Exploratory Data Analysis of CREDIT Dataset

Antarlin Chanda

Contents

Introduction

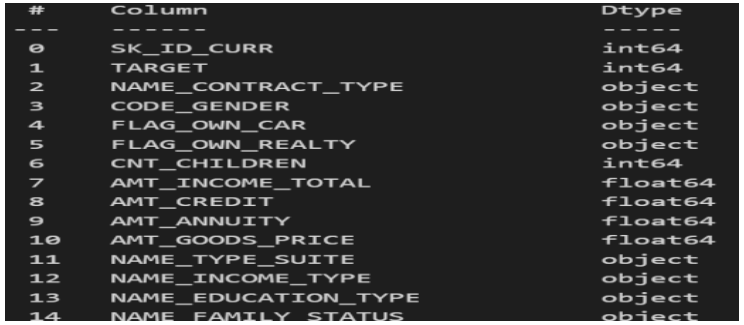
Businesses find it hard to decide who to give loan to and who not to. They have to maintain a delicate balance to ensure that they are ale to segregate customers who can repay loan and those who cannot so that they can refine their business strategy so as not to run into loss from NPA and such

Our objective in this project will be to generate insights and strategy so as business gets an idea of whom to give loan to and whom not to through Exploratory Data Analysis of Current Loan Applications and previous Loan Applications

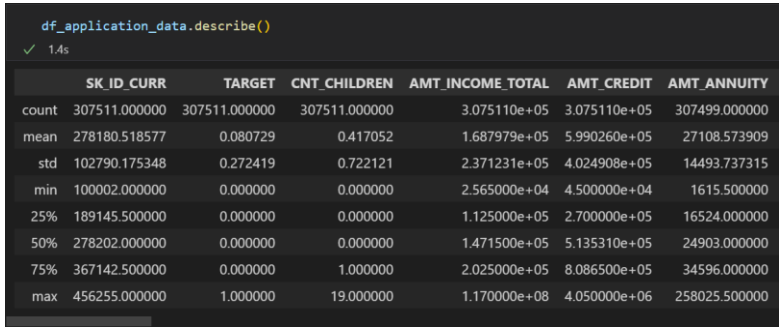
The outcomes will help refine business strategy for marketing campaigns, targeting appropriate customers to give loan to and help generate more profit.

Processes Steps

Dataset Structure Inspection- Here we inspected various attrbutes of the dataset like its columns data types, dataframe shape and descriptive statistics to get and idea of what we are dealing with. For example we had 121 columns in loan application dataset



#	Column	Dtype
0	SK_ID_CURR	int64
1	TARGET	int64
2	NAME_CONTRACT_TYPE	object
3	CODE_GENDER	object
4	FLAG_OWN_CAR	object
5	FLAG_OWN_REALTY	object
6	CNT_CHILDREN	int64
7	AMT_INCOME_TOTAL	float64
8	AMT_CREDIT	float64
9	AMT_ANNUITY	float64
10	AMT_GOODS_PRICE	float64
11	NAME_TYPE_SUITE	object
12	NAME_INCOME_TYPE	object
13	NAME_EDUCATION_TYPE	object
14	NAME_FAMILY_STATUS	object



```
df_application_data.describe()
```

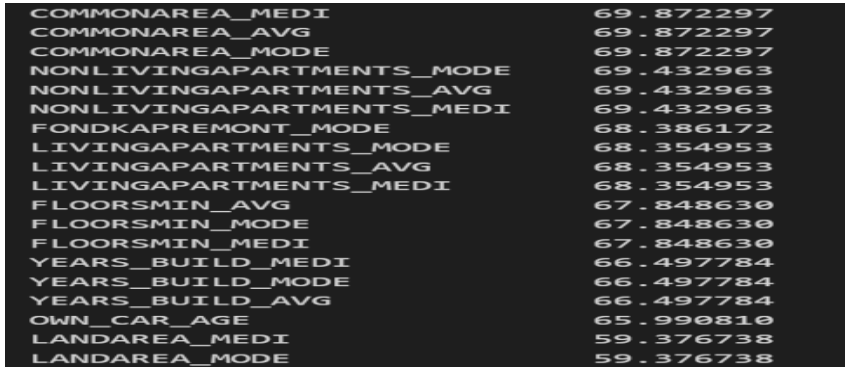
✓ 1.4s

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	307499.000000
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	27108.573909
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	14493.737315
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	1615.500000
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16524.000000
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000

Handling Incorrect/Invalid Datatypes- We checked the dataset to see if any variable datatypeis wrong like integer variable present as categorical variable and vice versa. No such problem was found.

P.S-Due to paucity of space we have shown pictorial representation of only a few inspections done. Detailed graphs charts available in notebook.

Handling Missing Values-We checked amount of data missing in each columns if more than 35 percent of data is missing we have removed the columns since using imputation may introduce bias. For the remaining columns with missing values if datatype was categorical we imputed missing if numerical imputed by median value(median less effected by outliers) .Figure shows % of missing values in the respective columns(some of them)



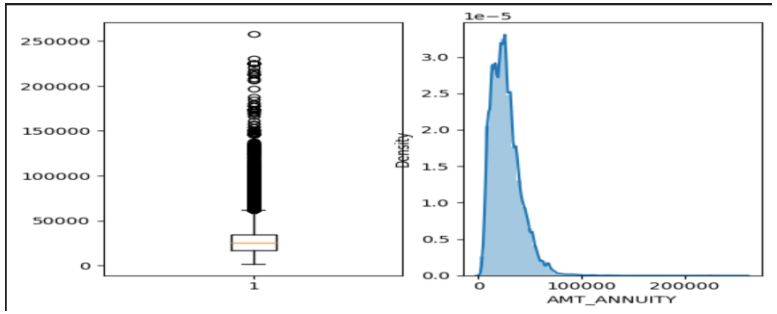
COMMONAREA_MEDI	69.872297
COMMONAREA_AVG	69.872297
COMMONAREA_MODE	69.872297
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAPARTMENTS_AVG	69.432963
NONLIVINGAPARTMENTS_MEDI	69.432963
FONDKAPREMONT_MODE	68.386172
LIVINGAPARTMENTS_MODE	68.354953
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAPARTMENTS_MEDI	68.354953
FLOORSMIN_AVG	67.848630
FLOORSMIN_MODE	67.848630
FLOORSMIN_MEDI	67.848630
YEARS_BUILD_MEDI	66.497784
YEARS_BUILD_MODE	66.497784
YEARS_BUILD_AVG	66.497784
OWN_CAR_AGE	65.990810
LANDAREA_MEDI	59.376738
LANDAREA_MODE	59.376738

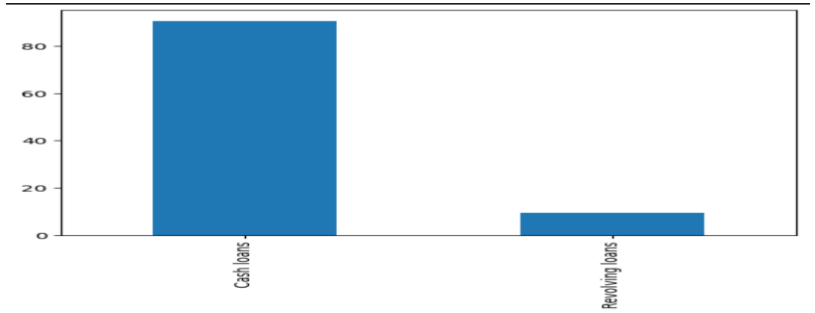
Feature engineering- We created new features like converting variables in day time period to years for getting a better understanding of underlying long term trends.

Data Balance Checking-Checked balance between categories present in the target variable.Datasets should have similar data balance for data modelling and such.We found high data imbalance.0 or no default has almost 92% presence while defaulters are present only in 8% cases.

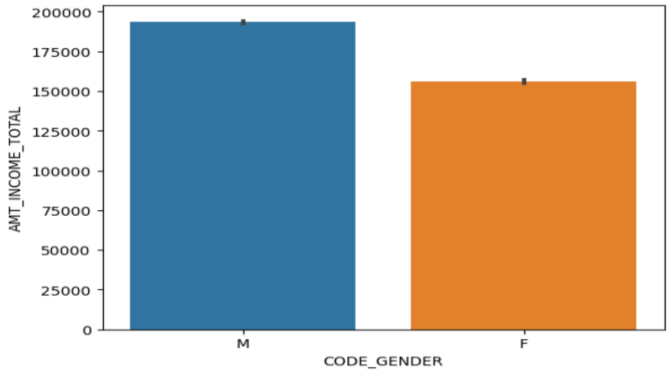
```
df_application_data.TARGET.value_counts(normalize=True)*100
✓ 0.0s

0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

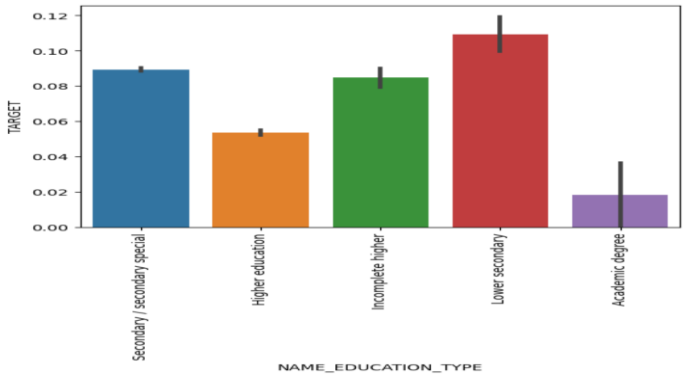
Univariate Analysis- Checked Distribution of variables through descriptive statistics and through boxplots and histograms with density. Derived individual variable level insights. Below we can see distribution of Annuity amount of loans.We see how right skewed the data is.Also we can find the outlier cutoff points at both sides.For categorical we can variables shares like shown below for loan contract type.



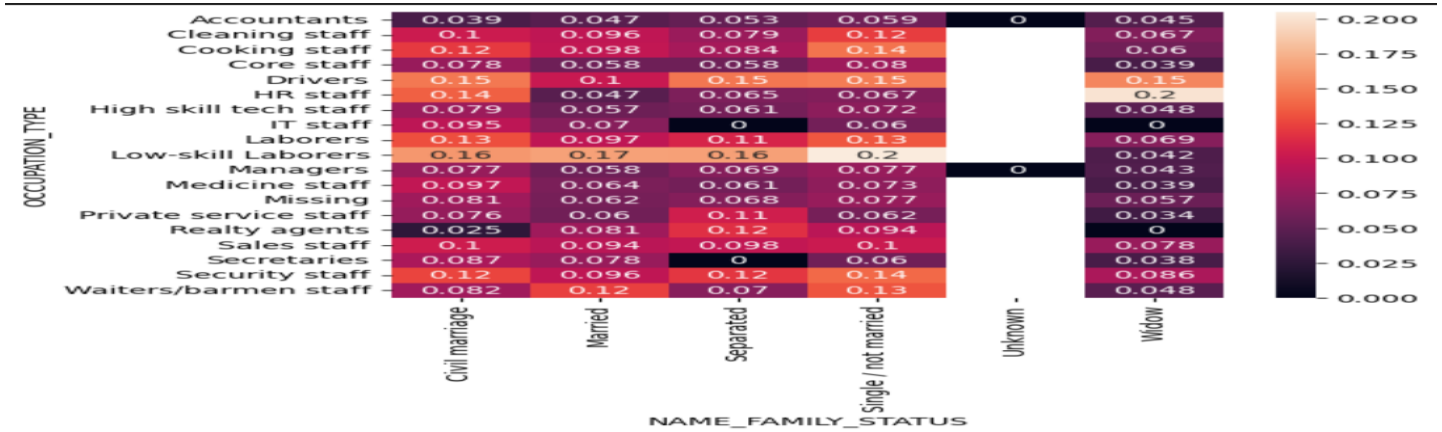


Bivariate Analysis-In bivariate Analysis we compare 2 variables to check their relationships and find related insights.Be low we have compare income amount between Males and Females. We found how education levels affect default chances.





Multivariate Analysis- Here we find out relationships between 2 or more variables. In example shown below we have checked default rates based on occupation type and family status.We found how accountants are least likely to default and how low skilled labourers have high chance to default



Previous Application Analysis- We have done similar analysis steps as shown above for the previous loan application datasets and derived insights.

Merged Dataset Analysis- We also merged the loan application and previous loan dataset to inspect both datasets side by side and derive insights.

Insights

1. Small amount Loans of the order of 500000 are very popular
2. People around middle age 35-55 are highest loan applicants
3. People employed for more than 18 years rarely apply for loans
4. Cash Loans are very much more popular compared to revolving loans almost(5 to 6 times)
5. Females comprise almost 65% of loan applicants while males 35%
6. Managers are the highest paid among professions
7. Working People comprise more than 50 percent of loan applicants
8. Secondary/Secondary Special Education level are the highest loan applicants(around 70%)
9. Married People Comprise of almost 65% of Loan Applicants
10. People Owning House/Apartment comprise of almost 90 percent of loan applicants
11. Young People(<38) have the highest chance to default followed by medium(38-55) and then old people(55+)

12. Medium Amount (350k to 700k) Loan Takers have more propensity to default compared to High(700k+) and Low Amount Applicants (45k to 350k)
13. Businessmen Apply for the largest size loans while students the lowest
14. Maternity Leave Individuals and Unemployed individuals have the highest chance to default
15. Students and Businessman have negligible chance to default on loans
16. Males have more chance to default compared to females(10% vs 7% chance) even though more females apply for loans
17. Cash Loans have higher chance to default compared to Revolving Loans (8.5% vs 5.5%)
18. Lower Secondary Education has the highest chance to default(around 11%) while Academic Degree (2%) and Higher Education(~5%) have the lowest default chance.
19. People living in Rented Apartments(12%) have the highest chance to default whereas Office Apartment(~6.5%) has lowest chance
20. Low Skilled Laborer have highest default chance at(17%) while Accountants have lowest chance(~5%)
21. Default chances become higher with more no. of children

22. Region Ratings show increasing default chance with increasing rating number
23. Recently employed people show the highest chance to default on loans
24. Recent ID Document change indicates more chance to default on loan
25. Widowed Hrs , Widow Drivers, Single Drivers, Single Security Staff, Single Laborer, Single Cooking and Cleaning Staff, Separated Laborer, Separated Drivers, Married Low Skilled Laborer, Married Waiters, Civil Marriage Security Staff, Civil Marriage Laborer and Civil Marriage HR Staff have difficulty in payments
26. Accountants, Core Staff, High Skill tech Staff and IT Staff they have comparatively lower default rates
27. Low Skilled Male and Female Laborer, Male Reality Agents, Male Secretaries have highest chances of default
28. Female and Male Accountants, Female Managers, Female Core Staff groups have lowest chances to default
29. Low Skilled Laborer in Industry Type 10,Medicine Staff Industry Type 7,It Staff in Legal Services, Security Staff in Realtor, Secretaries in Services,
30. Cooking Staff in Trade Type 2,Core Staff, Sales Stuff, and Waiters in Transport Type 3are showing almost a 100percent chance to default.
31. Other Subgroups have also shown higher default chances like Private Service Staff in Agriculture, Reality Agents in Business Entity Type 2 etc.

32. Unemployed who have secondary or incomplete higher education have a high chance to default.
33. Secondary education on maternity leaves have highest chances to default
34. Higher Education and Academic Degree groups have shown lower defaults
35. Civil Marriage and Single People show the highest chance to default while widows have lowest chance to default.
36. AP + Cash Loan have the highest chance to default
37. Animal and Weapon Loans have the lowest chance to default while Vehicle and Insurance Loans have the highest chance to default
38. SCOFR Rejection reason for previous loan application have the highest chances to default and systems reason have lowest chances
39. Cards portfolio for previous loan application have the highest chance to default while Car portfolio have the lowest chance to default
40. AP + Cash Loan Channel have the highest chance to default while Car Dealer Channel has lowest chance
41. Loans given by Tourism Industry have the lowest chance to be defaulted while Auto Technology Loans have the highest chance to be defaulted

More Risky Group Characteristics

Following List can be used as guidelines for excursing caution when disbursing loans(in terms of individuals and loan type)

1. **Young Recently Employed Individuals**
2. **Single or Civil Marriage**
3. **Recently changed ID Document before loan application**
4. **Low Skilled Male and Female Laborer, Male Reality Agents, Male Secretaries**
5. **Lower Secondary Education**
6. **Living in Rented Apartments**
7. **Unemployed or Maternity Leave Individuals**
8. **Individuals with very large families**
9. **Individuals coming in large group to apply for loan**
10. **Widowed Hrs , Widow Drivers, Single Drivers, Single Security Staff, Single Laborer, Single Cooking and Cleaning Staff, Separated Laborer, Separated Drivers, Married Low Skilled Laborer, Married Waiters, Civil Marriage Security Staff, Civil Marriage Laborer and Civil Marriage HR Staff**
11. **Vehicle and Insurance Loans**
12. **AP + Cash Loan Channel for customer Acquirement**
13. **Auto Technology Loans**

Less Risky Group Characteristics

Following List can be used as guidelines for customer targeting and loan disbursal strategy

1. **Businessman and Students and Pensioners**
2. **Female and Male Accountants, Female Managers, Female Core Staff groups**
3. **Academic Degree and Higher Education**
4. **Accountants, Core Staff, High Skill tech Staff and IT Staff**
5. **Employed for more than 10 years**
6. **Animal and Weapon Loans**
7. **Low Number of Children(1 or 2)**
8. **Widows**
9. **Living in Office Apartment or won Apartment**
10. **Car Dealer Channel for Customer Acquirement**
11. **Loans given by Tourism Industry**

Conclusion

The insights and recommendations provided above also the graphs and charts and comments provided in Notebooks

Together they can be used to get even deeper understanding of the analysis performed and the results.

Now along with the business stakeholders we can formulate a strategy to target certain segment of customers, how to spend advertising budgets and through the overall process we should be able to bring down the risk associated with disbursing loans thus reducing Non performing Assets and increasing profits

THANKS