Credit EDA Case Study

Exploratory Data Analysis

Team
Nistha Kumar & Gaurav Rana

Table of Content

1 Understanding of the problem statement

2 Overall EDA Approach

3 Data Quality Check

4 Data Analysis

5 EDA Summary (Key Patterns)

Problem Statement

Business Definition

→ A "consumer finance company" wants to minimise the risk of losing money while lending to customers

→ There are two risks:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Data Project definition

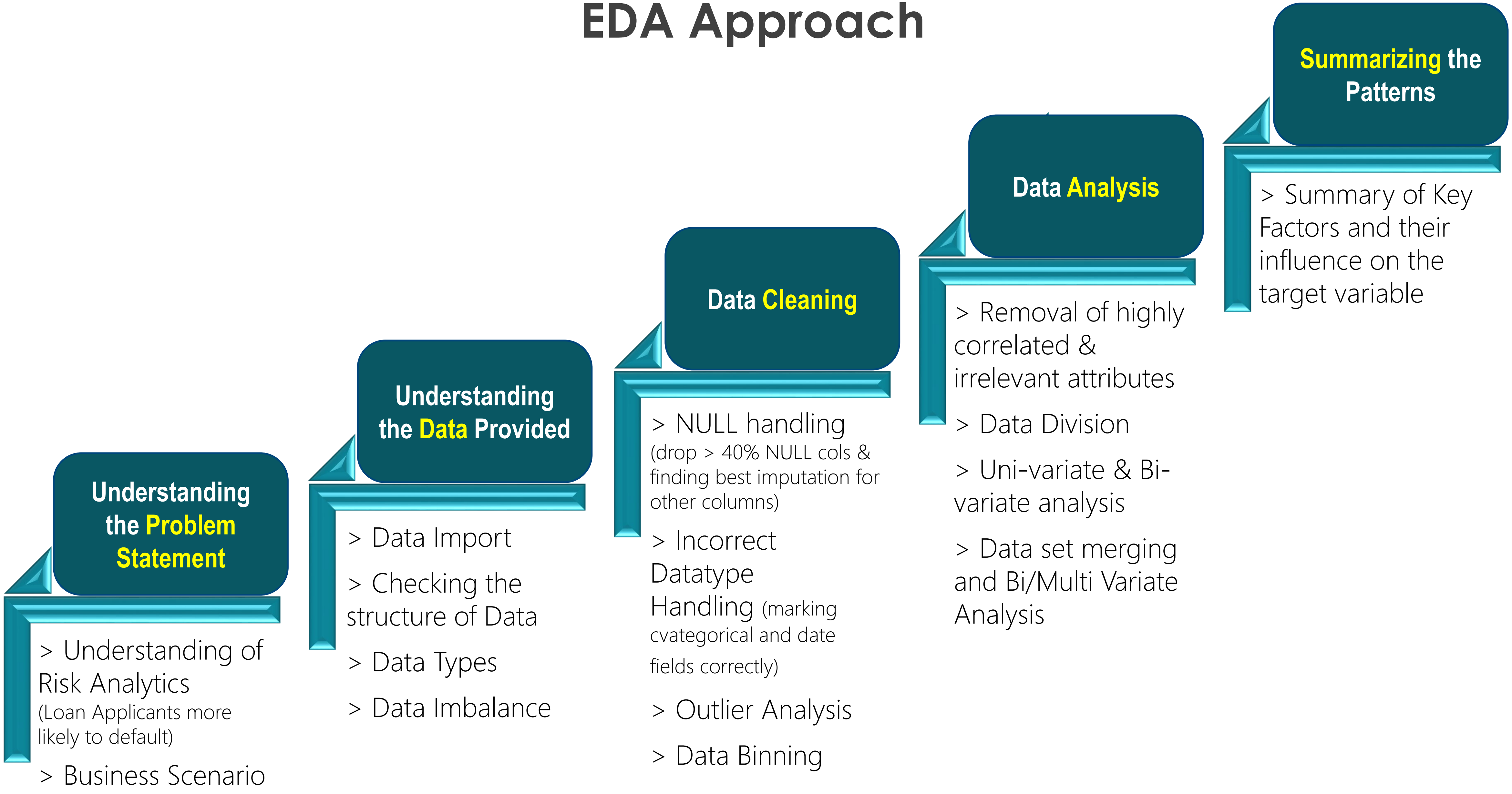
→ "Target" variable is what tells whether a customer has defaulted or not

→ Find all the "applicant", "loan", "previous application" related variables that have a correlation/influence on the target variable (strong indicators of default)

→ Provide a summary of all the variables that have a trend against the target variable, so that the company can utilise this knowledge for its portfolio and risk assessment

application_data			previous_application	
key	SK_ID_CURR	<-- join-->	key	SK_ID_CURR
target variable	TARGET		target variable	NAME_CONTRACT_STATUS
attribute1			attribute1	
attribute2			attribute2	
attribute3			attribute3	
…			…	
…			…	

Overall EDA Approach



Understanding the Problem Statement

> Understanding of Risk Analytics (Loan Applicants more likely to default)

> Business Scenario

Understanding the Data Provided

> Data Import

> Checking the structure of Data

> Data Types

> Data Imbalance

Data Cleaning

> NULL handling (drop > 40% NULL cols & finding best imputation for other columns)

> Incorrect Datatype Handling (marking cvategorical and date fields correctly)

> Outlier Analysis

> Data Binning

Data Analysis

> Removal of highly correlated & irrelevant attributes

> Data Division

> Uni-variate & Bi-variate analysis

> Data set merging and Bi/Multi Variate Analysis

Summarizing the Patterns

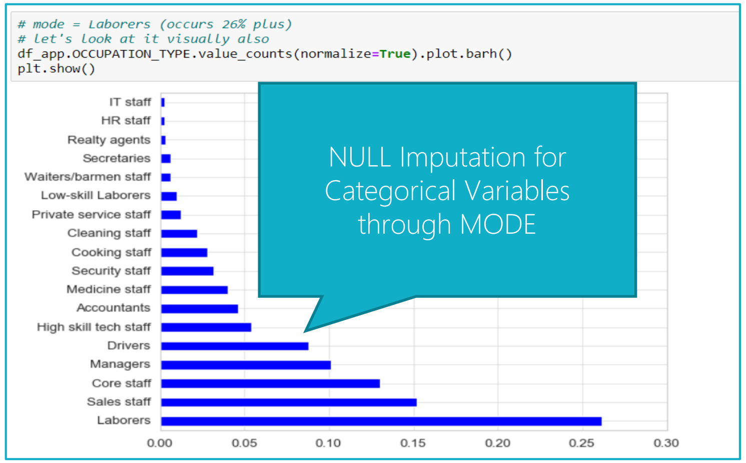
> Summary of Key Factors and their influence on the target variable

Data Quality check (1/2)

	NULL analysis	NULL Imputation	Datatype Correction	Date conversion	Outlier Analysis	Binning
Approach	• Find % age of NULLs in every column • Decided on a threshold that any column with more than 40% values as NULLs would be dropped	• For rest of the columns (NULL %age between 1% and 40%) – analysis needed on whether and how to impute • For Categorical : Mode imputation • For Continuous: Median or Mean	• Checking data types - all categorical columns (based on unique values) should be of object data type • All continuous columns should be integers/float)	• There were many dates with negative values • They were converted into datetime formats by adding the negative number to 01/01/2020 as a reference date.	• Found the topmost 5 most skewed columns for Outlier Analysis • Boxplot and distplot analysis on whether outliers need to be removed, binned or retained	• Binning some variables to make the continuous variable to categorical.
Application Data	• 49 columns found with > 40% NULLs and hence dropped	1 Categorical to be imputed by MODE **(OCCUPATION_TYPE)** 1 Continuous : imputed by MEAN **(EXT_SOURCE_3)** 6 continuous : imputed by MEADIAN **(AMT_REQ*)**	• Some columns converted to object type: TARGET, FLAG_MOBIL, FLAG_EMAIL, FLAG_DOCUMENT_*	• Columns converted : ['DAYS_BIRTH','DAYS_EMPLOYED','DAYS_REGISTRATION','DAYS_ID_PUBLISH','DAYS_LAST_PHONE_CHANGE']	> OBS_30/60_CNT_SOCIAL_CIRCLE : values over 24.0 should be deleted (99.99 percentile value) > AMT_REQ_CREDIT_BUREAU_H/D: have outliers but there is possibility of that data, no action > AMT_REQ_CREDIT_BUREAU_QRT : values over 8 should be deleted. (99.999 percentile value)	1.AMT_INCOME_TOTAL : binned it into :"Very Low", "Low", "Medium", "High", "Very High" categories. 2. HOUR_APPR_PROCESS_START_CAT binned into "Morning", "Afternoon", "Evening"
Previous Applications	• 11 columns found with > 40% NULLs and hence dropped	• Three columns to be imputed by median: • CNT_PAYMENT AMT_ANNUITY AMT_GOODS_PRICE	• One column converted to Object type: NFLAG_LAST_APPL_IN_DAY	• Column converted : DAYS_DECISION	➢ CNT_PAYMENT : no need to handle outliers, there is a possibility of this data ➢ SELLERPLACE_AREA : values over 120000 can be deleted. 99.999% values are within this range ➢ DAYS_DECISION : values over 2913 should be deleted. 99.90% values are within this range	1. HOUR_APPR_PROCESS_START_CAT binned into "Morning", "Afternoon", "Evening"

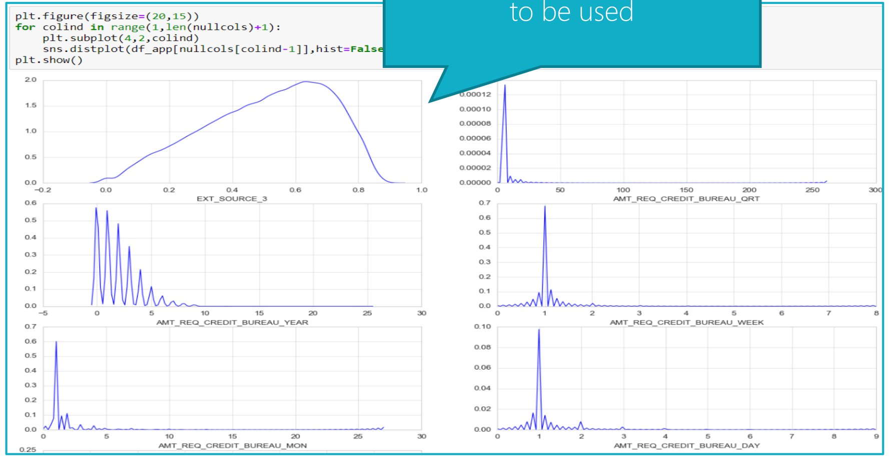
Data Quality check (2/2)

Some Sample work done for Data Quality
(refer to notebooks attached for details)



```
# mode = Laborers (occurs 26% plus)
# let's look at it visually also
df_app.OCCUPATION_TYPE.value_counts(normalize=True).plot.barh()
plt.show()
```

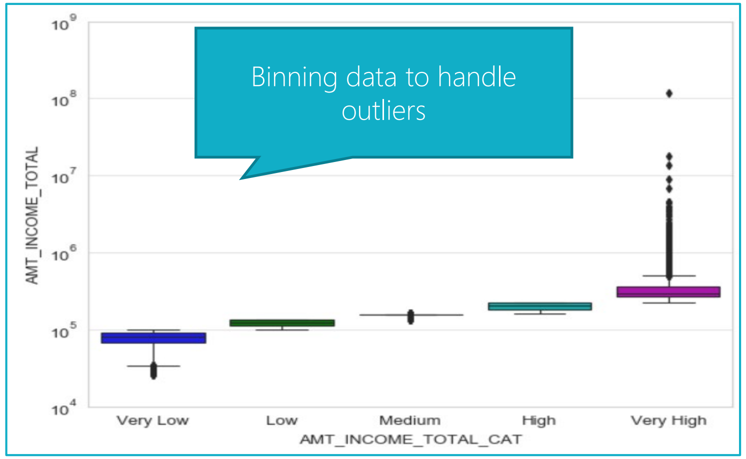
NULL Imputation for Categorical Variables through MODE



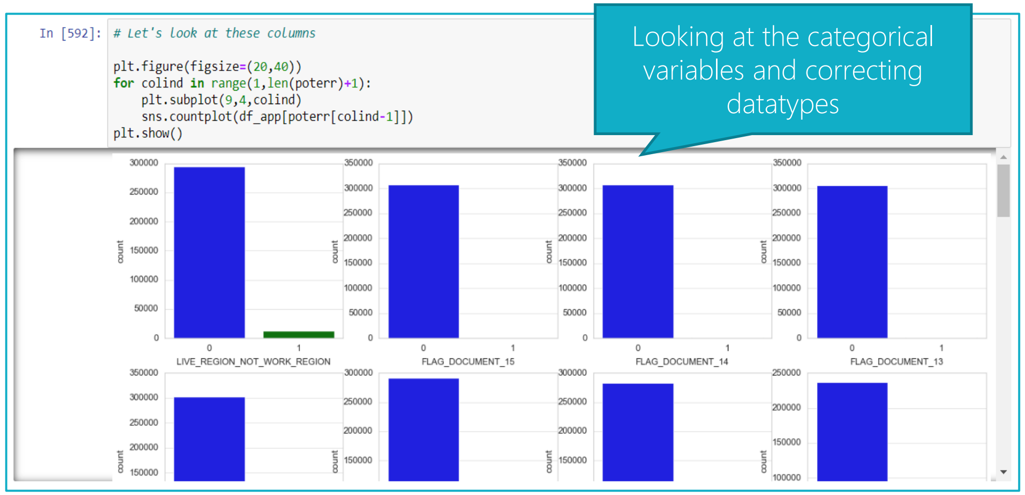
```
plt.figure(figsize=(20,15))
for colind in range(1,len(nullcols)+1):
    plt.subplot(4,2,colind)
    sns.distplot(df_app[nullcols[colind-1]],hist=False
plt.show()
```

EXT_SOURCE_3
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_DAY

Analyzing distributions & skewness to decide whether MEAN or MEDIAN statistic to be used



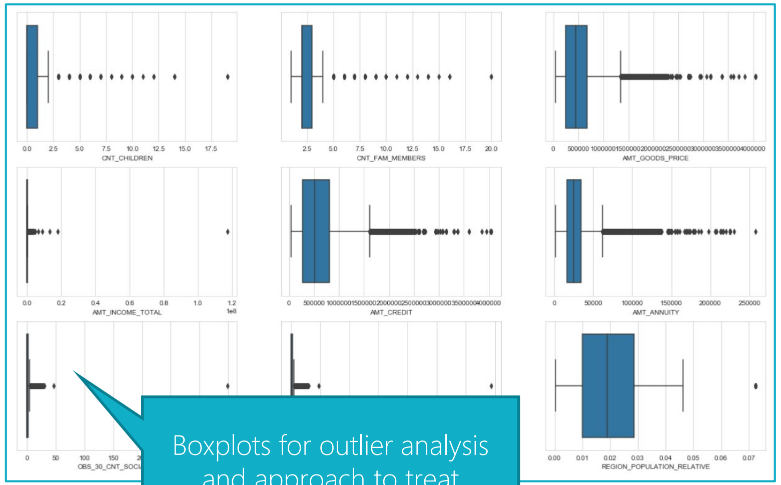
Binning data to handle outliers



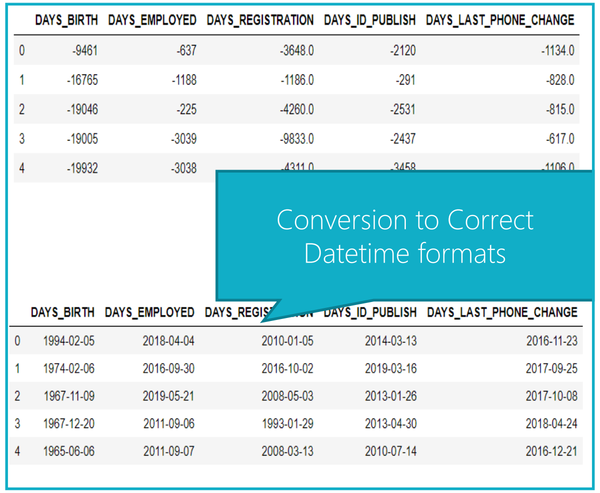
```
In [592]: # Let's look at these columns

          plt.figure(figsize=(20,40))
          for colind in range(1,len(poterr)+1):
              plt.subplot(9,4,colind)
              sns.countplot(df_app[poterr[colind-1]])
          plt.show()
```

Looking at the categorical variables and correcting datatypes



Boxplots for outlier analysis and approach to treat



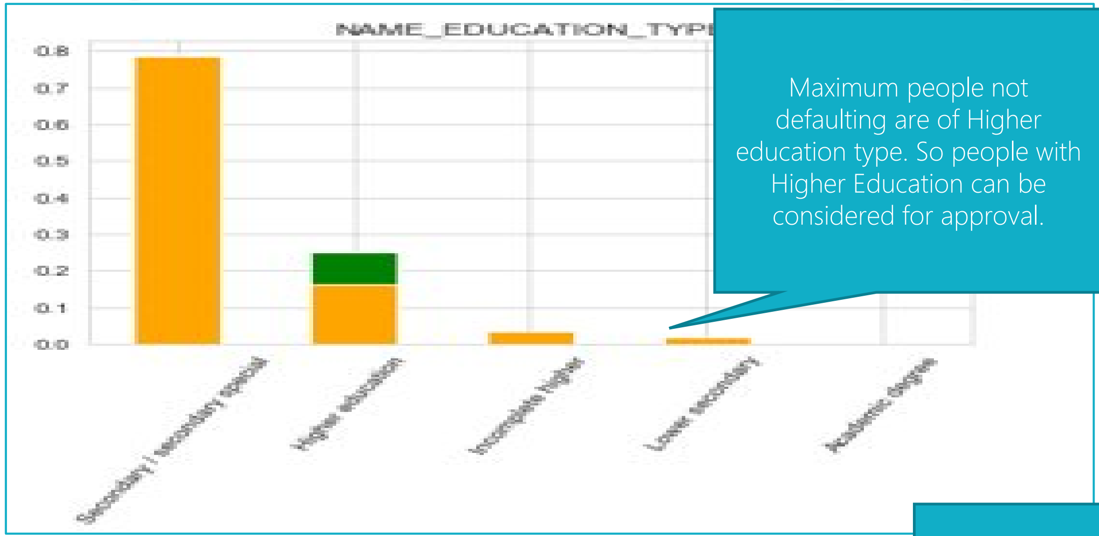
	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	-9461	-637	-3648.0	-2120	-1134.0
1	-16765	-1188	-1186.0	-291	-828.0
2	-19046	-225	-4260.0	-2531	-815.0
3	-19005	-3039	-9833.0	-2437	-617.0
4	-19932	-3038	-4311.0	-3458	-1106.0

Conversion to Correct Datetime formats

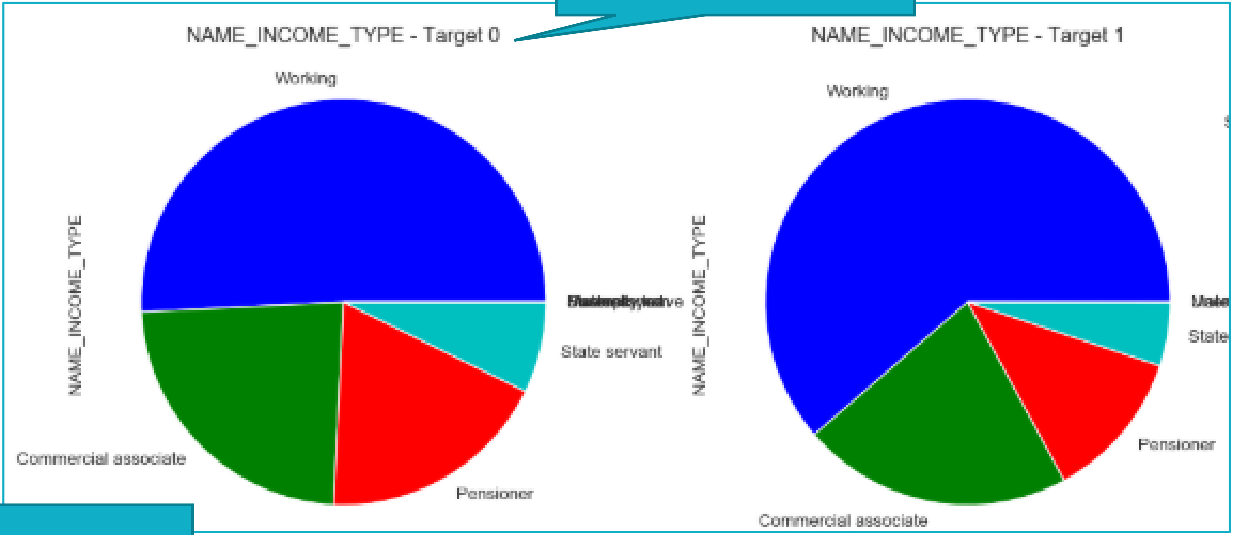
	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	1994-02-05	2018-04-04	2010-01-05	2014-03-13	2016-11-23
1	1974-02-06	2016-09-30	2016-10-02	2019-03-16	2017-09-25
2	1967-11-09	2019-05-21	2008-05-03	2013-01-26	2017-10-08
3	1967-12-20	2011-09-06	1993-01-29	2013-04-30	2018-04-24
4	1965-06-06	2011-09-07	2008-03-13	2010-07-14	2016-12-21

Data Analysis
Application data (1/3)

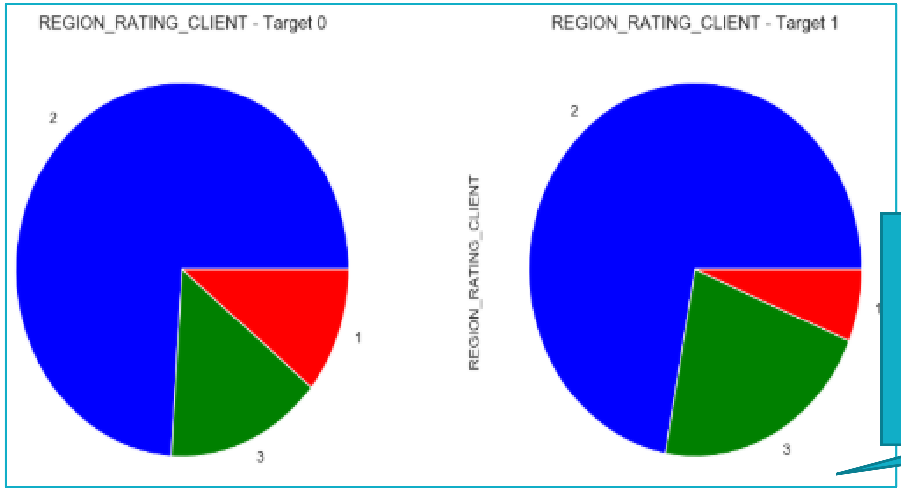
Categorical Variables influencing Target Variable



Maximum people not defaulting are of Higher education type. So people with Higher Education can be considered for approval.

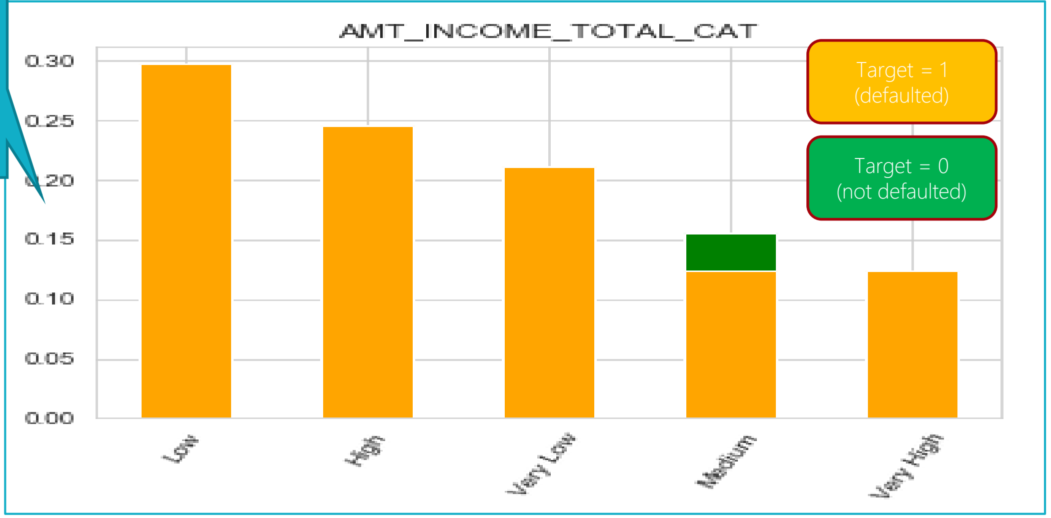


More Pensioners are less likely to default whereas Working applicants are expected to default more



Applicants are mostly staying in region with rating 2 for both target 0 and 1 . Also people staying in region 3 are more likely to default compared to region 1

Applicants with Medium Income Category are most likely to not default.

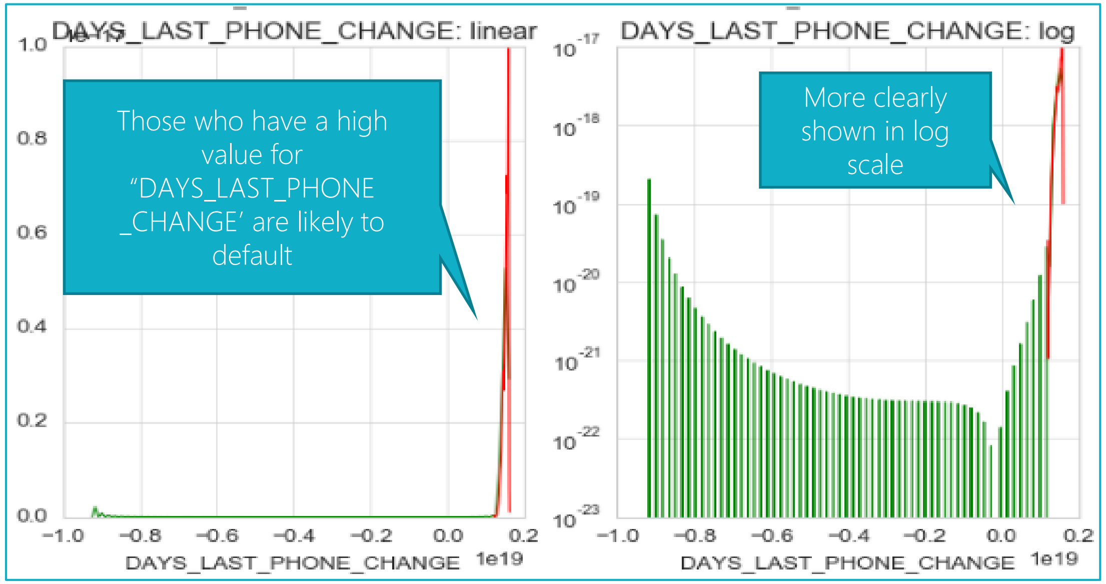


Data Analysis
Application data (2/3)

Target = 1 (defaulted)

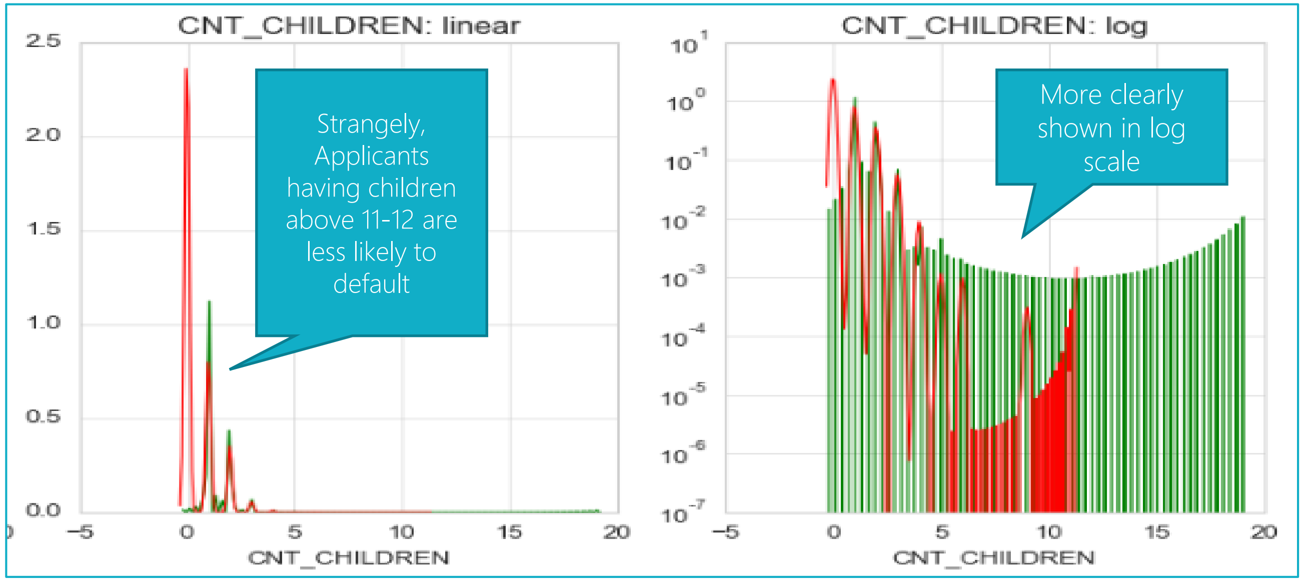
Target = 0 (not defaulted)

Continuous Variables influencing Target Variable (both liner and logarithmic trend shown)



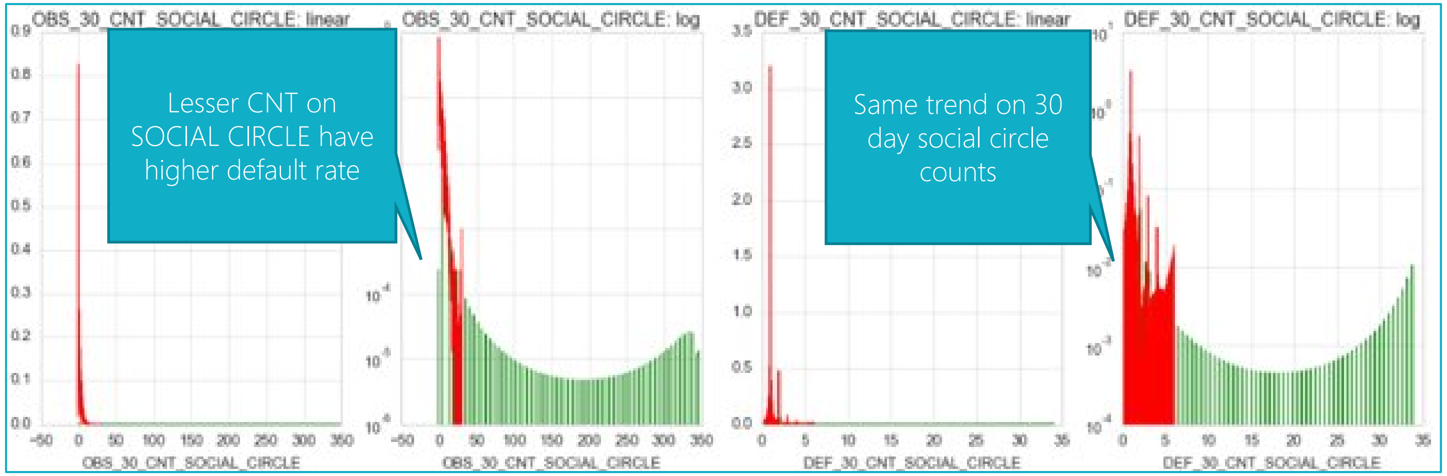
Those who have a high value for "DAYS_LAST_PHONE _CHANGE' are likely to default

More clearly shown in log scale



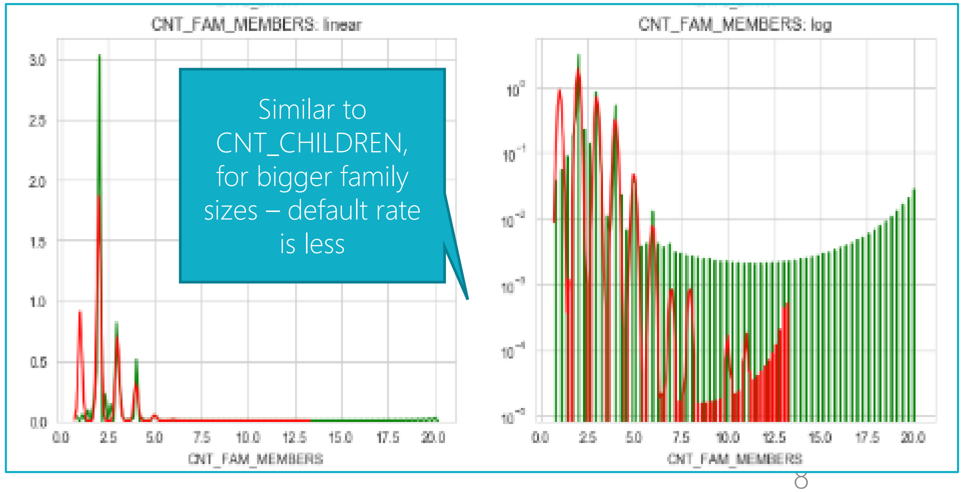
Strangely, Applicants having children above 11-12 are less likely to default

More clearly shown in log scale



Lesser CNT on SOCIAL CIRCLE have higher default rate

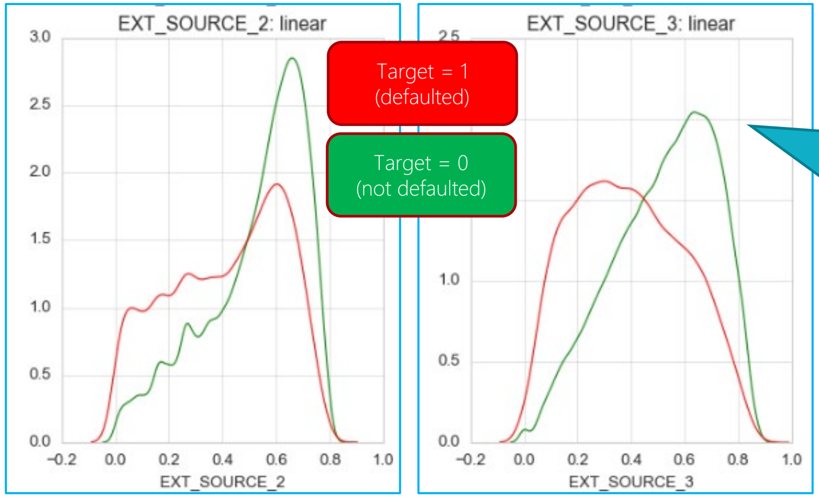
Same trend on 30 day social circle counts



Similar to CNT_CHILDREN, for bigger family sizes – default rate is less

Data Analysis
Application data (3/3)

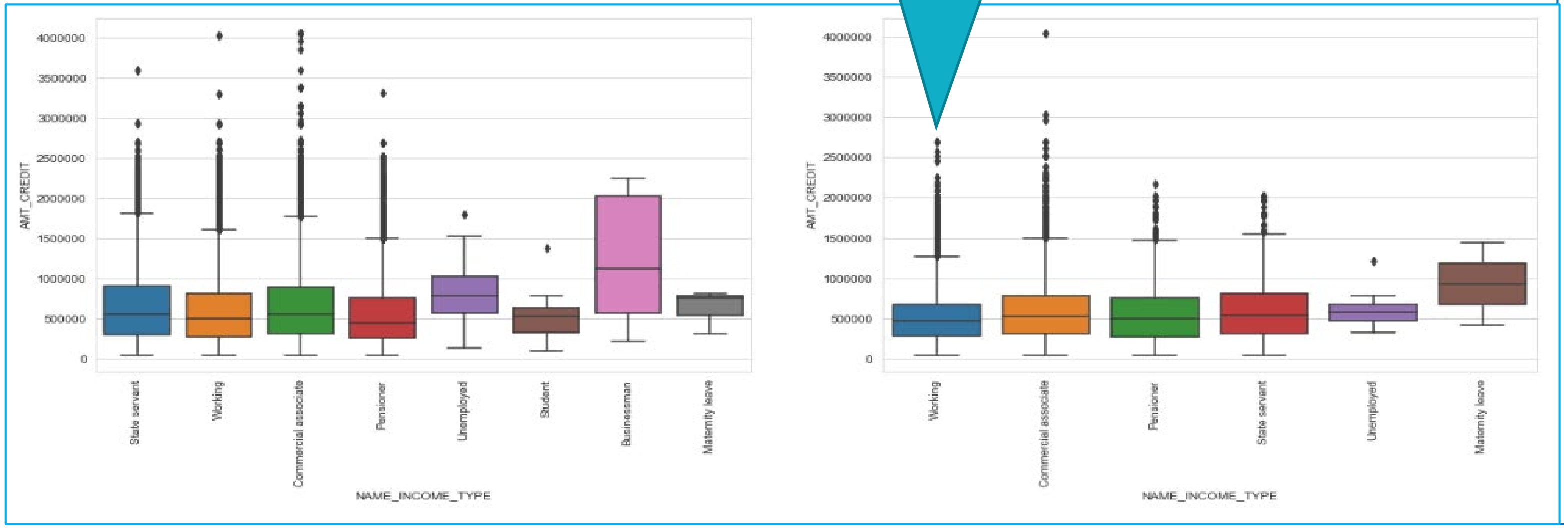
Continuous Variables influencing Target Variable (contd.)

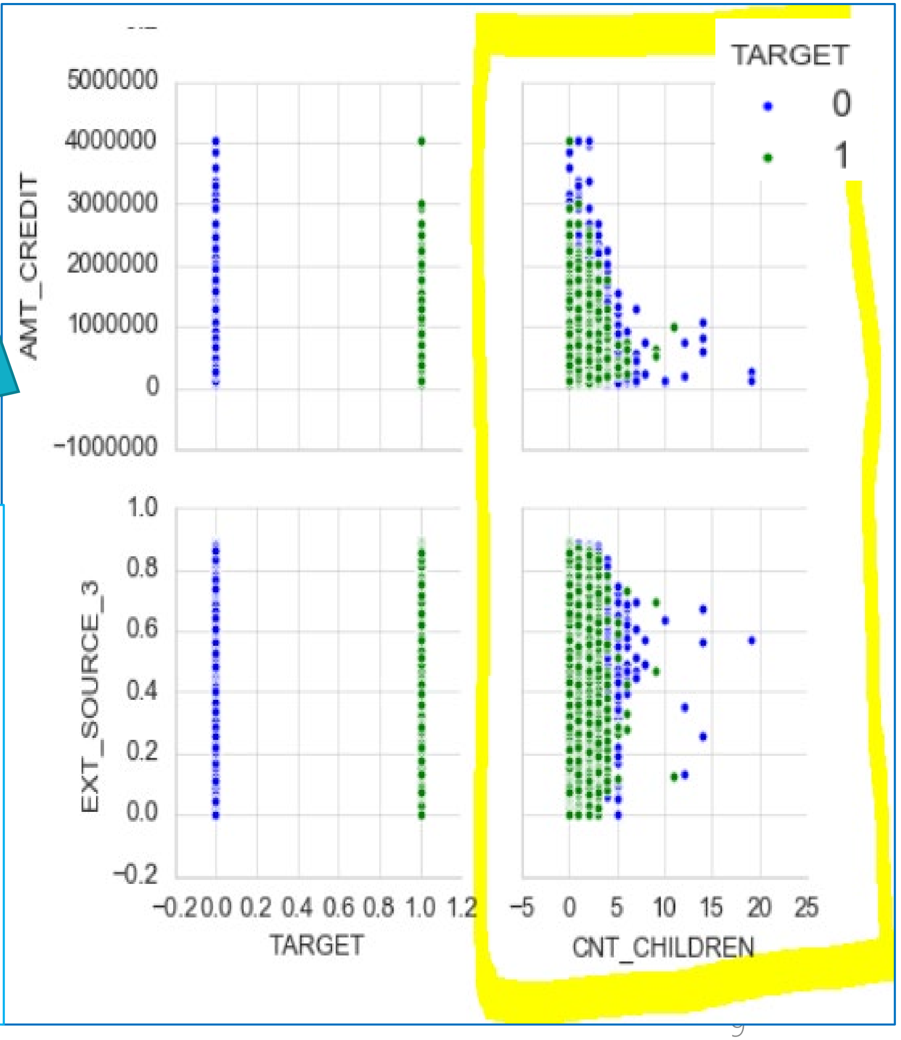


Both SOURCE 2 & SOURCE 3 have influence – higher values have lesser default rate

Slide # 7 shows working people have more chances to default. Here we see that working people who donot default (left) have higher AMT_CREDIT values than one ones who do(right)

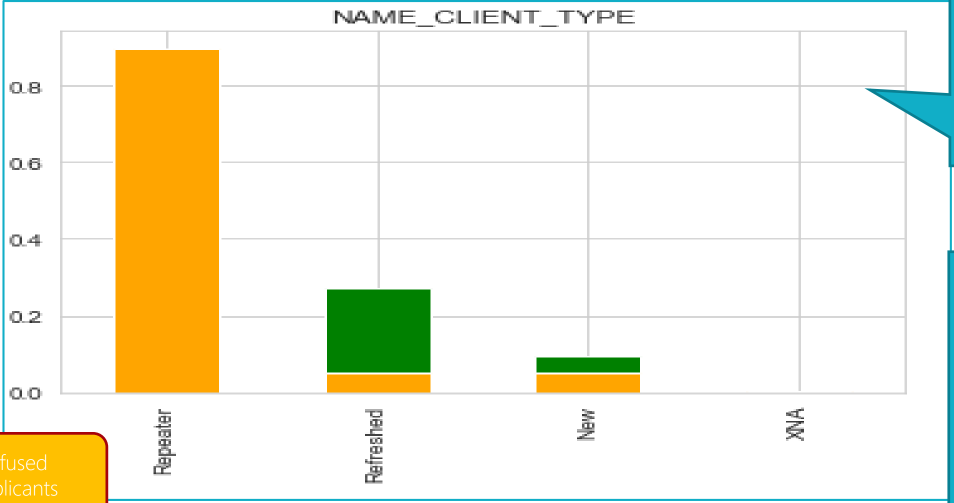
If you see the PAIRPLOT analysis, combination of AMT_CREDIT and CNT_CHILDREN has an impact i.e.
Lower children & higher credit
Or higher children & lower credit – both have a tendency to default





Data Analysis
Previous Applications (1/2)

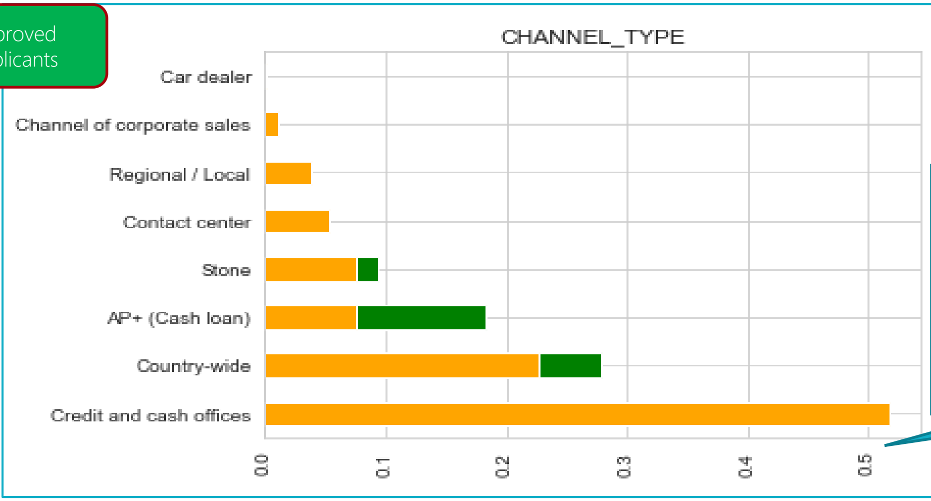
Categorical Variables influencing NAME_CONTRACT_STATUS Variable



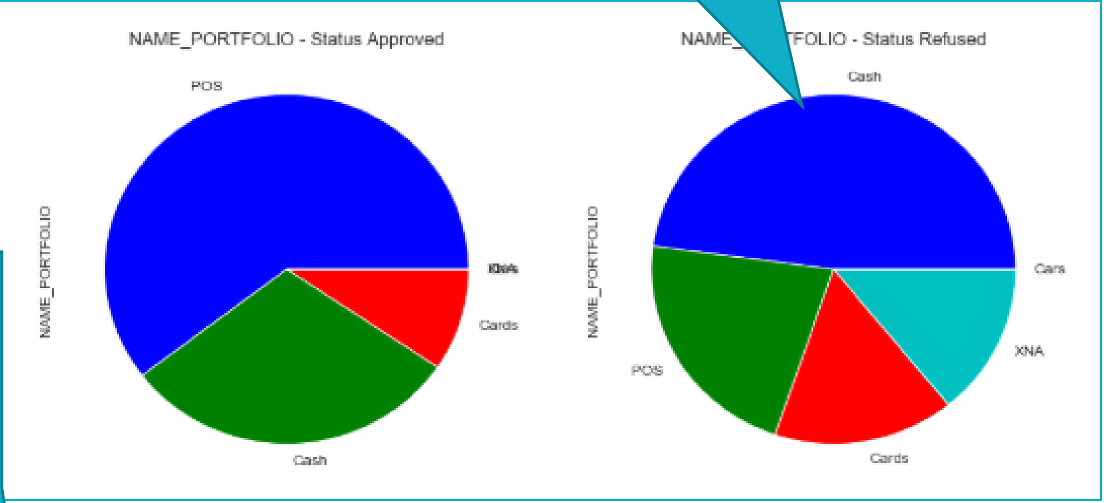
Clients of type "Refreshed" have most trend of being approved

Refused Applicants

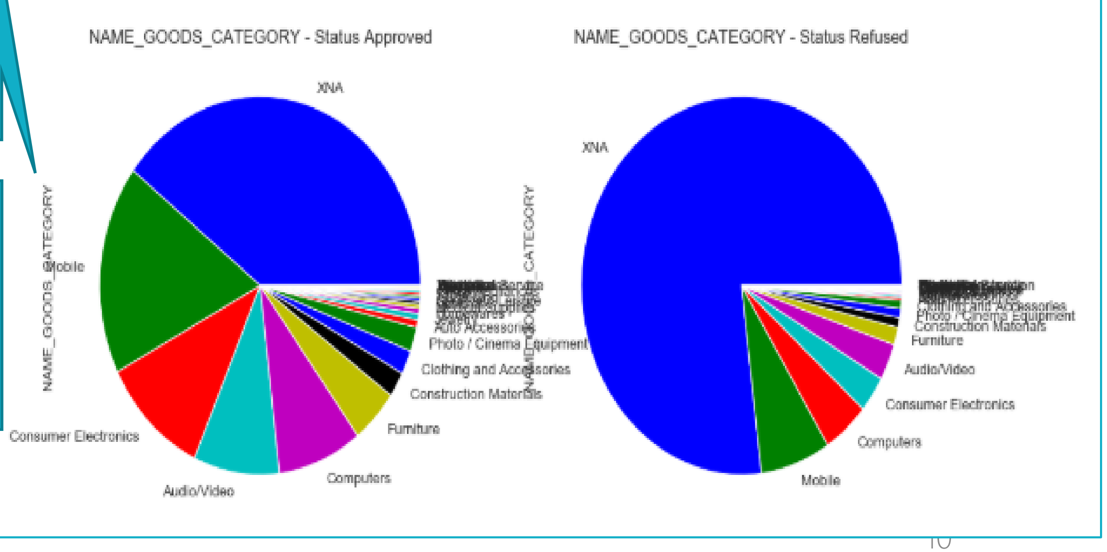
Approved Applicants



Channel type AP+(Cash Loan) have the highest value of loans getting approved. So this can be a good parameter for the bank to decide on the loan status



- Applicants with portfolio status Cash are more likely to be refused for the loan whereas those of type POS are mostly approved.



- Chances of Loan getting rejected are more if the applicant dosent specify the Goods_Category(XNA).
- Loans for Goods_Category Mobile are maximum for both approved and rejected loans

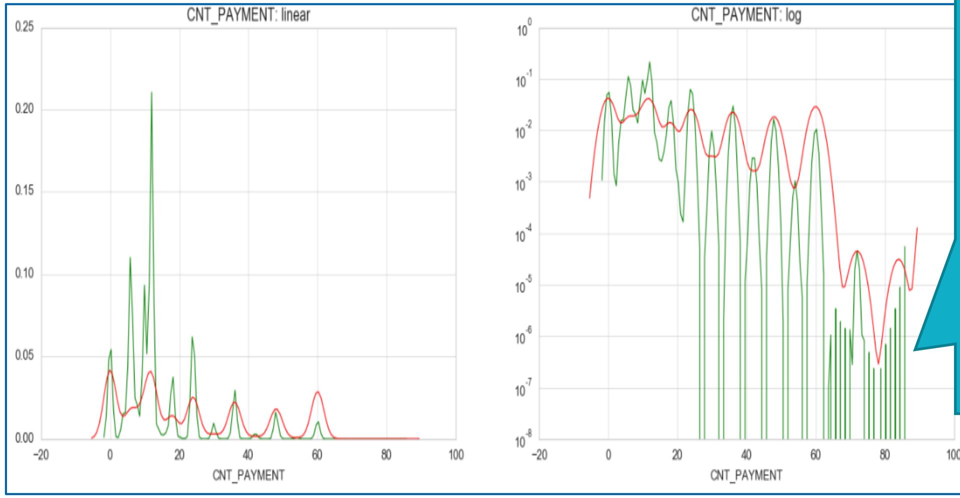
Data Analysis
Previous Applications (2/2)

Note: only Refused and Approved records have considered for this analysis (cancelled, not used have been ignored)

🟥 Refused Applications

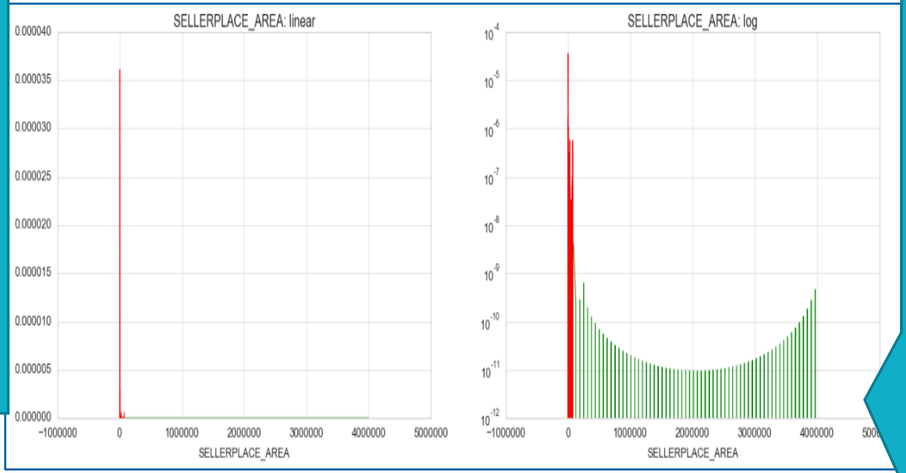
🟩 Approved Applications

Continuous Variables influencing NAME_CONTRACT_STATUS Variable



CNT_PAYMENT : Term of previous credit at application of the previous application

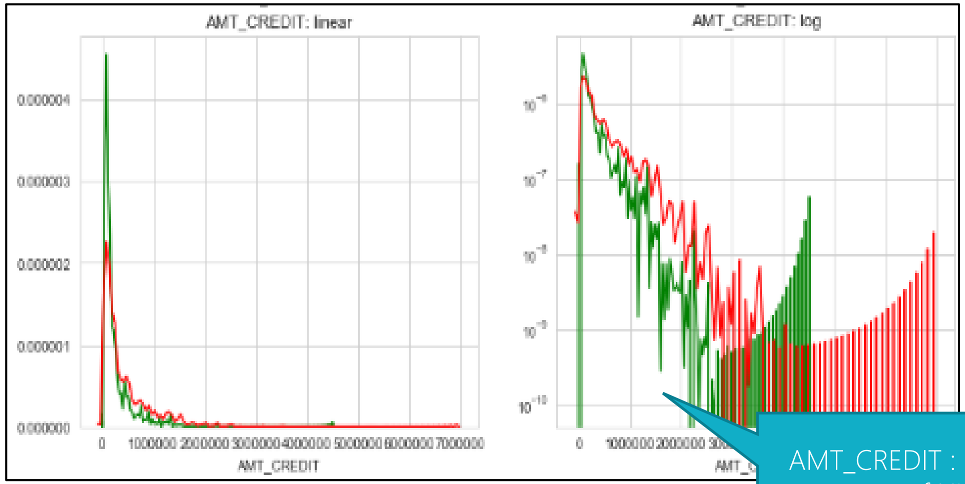
- ➢ Maximum approved previous applications are there where the term is between 10-14



SELLERPLACE_AREA : this is "Selling area of seller place of the previous application"

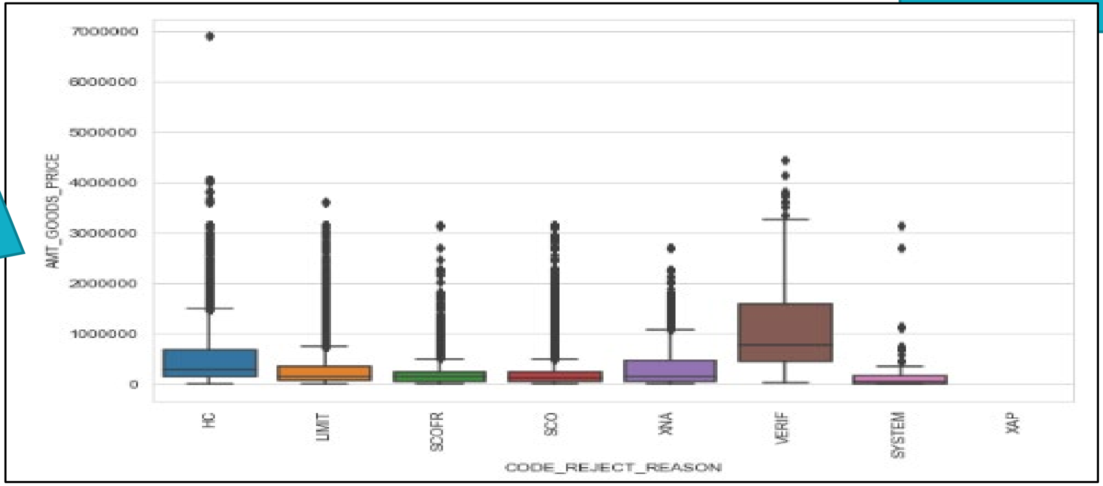
As can be seen (graph on the left is liner view and on the right is the logarithmic view)

- ➢ Records with Lower SELLERPLACE_AREA have a high REFUSAL RATE



CODE_REJECT_REASON : Maximum of the rejected loans are because of the reason: "HC" but "VERIF" have the highest AMT_GOODS_PRICE and high AMT_GOODS_PRICE results in more rejection of the loan.

AMT_CREDIT : Maximum rejected loans are of Higher AMT_CREDIT



Data Analysis
Merged Data (1/2)

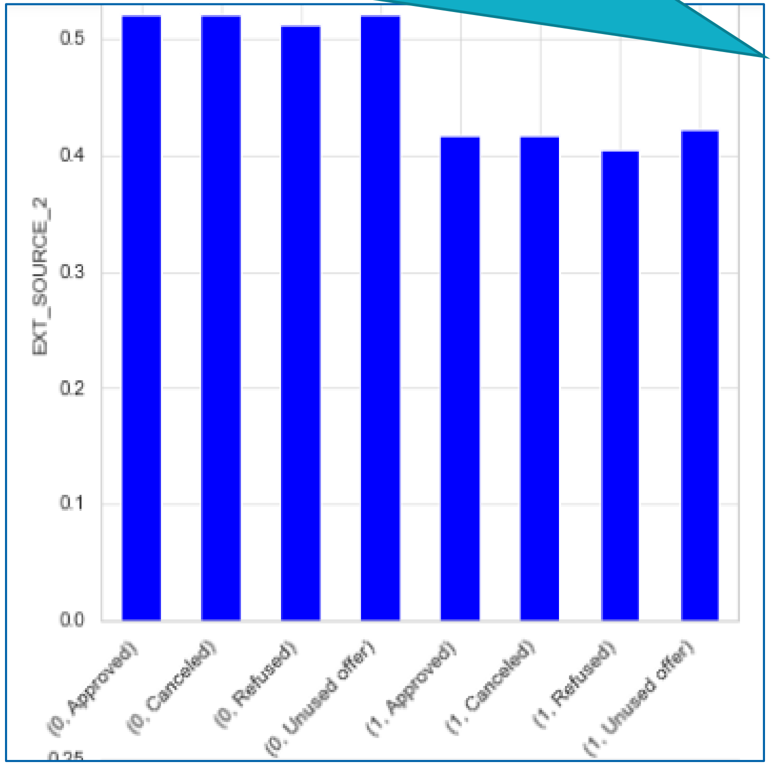
Impact of variables – on a combination of “TARGET” and “NAME_CONTRACT_STATUS”

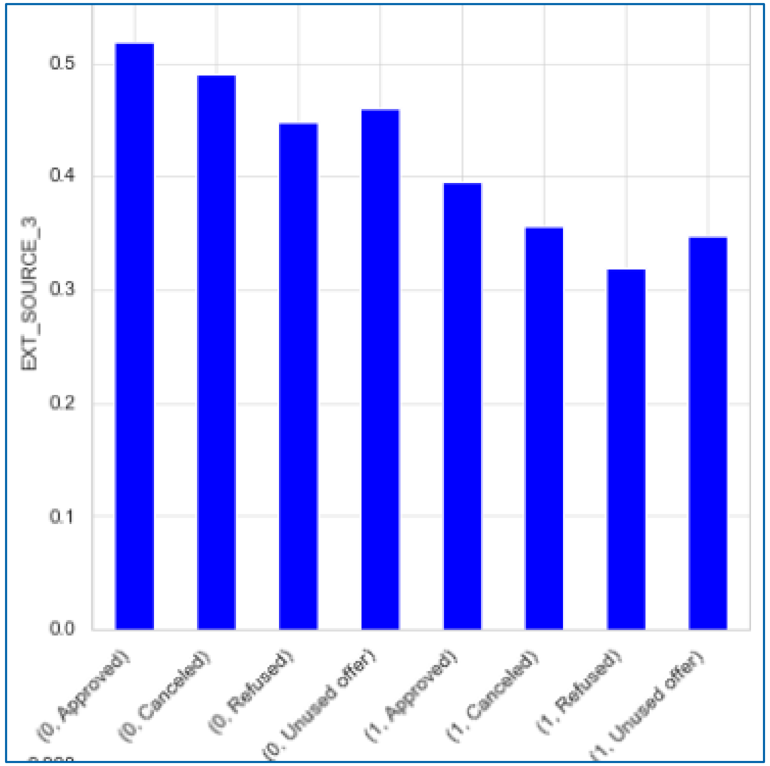
Note: first 4 bars show ‘non defaults’ – different statuses, and last 4 bars show “default” – different statuses. Bar height is the MEAN value

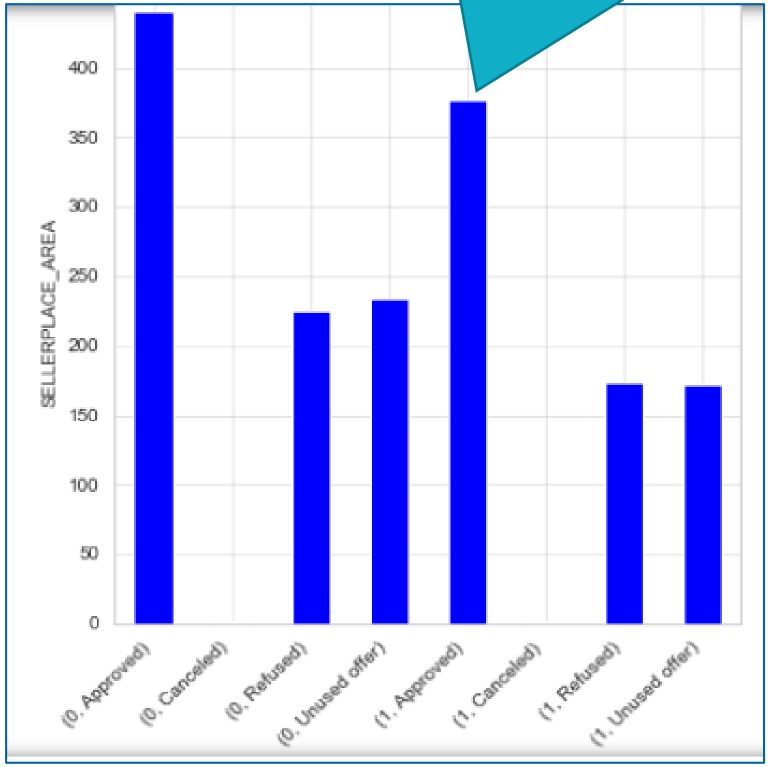
- EXT source 2 and 3 have an impact on the default rate.
- As can be seen, default cases have consistently lower values for both these variables

- As seen from this analysis also, Lower SELLERPLACE_AREA has a higher tendency for default

(comparison between approval – default and non-default cases)





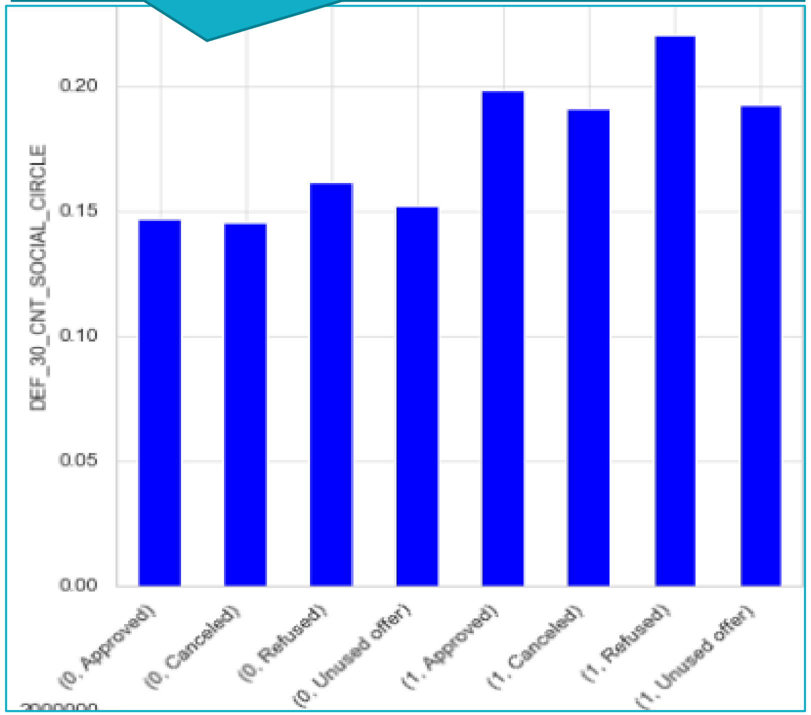


Data Analysis
Merged Data (2/2)

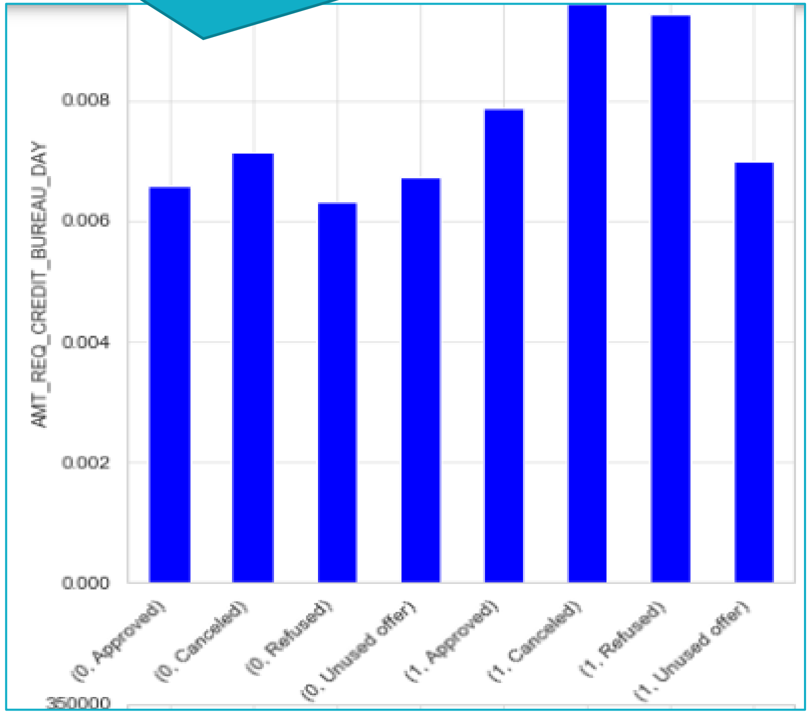
Impact of variables – on a combination of "TARGET" and "NAME_CONTRACT_STATUS"

Note: first 4 bars show 'non defaults' – different statuses, and last 4 bars show "default" – different statuses. Bar height is the MEAN value

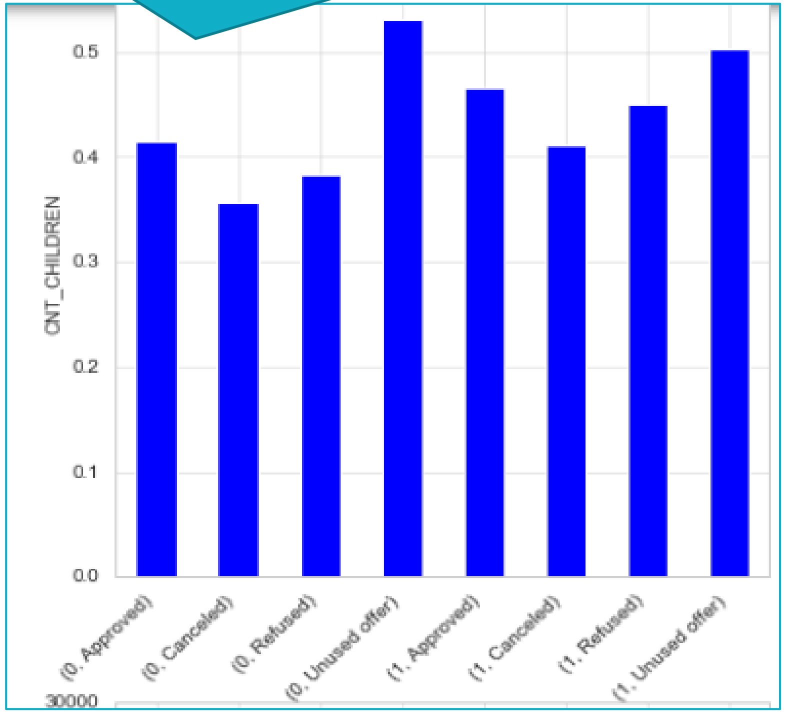
- ➢ Social status count (def 30) – mean value for non defaults is lower than mean value for defaults



- ➢ It seems when NUMBER of enquiries to credit bureau is higher, then default chances are also higher

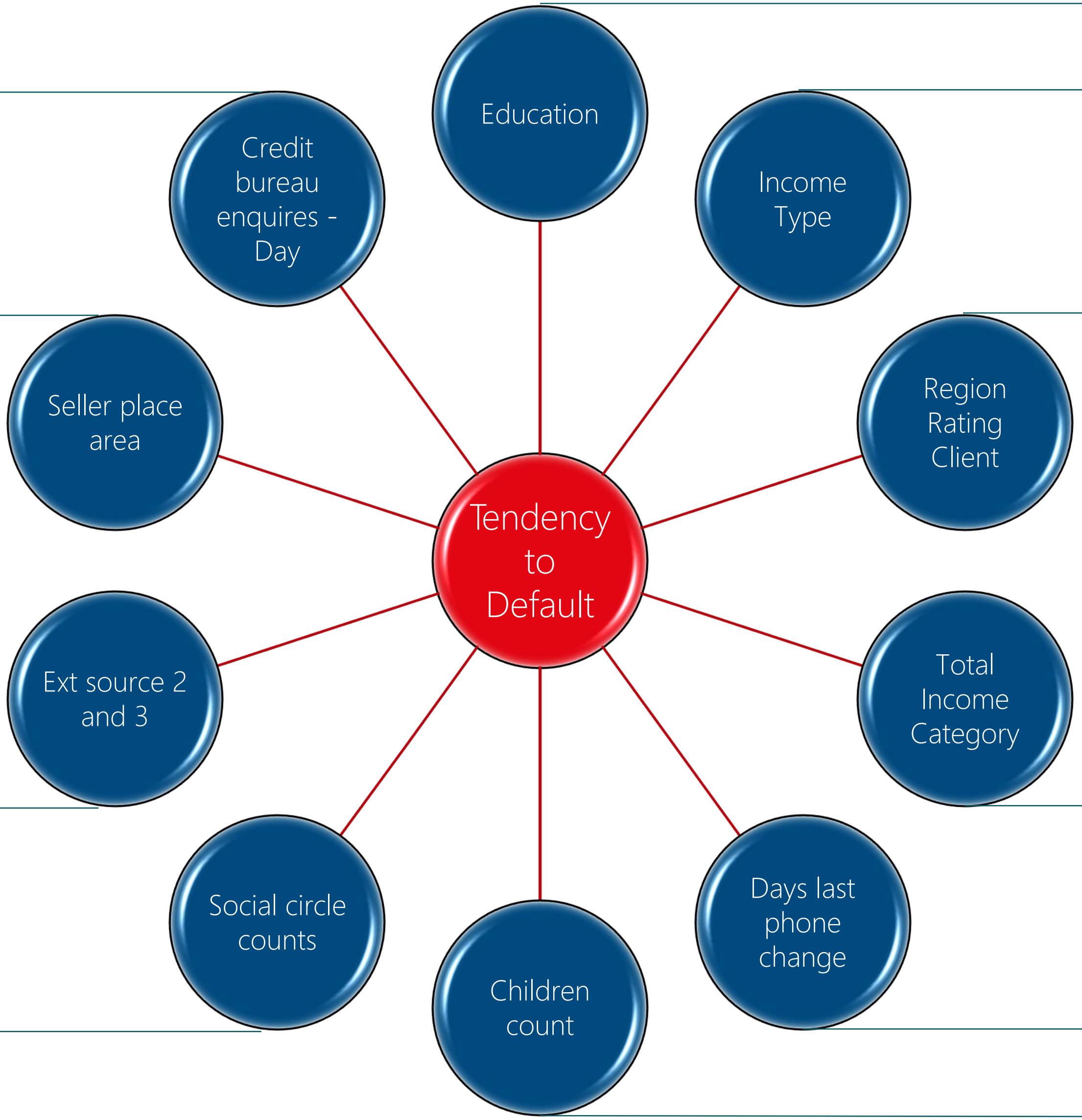


- ➢ Status to status comparison : mean of number of children is higher for defaulters than non-defaulters



Columns impacting the bank's decision:
AMT_CREDIT, AMT_ANNUITY, AMT_APPLICATION, AMT_GOODS_PRICE, CNT_FAMILY_MEMBERS, CNT_CHILDREN, OBS_60_CNT_SOCIAL_CIRCLE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, NAME_EDUCATION_TYPE, NAME_INCOME_TYPE, REGION_RATING_CLIENT, AMT_INCOME_TOTAL, DAYS_LAST_PHONE_CHANGE, EXT_SOURCE_2, EXT_SOURCE_3, SELLERPLACE_AREA, AMT_REQ_CREDIT_BUREAU_DAY, FLAG_OWN_CAR, FLAG_OWN_REALTY

Final EDA Summary



Maximum people not defaulting are of Higher education type

Seems when NUMBER of enquiries to credit bureau is higher, then default chances are also higher

Working people are expected to default more than Pensioners

Lower SELLERPLACE_AREA has a higher tendency for default

People staying in region 3 are more likely to default compared to region 1

Both SOURCE 2 & SOURCE 3 have influence – higher values have lesser default rate

Applicants with Medium Income Category are most likely to not default.

Lesser CNT on SOCIAL CIRCLE have higher default rate

Those who have a high value for "DAYS_LAST_PHONE _CHANGE' are likely to default

applicants having children above 11-12 are less likely to default

Thank You