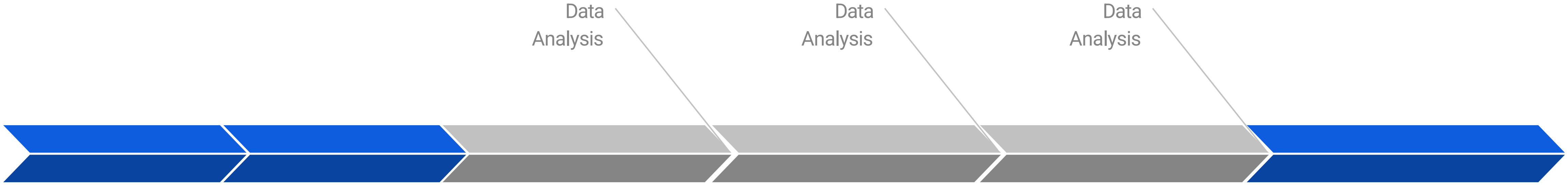
Credit EDA Assignment

Submitted by:
Princy Judson

Abstract

- Finance company which specializes in lending various types of loans to urban customers
- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.
- This makes it hard to identify possible defaulters
- The objective of analysis is to use the information about previous application and current application to find the top drivers leading to defaulters

Problem solving methodology



Data Analysis

Data Analysis

Data Analysis

Data Understanding

Going through the metadata to understand each columns

Data Cleaning and manipulation

Removing the columns having more than 40% null values. Dealing with outliers and converting data to proper format

Univariate Analysis

Analysing each column both categorical and numerical, plotting the distributions of each column.

Segmented Univariate Analysis

Analysing the continuous data after segregating the data based on categorical target column

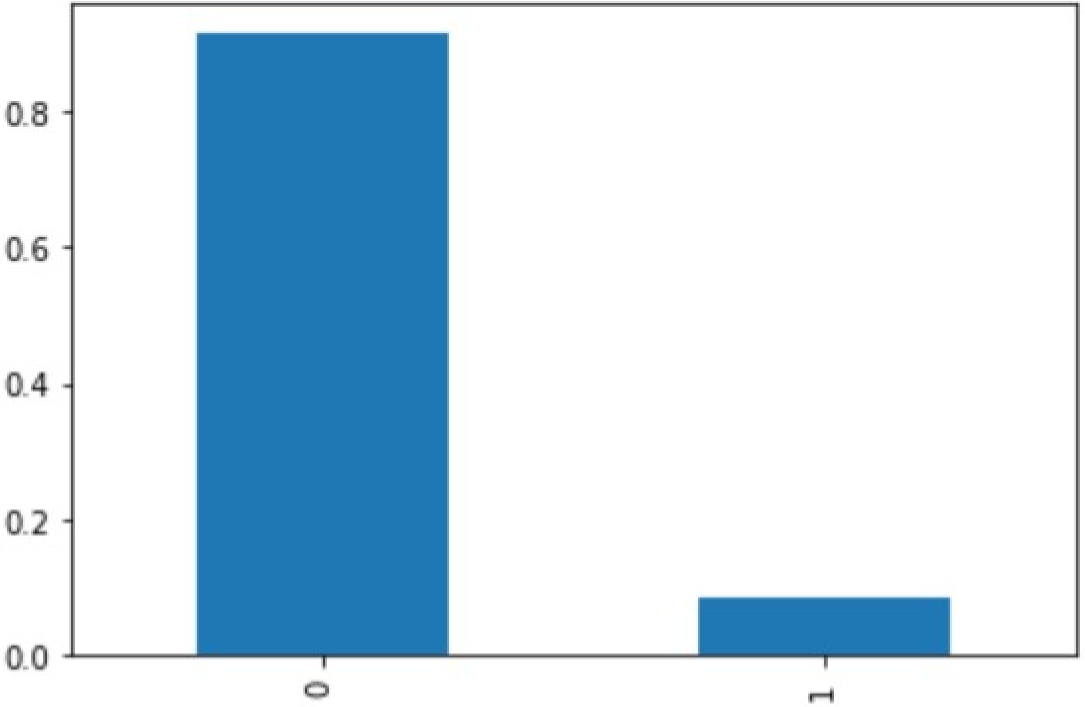
Bivariate Analysis

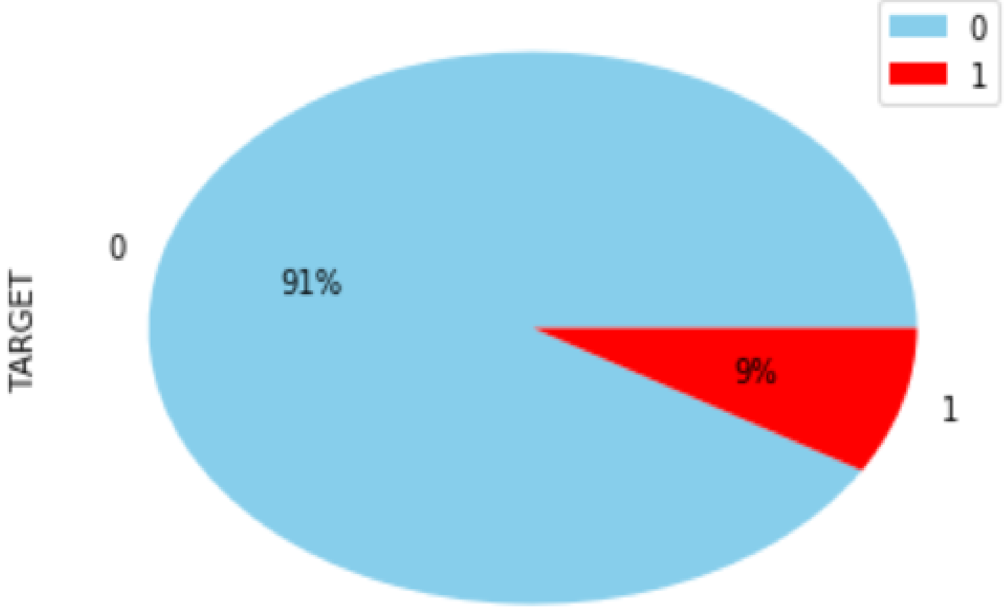
Analysing the two variable behaviour like income and age with respect to target variable.

Recommendations

Analysing all plots and give recommendation to reduce the rate of defaulters.

Understanding Data

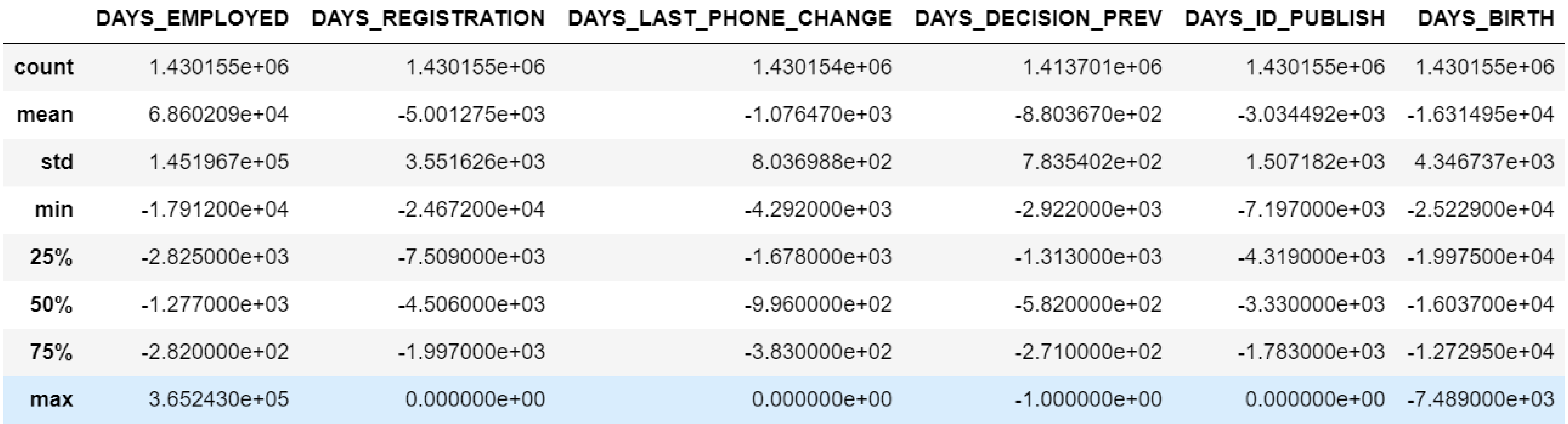




- The Data is heavily imbalanced
- The Target variable where target=1 (defaulters) is around 9% of whole data and category ‘other cases’ (target==0) is 91%

Data Cleaning

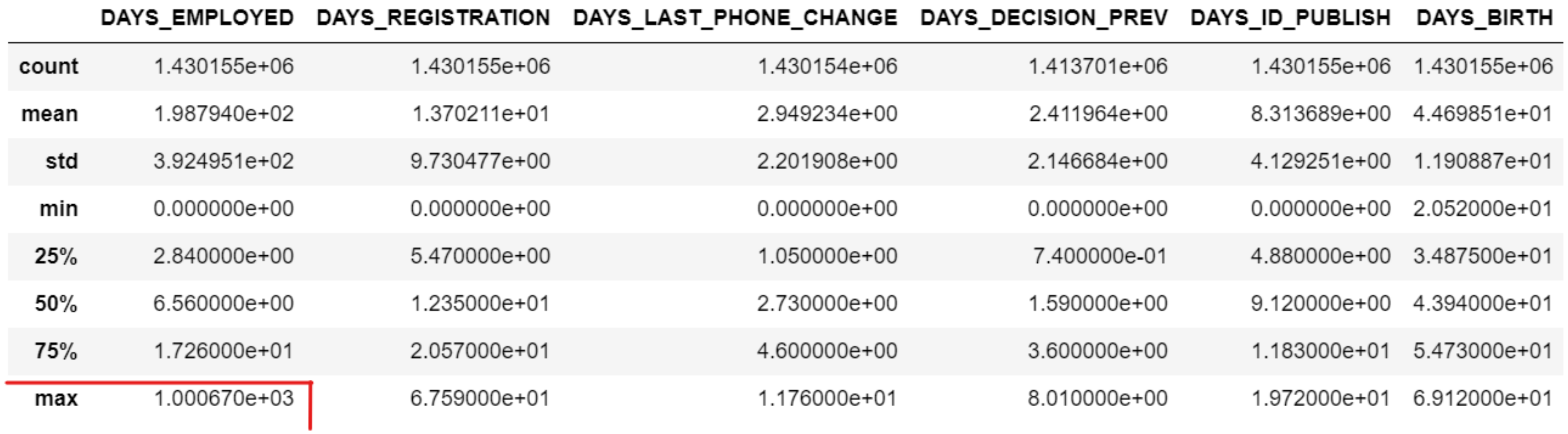
- Removed columns having more than 40% NaN values
- Observed certain column with days data having negative values. Took absolute value and converted this to years.



	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_LAST_PHONE_CHANGE	DAYS_DECISION_PREV	DAYS_ID_PUBLISH	DAYS_BIRTH
count	1.430155e+06	1.430155e+06	1.430154e+06	1.413701e+06	1.430155e+06	1.430155e+06
mean	6.860209e+04	-5.001275e+03	-1.076470e+03	-8.803670e+02	-3.034492e+03	-1.631495e+04
std	1.451967e+05	3.551626e+03	8.036988e+02	7.835402e+02	1.507182e+03	4.346737e+03
min	-1.791200e+04	-2.467200e+04	-4.292000e+03	-2.922000e+03	-7.197000e+03	-2.522900e+04
25%	-2.825000e+03	-7.509000e+03	-1.678000e+03	-1.313000e+03	-4.319000e+03	-1.997500e+04
50%	-1.277000e+03	-4.506000e+03	-9.960000e+02	-5.820000e+02	-3.330000e+03	-1.603700e+04
75%	-2.820000e+02	-1.997000e+03	-3.830000e+02	-2.710000e+02	-1.783000e+03	-1.272950e+04
max	3.652430e+05	0.000000e+00	0.000000e+00	-1.000000e+00	0.000000e+00	-7.489000e+03

Data Cleaning and Manipulation

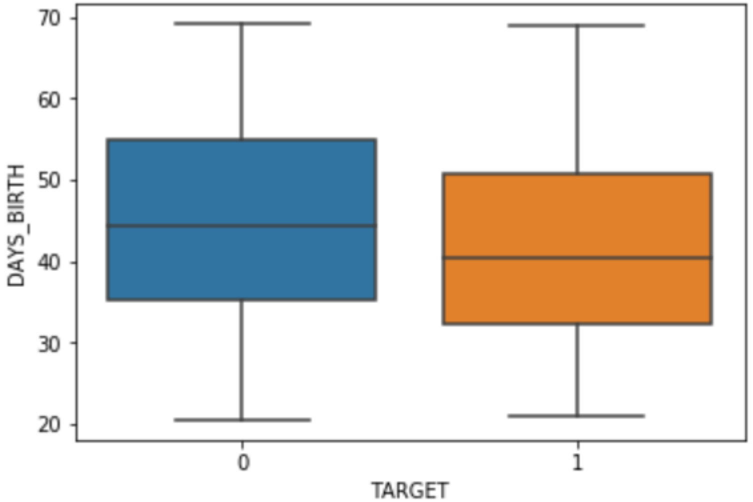
- Post converting Days to years, observed very large values as 1000 years.
- DAYS_EMPLOYED column had values which are greater than corresponding age, imputed such values with NaN. Any other approach does not ensure that employed years will be less than age.

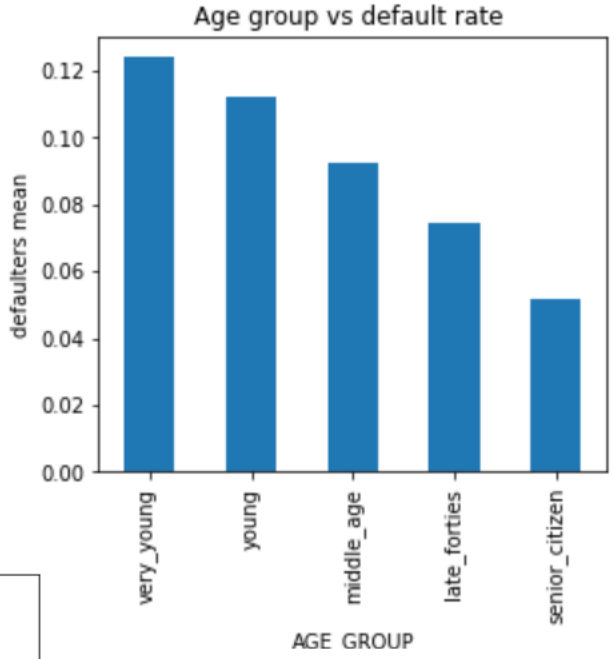


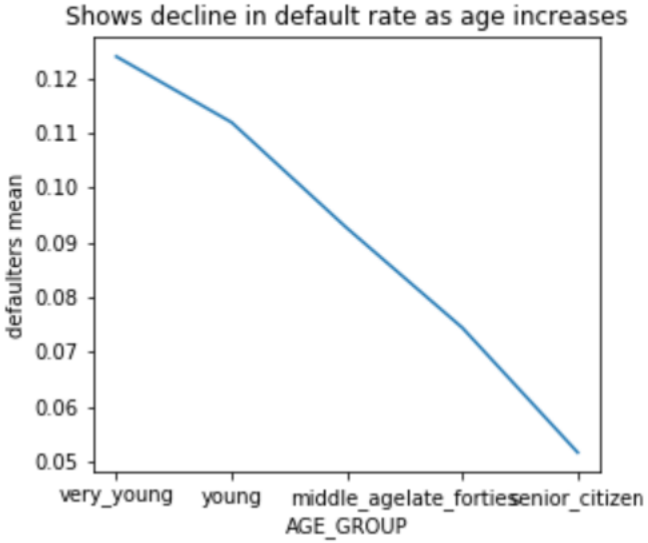
	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_LAST_PHONE_CHANGE	DAYS_DECISION_PREV	DAYS_ID_PUBLISH	DAYS_BIRTH
count	1.430155e+06	1.430155e+06	1.430154e+06	1.413701e+06	1.430155e+06	1.430155e+06
mean	1.987940e+02	1.370211e+01	2.949234e+00	2.411964e+00	8.313689e+00	4.469851e+01
std	3.924951e+02	9.730477e+00	2.201908e+00	2.146684e+00	4.129251e+00	1.190887e+01
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.052000e+01
25%	2.840000e+00	5.470000e+00	1.050000e+00	7.400000e-01	4.880000e+00	3.487500e+01
50%	6.560000e+00	1.235000e+01	2.730000e+00	1.590000e+00	9.120000e+00	4.394000e+01
75%	1.726000e+01	2.057000e+01	4.600000e+00	3.600000e+00	1.183000e+01	5.473000e+01
max	1.000670e+03	6.759000e+01	1.176000e+01	8.010000e+00	1.972000e+01	6.912000e+01

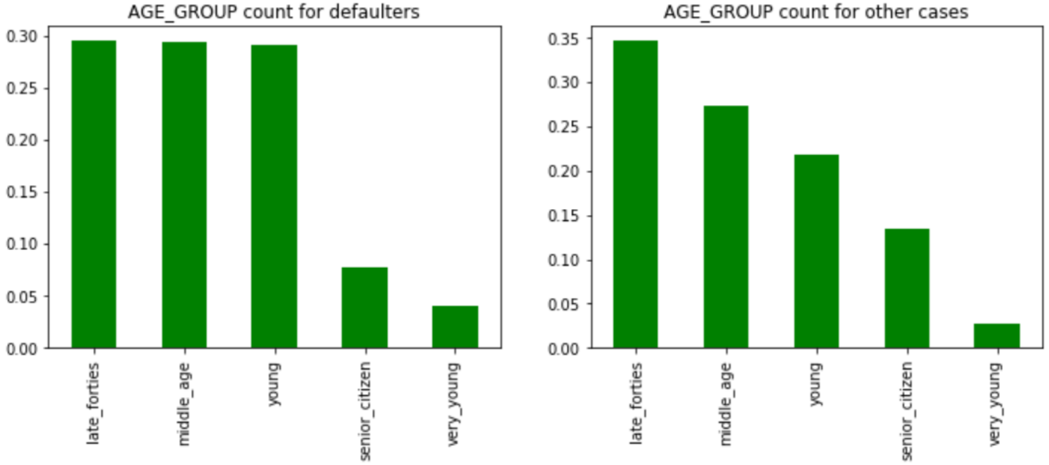
- Converted age and years employed to categorical columns.
- CNT_CHILDREN column had few outliers, the difference between 99th percentile and max value was very high, so capped the values to 5 children.

Analysis- Age of Customer




Age group vs default rate


Shows decline in default rate as age increases

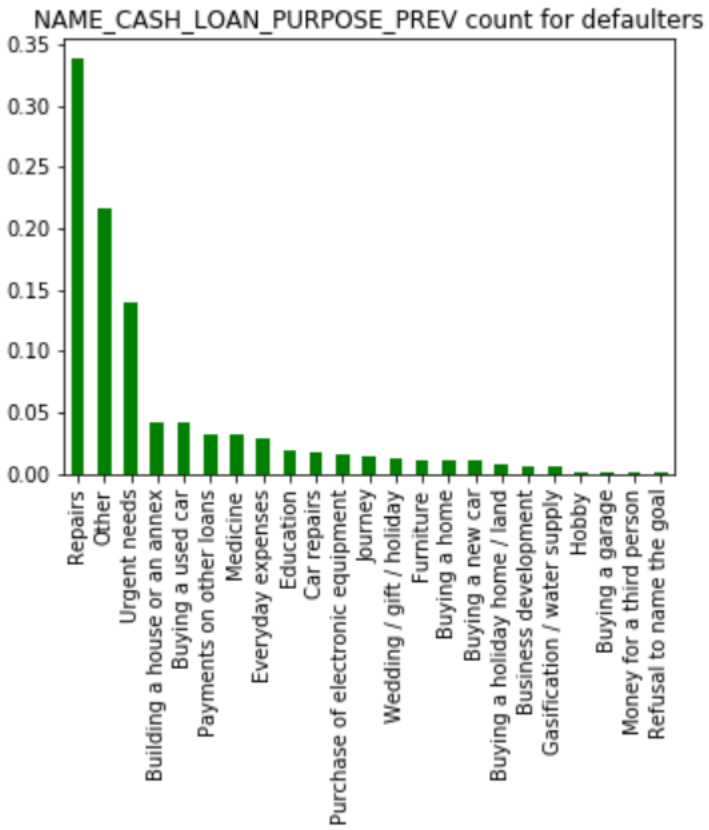

AGE_GROUP count for defaulters

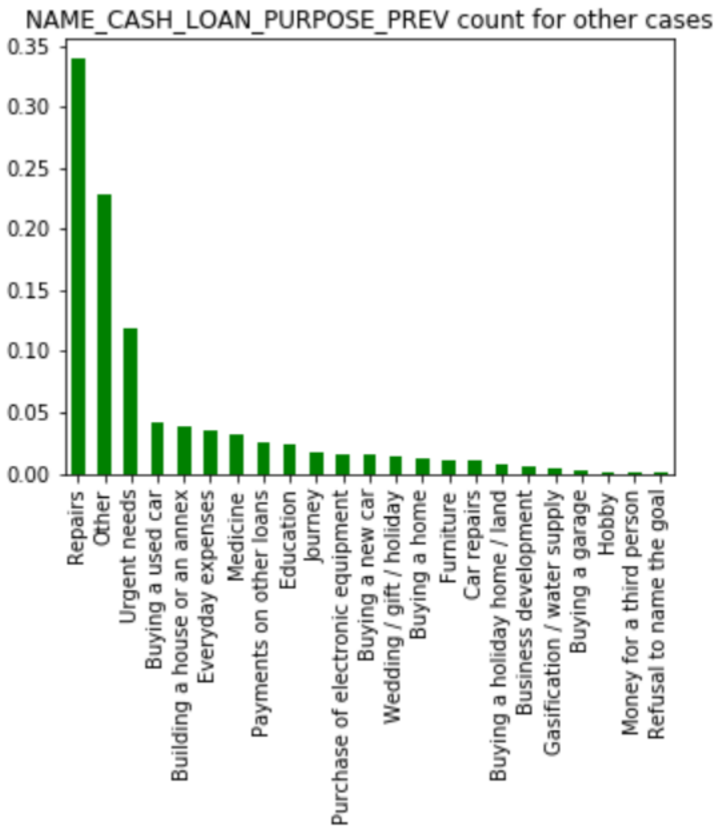
AGE_GROUP count for other cases

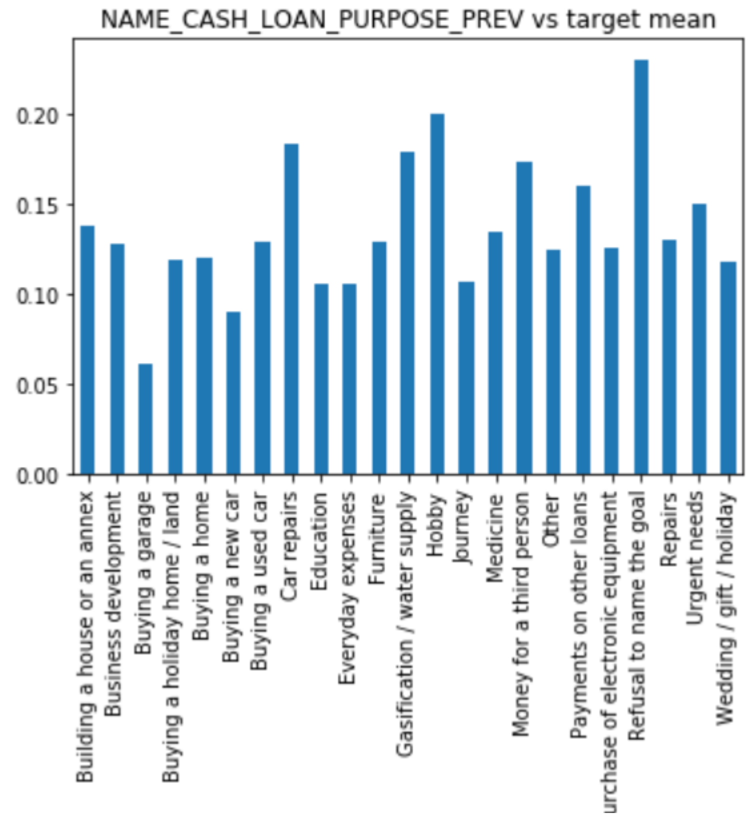
AGE_GROUP vs target mean

As age increases, the rate of defaulters is reducing. As this is not a strong indicator, it is recommend to not deny loan to young customers.

Analysis-Cash loan purpose

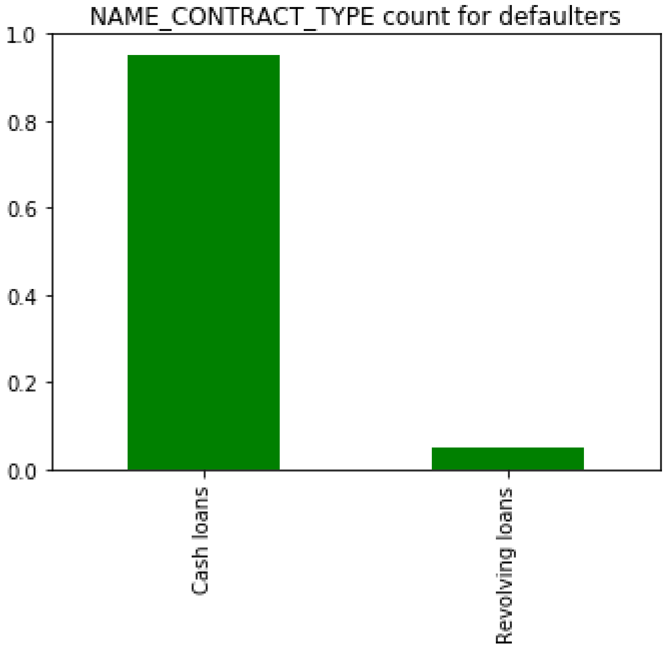


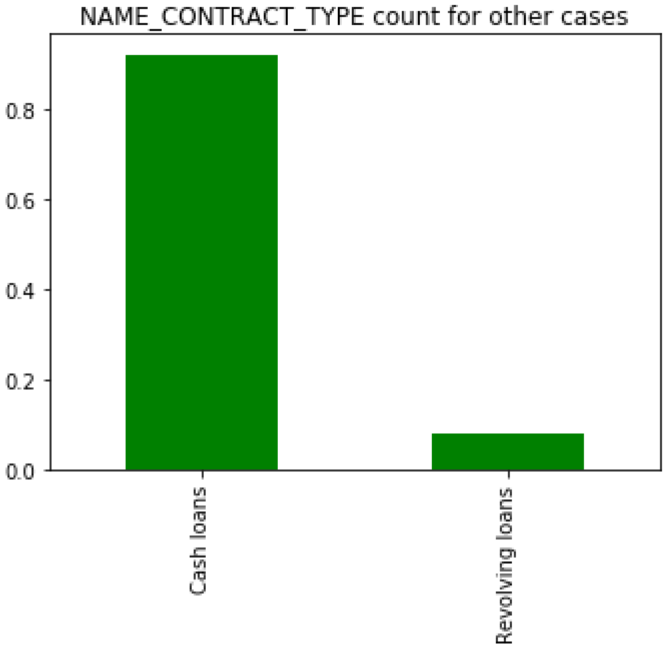


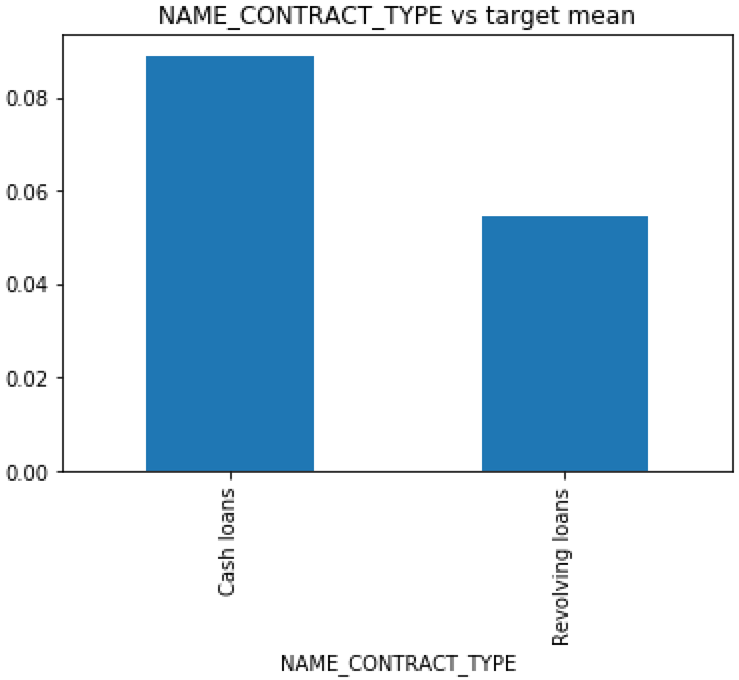


NAME_CASH_LOAN_PURPOSE_PREV, customers taking cash loan to pay other loans, refusal to name the purpose, have high default rates. Since records are very few in these categories, relying on these is not right. 'Taking money for third party' also has high defaulters. Recommendation would be to increase the interest rate or consider other parameters as well.

Analysis-Cash loan type

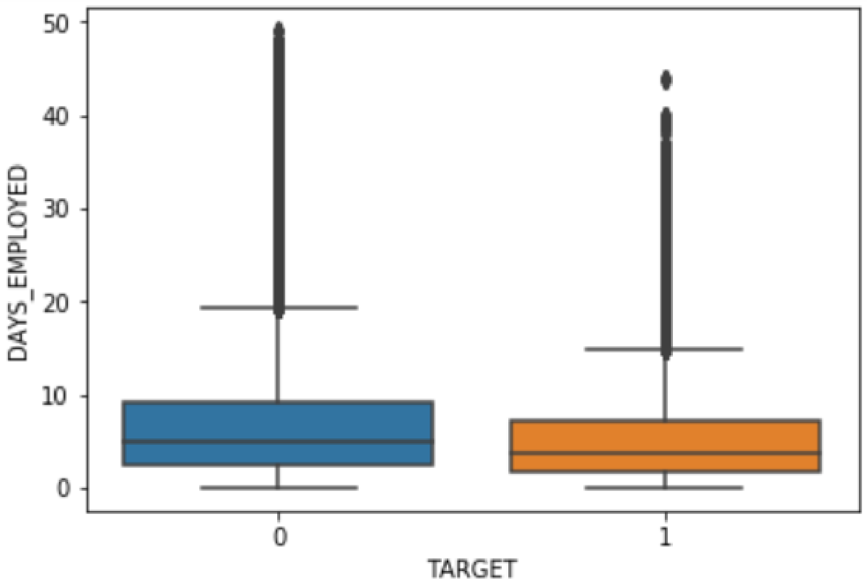

NAME_CONTRACT_TYPE count for defaulters

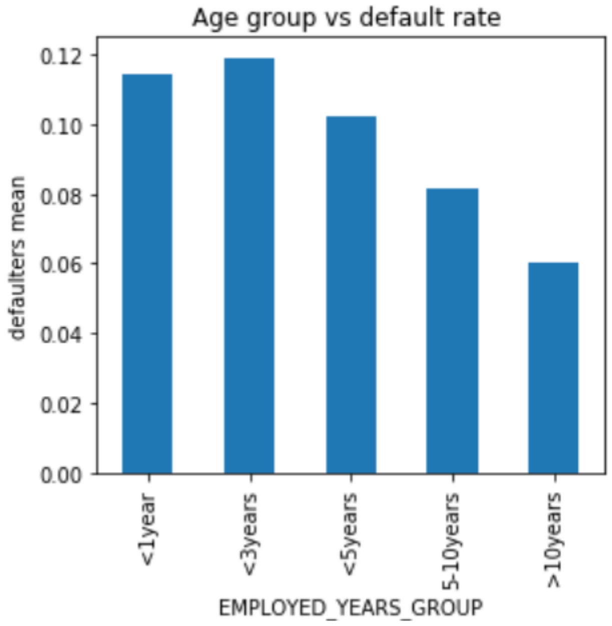

NAME_CONTRACT_TYPE count for other cases

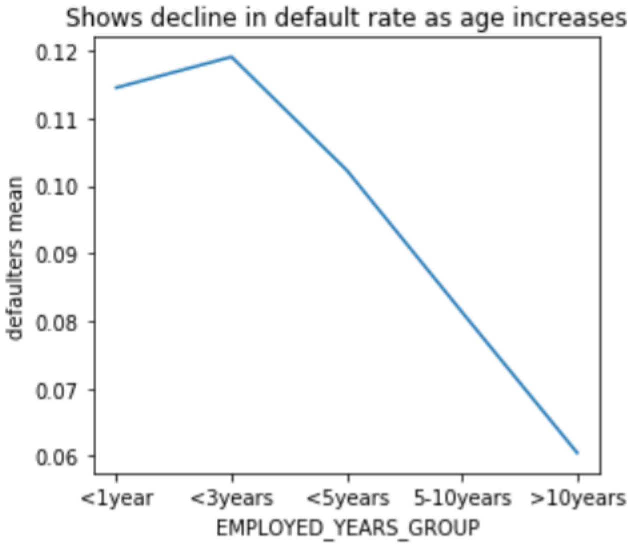

NAME_CONTRACT_TYPE vs target mean

Cash loans has more defaulters compared to that of revolving loans.

Analysis - Years of employment

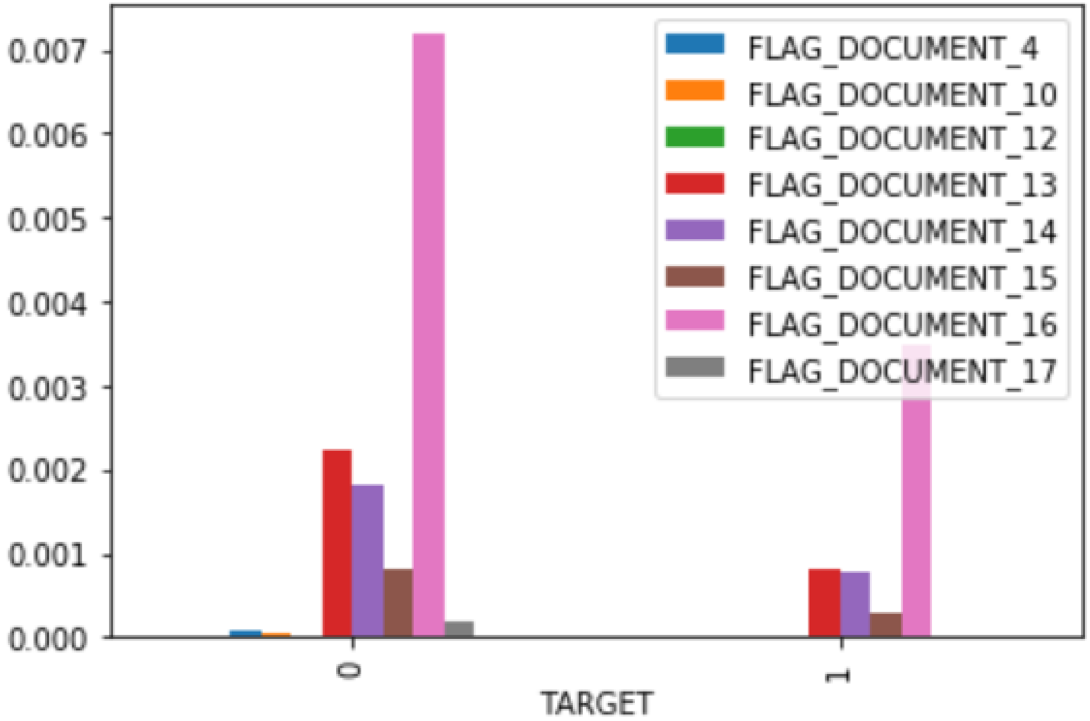






Customers with more than 10 Years of employment are less likely be defaulters compared to other groups. Above 40 years experience clients have no defaulters.

Analysis-Flag document columns

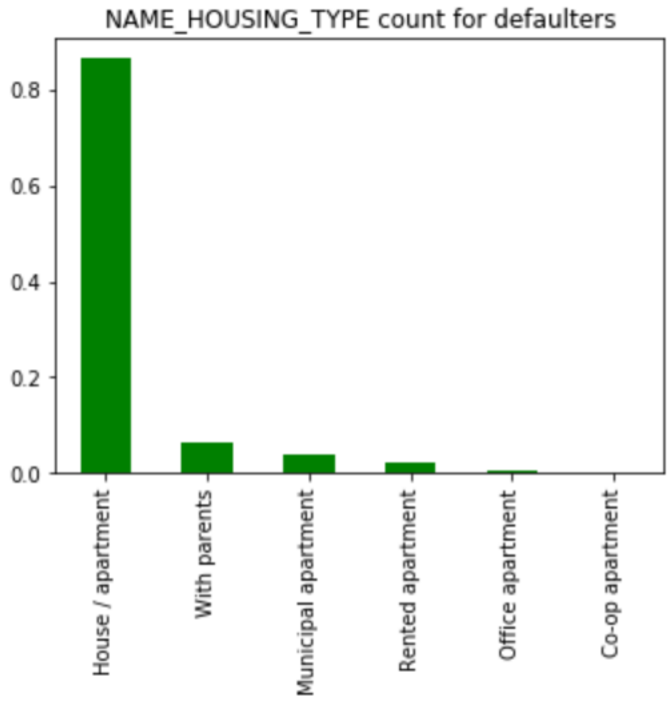


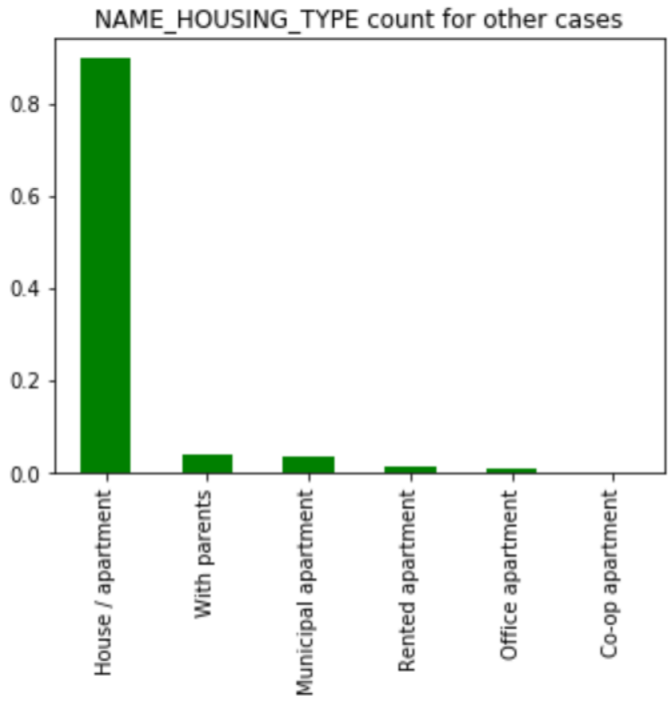
Documents 2 to 21 is not clear on the content of the document, on analysis found those who did not submit the documents has higher chance of defaulters.

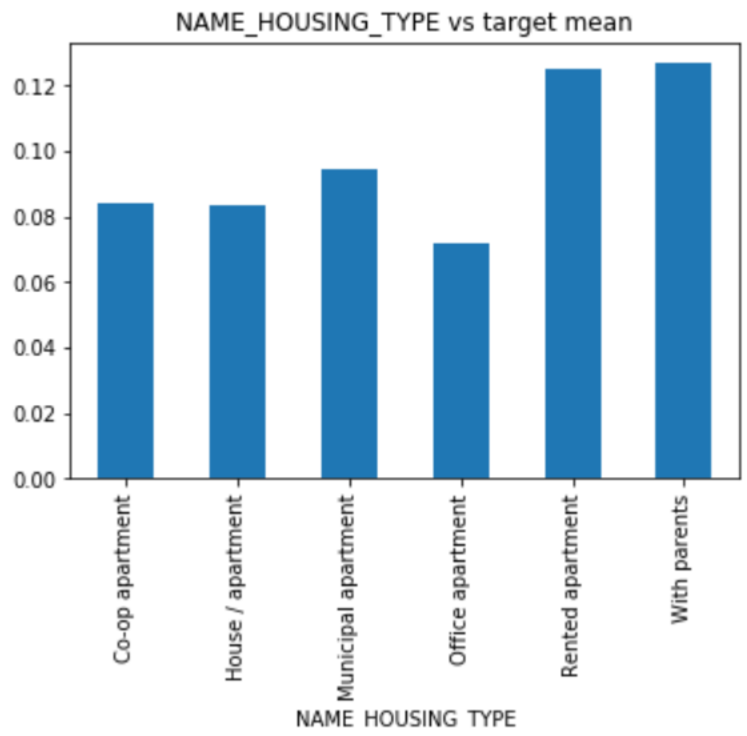
The Above graph shows the defaulters (x axis 1) have not submitted as many documents as other cases.

Some documents which are important among them are document 4, 10, 12, 13,14, 15, 16, 17

Analysis-housing type

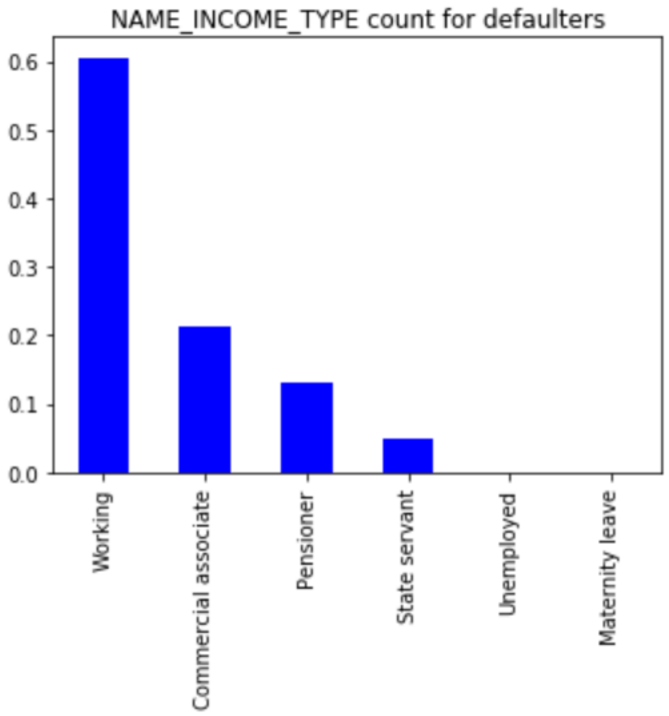

NAME_HOUSING_TYPE count for defaulters


NAME_HOUSING_TYPE count for other cases

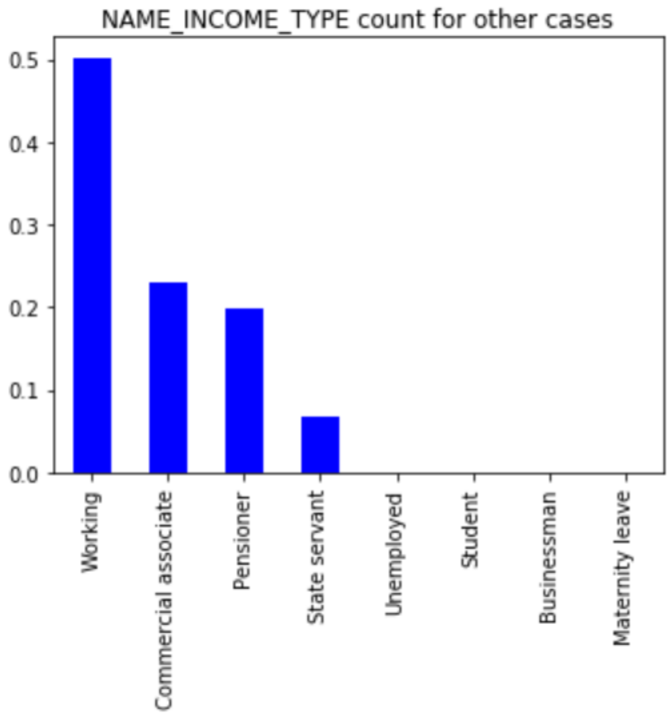

NAME_HOUSING_TYPE vs target mean

- NAME_HOUSING_TYPE, Rented apartments and people living with parents have increased default rates compared to other categories

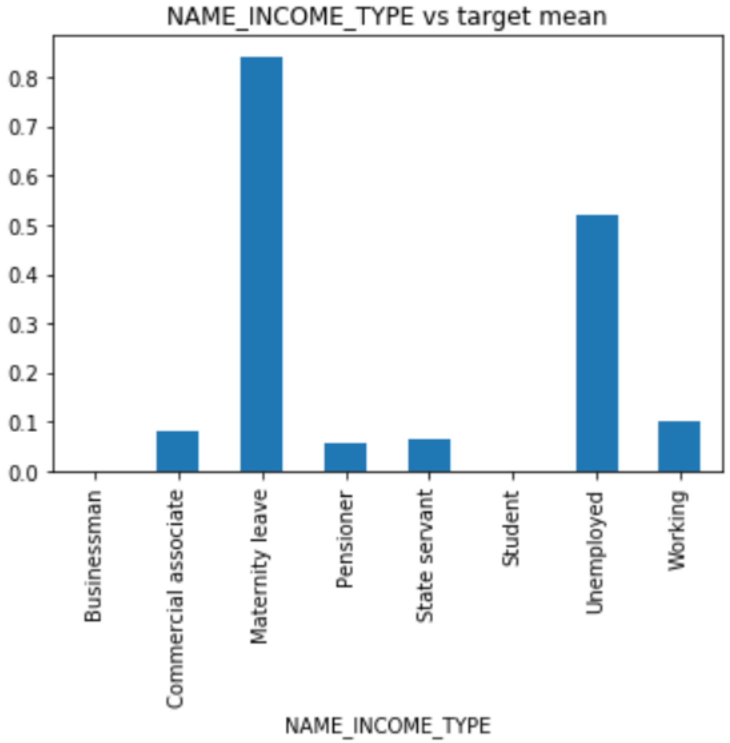
Analysis- Income type



NAME_INCOME_TYPE count for defaulters



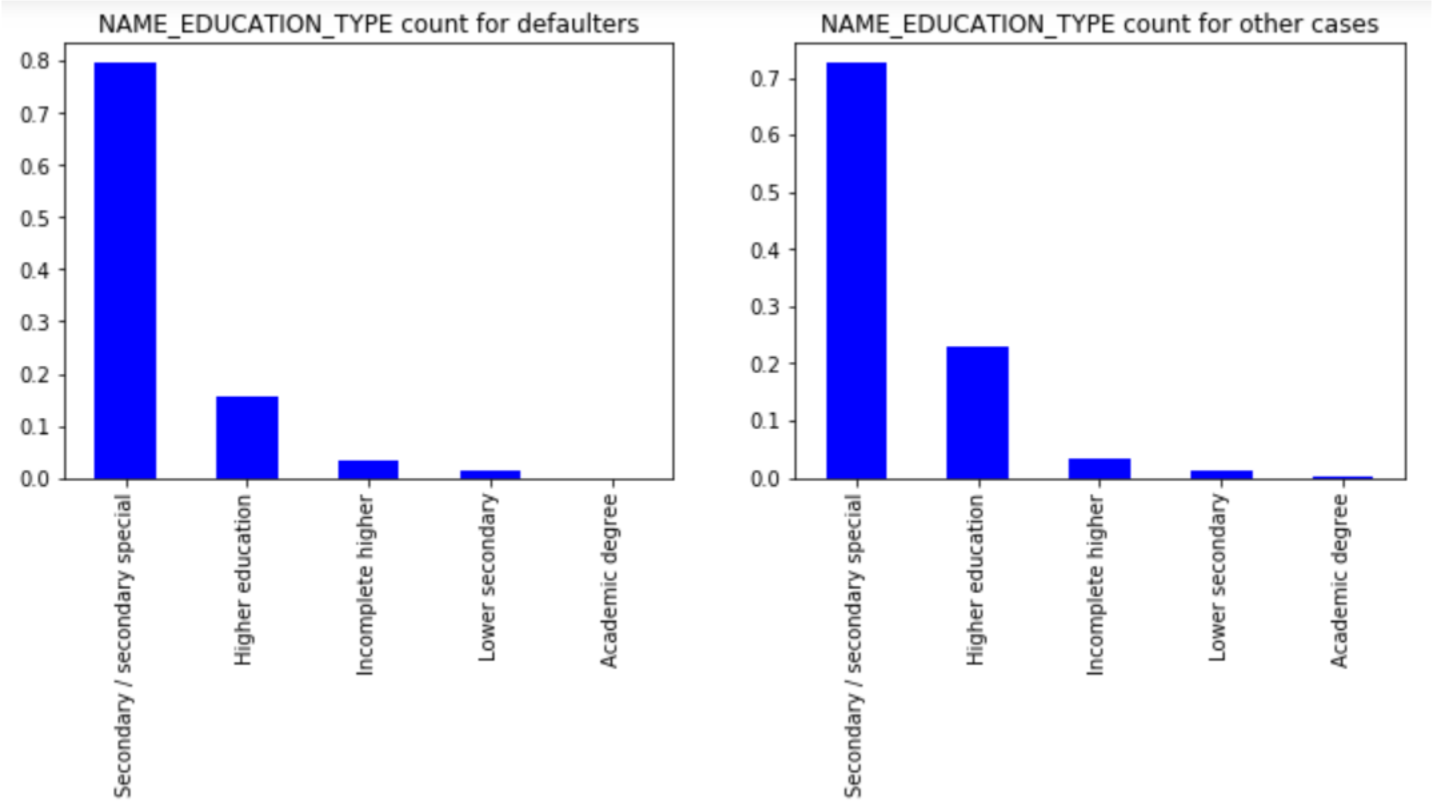
NAME_INCOME_TYPE count for other cases

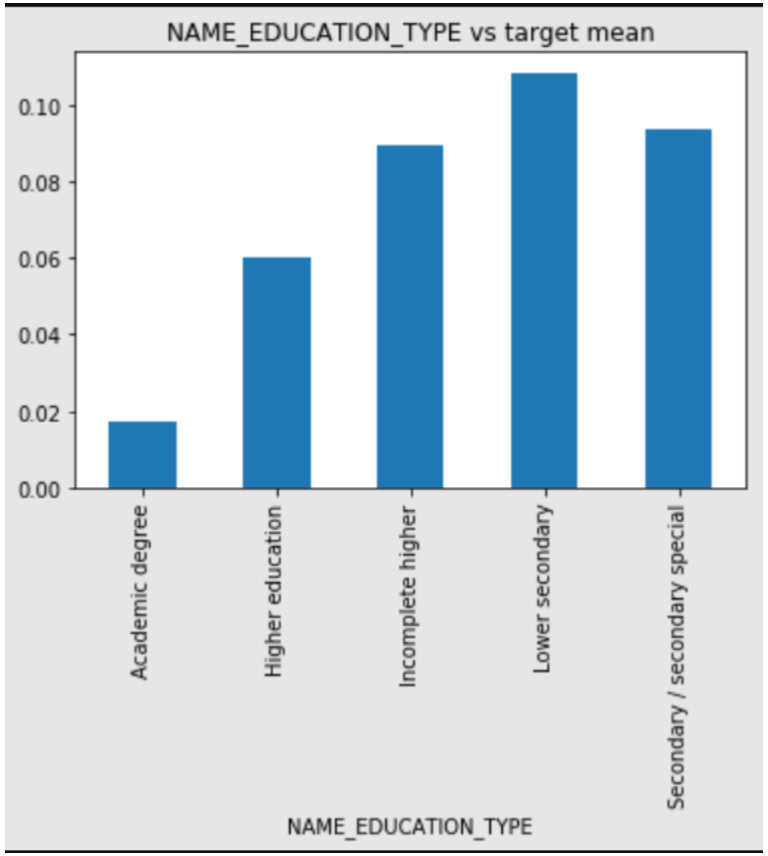


NAME_INCOME_TYPE vs target mean

Income type column, Maternity leave and Unemployed have high defaulters

Analysis- Education type

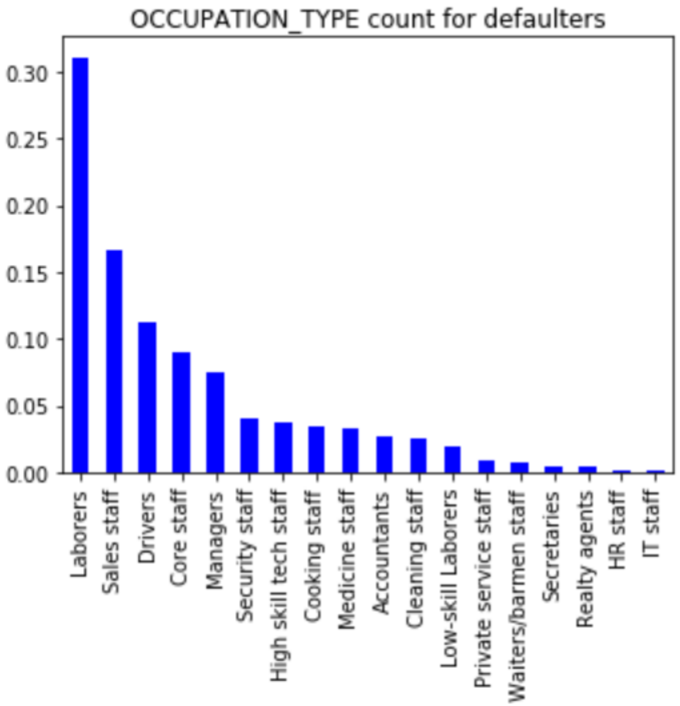


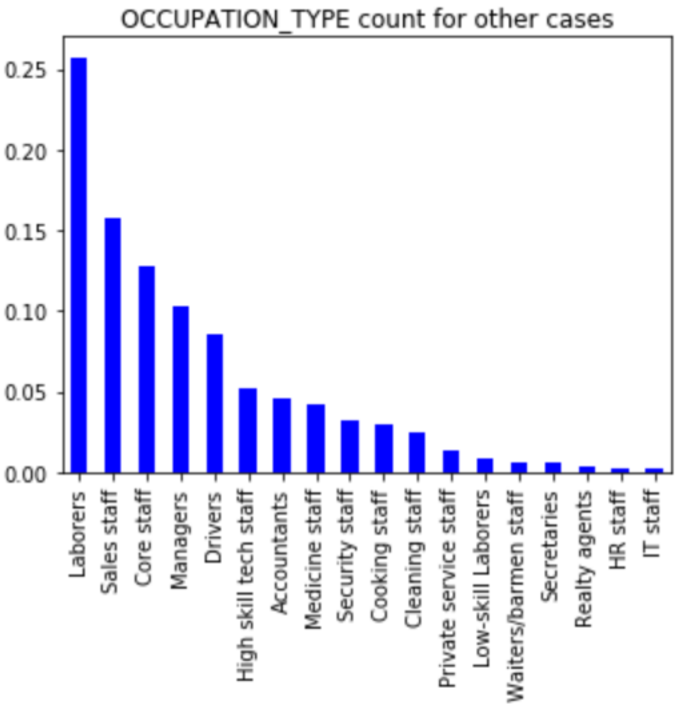


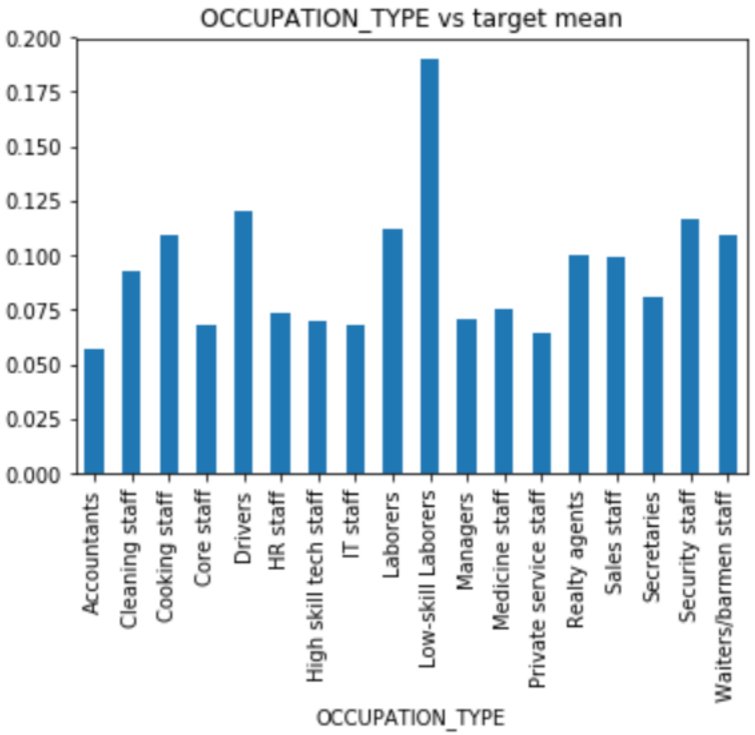
customers with education ‘lower secondary’ and ‘Secondary / secondary special’ have higher default rates

Borrower’s who took loans for small business purpose have defaulted more.

Analysis- Occupation type

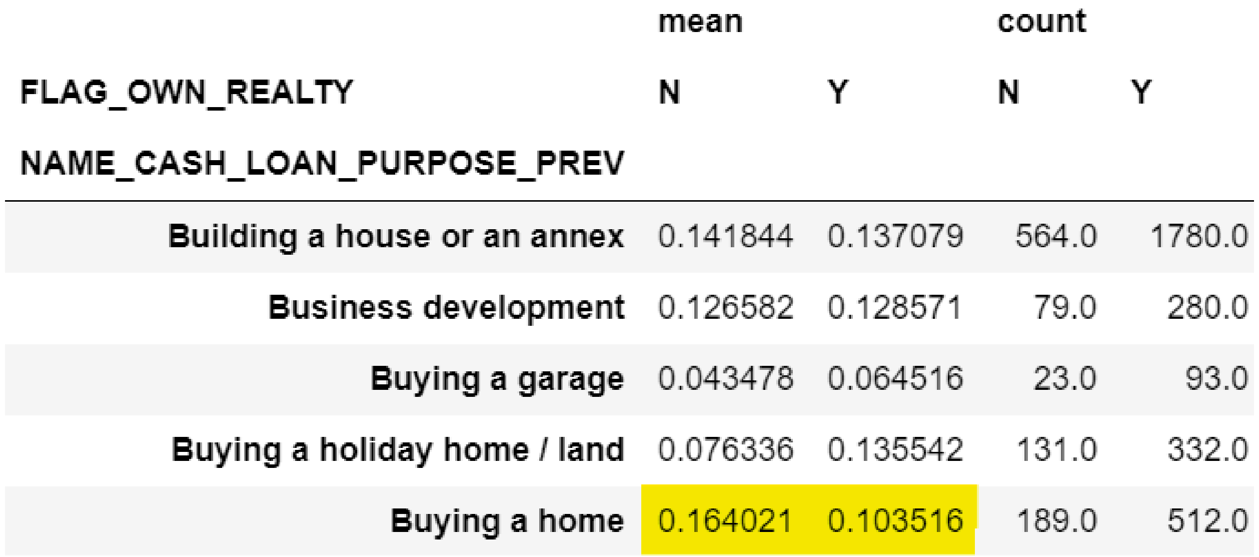




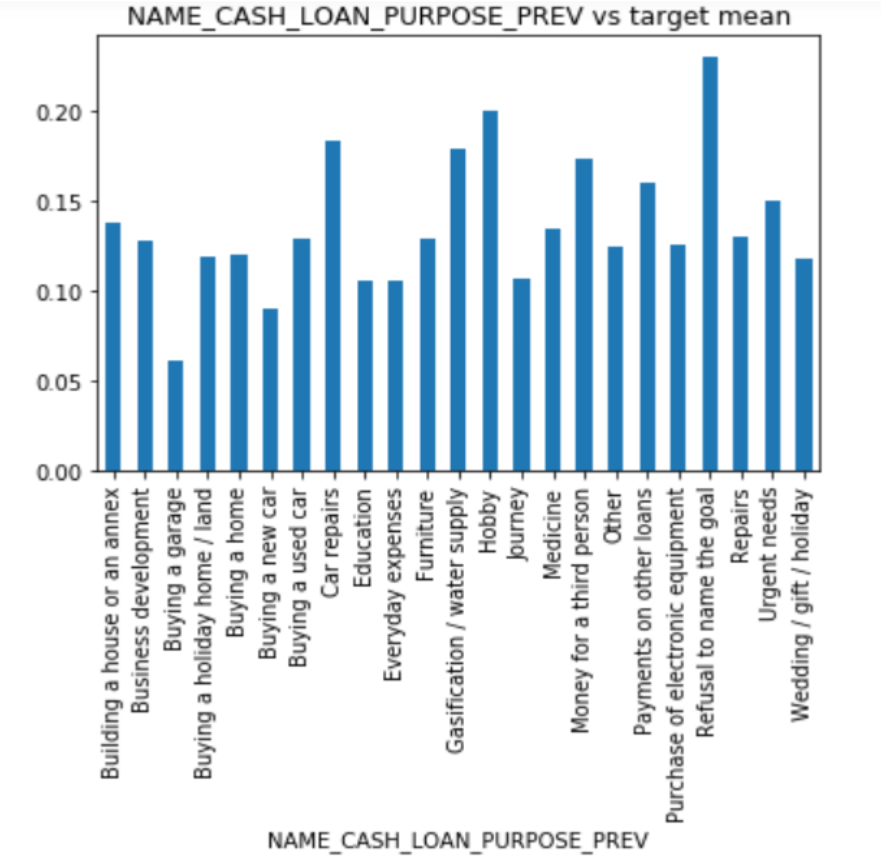


- Column Occupation type, low skill laborers have higher default rate.

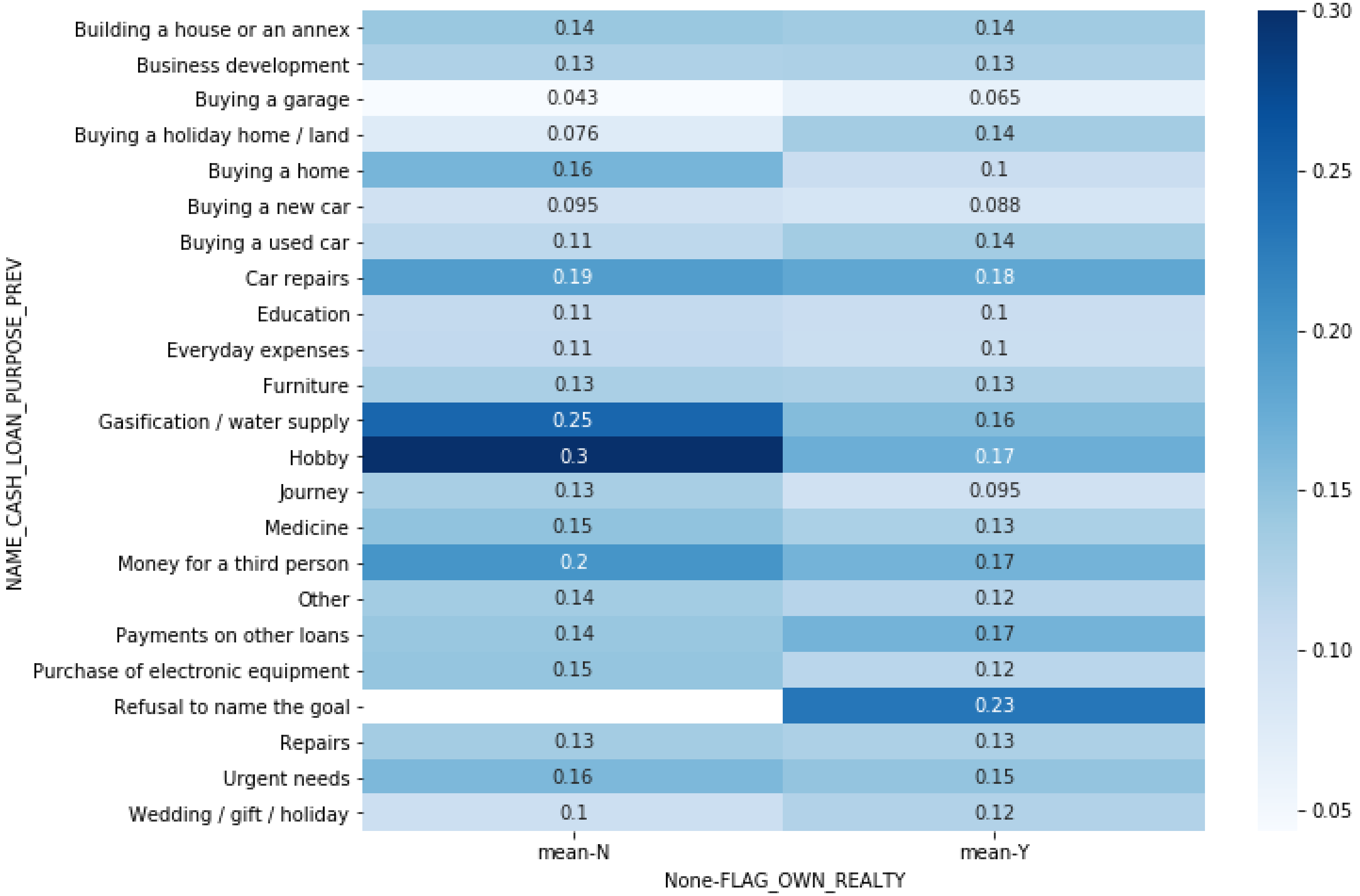
Analysis-Cash loan purpose vs flag own realty



	mean		count	
FLAG_OWN_REALTY	N	Y	N	Y
NAME_CASH_LOAN_PURPOSE_PREV				
Building a house or an annex	0.141844	0.137079	564.0	1780.0
Business development	0.126582	0.128571	79.0	280.0
Buying a garage	0.043478	0.064516	23.0	93.0
Buying a holiday home / land	0.076336	0.135542	131.0	332.0
Buying a home	0.164021	0.103516	189.0	512.0

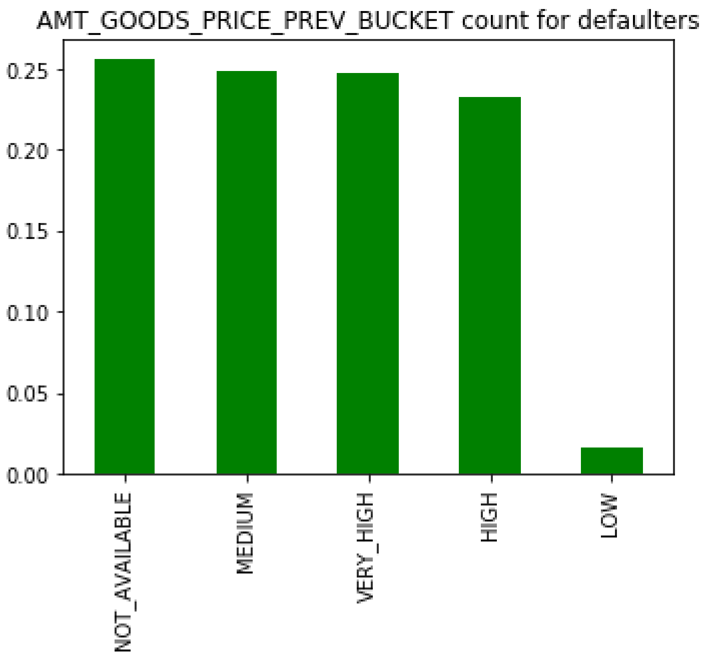


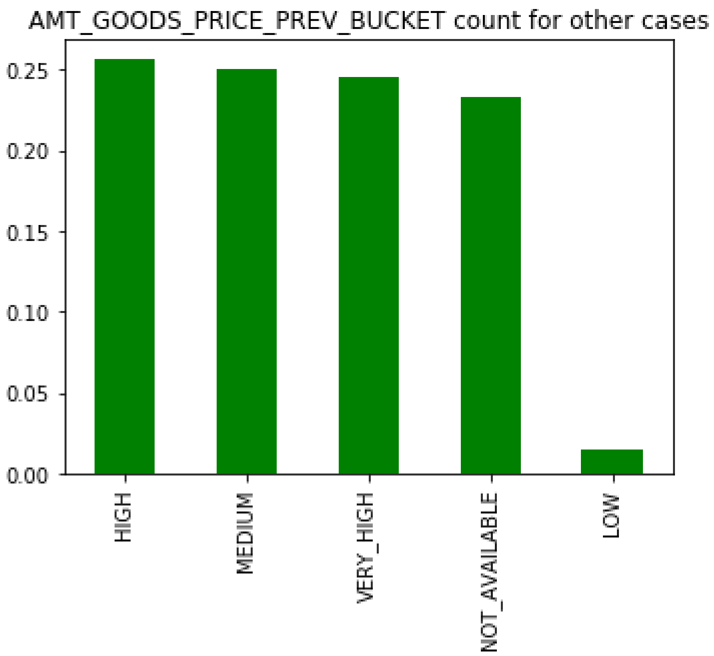
- If a customer owns a realty and then ‘buying house’ their defaulter percentage is lower compared to customers who don't owns a realty.

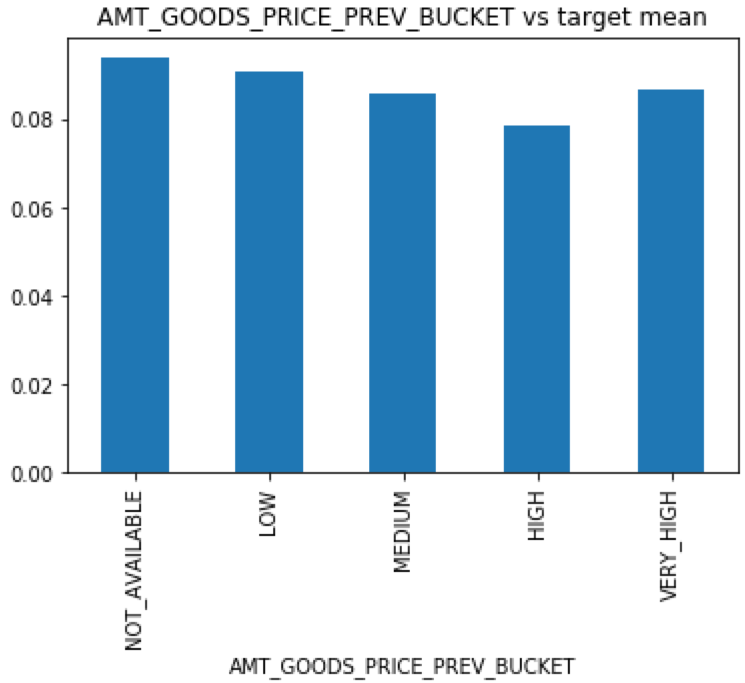


- 'Urgent needs', 'Hobby' also have high defaulters compared to other categories.

Analysis- Goods price bucket

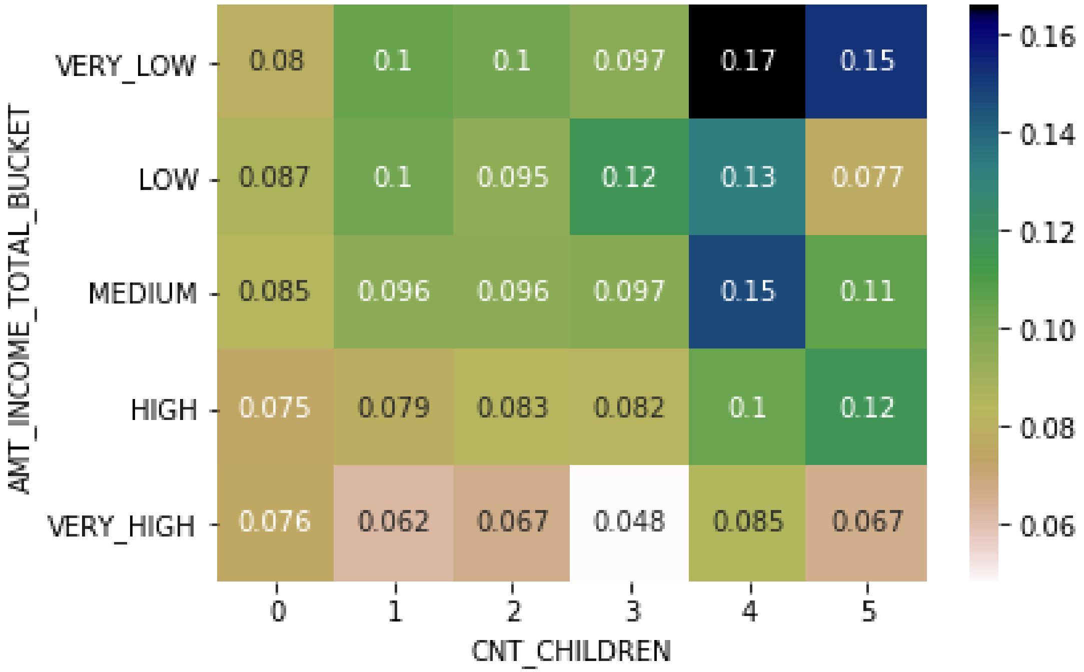

AMT_GOODS_PRICE_PREV_BUCKET count for defaulters

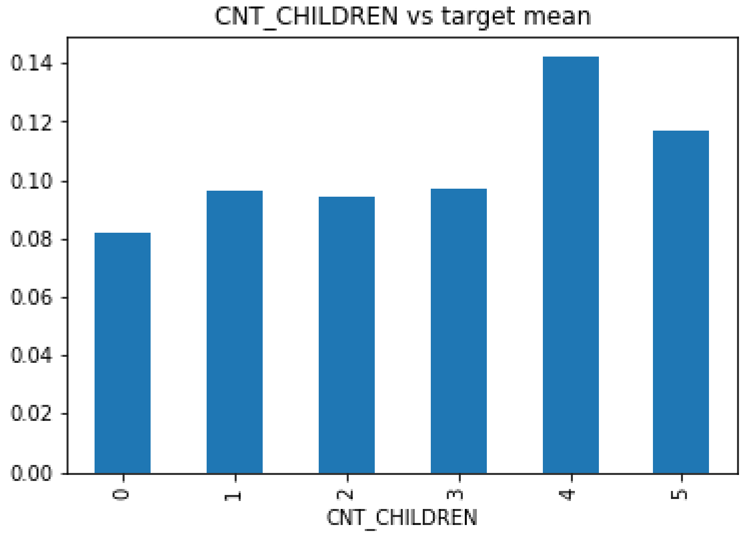

AMT_GOODS_PRICE_PREV_BUCKET count for other cases


AMT_GOODS_PRICE_PREV_BUCKET vs target mean

- The rate of defaulters are high where the goods price is not mentioned, followed by medium and very high price.

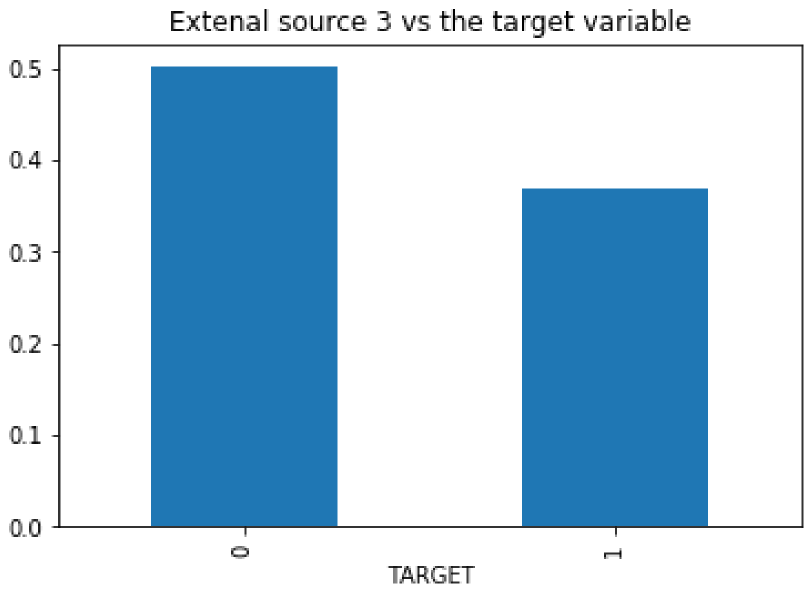
Analysis- No of children vs income bucket

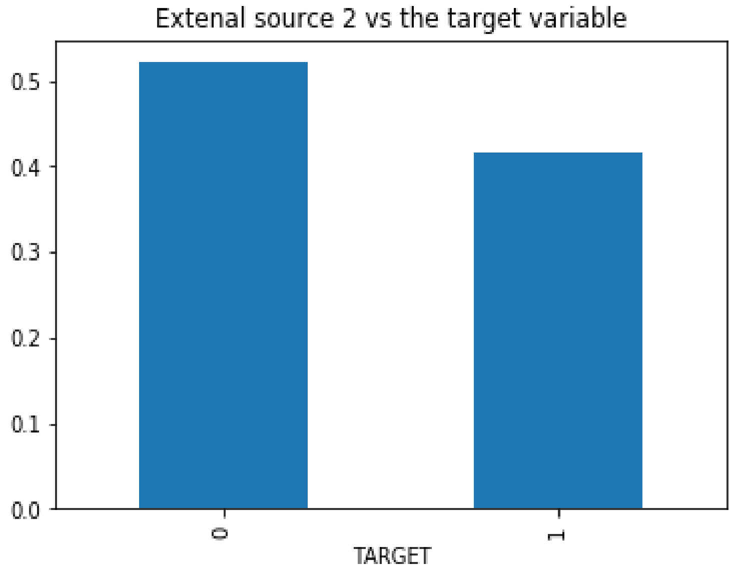




- CNT_CHILDREN column is insightful, we can see that as number of children increases the default rate increases.
- Analysis on CNT_CHILDREN with INCOME_BUCKETS, customers with low income and have many children are likely to be defaulters, as we can visualize on the top right of the heat map

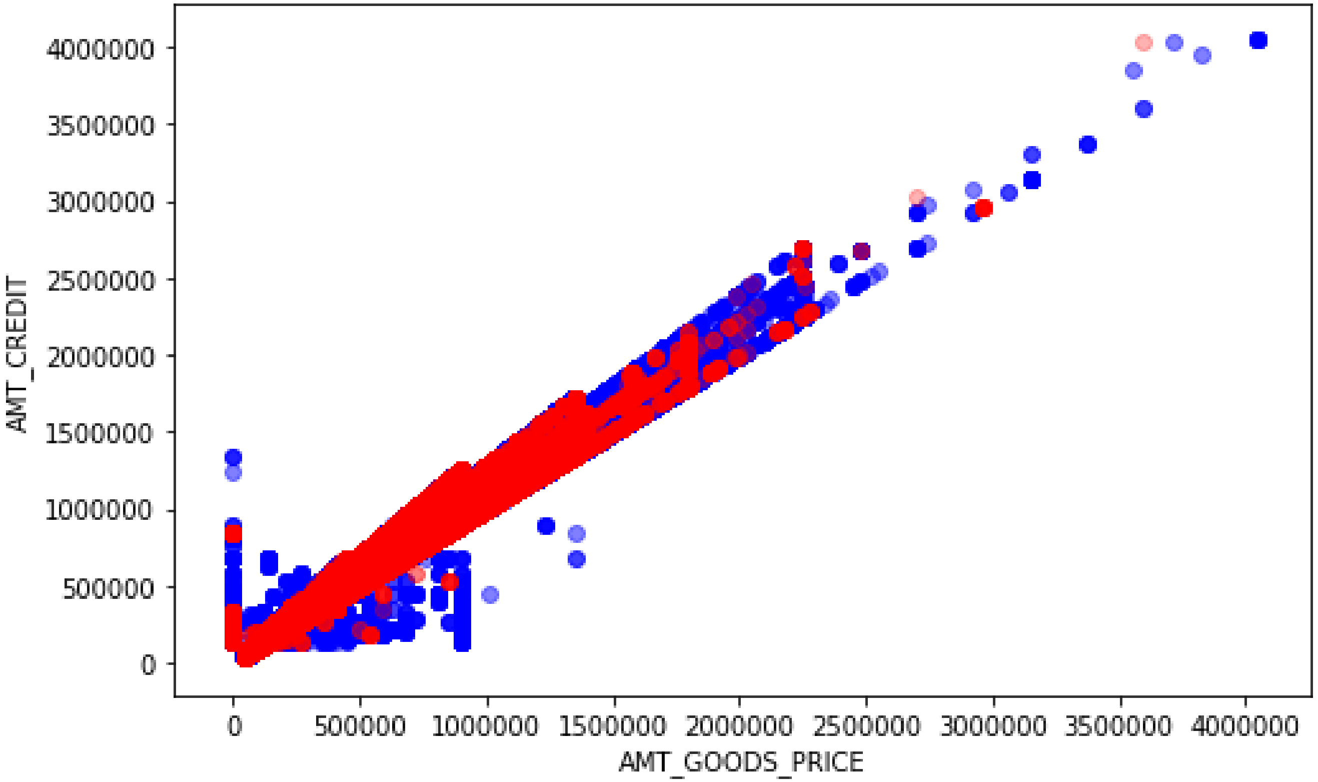
Analysis- External Data Source





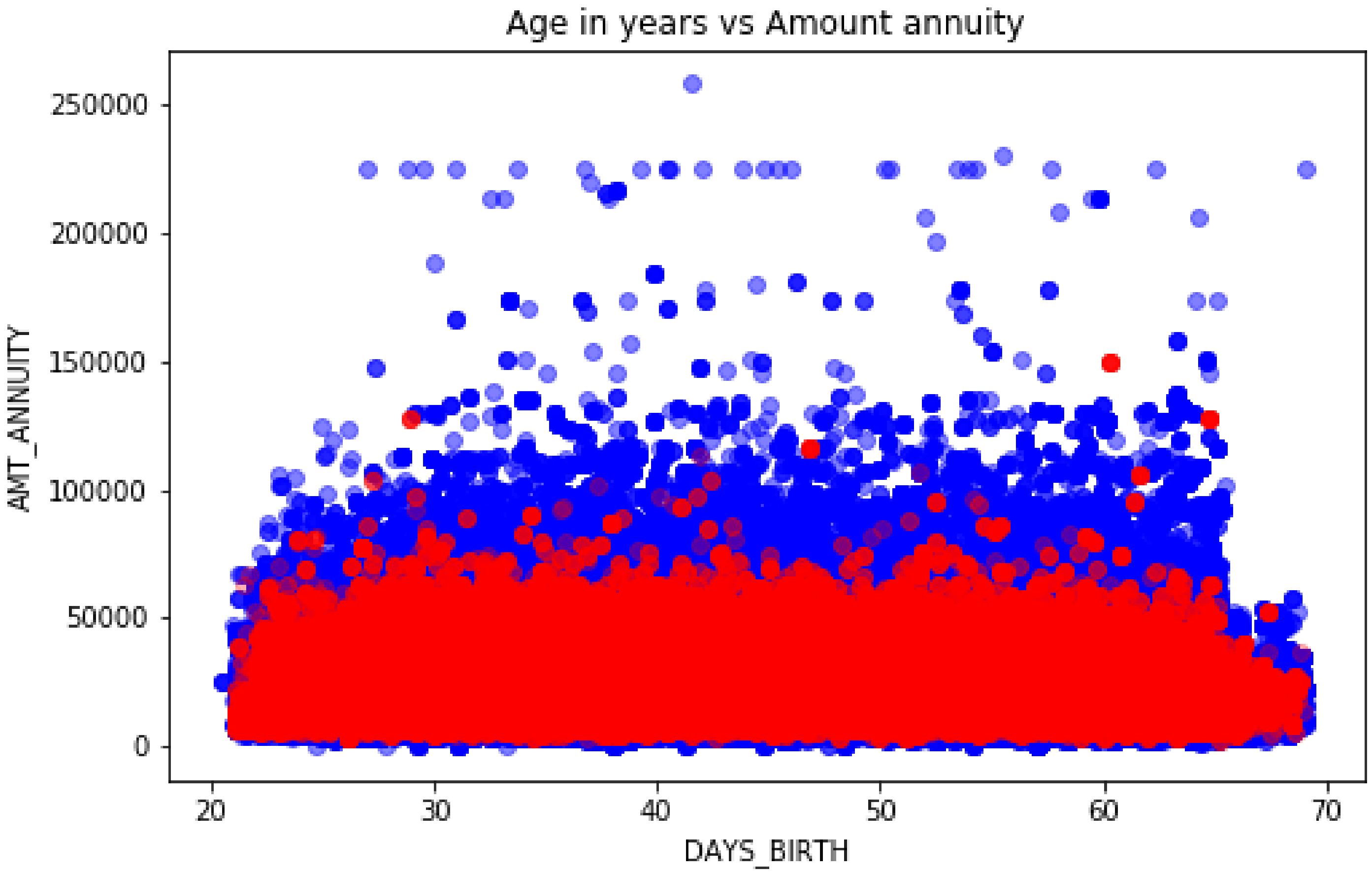
- In the External source data, if the normalised value is low, then the default rate is higher.

Analysis- Credit amount vs goods price



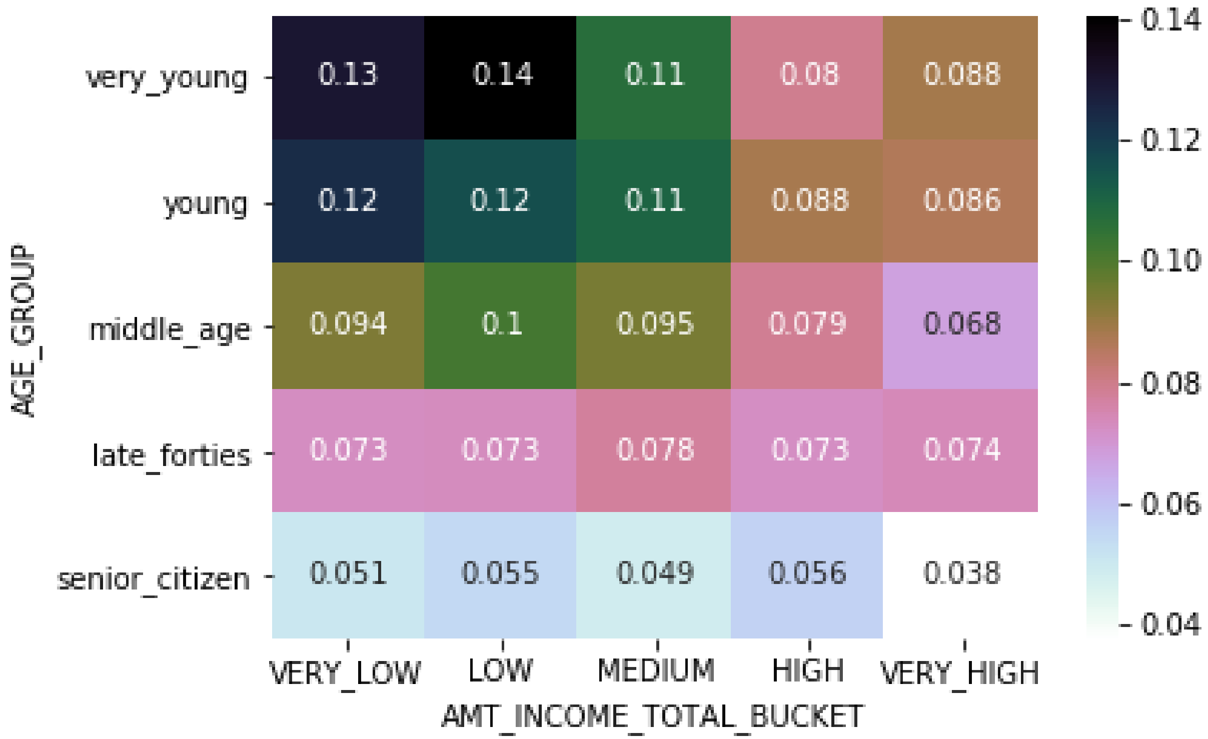
- If Credit amount more than 2,50,0000 the number of defaulters reduced by more than 10 times.

Analysis- Age vs annuity amount



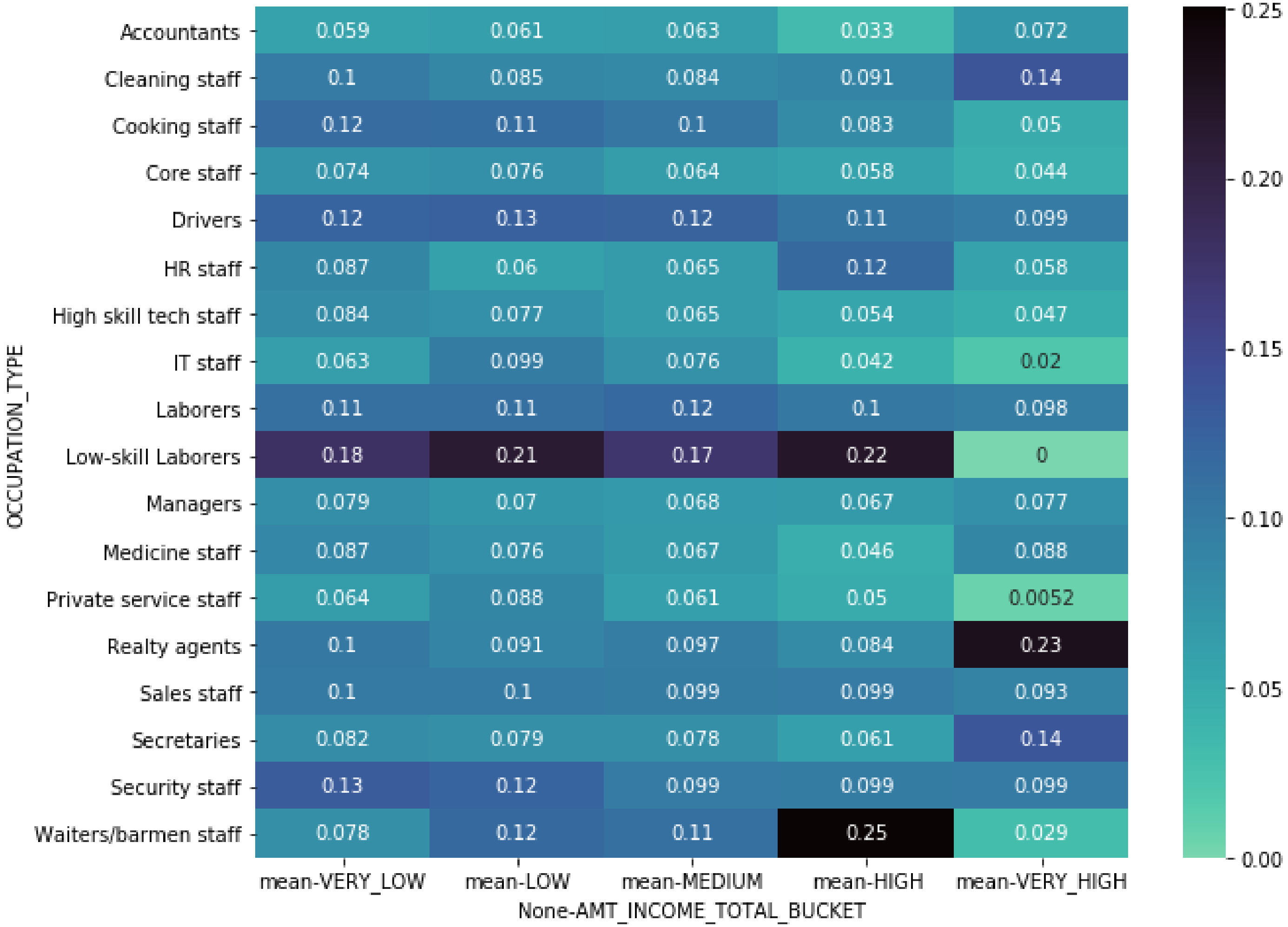
- Young and Middle age for low annuity has more default rate. Seniors have low default rate for all annuity buckets.
- For Annuity more than 15000, defaulters are very few almost 10 times lesser.

Analysis- Age vs Income



- Top left has higher default rates.
- Younger and less salary means higher default rate

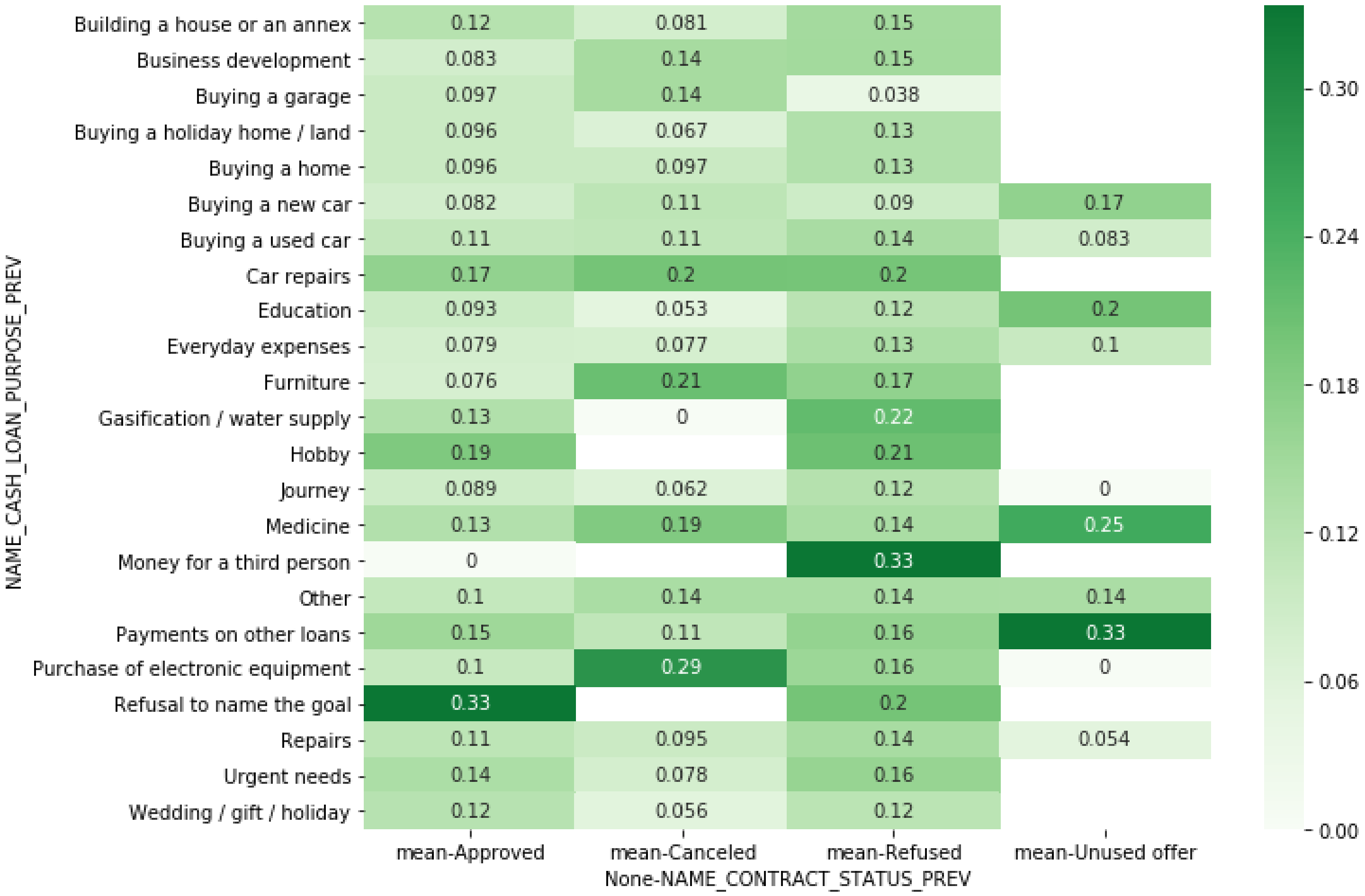
Analysis- AMT_INCOME vs Occupation



OCCUPATION_TYPE	mean-VERY_LOW	mean-LOW	mean-MEDIUM	mean-HIGH	mean-VERY_HIGH
Accountants	0.059	0.061	0.063	0.033	0.072
Cleaning staff	0.1	0.085	0.084	0.091	0.14
Cooking staff	0.12	0.11	0.1	0.083	0.05
Core staff	0.074	0.076	0.064	0.058	0.044
Drivers	0.12	0.13	0.12	0.11	0.099
HR staff	0.087	0.06	0.065	0.12	0.058
High skill tech staff	0.084	0.077	0.065	0.054	0.047
IT staff	0.063	0.099	0.076	0.042	0.02
Laborers	0.11	0.11	0.12	0.1	0.098
Low-skill Laborers	0.18	0.21	0.17	0.22	0
Managers	0.079	0.07	0.068	0.067	0.077
Medicine staff	0.087	0.076	0.067	0.046	0.088
Private service staff	0.064	0.088	0.061	0.05	0.0052
Realty agents	0.1	0.091	0.097	0.084	0.23
Sales staff	0.1	0.1	0.099	0.099	0.093
Secretaries	0.082	0.079	0.078	0.061	0.14
Security staff	0.13	0.12	0.099	0.099	0.099
Waiters/barmen staff	0.078	0.12	0.11	0.25	0.029

- 'Waiters/barmen' with high salary have higher default rate
- 'Realty agents' with very high salary have higher default rate

Analysis- cash loan type vs contract status



- 'Medicine' generally have a higher value but under previously 'canceled' and 'unused' offers have high default rate but since the records are few its not made this a important deciding factor.
- 'Hobby' has highest default rate for previously 'Approved' loans, it has high default rate
- 'gassification/water supply' generally have high values as high under 'approved' and high under 'Refused'

Top 10 correlations with recommendations

- Very low income and if number of children are more than 3, the default rate is high.
- If the customer already owns a realty, then the default rate for buying a house is less compared to the one who don’t own a realty. **The interest rate can be increased for buying a house if they don’t already own a realty.**
- If the education is lower secondary and Secondary / secondary special have higher default rates, then the default rate is high. **The interest rate can be increased for those with these education background.**
- Low-skill labourers at every income bucket and waiters/barmen with high pay, are high likely to be defaulters. **Higher amount loan can be refused for low-skill labourers and interest rate can be increased for barmen and waiter with high pay**
- Customers taking cash loans are high default rate compared to other type loans.
- **Taking cash loans to pay other loans should be rejected. Taking cash loans to pay third party should have high interest rate**

Top 10 correlations with recommendations

- Annuity more than 15000, defaulters are very few almost 10 times lesser. **So as the annuity amount reduces the interest rate should go high.**
- If Credit amount is more than 2,50,000, then the number of defaulters are less. **So interest rate should be increased as credit amount reduces.**
- Those who did not submit document 4, 10, 12, 13,14, 15, 16, 17 had more defaulters compared to customers who submitted. **Submitting some of these documents should be made compulsory.**
- As age decreases and if the customer income is low then **the interest rates should increase**.
- Customers who have <10 Years of employment have higher default rates. Hence loans at higher interest rates could be given these clients (<10 years employment)