EDA Case Study

Sandeep Srikanti

Problem Statement

- Perform EDA on the “Credit Loan Dataset” to understand the patterns leading to a consumer default
- Understand the key variables that drive “Loan Default” – Factors that differentiate between a good customer vs a customer with high likelihood of default

Overall Approach For Data Analysis

- Data Understanding, Cleansing & Manipulation
- Dealing With Columns Having Missing Data, Outlier Analysis & Treatment
- Univariate Analysis: Understanding the various columns & drawing insights
- Segmented Univariate Analysis: Segmenting According To Target Variable
- Bivariate Analysis: Understanding the relationship between multiple variables to see if there is a pattern or a trend - Categorical Variables & Numerical Variables
- Merging With Previous Data – Final Analysis
- Recommendations

Note: Only Few Snippets have been shown for each category; to save on no of slides. Full analysis available in jupyter notebook

Data Understanding, Cleansing & Manipulation

- Understanding the various columns & drawing insights
- Outlier Analysis

Dealing With Columns Having Missing/ Incorrect Values | Application Data Approach

Application_Data

- Dropped Columns having 40% or More Missing Values

Columns For Which Missing Data between 0-40%:

- OCCUPATION_TYPE: Set Missing Values to “Others”

- " EXT_SOURCE_3“: Almost 19% of data is missing for this external source 3. This seems like a score data that has been retrieved from external source. Better to impute it with a NaN value. We cannot use the mode or median to replace the value ; as it could lead to skewed analysis. At this stage, not sure if this score is important or not.

- #AMT_ANNUITY has around 12 rows where the data is missing. Better to drop as the % is quite less. or 0.004%

- #NAME_TYPE_SUITE has around 1292 rows where the data is missing.Since the % is only 0.4%, we can consider replacing by# mode or the most common value or we could choose to drop the records. Replaced with “Unaccompanied”

Dealing With Columns Having Missing/Incorrec Values | Application Data Approach

Columns For Which Missing Data between 0-40%:

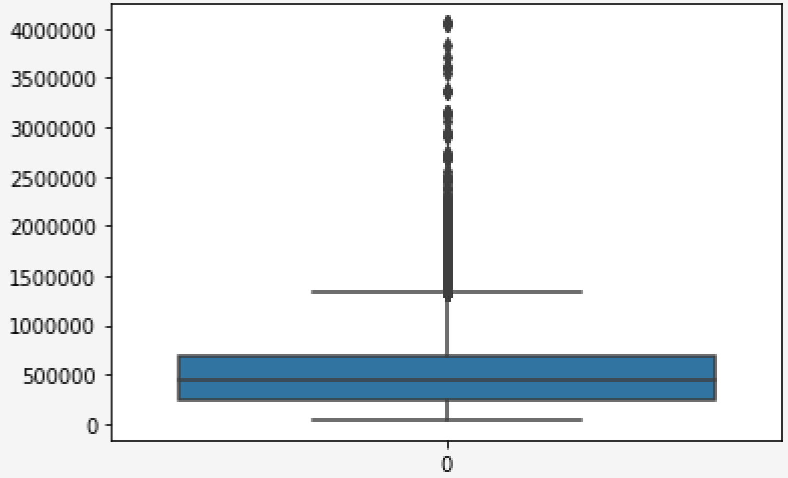
- OBS__30/60_CNT, DEF_#30/60_CNT columns have 0.33% null values. Replaced with np.NaN to make sure that the data doesn’t affect our analysis
- Retained the null values in AMT_REQ_CREDIT_BUREAU columns as this data might be important. It is possible that no enquiry was raised against a client.

Data Manipulation

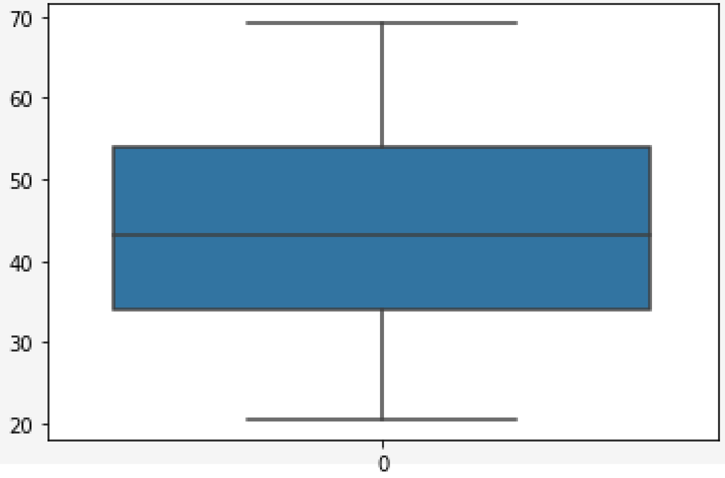
- Converted DAYS _* columns to absolute values as they are negative numbers.
- Converted DAYS_EMPLOYED, DAYS_BIRTH to Years by adding 2 new columns
- Created Bins for Salary, Loan etc. using new variables like salary grp, age grp, loan grp – Will be useful for categorical analysis later on
- Created new variables for credit risk default analysis: Loan as % of salary, Annuity as % of Salary

Dealing With Outliers – Application Data | Approach

AMT_GOODS Price: The data seems ok. Most of the goods price seems to lie between 2.4 to 6.8 Lakhs. There are outliers. But they seem reasonable

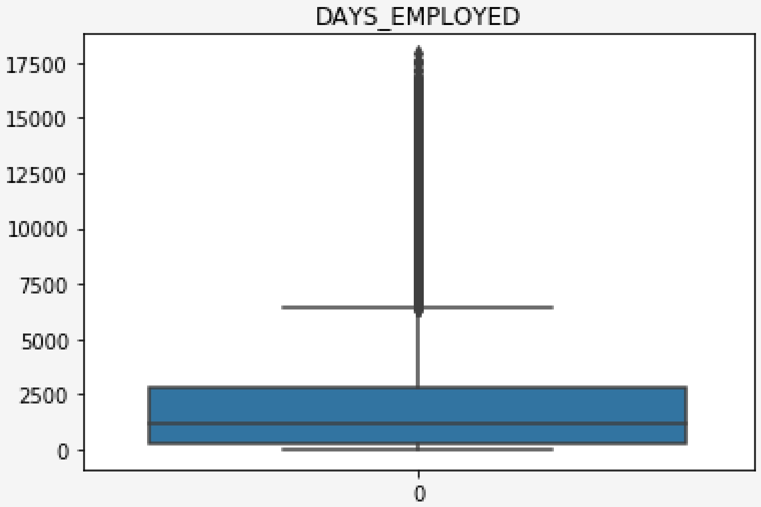


YEARS_BIRTH (derived column): The data seems ok. No outliers in this.



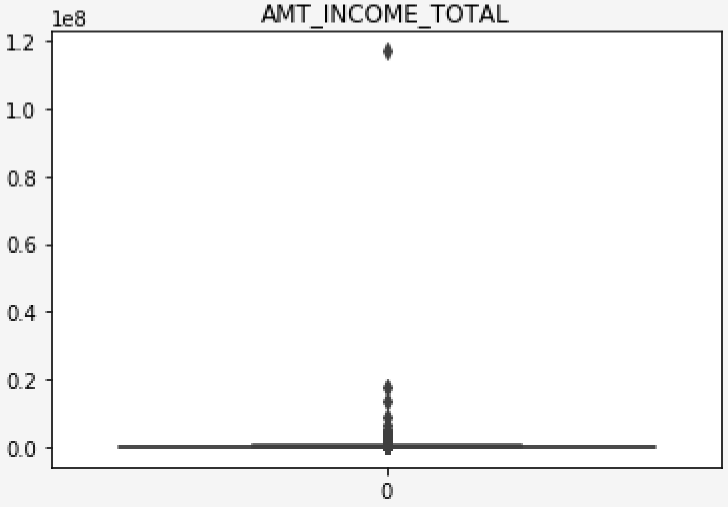
Dealing With Outliers – Application Data | Approach

DAYS_EMPLOYED (YEARS_EMPLOYED) : Also, there is a number 365243 which is appearing 55374 times. This 365243 number seems to be used for pensioners or unemployed people. This seems like an outlier as it is not realistic. We can bin them or cap them. Without the outliers, the max is around 50 years or equivalent no of days. So, its better to impute records with DAYS_EMPLOYED == 354243 with 0. 0 makes sense as Pensioners and unemployed people are not currently working



AMT_INCOME: It also has outliers. Created Bins to deal with these outliers. Created a new variable called salary_range to bucket these outliers.

Similarly, Checked for other numerical columns (Refer to the jupyter notebook).



Dealing With Columns Having Missing/Incorrect Values | Previous Application Data Approach

Previous Application_Data

- Dropped Columns having 40% or More Missing Values

- Dropped all columns (except NFLAG_INSURED_ON_APPROVAL). Hypothesis: Guess this might be an important variable which depicts the risk profile of the customer – where the customer is requesting for insurance

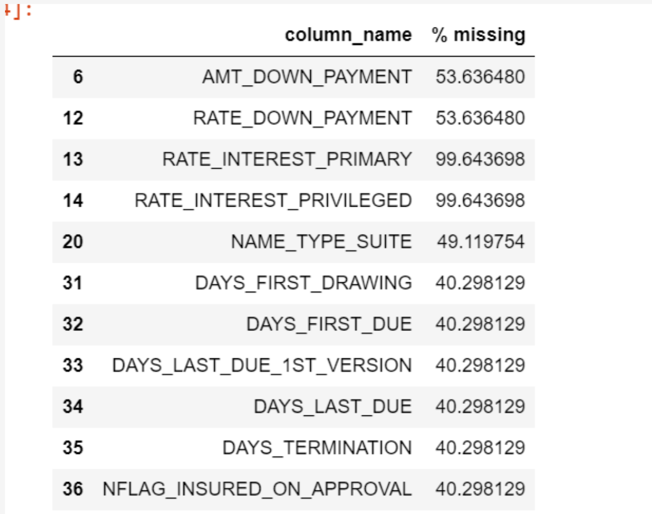
Columns For Which Missing Data between 0-40%:

- AMT_ANNUITY, CNT_PAYMENT, AMT_GOODS_PRICE: Didn’t Impute any values as the missing percentage is higher and imputing might skew our analysis
- PRODUCT COMBINATION: Impute with Cash as Cash has the highest no of applicants
- AMT_CREDIT: Missing values is only 0.0006%. Left it and didn’t impute it. It wont affect our analysis

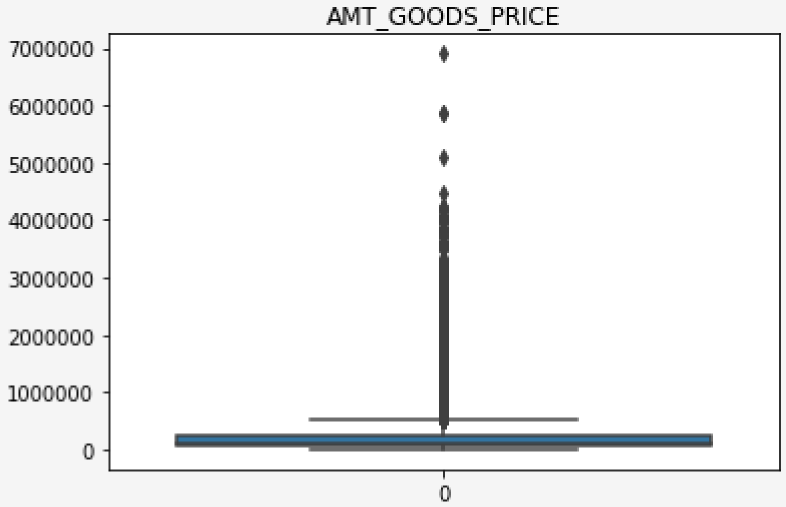
Data Manipulation

- Converted DAYS _* columns to absolute values as they are negative numbers.

- Replaced XNA in CLIENT_TYPE with most frequent occurance : “Repeater”



	column_name	% missing
6	AMT_DOWN_PAYMENT	53.636480
12	RATE_DOWN_PAYMENT	53.636480
13	RATE_INTEREST_PRIMARY	99.643698
14	RATE_INTEREST_PRIVILEGED	99.643698
20	NAME_TYPE_SUITE	49.119754
31	DAYS_FIRST_DRAWING	40.298129
32	DAYS_FIRST_DUE	40.298129
33	DAYS_LAST_DUE_1ST_VERSION	40.298129
34	DAYS_LAST_DUE	40.298129
35	DAYS_TERMINATION	40.298129
36	NFLAG_INSURED_ON_APPROVAL	40.298129



Univariate Analysis – Complete Dataset

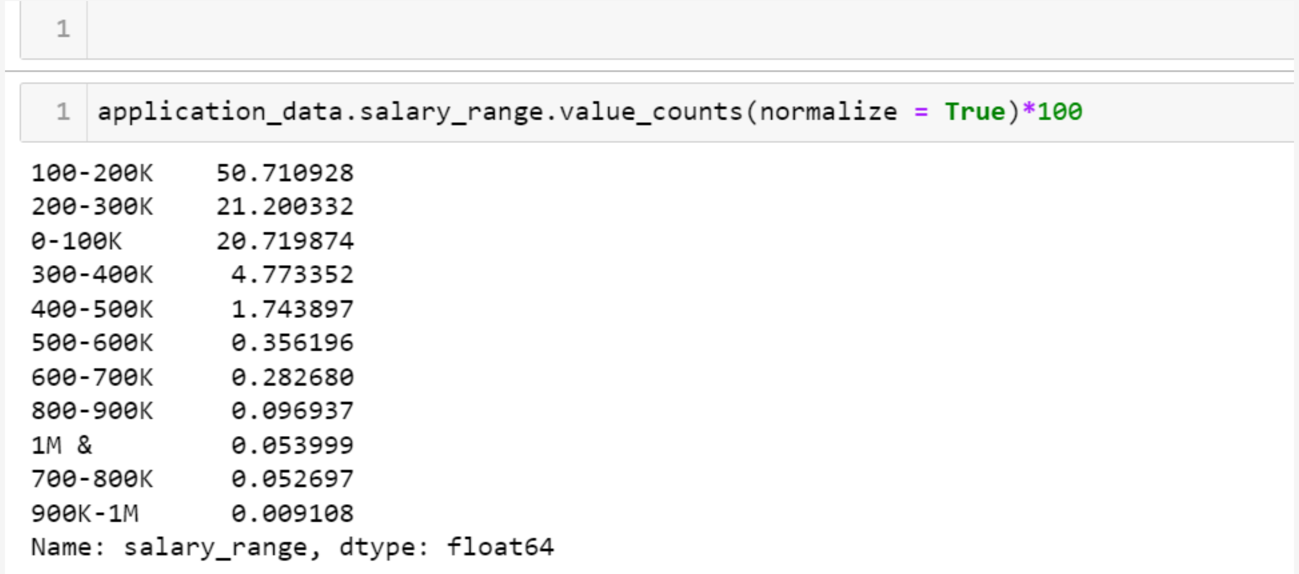
- Understanding the various columns & drawing insights

- Outlier Analysis

Univariate Analysis
Salary

- ~90% of Loan Applicants seem to be in the income range of 300K and below(3 Lakhs and below)

- There are some outliers in the salary column. But this seems perfectly fine for our analysis. No need to drop them



```
application_data.salary_range.value_counts(normalize = True)*100
```

```
100-200K    50.710928
200-300K    21.200332
0-100K      20.719874
300-400K     4.773352
400-500K     1.743897
500-600K     0.356196
600-700K     0.282680
800-900K     0.096937
1M &         0.053999
700-800K     0.052697
900K-1M      0.009108
Name: salary_range, dtype: float64
```

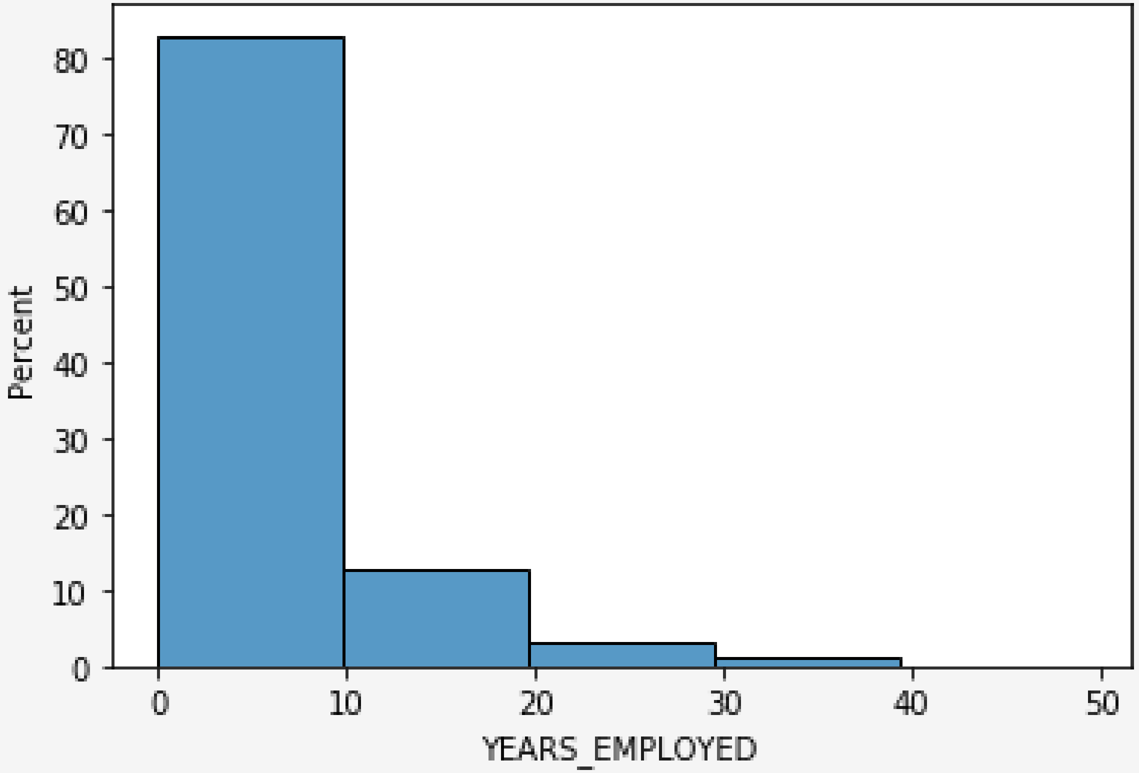
Univariate Analysis
Employment In No. Of Years (DAYS_EMPLOYED converted to Years)

- ~80% of Loan Applicants seem to be in the Employment Year Bucket of 0-10 Years

- There are some outliers in the Employment Year (After Converting from Days to Years)

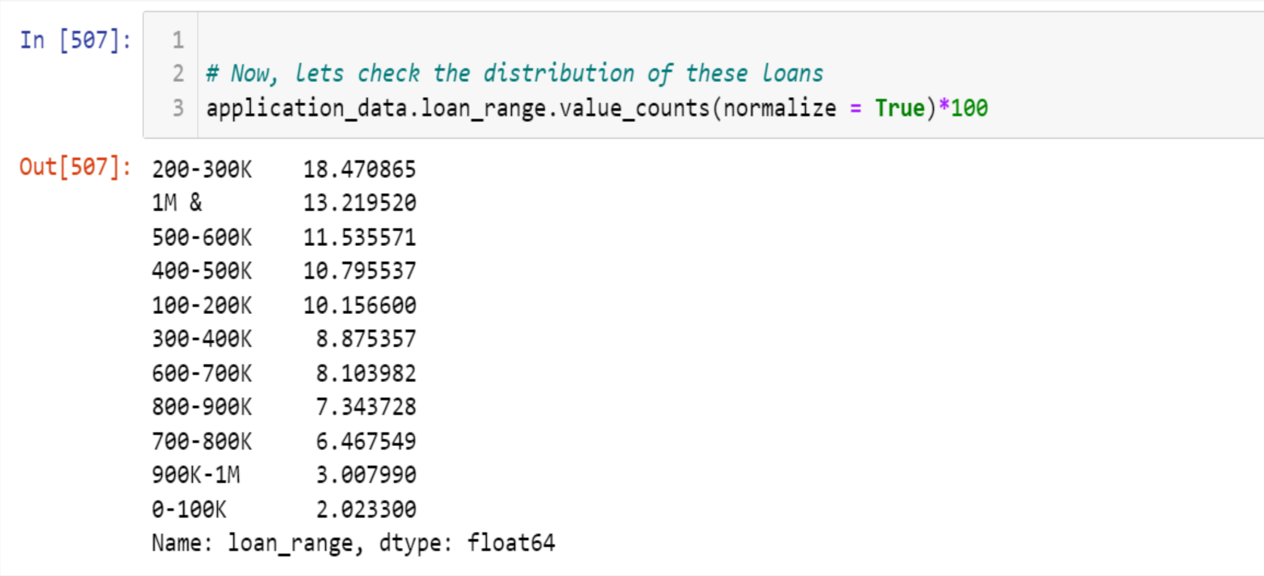
  ✓ # This 1000+ number for years (365243 for DAYS) seems to be used for pensioners or unemployed people. This seems like an outlier. We can replace it with something

  ✓ ### For pensioners and unemployed people, we can replace the DAYS_EMPLOYED and YEARS_EMPLOYED by 0. As they are not currently employed.



Univariate Analysis
AMT_CREDIT (Converted to Loan Range)

- Around 46% of Loans Are in the range of 100K-500K



```
In [507]:
# Now, lets check the distribution of these loans
application_data.loan_range.value_counts(normalize = True)*100
```

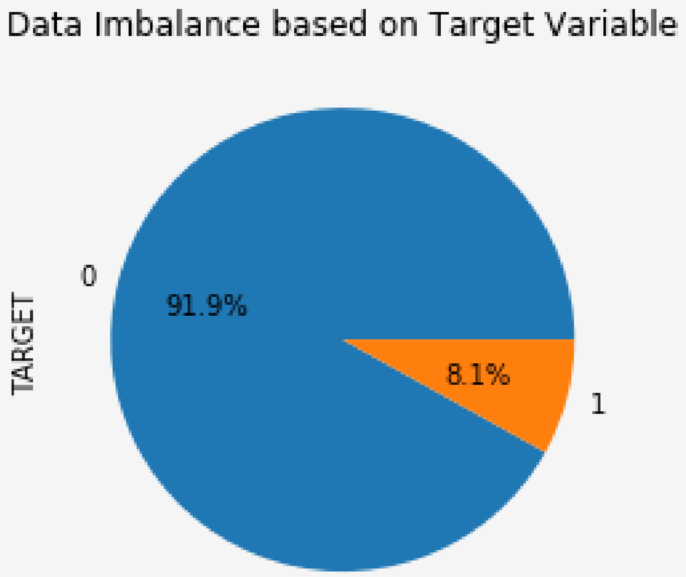
```
Out[507]: 200-300K    18.470865
1M &        13.219520
500-600K    11.535571
400-500K    10.795537
100-200K    10.156600
300-400K     8.875357
600-700K     8.103982
800-900K     7.343728
700-800K     6.467549
900K-1M      3.007990
0-100K       2.023300
Name: loan_range, dtype: float64
```

Univariate Analysis – Segmented Dataset

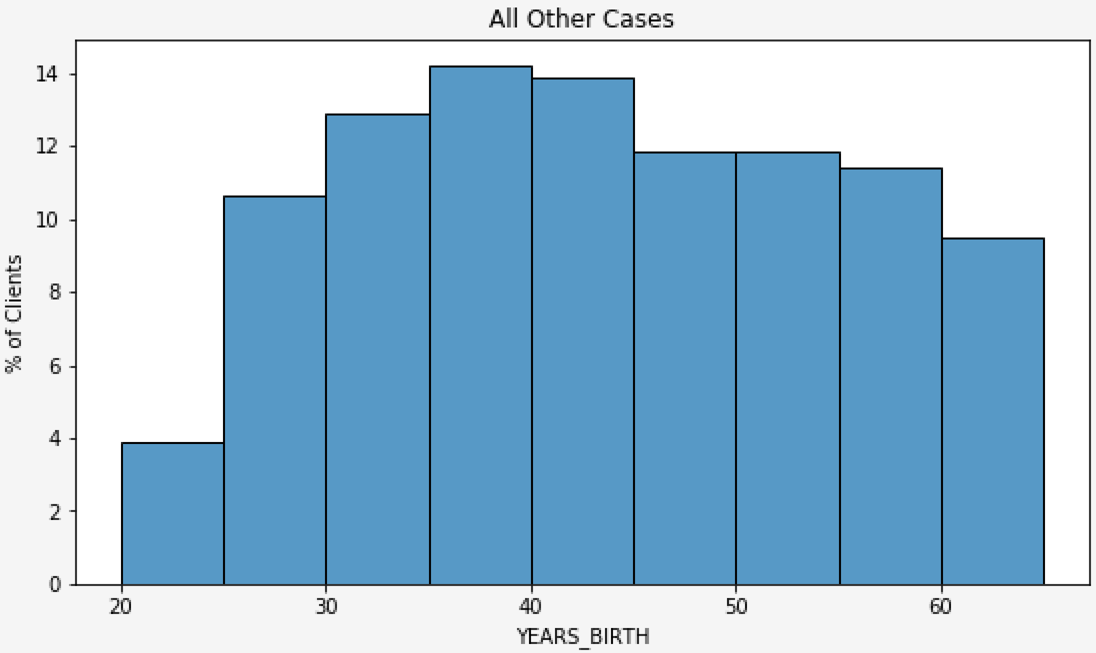
- Understanding the various columns & drawing insights By Segmenting According To Target Variable

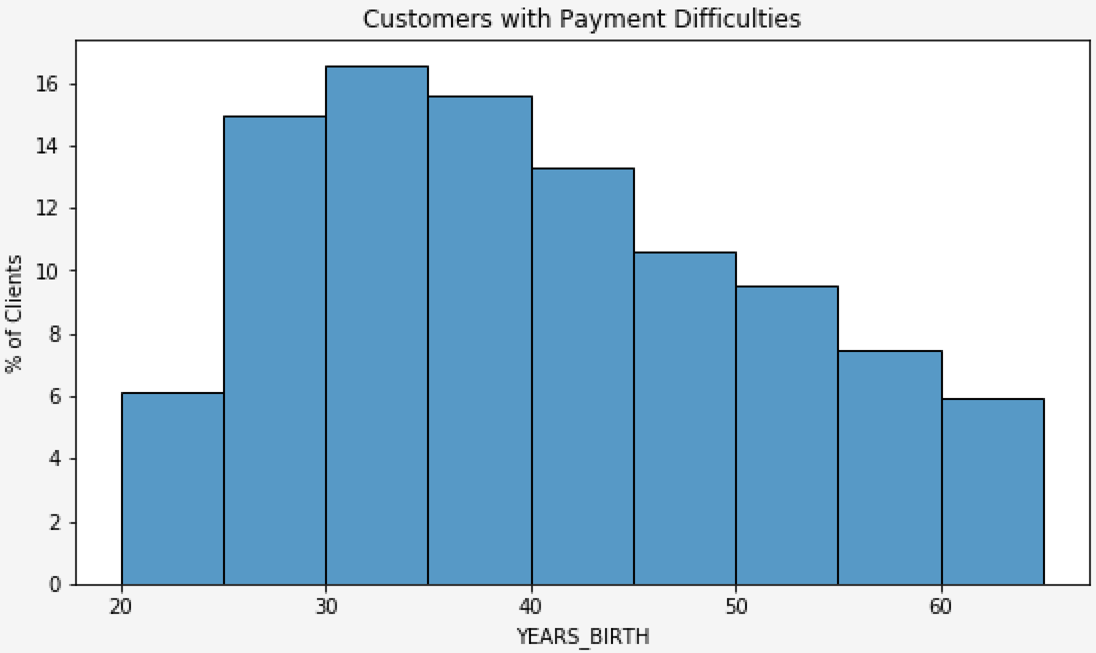
Data Imbalance

- We do see that there is a huge data imbalance – based on the Target Variable.
- Only 8% of Data is for "Customers with Payment Difficulties"



Segmented Univariate Analysis
Age Group

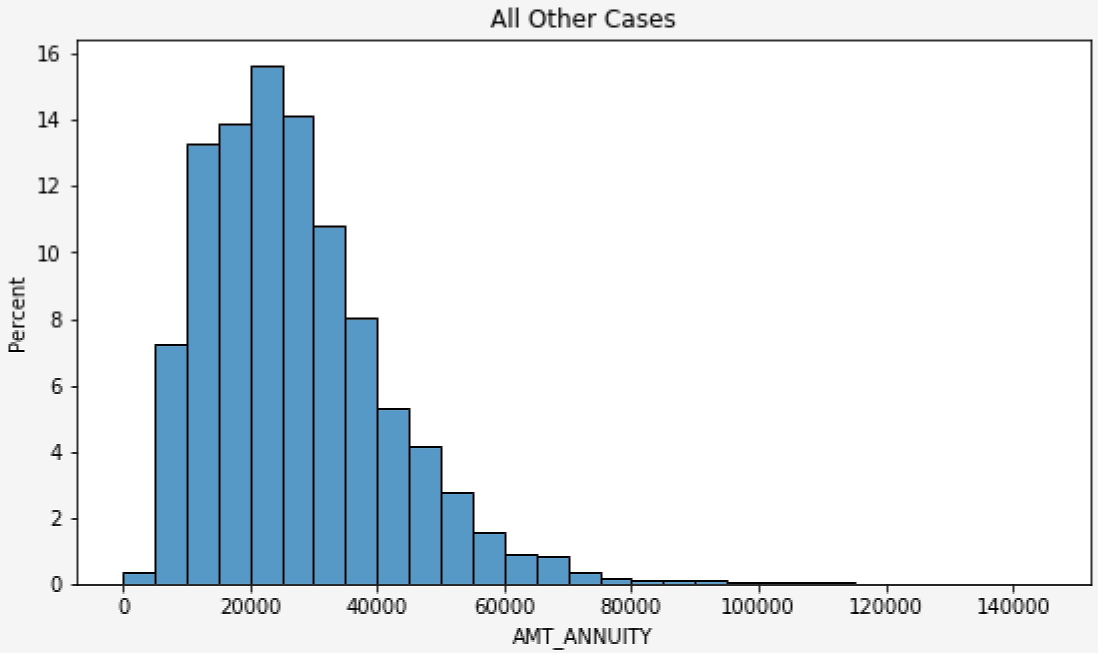


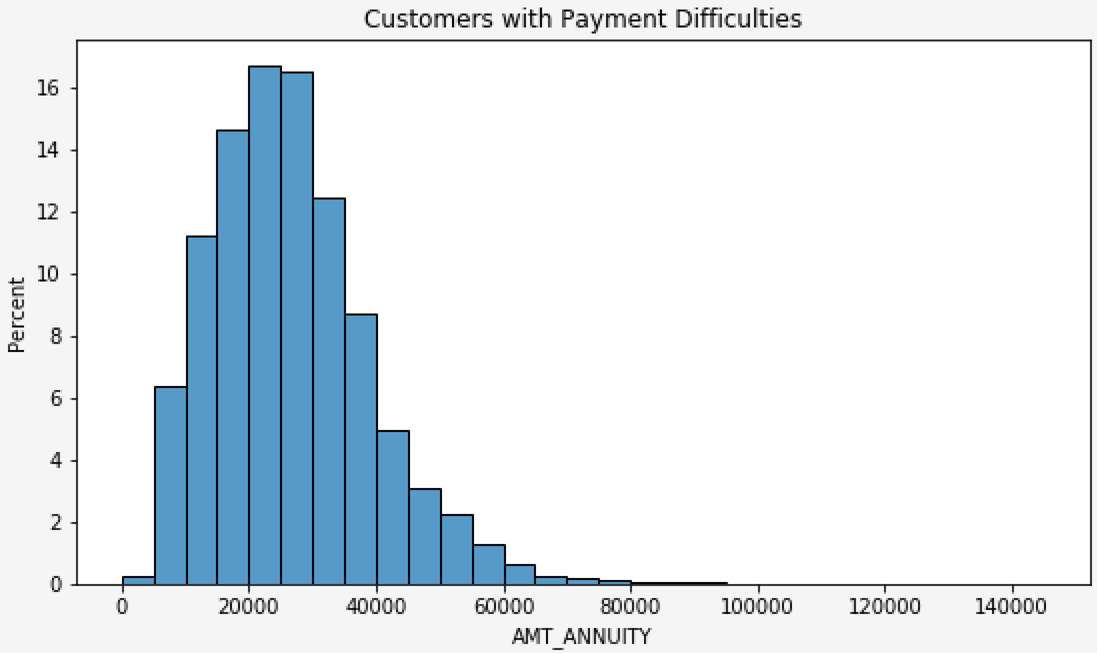


Key Insight

- Most of the “Customers with Payment Difficulties” seem to lie between 25-40 years with almost 45% of them distributed in this range
- Whereas, for “All Other Cases” , the customers age seems to be distributed across the buckets.

Segmented Univariate Analysis
Annuity Amount





Key Insight

- Most of the "Customers with Payment Difficulties" seem to have an annuity amount between 25k-35k years with almost 45% of them distributed in this range
- Whereas, for "All Other Cases" , the customers age seems to be distributed across the buckets.

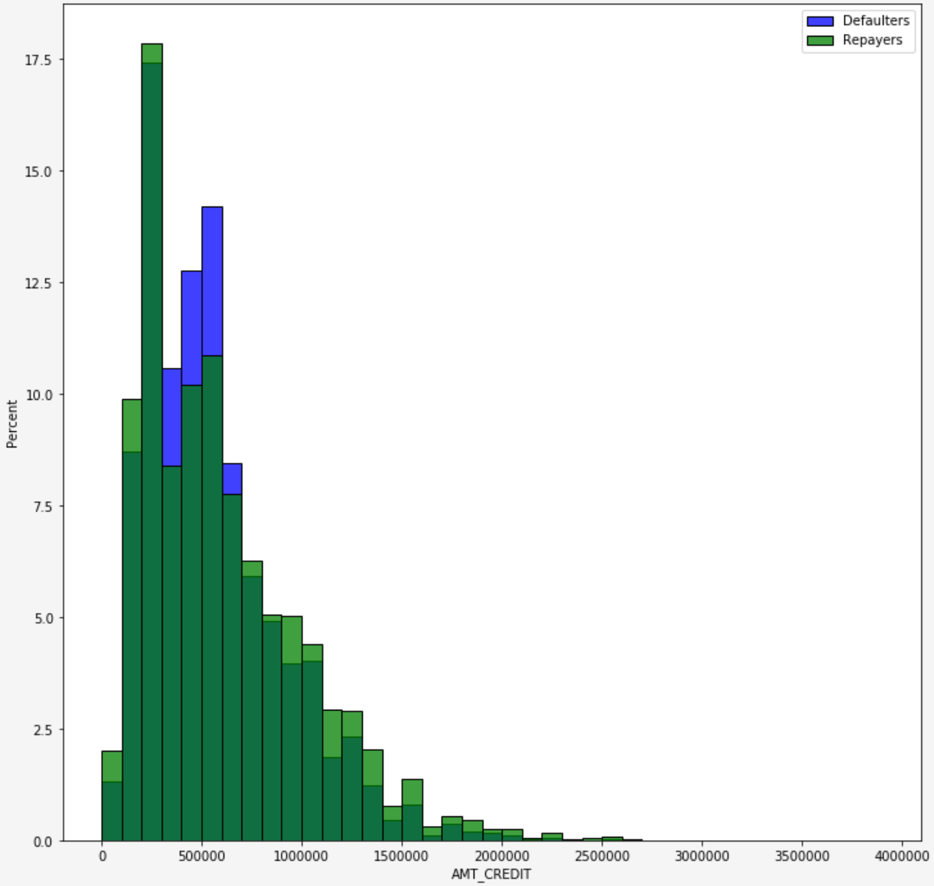
Segmented Univariate Analysis
Credit Amount

Key Insight

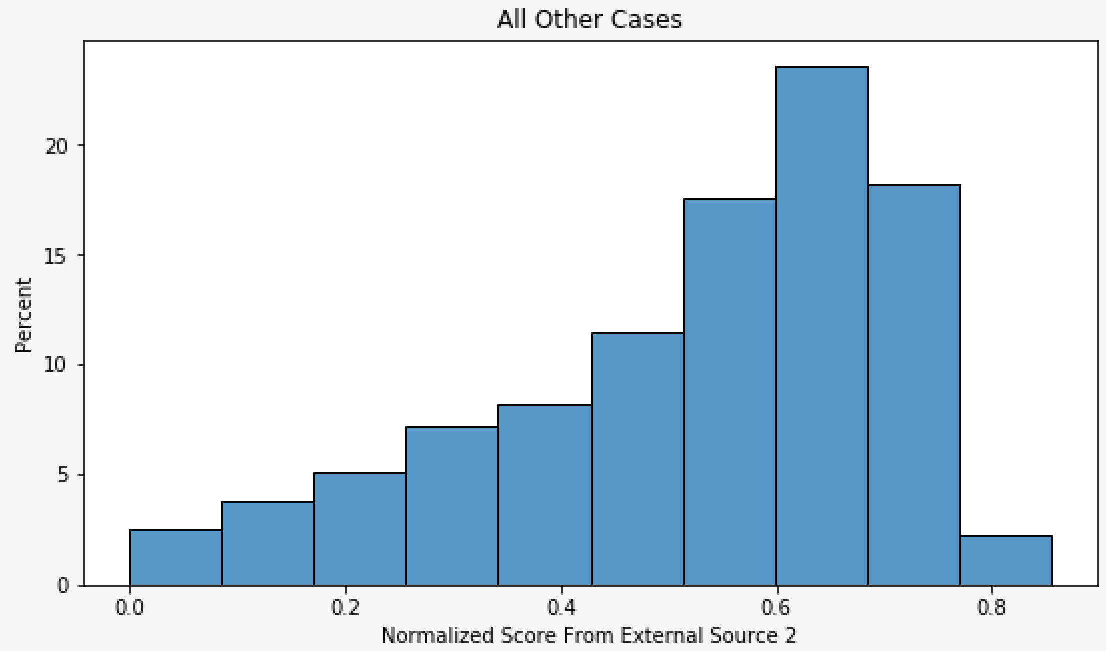
- The % of Customers Defaulting Seems To Spike Once The Loan Amount crosses 300K, i.e. We can see a spike between 300K-600K

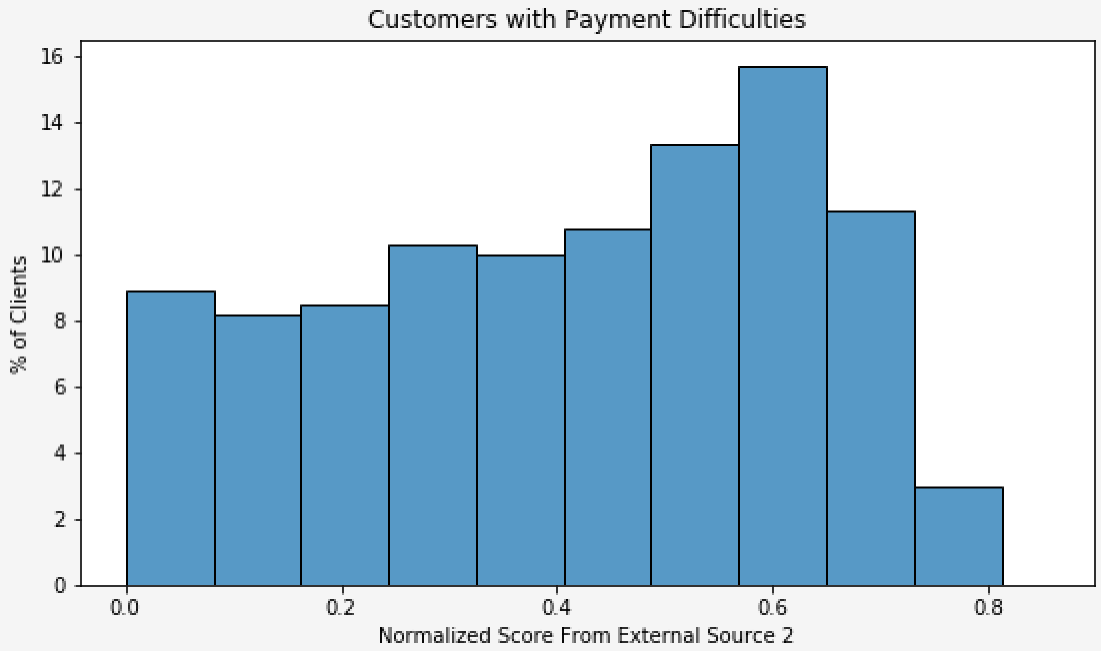
- Blue : Defaulters

- Green: Customers With Good Payment Record



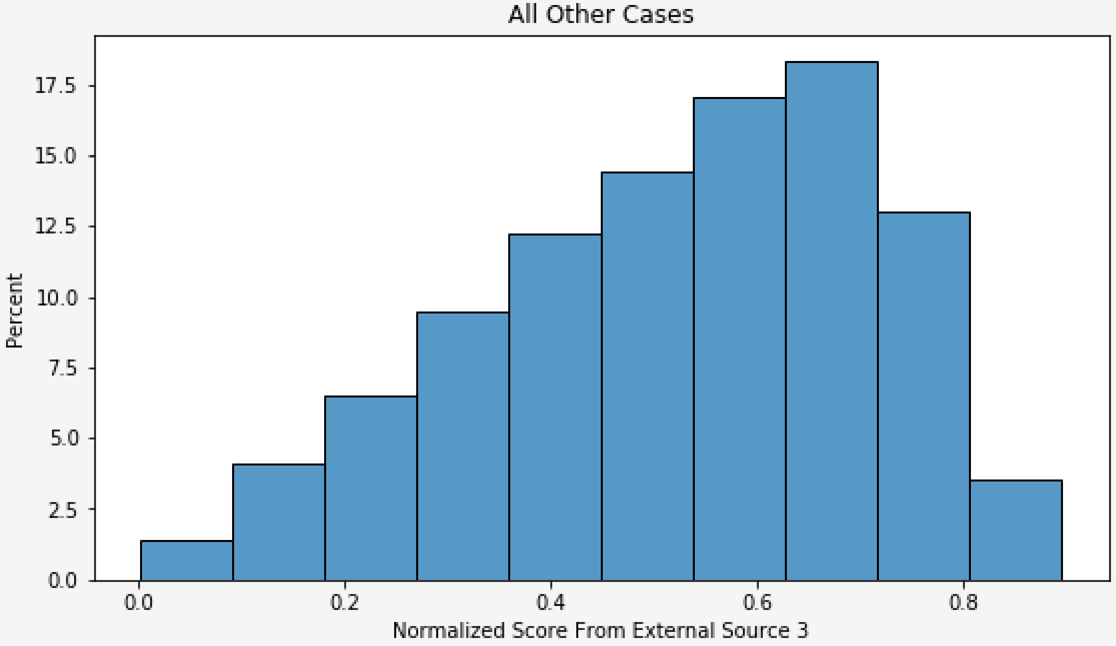
Segmented Univariate Analysis
Normalized Score From External Source 2

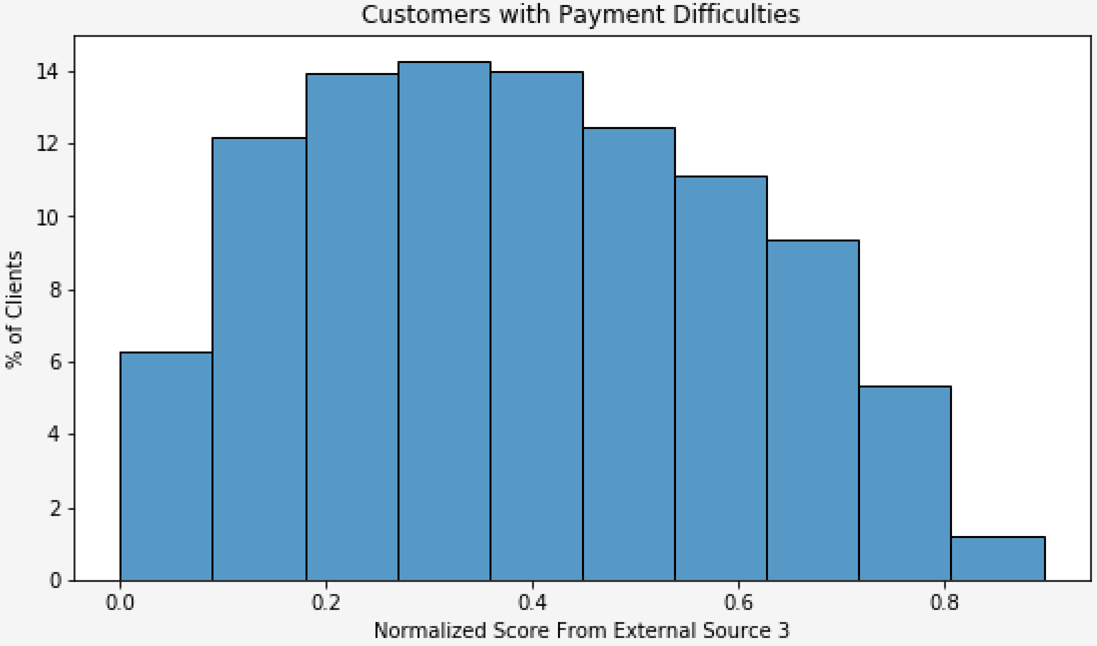




- 50-60% of the clients with payment difficulties seem to have "Normalized Score from External Source" <0.5
- 60% of clients with good record – seem to have a score of 0.5-0.8
- A higher score could possibly mean lesser chances of default

Segmented Univariate Analysis
Normalized Score From External Source 3





- Close to 70% of the clients with payment difficulties seem to have “Normalized Score from External Source” <=0.5
- 60% of clients with good record – seem to have a score of 0.5-0.8
- A higher score could possibly mean lesser chances of default

Top Correlation Between The Variables – Customers With Target Variable 0

- Credit Amount has a high correlation with

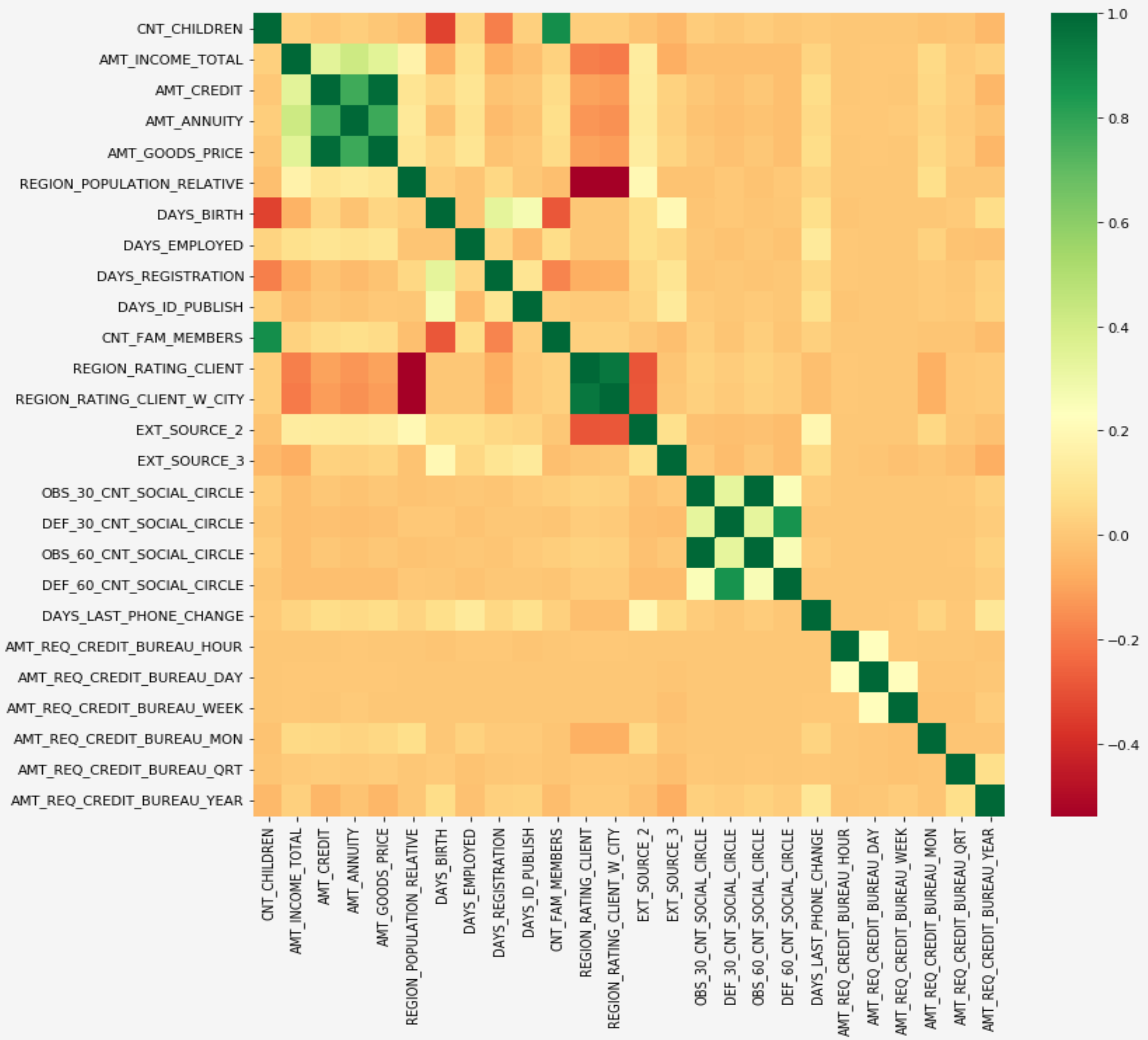
  ✓ Amount of Income

  ✓ Amount of Goods Price

  ✓ Annuity Amount

  Didn’t Include Annuity Amount as it could be linked to “Credit Amount”

  Observable Defaults seems to have a high correlation with Actual no of defaults (30 or 60 days)



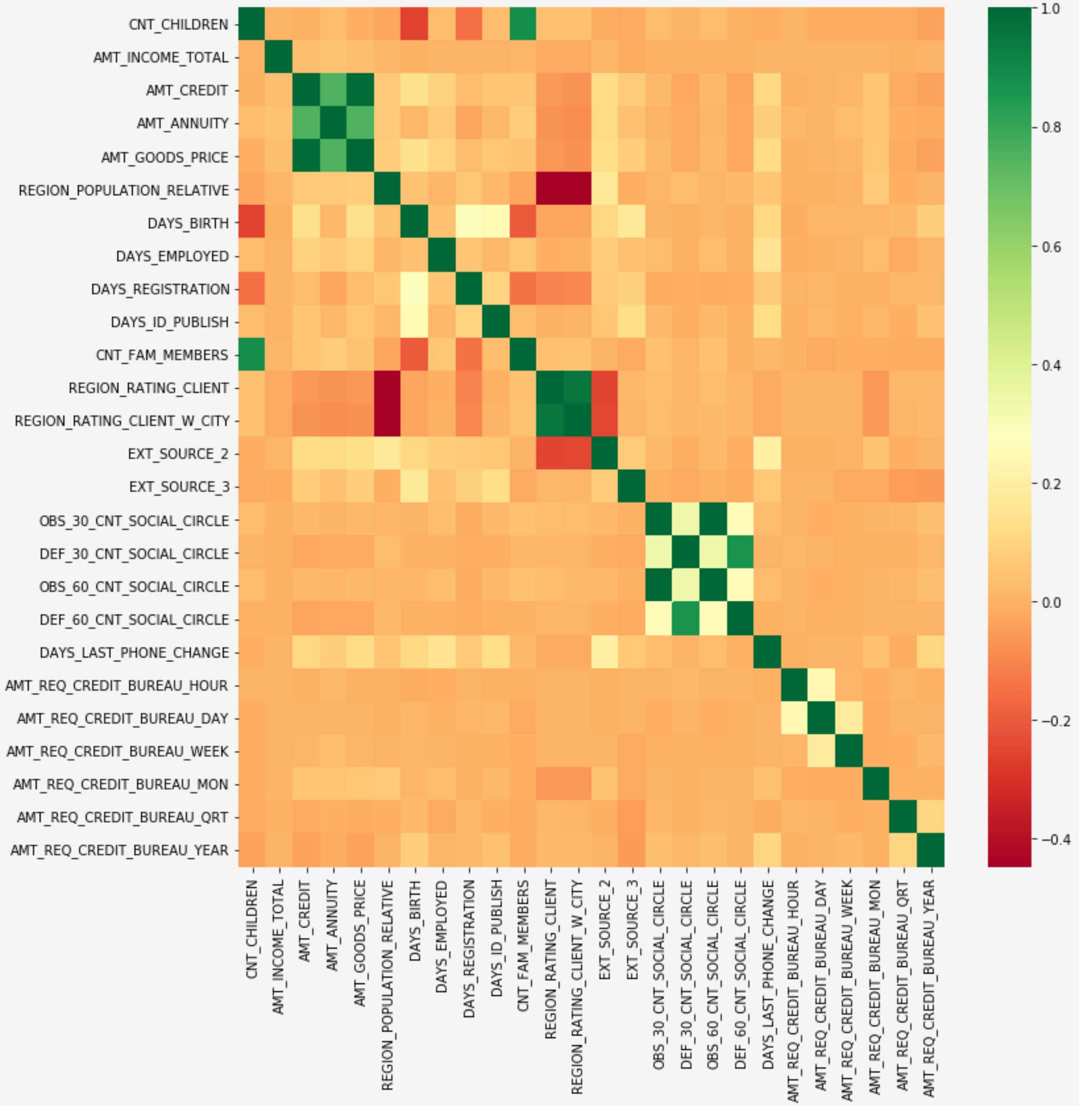
Top Correlation Between The Variables – Customers With Target Variable 1 (Defaulters)

- Credit Amount has a high correlation with

  ✓ Amount of Goods Price

  ✓ Annuity Amount

Key Difference: Drop In Correlation between AMT_INCOME vs "Credit Amount, Annuity Amount, Goods Price" etc. AMT_INCOME_TOTAL doesn't seem to have any correlation; whereas it was higher in the case of Repayers



Bivariate Analysis – Application Data (without Previous Dataset)

- Understanding the relationship between multiple variables to see if there is a pattern or a trend

  ✓ Categorical Variables

  ✓ Numerical Variables

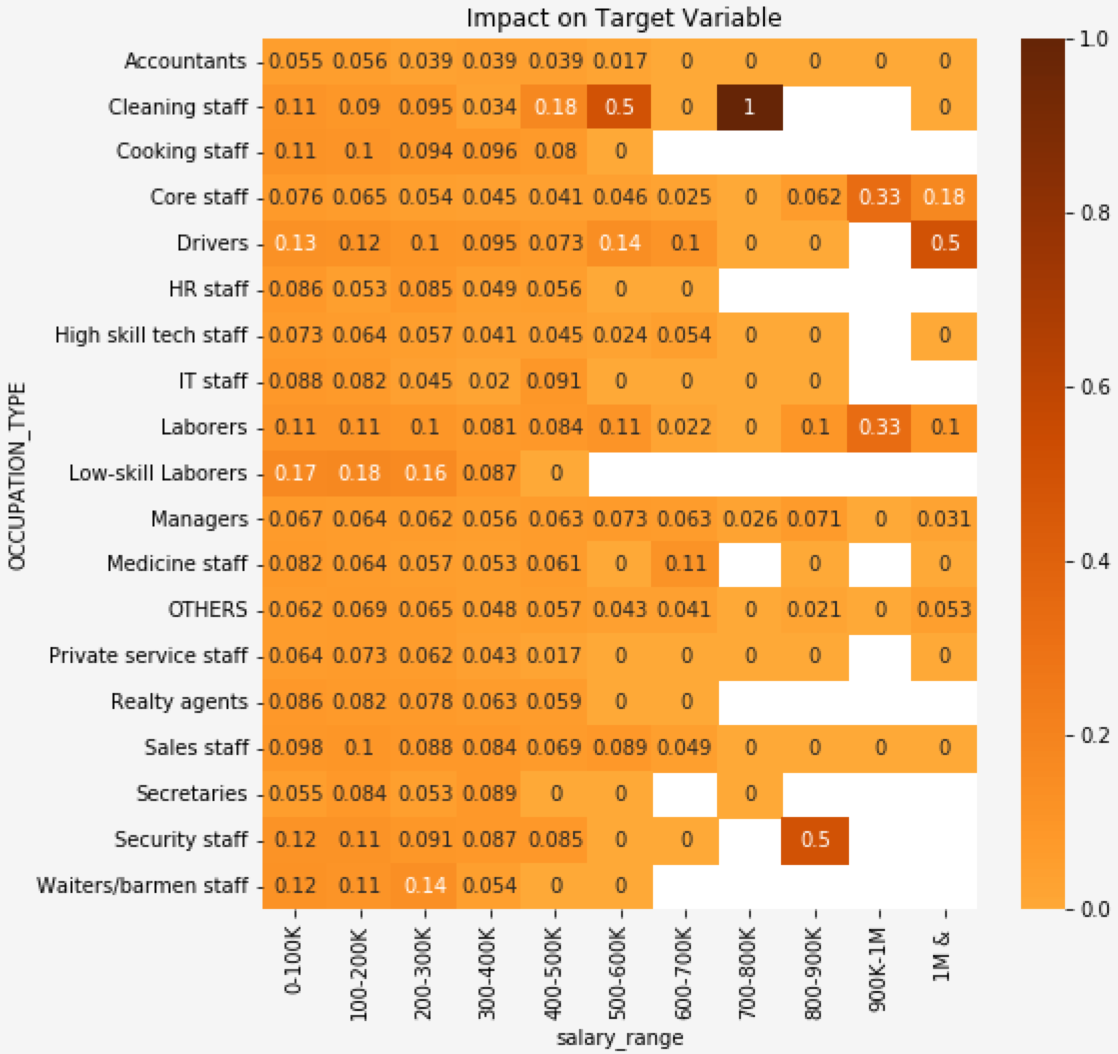
Occupation Type & Salary Range Impact on Target Variable

High Risk Of Default

- “Low Skill Laborers”, ”Laborers”, ”Waiters/Barmen Staff”, “Drivers”, “Security staff” with salary less than <400k
- Cleaning Staff with Salaries between “400-800K”

Low Risk Of Default:

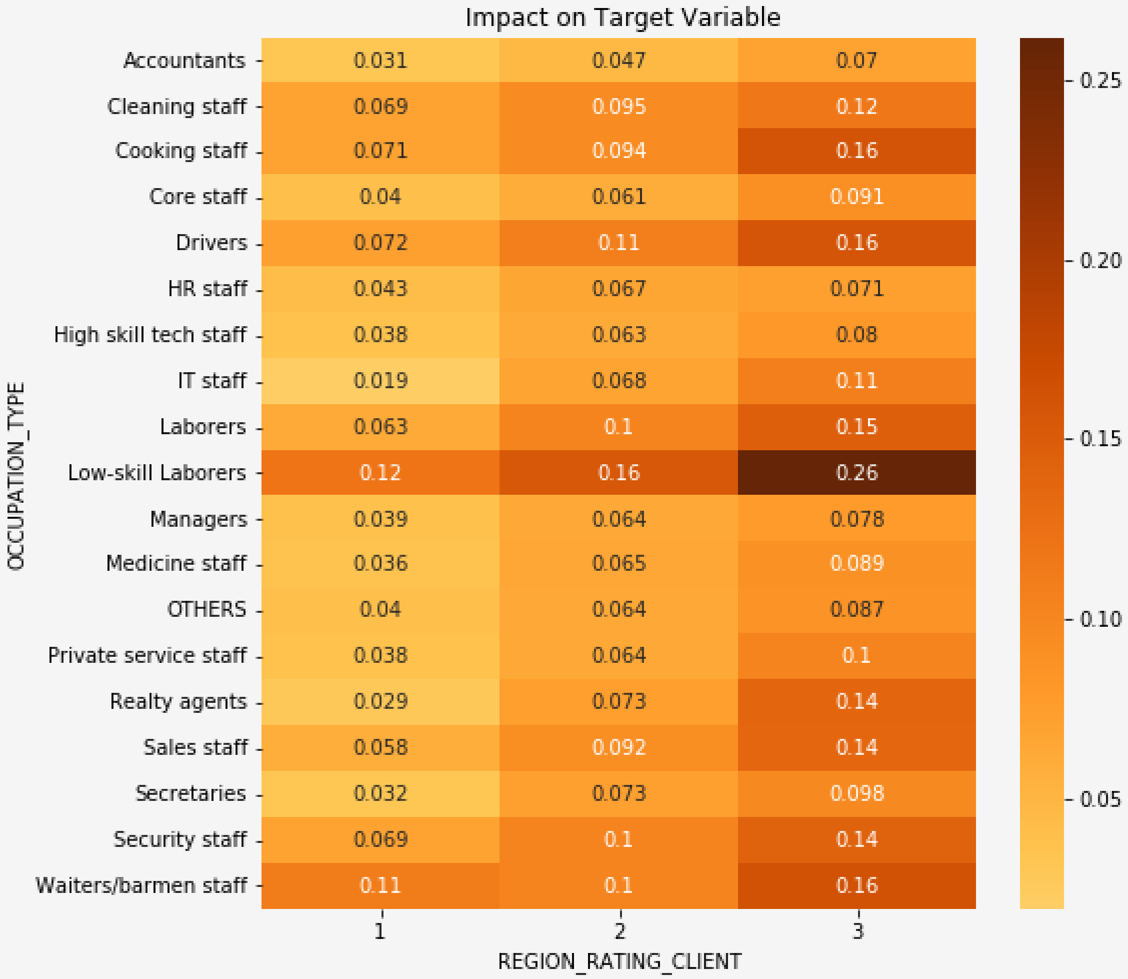
- Accountants
- High Skill Tech Staff
- Managers
- Medicine Staff



Occupation Type & Rating of Region Impact on Target Variable

High Risk Of Default

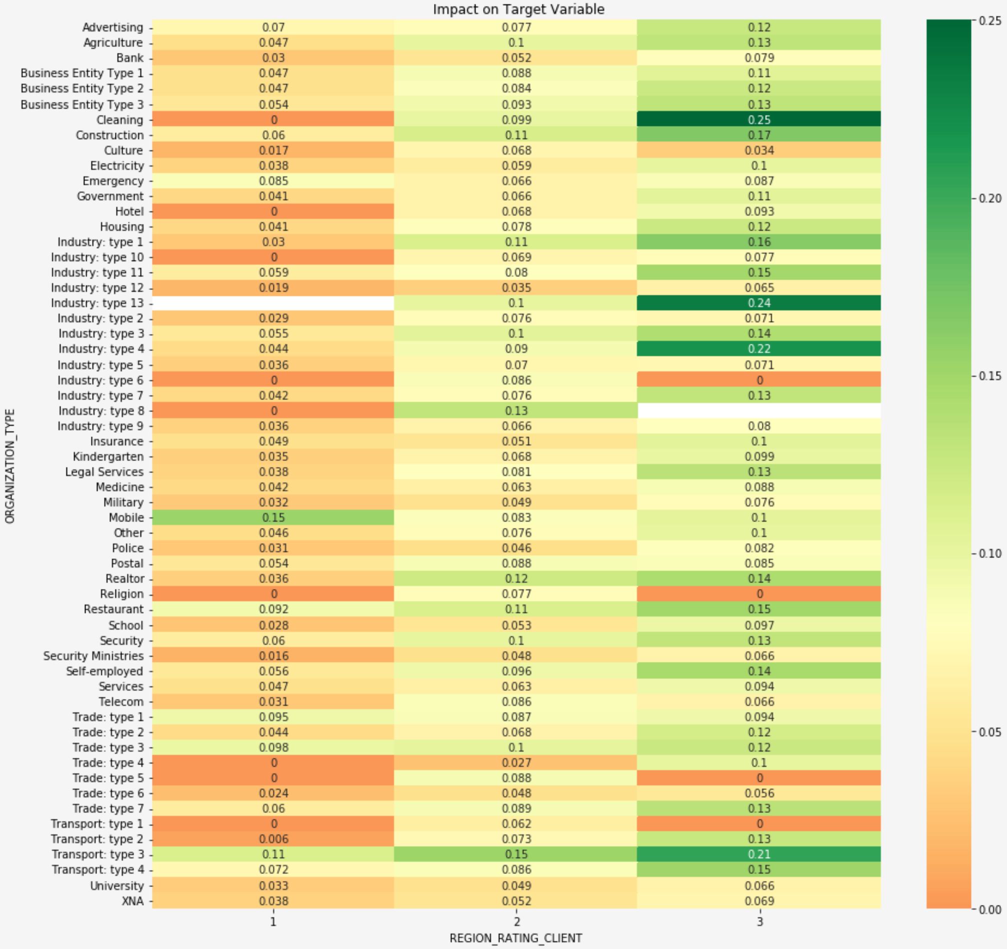
- "Low Skill Laborers" Across 1,2 and 3 Type of Cities have a high risk of default
- Tier3 Clients seem to have a high risk of default – Overall
- Clients with 1 Region Rating – Seem To Have Lesser Risk of Default



Rating of City & Org Type Impact on Target Variable

High Risk Of Default

- Cities with 3 Rating
- Transport Type 3,4, Trade Type 2, Self-Employed, Industry Type 1, 4 and 13, Cleaning, Construction



ORGANIZATION_TYPE	1	2	3
Advertising	0.07	0.077	0.12
Agriculture	0.047	0.1	0.13
Bank	0.03	0.052	0.079
Business Entity Type 1	0.047	0.088	0.11
Business Entity Type 2	0.047	0.084	0.12
Business Entity Type 3	0.054	0.093	0.13
Cleaning	0	0.099	0.25
Construction	0.06	0.11	0.17
Culture	0.017	0.068	0.034
Electricity	0.038	0.059	0.1
Emergency	0.085	0.066	0.087
Government	0.041	0.066	0.11
Hotel	0	0.068	0.093
Housing	0.041	0.078	0.12
Industry: type 1	0.03	0.11	0.16
Industry: type 10	0	0.069	0.077
Industry: type 11	0.059	0.08	0.15
Industry: type 12	0.019	0.035	0.065
Industry: type 13		0.1	0.24
Industry: type 2	0.029	0.076	0.071
Industry: type 3	0.055	0.1	0.14
Industry: type 4	0.044	0.09	0.22
Industry: type 5	0.036	0.07	0.071
Industry: type 6	0	0.086	0
Industry: type 7	0.042	0.076	0.13
Industry: type 8	0	0.13	
Industry: type 9	0.036	0.066	0.08
Insurance	0.049	0.051	0.1
Kindergarten	0.035	0.068	0.099
Legal Services	0.038	0.081	0.13
Medicine	0.042	0.063	0.088
Military	0.032	0.049	0.076
Mobile	0.15	0.083	0.1
Other	0.046	0.076	0.1
Police	0.031	0.046	0.082
Postal	0.054	0.088	0.085
Realtor	0.036	0.12	0.14
Religion	0	0.077	0
Restaurant	0.092	0.11	0.15
School	0.028	0.053	0.097
Security	0.06	0.1	0.13
Security Ministries	0.016	0.048	0.066
Self-employed	0.056	0.096	0.14
Services	0.047	0.063	0.094
Telecom	0.031	0.086	0.066
Trade: type 1	0.095	0.087	0.094
Trade: type 2	0.044	0.068	0.12
Trade: type 3	0.098	0.1	0.12
Trade: type 4	0	0.027	0.1
Trade: type 5	0	0.088	0
Trade: type 6	0.024	0.048	0.056
Trade: type 7	0.06	0.089	0.13
Transport: type 1	0	0.062	0
Transport: type 2	0.006	0.073	0.13
Transport: type 3	0.11	0.15	0.21
Transport: type 4	0.072	0.086	0.15
University	0.033	0.049	0.066
XNA	0.038	0.052	0.069

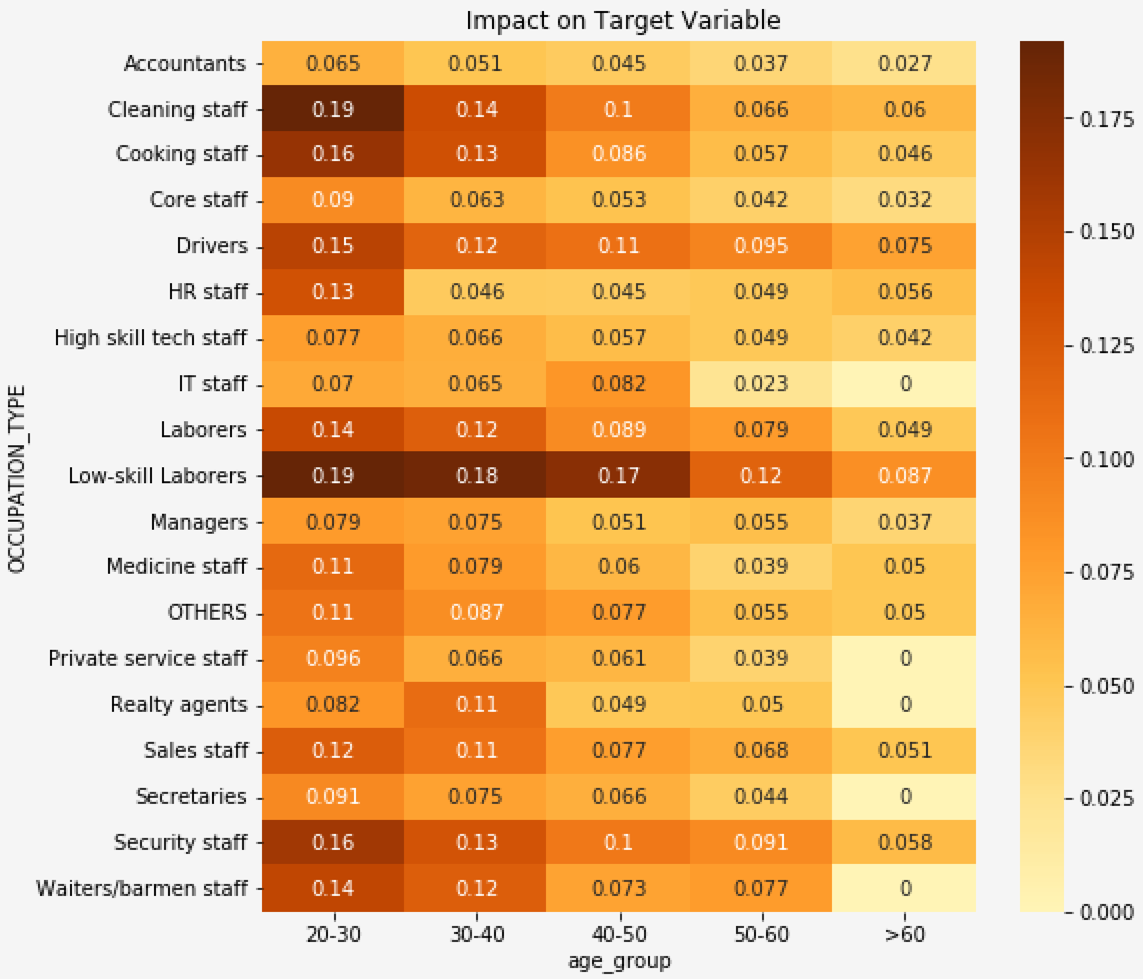
Occupation Type & Age Group Impact on Target Variable

High Risk Of Default - Combinations

- 20-40 Age Group & Cleaning Staff, Cooking Staff, Drivers, Laborers, Low Skill Laborers, Security Staff, Waiters/Barmen Staff

Low Risk Of Default

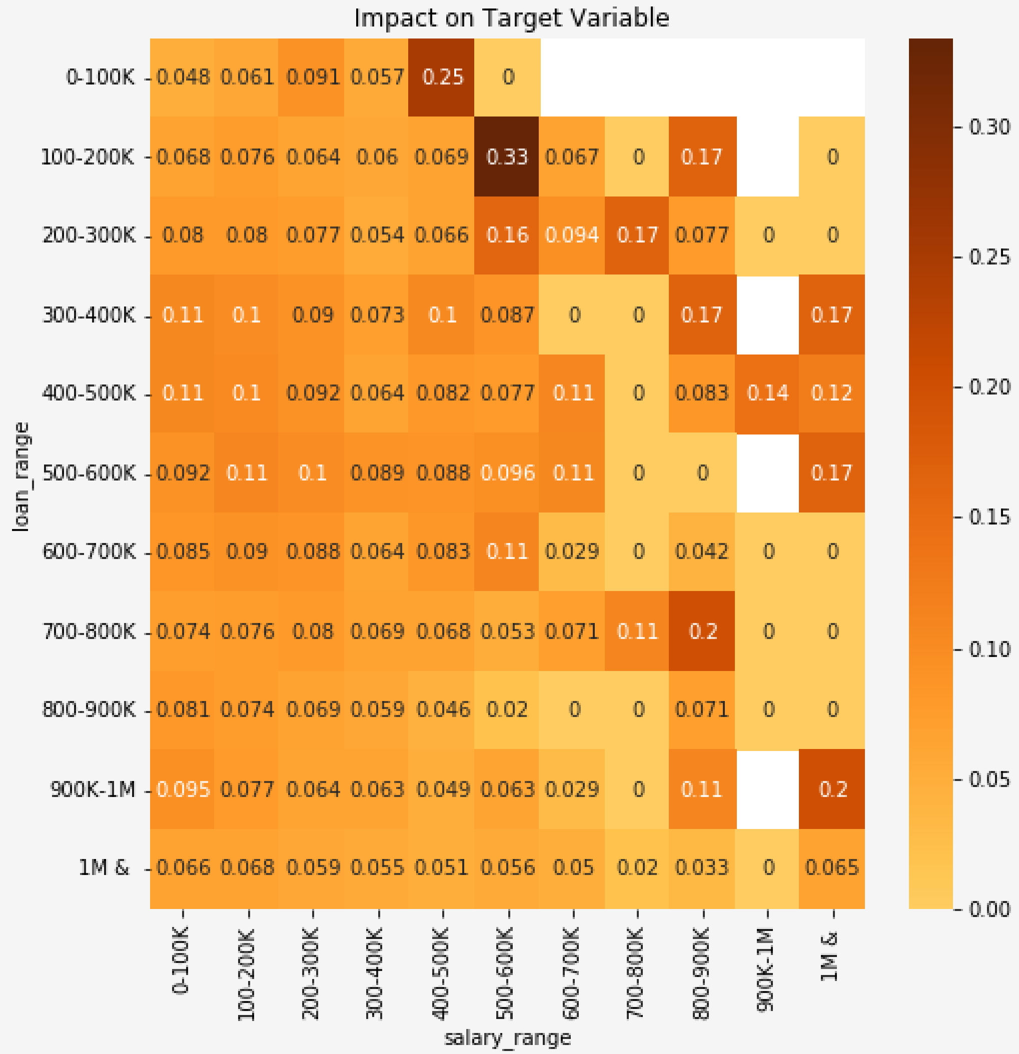
- >40 Age Group



Loan Amount & Salary Range Impact on Target Variable

High Risk Of Default - Combinations

- Salary 400-500K & Loan of <100K seems to have a high rate of default (0.25%)
- Salary 500-600K and Loan of 100-300K (0.16-0.33 rate of default)
- Salary 800K and Above and Loan of 100-400K



Impact on Target Variable

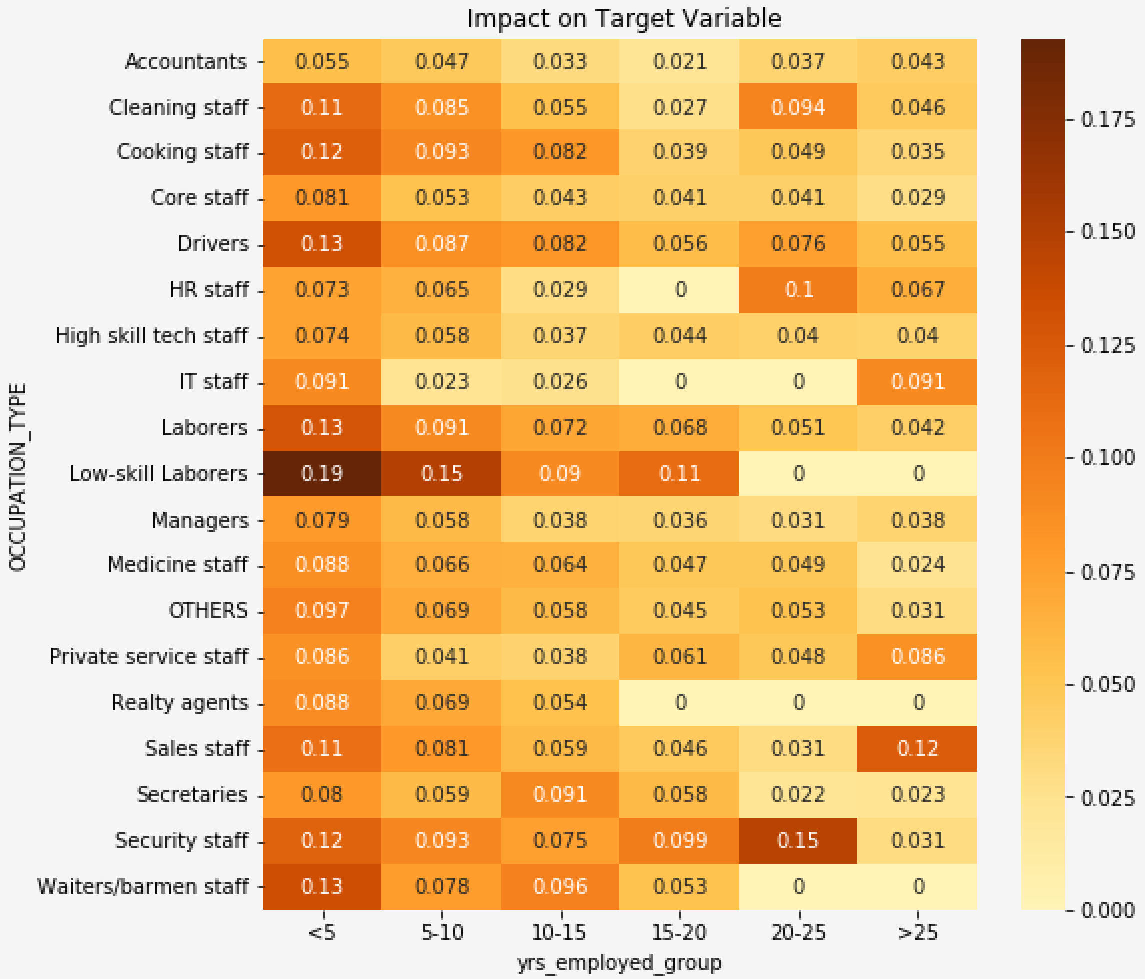
loan_range	0-100K	100-200K	200-300K	300-400K	400-500K	500-600K	600-700K	700-800K	800-900K	900K-1M	1M &
0-100K	0.048	0.061	0.091	0.057	0.25	0					
100-200K	0.068	0.076	0.064	0.06	0.069	0.33	0.067	0	0.17		0
200-300K	0.08	0.08	0.077	0.054	0.066	0.16	0.094	0.17	0.077	0	0
300-400K	0.11	0.1	0.09	0.073	0.1	0.087	0	0	0.17		0.17
400-500K	0.11	0.1	0.092	0.064	0.082	0.077	0.11	0	0.083	0.14	0.12
500-600K	0.092	0.11	0.1	0.089	0.088	0.096	0.11	0	0		0.17
600-700K	0.085	0.09	0.088	0.064	0.083	0.11	0.029	0	0.042	0	0
700-800K	0.074	0.076	0.08	0.069	0.068	0.053	0.071	0.11	0.2	0	0
800-900K	0.081	0.074	0.069	0.059	0.046	0.02	0	0	0.071	0	0
900K-1M	0.095	0.077	0.064	0.063	0.049	0.063	0.029	0	0.11		0.2
1M &	0.066	0.068	0.059	0.055	0.051	0.056	0.05	0.02	0.033	0	0.065

Occupation Type & Yrs. Employed Impact on Target Variable

High Risk Of Default - Combinations

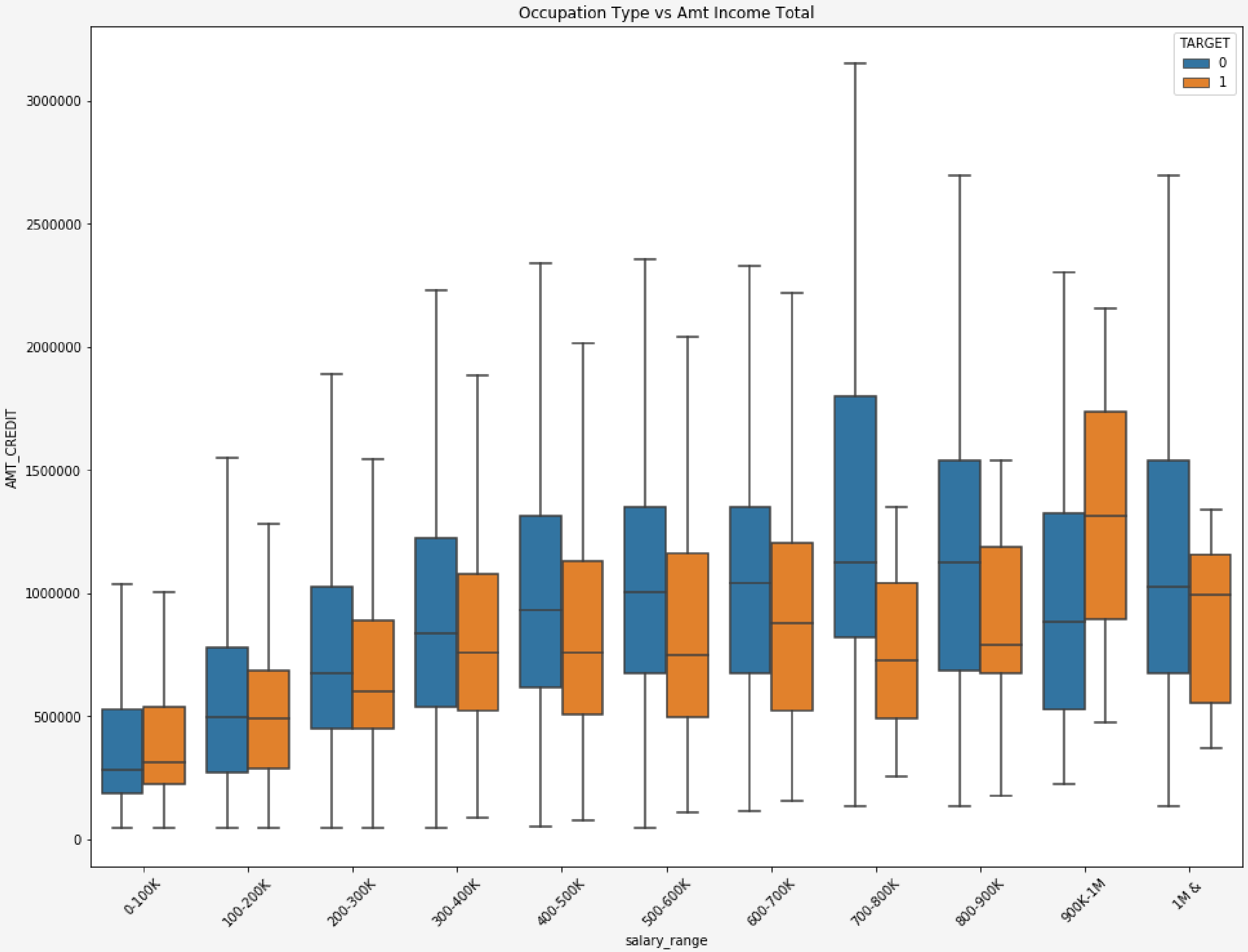
- <5 Years of Experience in Current Employment

Low Risk : Overall ; >10 years seem to have lesser risk of default



Occupation Type & Yrs. Employed Impact on Target Variable

For 700-800K Salaries, we can see a clear difference in the distribution between repayers and defaulters. The Defaulters seem to lie for loans between 500K to 1000K



Bivariate Analysis – Application Data (After Merging Previous Dataset)

- Understanding the relationship between multiple variables to see if there is a pattern or a trend

  ✓ Categorical Variables

  ✓ Numerical Variables

Channel Type, Previous Application Status Vs Default Rate

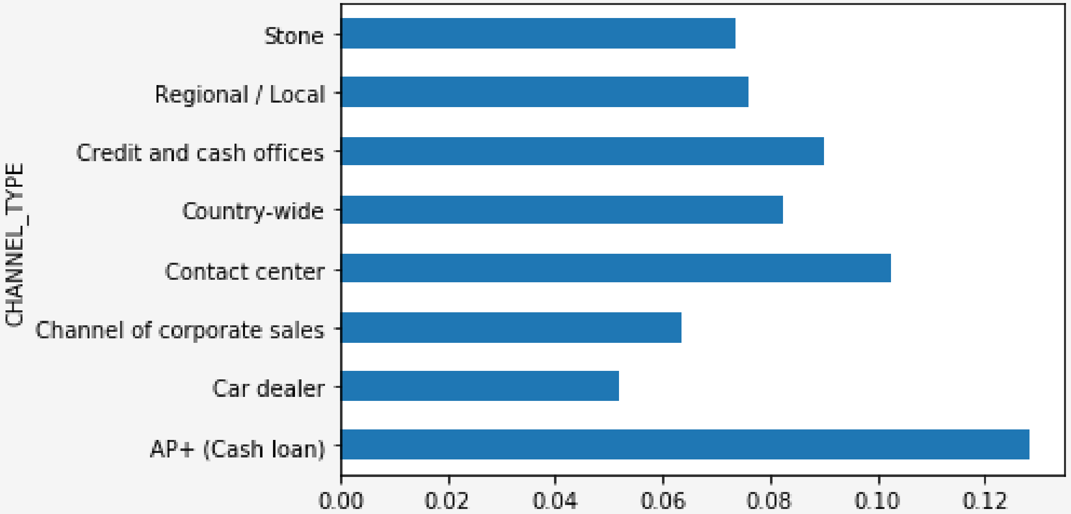
Channel Type

- AP+(Cash Loan) and Contact Center Acquried Clients Seem To have Higher Default Percentage
- Clients Acquired Through Car Dealer Have the least Default Percentage

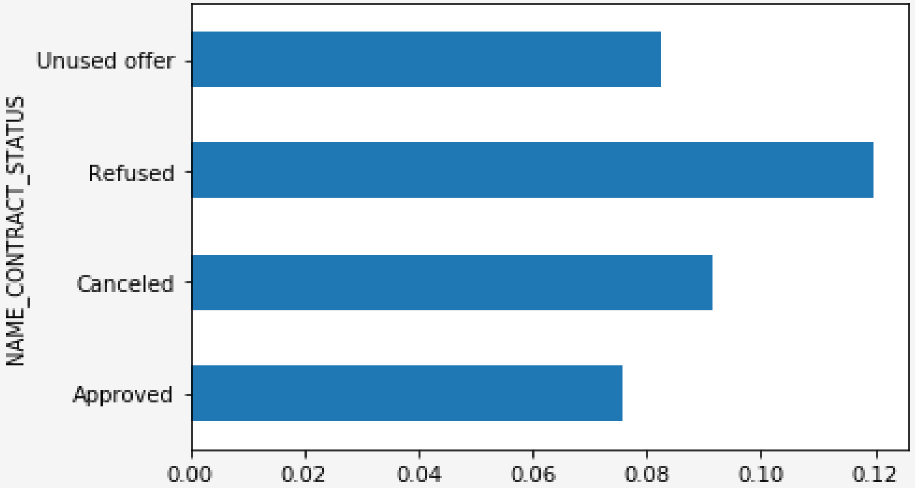
Based on Previous Loan Status

- 90% of Clients whose loan was Canceled had repaid their loan in the current application – Offers Scope for negotiation of interest rates

% of Defaulters based on Channel Type



% of Defaulters based on Previous Application Status



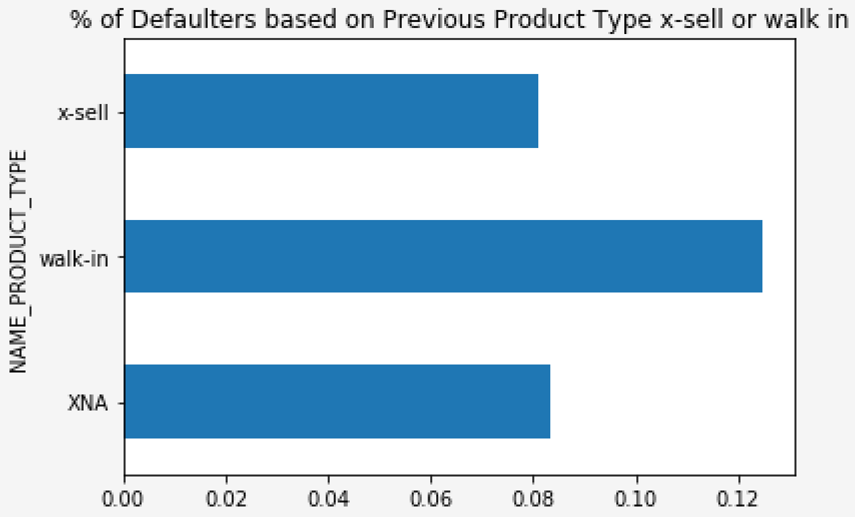
Channel Type, Previous Application Status Vs Default Rate

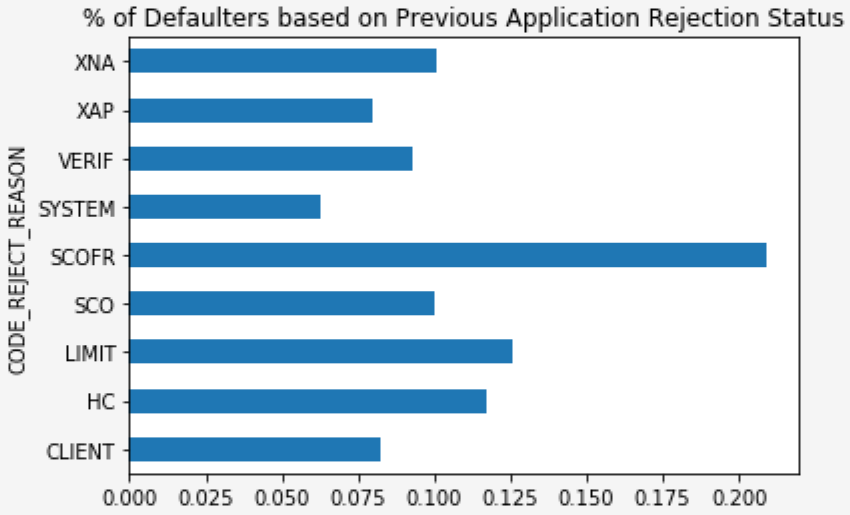
X-Sell or Walk In

- Walk In Customers tend to have a higher default rate

Based on Previous Application Rejection Status

- Clients whose application had a “SCOFR” rejection code have a higher default ~20%

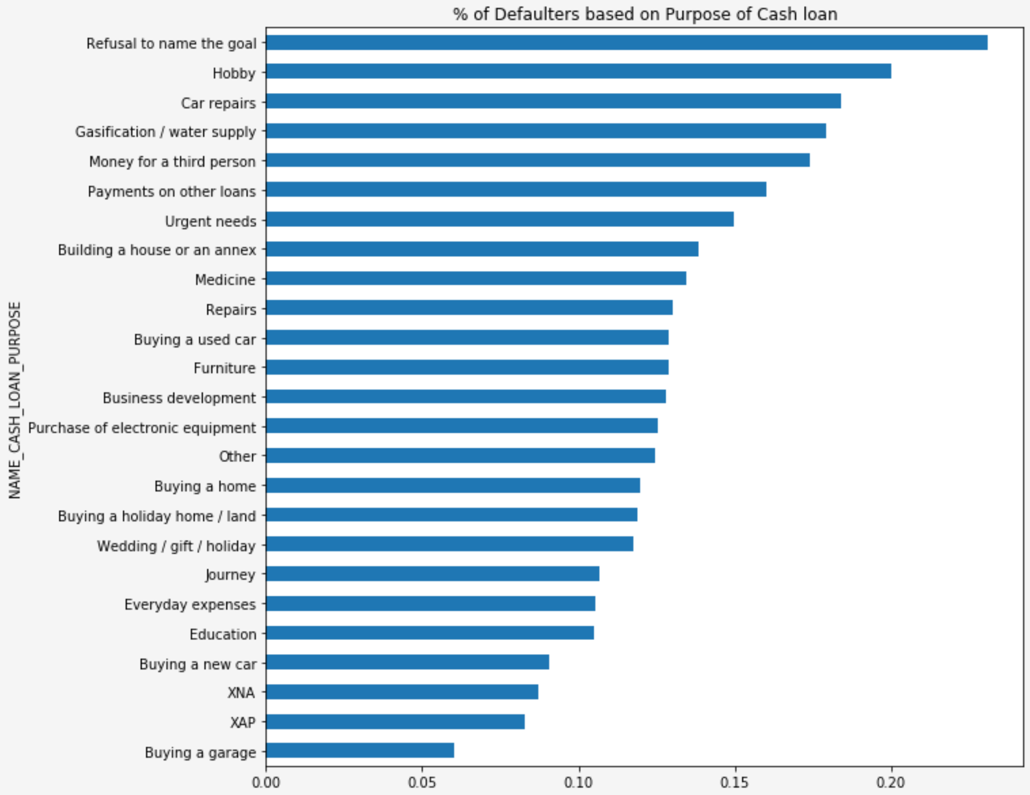




Previous Loan Purpose Vs Default Rate

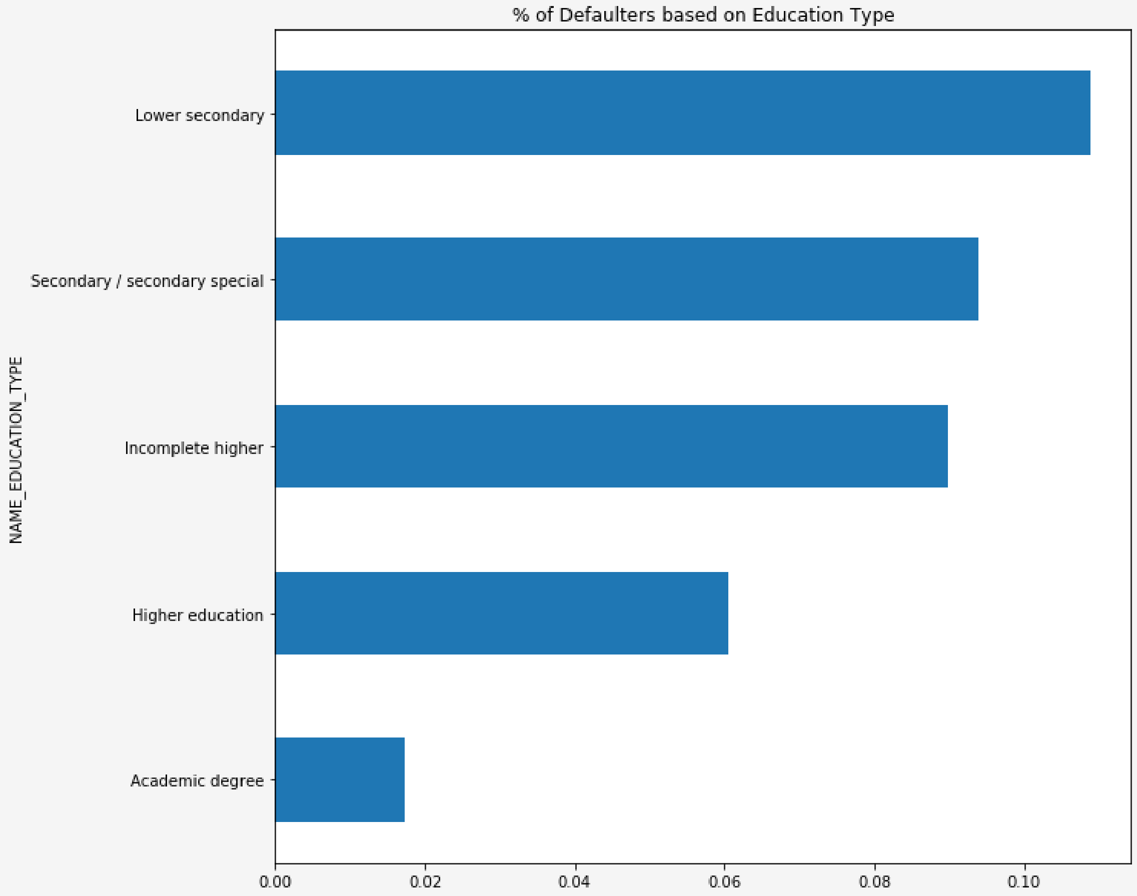
- Customers Who refuse to state the purpose of the loan, or take it for Hobby, Car Repairs, Urgent Needs, Payment on Other Loans, have a higher default rate

- If the purpose is Buying a garage, Education, New Car ; the possibility of default is lesser



Education Type vs Rate of Default

If the education is “Academic degree”, the chances of default are lesser



Recommendation

Attributes of “Good Customers or Clients with Less Chances of Default Rate”

- **NAME_EDUCATION_TYPE:** ACADEMIC Degree
- **NAME_INCOME_TYPE:** Students, Pensioners, State Servants
- **REGION_RATING_CLIENT:** 1
- **OCCUPATION TYPE**: Accountants, High Skill Tech Staff, Managers, Medicine Staff
- **ORGANIZATION TYPE**: Trade Type 4, 5, Industry Type 12
- **AGE GROUP**: 40 and Above years
- **EXPERIENCE YEARS**: >10 years in current jobs
- **CHANNEL_TYPE:** Clients acquired through Car Dealers , Corporate Sales
- **NAME_CASH_LOAN_PURPOSE:** If the purpose is Buying a garage, Education, New Car ; the possibility of default is lesser
- **INCOME_LEVEL:** 600K and above salaries
- **EXT_SOURCE_3:** Scores of 0.5-0.8 and above

Attributes of “Bad Customers or Clients with High Chances of Default Rate”

- **NAME_EDUCATION_TYPE:** Lower secondary, Secondary Special
- **NAME_INCOME_TYPE:** Maternity Leave, Unemployed
- **REGION_RATING_CLIENT:** 3
- **OCCUPATION TYPE**: Low Skill Laborers”, ”Laborers”, ”Waiters/Barmen Staff”, “Drivers”, ”Security staff, “Cleaners”
- **ORGANIZATION TYPE**: Transport Type 3, Industry Type 1,3, Realtors, Construction, Restaurants, Trade Type 3
- **AGE GROUP**: 20-40 Years
- **EXPERIENCE YEARS**: <5 years in current jobs
- **CHANNEL_TYPE:** Clients acquired through AP+ Cash Loan and Contact Centers
- **NAME_CASH_LOAN_PURPOSE:** Customers Who refuse to state the purpose of the loan, or take it for Hobby, Car Repairs, Urgent Needs, Payment on Other Loans, have a higher default rate
- **INCOME_LEVEL:** 400K and below salaries
- **CODE_GENDER:** Male
- **EXT_SOURCE_3**: Scores of 0.5 and Lower
- **LOAN_AMOUNT:** 300K and Above
- **GOODS_PRICE_RANGE:** 300-500K
- **NAME_GOODS_CATEGORY:** Insurance, Education, Direct Sales