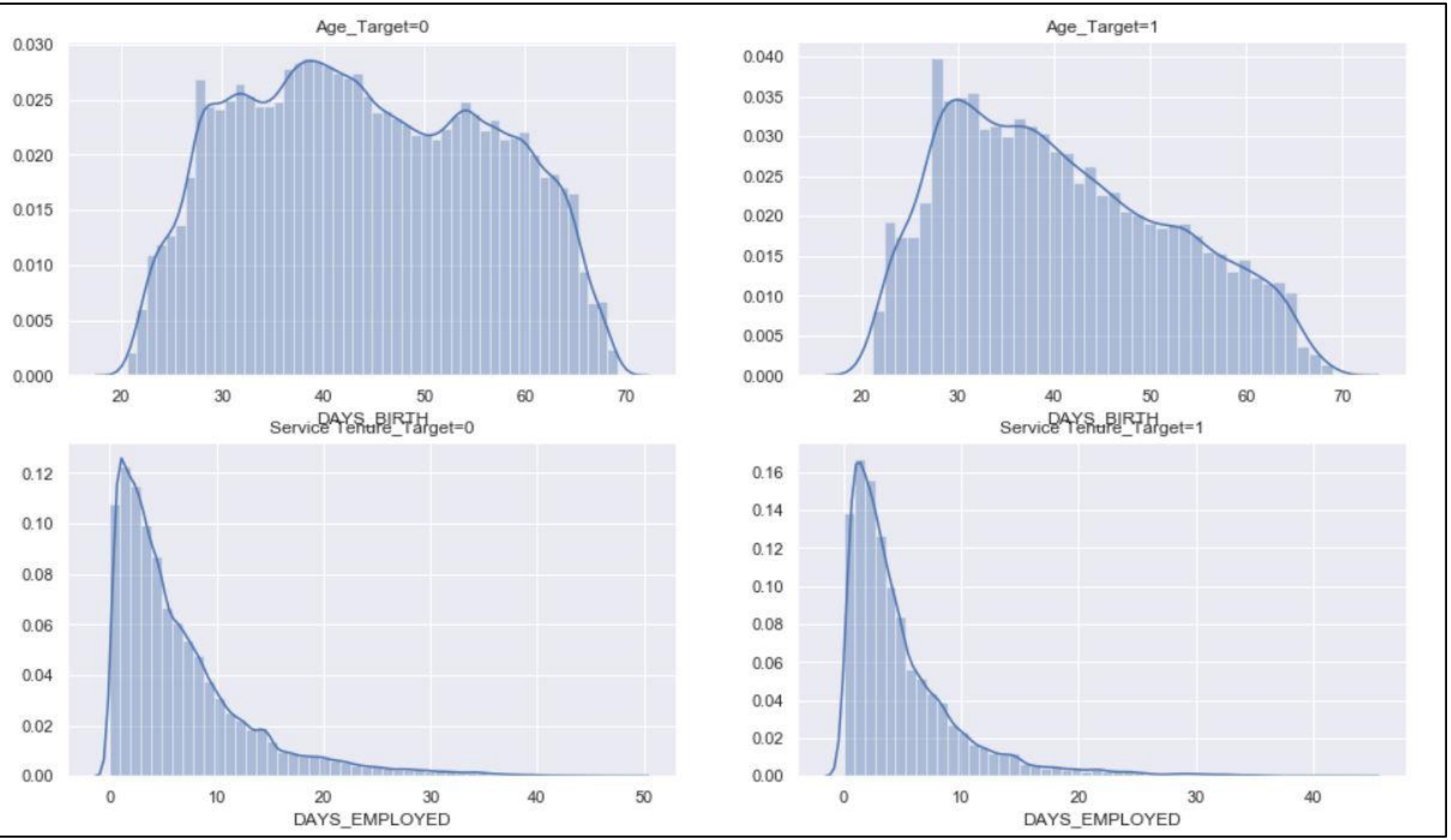
Default Driving Factor :: Strong Indicators



3 & 4.
External Agency Scores(Combined)

In the scatter plot its is clearly indicated that for loan default cases(TARGET=1) the both external agency scores are accumulated at the left bottom corner of the scatter plot. This again further strengthen our claim that lower the external agency score , higher the chances of applicant defaulting on loan repayment.

Default Driving Factor :: Strong Indicators



5. Age

While comparing the plots, it is indicated that for loan default cases there is an increase in the population density for applicant below age of 30 years, So, the young applicants of 20 to 30 years of age more likely to default.

6. Service Tenure

While comparing the plots, it is indicated that for loan default cases there is an increase in the population density for applicant with service tenure below 2 years So, the applicants with less work experience more likely to default.

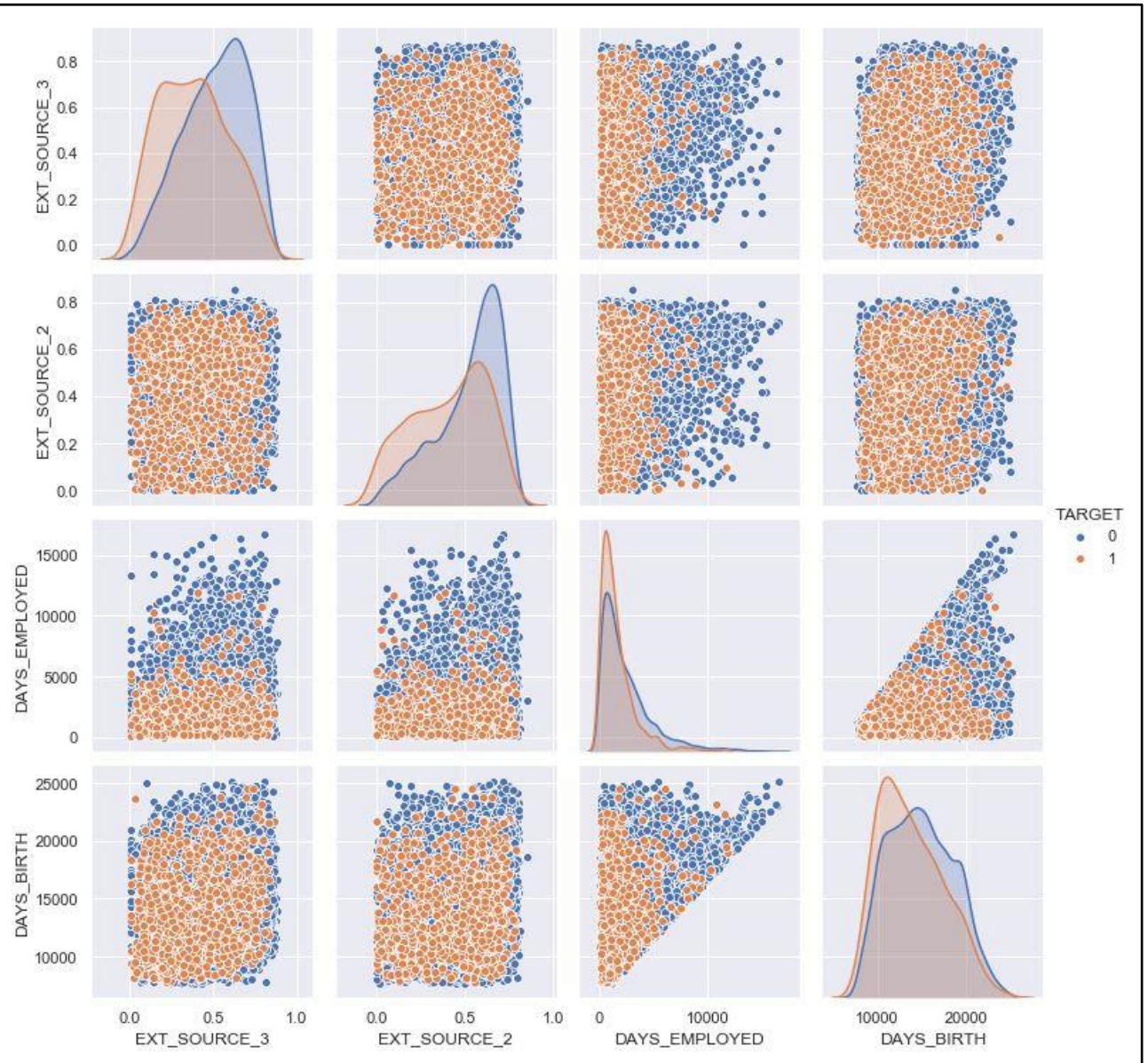
Default Driving Factor :: Strong Indicators

Correlation of continuous variables with TARGET Variable

-ve Correlation :- This means with decrease in the value of these variables the chances of client defaulting on loan increases

1. EXT_SOURCE_3 -0.178898
2. EXT_SOURCE_2 -0.160453
3. DAYS_BIRTH -0.078232
4. DAYS_EMPLOYED -0.074952

Default Driving Factor :: Strong Indicators

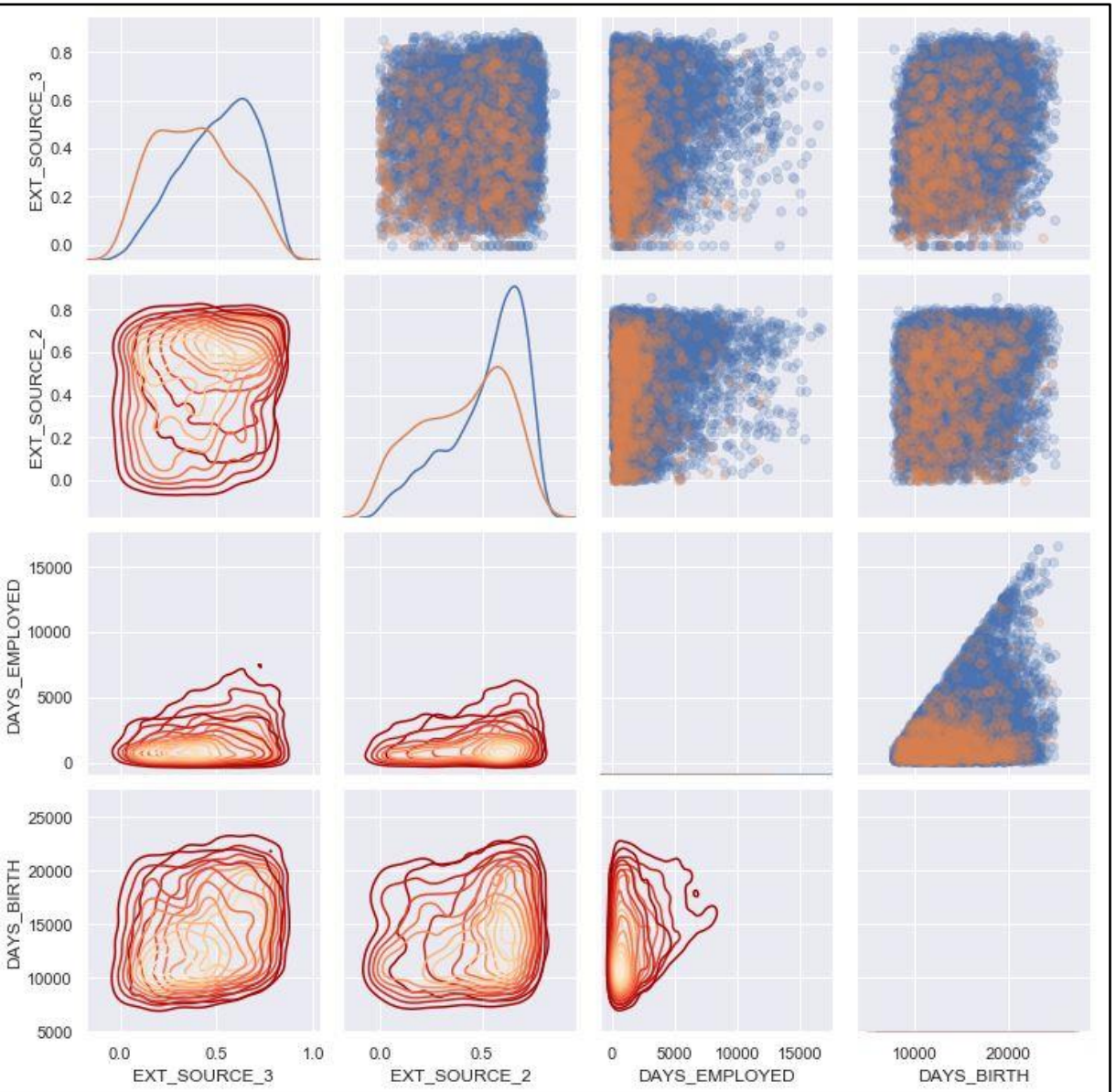


Correlation of continuous variables with TARGET Variable: Pair Grid

In the pair grid, it is visibly clear that for the cases of loan default all the continuous variables are accumulated a the left bottom corner of the scatter plots. This essentially means that lower the values odd continuous variables higher are chances that the respective applicant will default on loan repayment.

Further, to strengthen our claim, the population density of the distribution plots also confirms that for cases of loan default the density population shifts tower lower values.

Default Driving Factor :: Strong Indicators

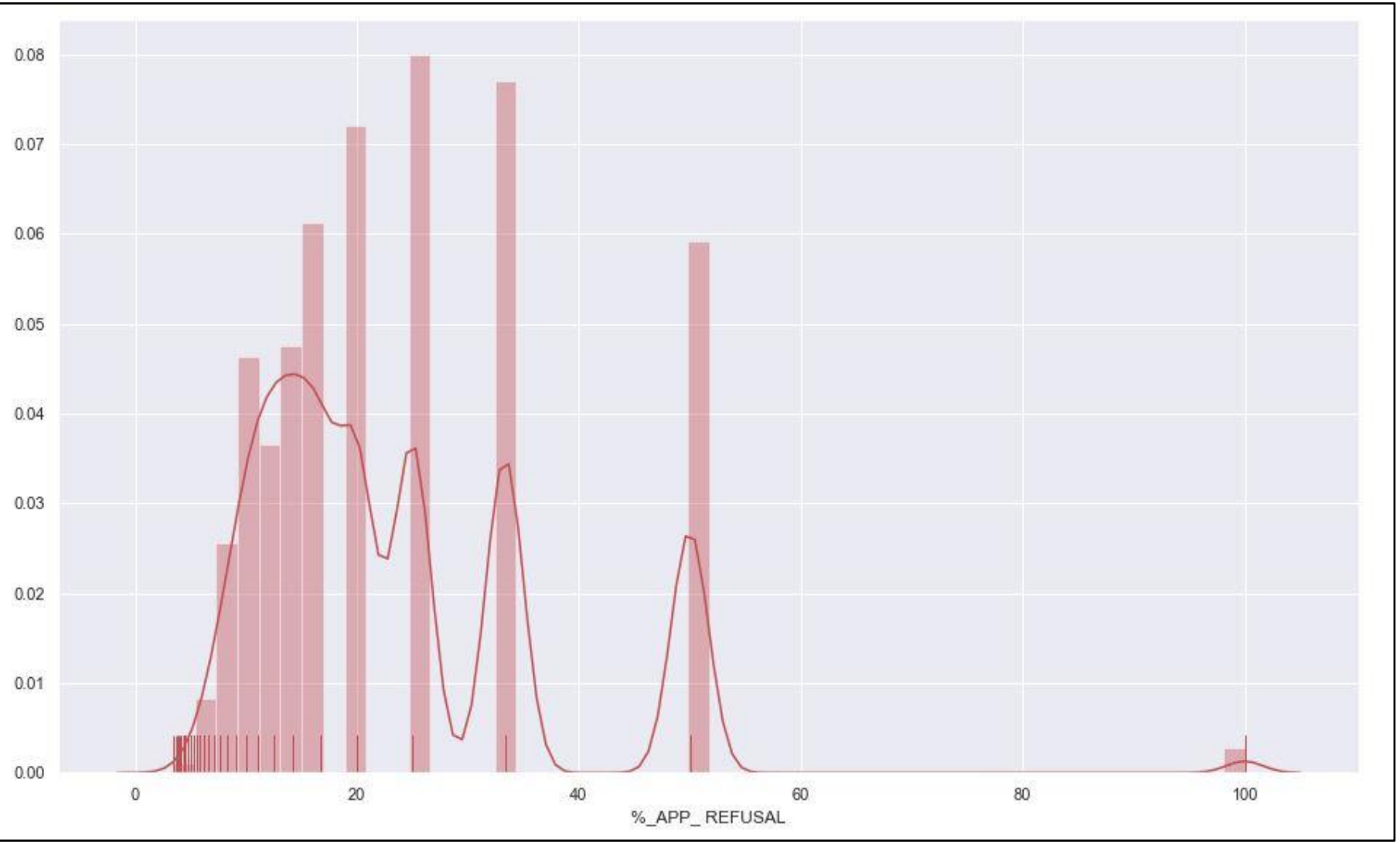


Correlation of continuous variables with TARGET Variable: Density Plots

In the Density Plots, it is visibly clear that for the cases of loan default all the continuous variables are accumulated a the left bottom corner of the scatter plots. This essentially means that lower the values odd continuous variables higher are chances that the respective applicant will default on loan repayment.

Further, to strengthen our claim, the kde plots also confirms that for cases of loan default the density population shifts tower lower values.

Default Driving Factor :: Strong Indicators



7. Rejection Rate of Previous Applications

The %_APP_Refusal variable derived from previous application data and mapped with the current application data. It has a +ve correlation with TARGET variable (0.054). We understand that if out of the total loan applications made by applicant in the past, if more than 20% has been rejected by the bank, then the applicant is more likely to default on loan repayment

Default Driving Factor :: Mild Indicators

Based on trends & distributions identified during Univariate Analysis

6

Family Size

Var-CNT_FAM_MEMBERS

People with smaller families are more likely to default

1

Loan Type

Var- NAME_CONTRACT_TYPE

Cash Loan more prone to default

5

Children

Var-CNT_CHILDREN

People who don't have children are more likely to default

2

Gender

Var-CODE_GENDER

Male borrowers more prone to default

4

Marital Status

Var-NAME_FAMILY_STATUS

Single people more likely to default

3

Education

Var-NAME_EDUCATION_TYPE

clients with secondary education are more prone to default

Recommendations to the Bank::To Minimize Loan Default

Based on the extensive EDA conducted on the data set and our professional judgement based on the finding of the EDA, we recommend the Bank implement a 4 layer scrutiny system in order to minimized loan default. These layers will refine the applicants profile based on the following 4 checks and the applicants profile matching all 4 checks can be rejected.

CHECK 1

Age & Service Tenure

Avoid giving loan to applicants with age<30 years & job tenure<2 year

CHECK 2

Region & City

Avoid giving loan to applicants who resides in region & city with rating 3 assigned by bank

CHECK 3

External Agency Scores

Avoid giving loan to applicants with external agency score less than 0.4

CHECK 4

Past Rejected Applications

Avoid giving loan to applicants with more than 20% past applications rejected by bank


Thank you
Ende