Credit EDA Case Study

Sumit Raghuvanshi
Gaurav Rai

Problem Statement: Profit vs Risk

Profit

If an applicant who is **‘likely’** to repay the loan is not sanctioned the loan would result in loss of business for the company.

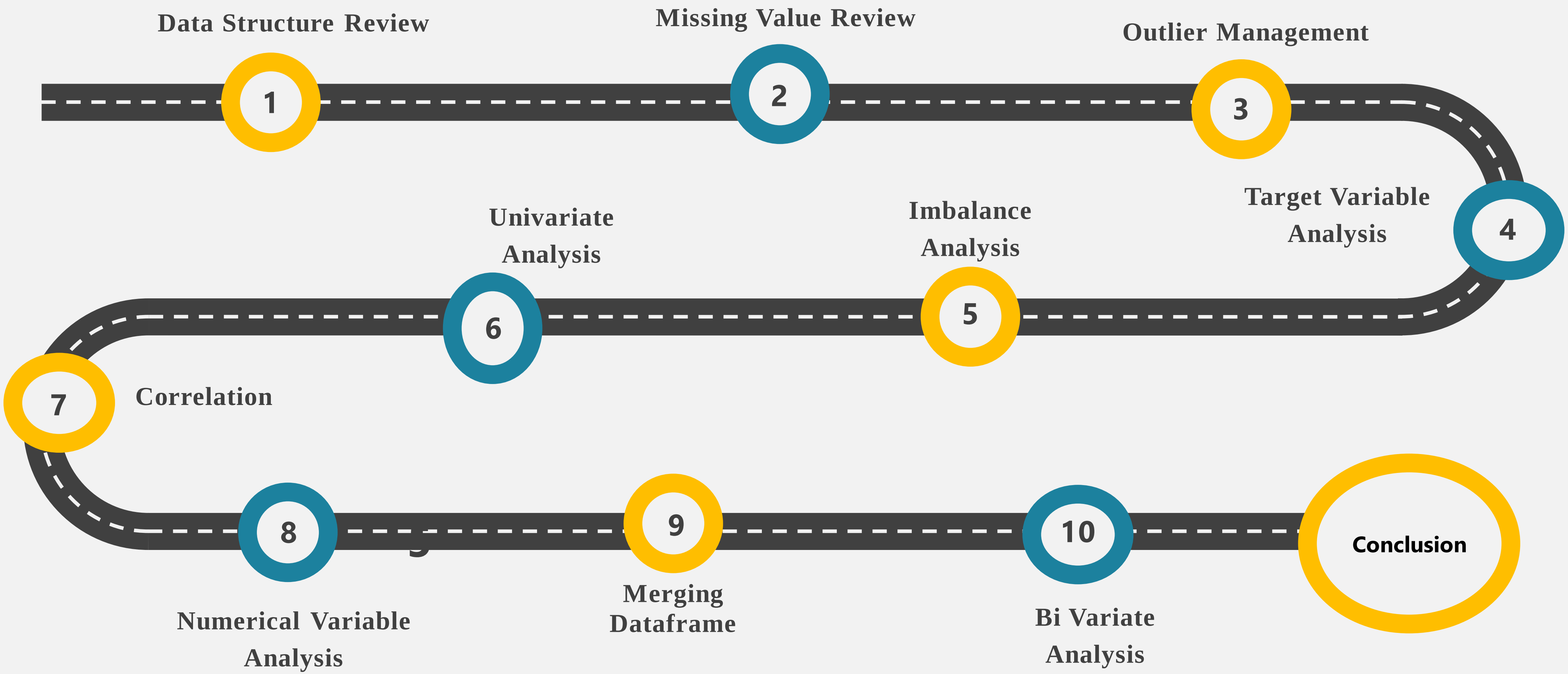
Business Objective

- ✓ To understand the driving factors (or driver variables) behind loan default.
- ✓ Identify the variables which are strong indicators of default.
- ✓ Utilize this for portfolio and risk assessment.

Risk

An applicant who is ‘**not likely**’ to repay the loan, (is likely to default) is sanctioned the loan may lead to a financial loss for the company.

Roadmap to case study



Data Analysis Approach and Data Cleaning

Data Analysis

We have conducted EDA on 2 Data sets provided to us :

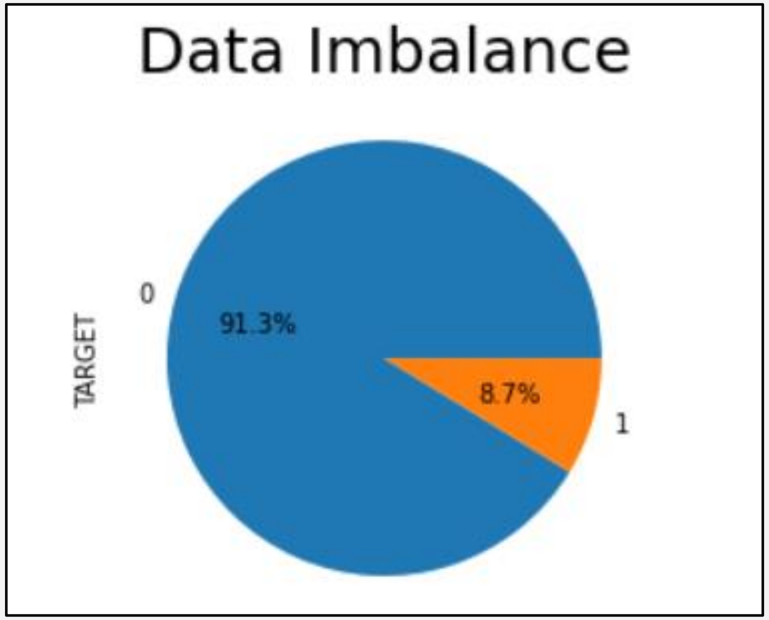
- **Application Data-** contains all the information of the client at the time of application. This data also has a variable (TARGET) stating whether the client has defaulted on any of his loan instalments. The data is about whether a client has payment difficulties.
- **Previous Application Data-** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Data Cleaning

- For this analysis we dropped the columns where missing values were greater than 30%.
- After that we drop columns which had null percentage greater than 30% in previous application data .
- Imputation of missing values was done on columns which have less missing values according to the column distribute and data type.
- For categorical type column 'NAME_TYPE_SUITE' we imputed missing values by mode.
- For numerical type column 'AMT_GOODS_PRICE' we imputed missing values by mean since the data was evenly distributed
- For numerical type column 'AMT_REQ_CREDIT_BUREAU_HOUR','AMT_REQ_CREDIT_BUREAU_DAY','AMT_REQ_CREDIT_BUREAU_WEEK' and few others we imputed missing values by 0, assuming missing values means no enquiry made for person.
- Converted the data type of DAYS_REGISTRATION, CNT_FAM_MEMBERS, OBS_30_CNT_SOCIAL_CIRCLE and few others from float to int
- Converted the Flag values of 0 & 1 to Y & N in categorical columns.
- Dropped variables in application data that weren't necessary.

Data Imbalance and Correlation

- High Data imbalance of 10.4
- 91% of loan applicant have never defaulted
- However ~9% of applicants have defaulted on their instalments at least once.
- We need to identify the reason for such a high imbalance



- Correlation for both Non-Defaulter (Tgt0) & Defaulter (Tgt1) are same

Top 10 Correlation list		
OBS_30_CNT_SOCIAL_CIRCLE	&	OBS_60_CNT_SOCIAL_CIRCLE
AMT_GOODS_PRICE	&	AMT_CREDIT
DEF_30_CNT_SOCIAL_CIRCLE	&	DEF_60_CNT_SOCIAL_CIRCLE
REG_REGION_NOT_WORK_REGION	&	LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_WORK_CITY	&	LIVE_CITY_NOT_WORK_CITY
AMT_ANNUITY	&	AMT_CREDIT
AMT_ANNUITY	&	AMT_GOODS_PRICE
REG_REGION_NOT_WORK_REGION	&	REG_REGION_NOT_LIVE_REGION
REG_CITY_NOT_WORK_CITY	&	REG_CITY_NOT_LIVE_CITY
AMT_ANNUITY	&	AMT_INCOME_TOTAL
AMT_INCOME_TOTAL	&	AMT_GOODS_PRICE

Univariate Analysis of Target Variable

- Used Application Data to analysis impact of various factors on Target Variables i.e Non-Defaulter (T0) and Defaulter (T1)
- Conducted Univariate Analysis to conclude the following :

Parameter	Conclusion
Gender	1. Female loan applicants (approx 63%) are more than male applicants (~37%). 2. However, proportionate default applicants are more amongst male applicants compared to female applicants. 3. We can say here that 'Males' are less credit worthy than females.
Profession	1. While 'Working' type are the highest loan applicant at 63.5% followed by 'Commercial associate' and 'State servant' however working population has higher defaulter proportion compared to non-defaulter population. 2. 'Commercial associate' and 'State servant' are better prospect from lending underwriting purpose.
Type of Loan	1. Loan applicants have requested for cash loans the most with ~90% of disbursed loan being cash loans. Albeit cash loan segment has a higher default proportion when compared to non-default proportion. 2. Hence we can infer that revolving loans are comparatively safer. 3. This may be attributed to the Nature of revolving loan as it is considered a flexible financing tool due to its repayment and re-borrowing flexibility hence people avoid defaulting on these loans.
Family	1. Single/ Unmarried people have higher defaulter proportion compared to its non-defaulter universe.
Education	1. Applicants with Secondary education have higher default proportion compared to non-defaulter. 2. People with higher education are more reliable from lending prospective.

- Pls see Appendix –> Slide 14, click

Types of Loan Decision

Refused

The company rejected the loan because the client does not meet their loan requirements.

Cancelled

Client cancelled the loan request due to plan change or Adverse Pricing being offered by the company.

Unused offer

Loan has been cancelled by the client but on different stages of the process

Approved

Company has approved the Loan Application

Bivariate Analysis-Types of Loan Decision

- Created New Merged data by combining - Application Data and Previous Application Data
- New merged dataframe separated based 4 types of loan decision- Approved, Refused, Cancelled, Unused
- Conducted Bivariate Analysis to conclude the following :

Parameter	Conclusion
PRODUCT_COMBINATION	1. Most number of loans were approved for POS household with interest. 2. Most number of refused loans were of Cash X-Sell. 3. Most Canceled loans were Cash loan.
CHANNEL_TYPE	1. Most approved loans were from Country-wide Channel. 2. Most refused loans were from Credit and Cash Offices Channel
FAMILY_STATUS	1. Married segment accounted for highest number of loans Approved, Refused and Rejected. 2. This is in line with this segment being the highest applicant of loans. Hence we can't conclude anything from this
HOUSING	1. Housing/Apartment segment accounted for highest number of loans Approved, Refused and Rejected. 2. This is in line with highest number loans being applied for Housing segment only. Hence we can't conclude anything from this .
Education	1. Secondary/Secondary Special segment accounted for highest number of loans Approved, Refused and Rejected. 2. This is in line with highest number loans being applied by people with Secondary/Secondary special education. Hence we can't conclude anything from this .

- Pls see Appendix –> Slide 15, click. Pls note that we have not included Family Status, Housing and Education charts in this slidedeck as these three were not reflecting any important analysis . These three graphs can be viewed in the python file for reference if needed.

Bivariate Analysis (contd..)

- Created New Merged data by combining - Application Data and Previous Application Data.
- Used variables from both datasets.
- Conducted Bivariate Analysis to conclude the following :

Parameter	Conclusion
Application Amount vs Housing Type	1. It has been observed that Credit/loan request by applicants was highest for ‘Office Apartment’ followed by ‘municipal apartment’ and ‘house/apartment’. 2. However, both ‘Office apartment’ and ‘municipal apartment’ have reported higher defaulter (target1) proportion than non defaulter(target0). 3. So, we can conclude that bank should avoid giving loans for office and municipal apartment as they are having difficulties in payment. 4. Bank can focus mostly on disbursing loans for house/apartment , co-op apartment.
Loan Purpose vs Income type	1. The credit amount of Loan purposes like ‘Buying a new car’, Purchasing Electronic’, ’Buying a house’, ‘Medicine' and ‘Building a house' is higher. 2. Income of state servants and commercial associate who have applied for loan is significantly higher than other loan applicants. 3. Loan applied for 'Hobby' & 'garage buying' is significantly low.

- Pls see Appendix –> [Slide 16, click]

Conclusion

Safety

Revolving loans are comparatively safe loans to be disbursed.

Education

Clients with Low Education have high default rate. High Educated individuals should be target customers.

Product Combination

'POS household with interest' is a successful product offering with highest approval rate. Hence this can be a good product offering

Channel Distribution

Loans routed via 'Country-wide Channel' were most approved loans. Hence clients originated via Country wide channel should be emphasised.

Portfolio Mix

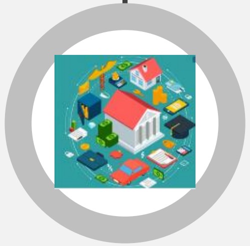
Avoid giving loans for 'Office' and "Municipal apartment" as they are having difficulties in payment.

Can focus mostly on disbursing loans for 'House/apartment, Co-op apartment

Contd…

Conclusion

Purpose



Loan amount of ‘Buying a new car’, ’Buying a house’, ‘Medicine' and ‘Building a house' is higher.

Gender

'Males' are less credit worthy than females. Bank can target female clients as they have lower default rate.

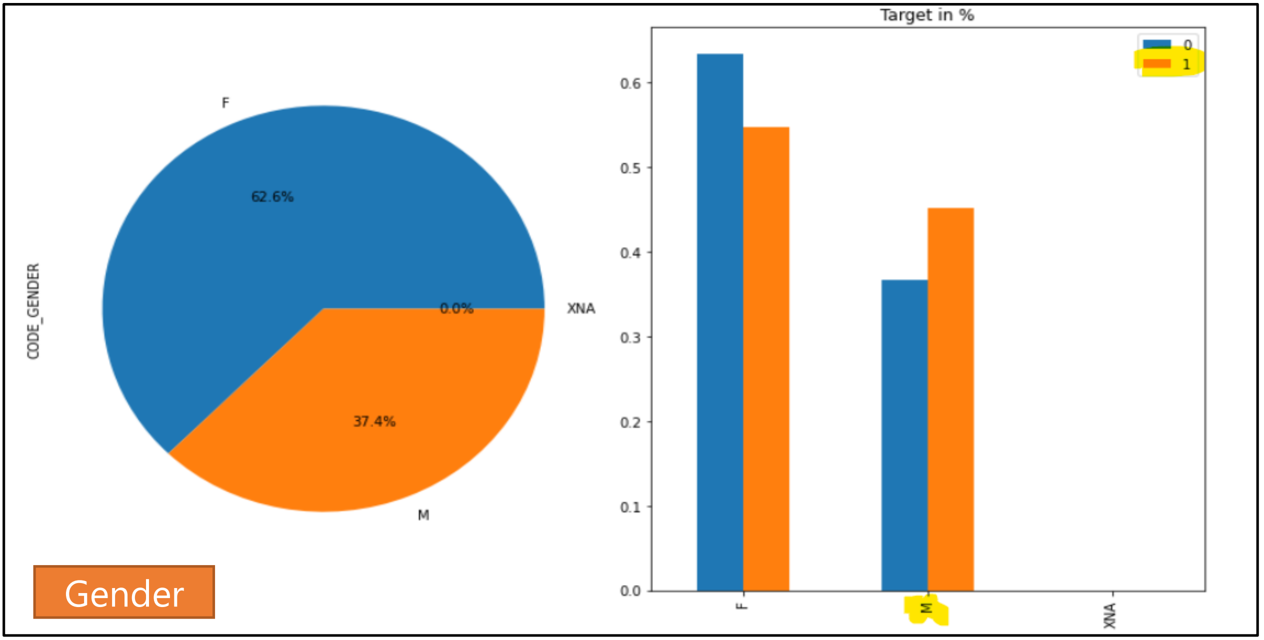
Profession

"'Commercial associate' and 'State servant' are better prospect from lending purpose as they have low default rate.

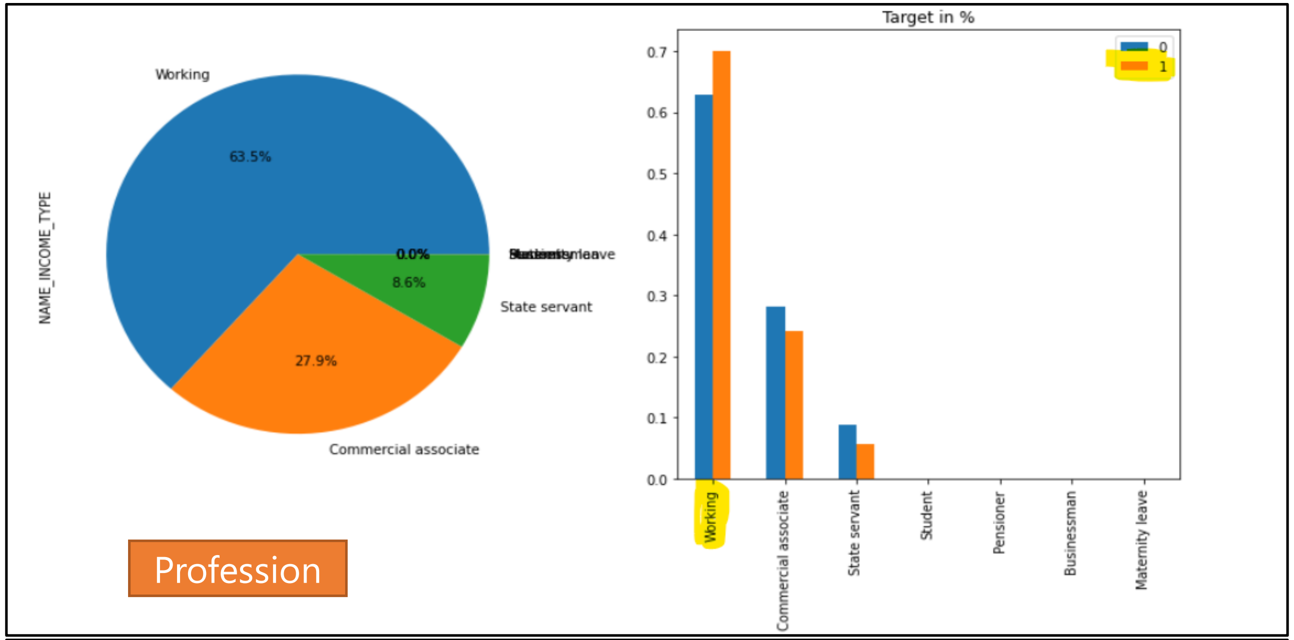
THANK YOU

Appendix

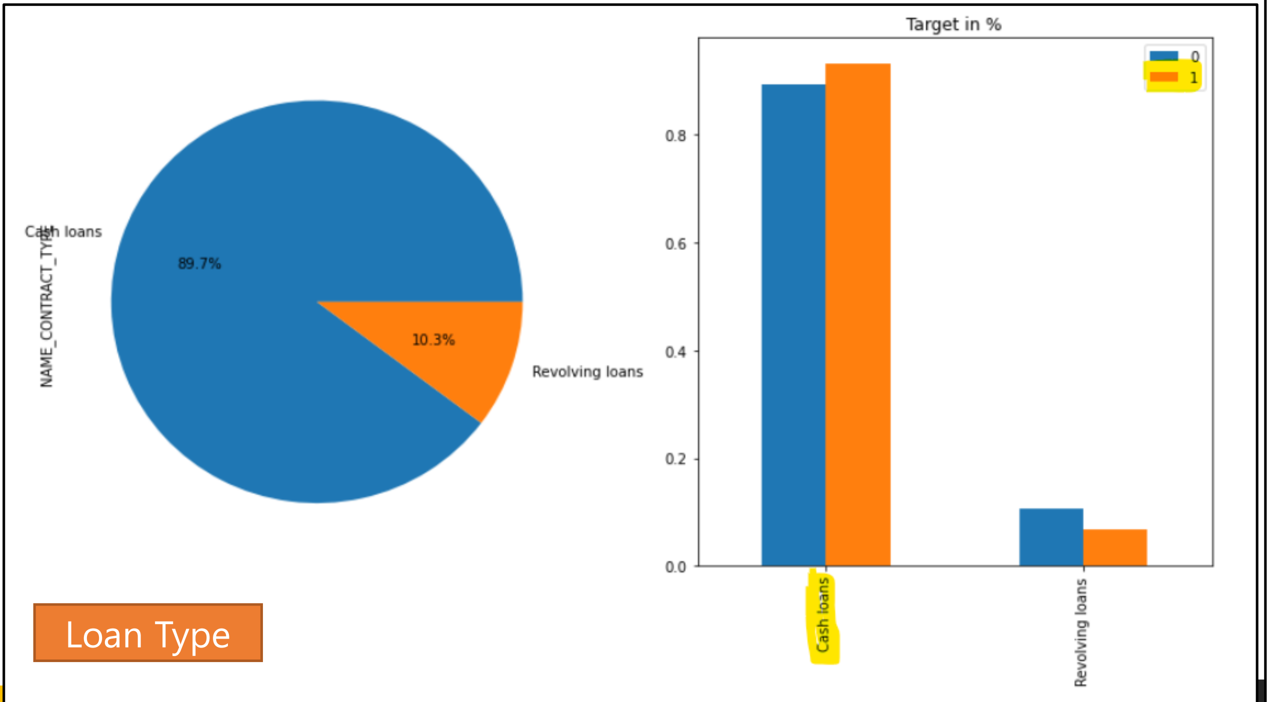
Univariate Analysis



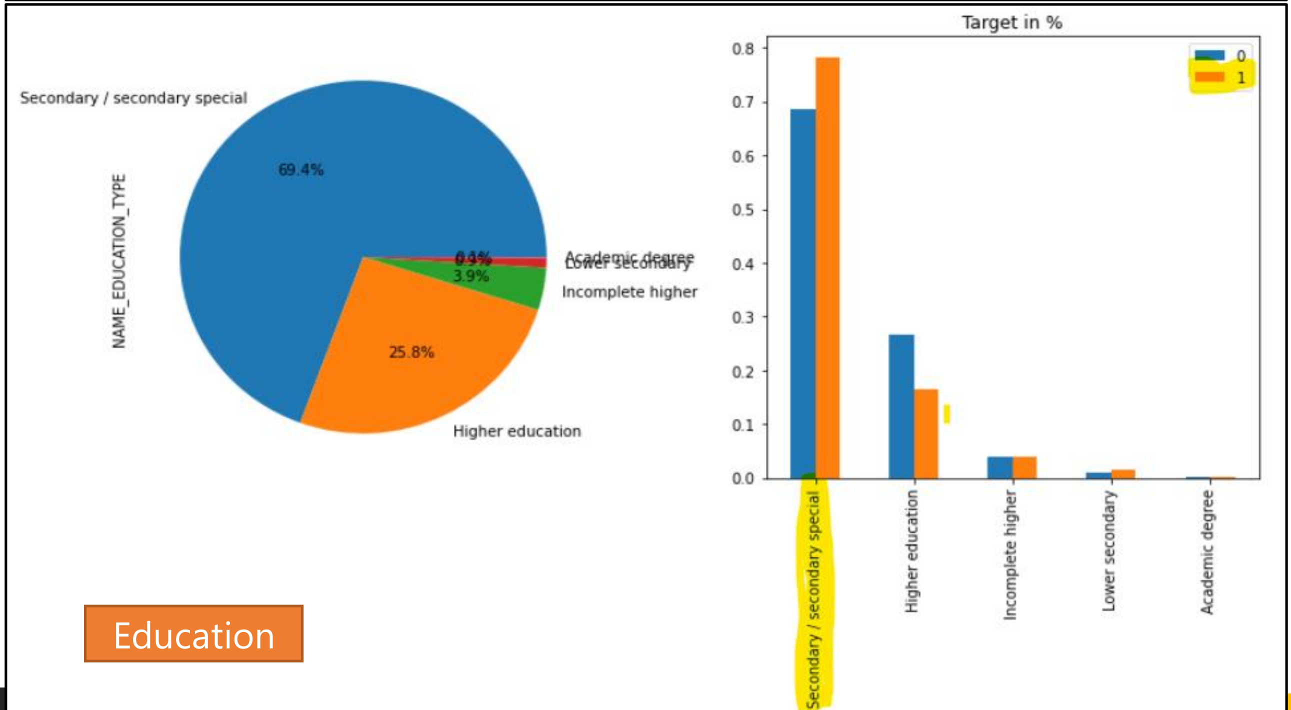
Gender



Profession

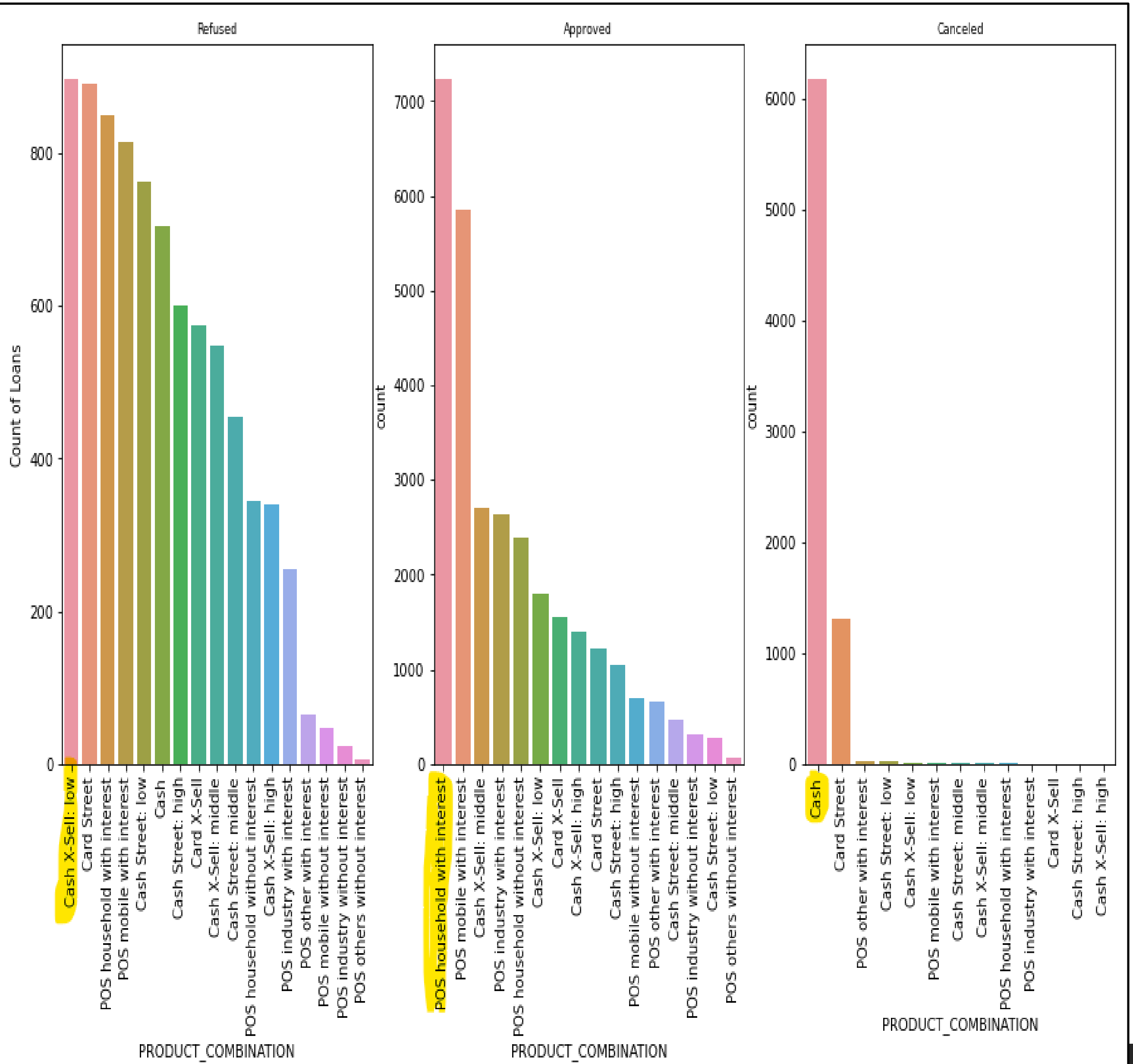


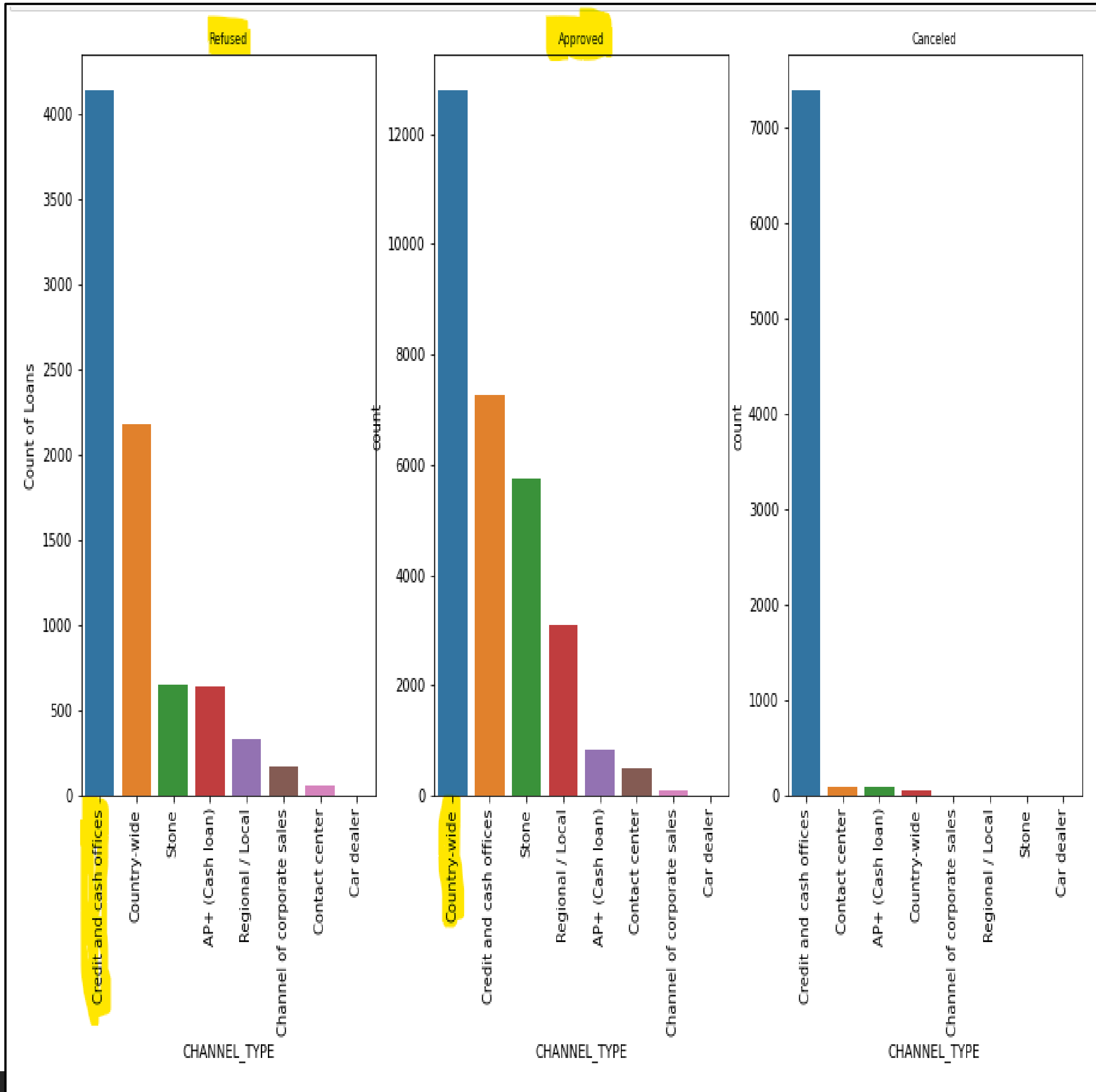
Loan Type



Education

Bivariate Analysis-Types of Loan Decision





Bivariate Analysis-Loan Purpose vs Income Type