Exploratory Data Analysis

Provided Data Shape:

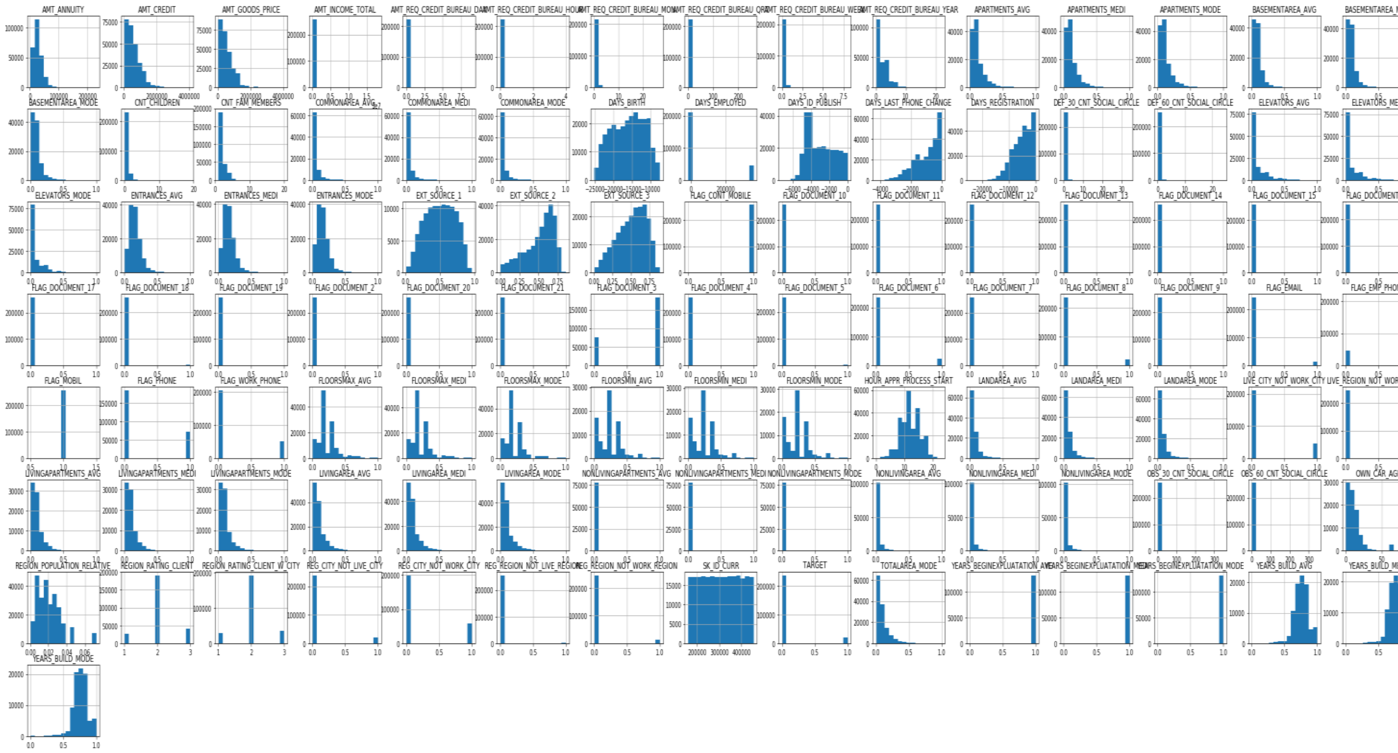
(257512, 122)->Train size

(49999, 121)->Test Size

Modelling Data Shape

(257512, 248)->Train, (49999, 247)->Test

1. Histogram



DAYS_EMPLOYED columns has ambiguous data as verified from the histogram ,which needed replacement with -999.other columns seems to have unambiguous data.

2. Statistical description of dataset

	count	mean	std	min	25%	50%	75%
SK_ID_CURR	257512	307143.1	86047.05	157876	232638.8	307140.5	381476
TARGET	257512	0.080769	0.272481	0	0	0	0
CNT_CHILDREN	257512	0.416509	0.721749	0	0	0	1

AMT_INCOME_TOTAL	257512	168415.5	110587.2	26100	112500	147600	20250
AMT_CREDIT	257512	598895	402506.1	45000	270000	513531	80865
AMT_ANNUITY	257501	27108.81	14480.29	1615.5	16542	24903	3459
AMT_GOODS_PRICE	257272	538267.3	369368	40500	238500	450000	67950
REGION_POPULATION_RELATIVE	257512	0.020882	0.013845	0.00029	0.010006	0.01885	0.0286
DAYS_BIRTH	257512	-16039.9	4364.494	-25229	-19689	-15753	-1242
DAYS_EMPLOYED	257512	63930.69	141369	-17912	-2756	-1212	-288
DAYS_REGISTRATION	257512	-4987.84	3522.374	-24672	-7483	-4506	-201
DAYS_ID_PUBLISH	257512	-2993.7	1509.495	-7197	-4299	-3253	-171
OWN_CAR_AGE	87533	12.06797	11.95599	0	5	9	15
FLAG_MOBIL	257512	1	0	1	1	1	1
FLAG_EMP_PHONE	257512	0.819581	0.384537	0	1	1	1
FLAG_WORK_PHONE	257512	0.199389	0.399542	0	0	0	0
FLAG_CONT_MOBILE	257512	0.998163	0.042819	0	1	1	1
FLAG_PHONE	257512	0.281715	0.449836	0	0	0	1
FLAG_EMAIL	257512	0.056926	0.231701	0	0	0	0
CNT_FAM_MEMBERS	257511	2.151446	0.910552	1	2	2	3
REGION_RATING_CLIENT	257512	2.052619	0.50924	1	2	2	2
REGION_RATING_CLIENT_W_CITY	257512	2.031676	0.502838	1	2	2	2
HOUR_APPR_PROCESS_START	257512	12.06553	3.268401	0	10	12	14
REG_REGION_NOT_LIVE_REGION	257512	0.015172	0.122237	0	0	0	0
REG_REGION_NOT_WORK_REGION	257512	0.050934	0.219862	0	0	0	0
LIVE_REGION_NOT_WORK_REGION	257512	0.040856	0.197958	0	0	0	0
REG_CITY_NOT_LIVE_CITY	257512	0.077825	0.267897	0	0	0	0
REG_CITY_NOT_WORK_CITY	257512	0.230121	0.420911	0	0	0	0
LIVE_CITY_NOT_WORK_CITY	257512	0.179526	0.383793	0	0	0	0
EXT_SOURCE_1	112306	0.502105	0.211072	0.014691	0.334085	0.505822	0.675
EXT_SOURCE_2	256978	0.514503	0.19104	1.32E-06	0.39262	0.565999	0.6636
EXT_SOURCE_3	206491	0.510653	0.194872	0.000527	0.37065	0.535276	0.6690
APARTMENTS_AVG	126836	0.117376	0.108196	0	0.0577	0.0876	0.148
BASEMENTAREA_AVG	106768	0.088344	0.082396	0	0.0442	0.0763	0.11
YEARS_BEGINEXPLUATATION_AVG	131899	0.977676	0.05974	0	0.9767	0.9816	0.986
YEARS_BUILD_AVG	86263	0.752633	0.113322	0	0.6872	0.7552	0.823
COMMONAREA_AVG	77607	0.044587	0.075449	0	0.0079	0.0211	0.051
ELEVATORS_AVG	120272	0.078993	0.134612	0	0	0	0.12
ENTRANCES_AVG	127879	0.149564	0.099891	0	0.069	0.1379	0.206
FLOORSMAX_AVG	129367	0.22644	0.144538	0	0.1667	0.1667	0.333
FLOORSMIN_AVG	82764	0.231941	0.161349	0	0.0833	0.2083	0.37
LANDAREA_AVG	104643	0.06633	0.081478	0	0.0186	0.0481	0.085
LIVINGAPARTMENTS_AVG	81539	0.10084	0.092536	0	0.0504	0.0756	0.12
LIVINGAREA_AVG	128299	0.107343	0.110516	0	0.0452	0.0744	0.129
NONLIVINGAPARTMENTS_AVG	78712	0.008753	0.047435	0	0	0	0.003
NONLIVINGAREA_AVG	115402	0.02837	0.069668	0	0	0.0036	0.027
APARTMENTS_MODE	126836	0.11419	0.107922	0	0.0525	0.084	0.14
BASEMENTAREA_MODE	106768	0.087452	0.084336	0	0.040675	0.0746	0.112

YEARS_BEGINEXPLUATATION_MODE	131899	0.977	0.065127	0	0.9767	0.9816	0.986
YEARS_BUILD_MODE	86263	0.759768	0.110172	0	0.6994	0.7648	0.823
COMMONAREA_MODE	77607	0.042532	0.073986	0	0.0073	0.0191	0.049
ELEVATORS_MODE	120272	0.074554	0.1323	0	0	0	0.120
ENTRANCES_MODE	127879	0.145036	0.10085	0	0.069	0.1379	0.206
FLOORSMAX_MODE	129367	0.222476	0.143597	0	0.1667	0.1667	0.333
FLOORSMIN_MODE	82764	0.228019	0.161068	0	0.0833	0.2083	0.375
LANDAREA_MODE	104643	0.06498	0.082014	0	0.0165	0.0458	0.084
LIVINGAPARTMENTS_MODE	81539	0.105707	0.0978	0	0.0542	0.0771	0.131
LIVINGAREA_MODE	128299	0.105949	0.111832	0	0.0425	0.0731	0.125
NONLIVINGAPARTMENTS_MODE	78712	0.008048	0.046218	0	0	0	0.003
NONLIVINGAREA_MODE	115402	0.027005	0.070283	0	0	0.0011	0.023
APARTMENTS_MEDI	126836	0.117781	0.109027	0	0.0583	0.0864	0.148
BASEMENTAREA_MEDI	106768	0.087844	0.082162	0	0.0437	0.0759	0.111
YEARS_BEGINEXPLUATATION_MEDI	131899	0.977698	0.060377	0	0.9767	0.9816	0.986
YEARS_BUILD_MEDI	86263	0.755901	0.112114	0	0.6914	0.7585	0.825
COMMONAREA_MEDI	77607	0.044573	0.07572	0	0.0079	0.0209	0.051
ELEVATORS_MEDI	120272	0.078124	0.134508	0	0	0	0.12
ENTRANCES_MEDI	127879	0.149053	0.10021	0	0.069	0.1379	0.206
FLOORSMAX_MEDI	129367	0.226055	0.144968	0	0.1667	0.1667	0.333
FLOORSMIN_MEDI	82764	0.231638	0.16191	0	0.0833	0.2083	0.375
LANDAREA_MEDI	104643	0.067181	0.082523	0	0.0186	0.0486	0.086
LIVINGAPARTMENTS_MEDI	81539	0.101997	0.093526	0	0.0513	0.0761	0.123
LIVINGAREA_MEDI	128299	0.108564	0.112259	0	0.0456	0.0749	0.130
NONLIVINGAPARTMENTS_MEDI	78712	0.008595	0.047187	0	0	0	0.003
NONLIVINGAREA_MEDI	115402	0.028223	0.070261	0	0	0.003	0.026
TOTALAREA_MODE	133229	0.102519	0.107368	0	0.0412	0.0688	0.127
OBS_30_CNT_SOCIAL_CIRCLE	256659	1.42253	2.419727	0	0	0	2
DEF_30_CNT_SOCIAL_CIRCLE	256659	0.143732	0.447885	0	0	0	0
OBS_60_CNT_SOCIAL_CIRCLE	256659	1.405608	2.398374	0	0	0	2
DEF_60_CNT_SOCIAL_CIRCLE	256659	0.100382	0.363259	0	0	0	0
DAYS_LAST_PHONE_CHANGE	257512	-962.58	826.289	-4292	-1569	-757	-275
FLAG_DOCUMENT_2	257512	4.27E-05	0.006536	0	0	0	0
FLAG_DOCUMENT_3	257512	0.70959	0.453952	0	0	1	1
FLAG_DOCUMENT_4	257512	6.21E-05	0.007882	0	0	0	0
FLAG_DOCUMENT_5	257512	0.015001	0.121558	0	0	0	0
FLAG_DOCUMENT_6	257512	0.088318	0.283758	0	0	0	0
FLAG_DOCUMENT_7	257512	0.000186	0.013652	0	0	0	0
FLAG_DOCUMENT_8	257512	0.081495	0.273595	0	0	0	0
FLAG_DOCUMENT_9	257512	0.003938	0.062627	0	0	0	0
FLAG_DOCUMENT_10	257512	2.33E-05	0.004827	0	0	0	0
FLAG_DOCUMENT_11	257512	0.003844	0.061885	0	0	0	0
FLAG_DOCUMENT_12	257512	7.77E-06	0.002787	0	0	0	0
FLAG_DOCUMENT_13	257512	0.003584	0.059762	0	0	0	0
FLAG_DOCUMENT_14	257512	0.002893	0.05371	0	0	0	0

FLAG_DOCUMENT_15	257512	0.001285	0.035829	0	0	0	0
FLAG_DOCUMENT_16	257512	0.00991	0.099056	0	0	0	0
FLAG_DOCUMENT_17	257512	0.00026	0.016128	0	0	0	0
FLAG_DOCUMENT_18	257512	0.008058	0.089403	0	0	0	0
FLAG_DOCUMENT_19	257512	0.000575	0.023967	0	0	0	0
FLAG_DOCUMENT_20	257512	0.000505	0.022463	0	0	0	0
FLAG_DOCUMENT_21	257512	0.000326	0.018058	0	0	0	0
AMT_REQ_CREDIT_BUREAU_HOUR	222727	0.006268	0.083078	0	0	0	0
AMT_REQ_CREDIT_BUREAU_DAY	222727	0.006901	0.111287	0	0	0	0
AMT_REQ_CREDIT_BUREAU_WEEK	222727	0.034747	0.20668	0	0	0	0
AMT_REQ_CREDIT_BUREAU_MON	222727	0.266833	0.913544	0	0	0	0
AMT_REQ_CREDIT_BUREAU_QRT	222727	0.266348	0.825488	0	0	0	0
AMT_REQ_CREDIT_BUREAU_YEAR	222727	1.903653	1.8701	0	0	1	3

3.Null %

Column Name	Null %
COMMONAREA_AVG	81%
COMMONAREA_MODE	81%
COMMONAREA_MEDI	81%
NONLIVINGAPARTMENTS_AVG	80%
NONLIVINGAPARTMENTS_MODE	80%
NONLIVINGAPARTMENTS_MEDI	80%
FONDKAPREMONT_MODE	79%
LIVINGAPARTMENTS_AVG	79%
LIVINGAPARTMENTS_MODE	79%
LIVINGAPARTMENTS_MEDI	79%
FLOORSMIN_AVG	78%
FLOORSMIN_MODE	78%
FLOORSMIN_MEDI	78%
YEARS_BUILD_AVG	77%
YEARS_BUILD_MODE	77%
YEARS_BUILD_MEDI	77%
OWN_CAR_AGE	76%
LANDAREA_AVG	69%
LANDAREA_MODE	69%
LANDAREA_MEDI	69%
BASEMENTAREA_AVG	68%
BASEMENTAREA_MODE	68%
BASEMENTAREA_MEDI	68%
EXT_SOURCE_1	65%
NONLIVINGAREA_AVG	64%
NONLIVINGAREA_MODE	64%

NONLIVINGAREA_MEDI	64%
ELEVATORS_AVG	62%
ELEVATORS_MODE	62%
ELEVATORS_MEDI	62%
WALLSMATERIAL_MODE	59%
APARTMENTS_AVG	59%
APARTMENTS_MODE	59%
APARTMENTS_MEDI	59%
ENTRANCES_AVG	58%
ENTRANCES_MODE	58%
ENTRANCES_MEDI	58%
HOUSETYPE_MODE	58%
LIVINGAREA_AVG	58%
LIVINGAREA_MODE	58%
LIVINGAREA_MEDI	58%
FLOORSMAX_AVG	58%
FLOORSMAX_MODE	58%
FLOORSMAX_MEDI	58%
YEARS_BEGINEXPLUATATION_AVG	56%
YEARS_BEGINEXPLUATATION_MODE	56%
YEARS_BEGINEXPLUATATION_MEDI	56%
TOTALAREA_MODE	56%
EMERGENCYSTATE_MODE	55%
OCCUPATION_TYPE	36%
EXT_SOURCE_3	23%
AMT_REQ_CREDIT_BUREAU_HOUR	16%
AMT_REQ_CREDIT_BUREAU_DAY	16%
AMT_REQ_CREDIT_BUREAU_WEEK	16%
AMT_REQ_CREDIT_BUREAU_MON	16%
AMT_REQ_CREDIT_BUREAU_QRT	16%
AMT_REQ_CREDIT_BUREAU_YEAR	16%
NAME_TYPE_SUITE	0%
OBS_30_CNT_SOCIAL_CIRCLE	0%
DEF_30_CNT_SOCIAL_CIRCLE	0%
OBS_60_CNT_SOCIAL_CIRCLE	0%
DEF_60_CNT_SOCIAL_CIRCLE	0%
EXT_SOURCE_2	0%
AMT_GOODS_PRICE	0%
AMT_ANNUITY	0%
CNT_FAM_MEMBERS	0%
SK_ID_CURR	0%
TARGET	0%
NAME_CONTRACT_TYPE	0%
CODE_GENDER	0%

FLAG_OWN_CAR	0%
FLAG_OWN_REALTY	0%
CNT_CHILDREN	0%
AMT_INCOME_TOTAL	0%
AMT_CREDIT	0%
NAME_INCOME_TYPE	0%
NAME_EDUCATION_TYPE	0%
NAME_FAMILY_STATUS	0%
NAME_HOUSING_TYPE	0%
REGION_POPULATION_RELATIVE	0%
DAYS_BIRTH	0%
DAYS_EMPLOYED	0%
DAYS_REGISTRATION	0%
DAYS_ID_PUBLISH	0%
FLAG_MOBIL	0%
FLAG_EMP_PHONE	0%
FLAG_WORK_PHONE	0%
FLAG_CONT_MOBILE	0%
FLAG_PHONE	0%
FLAG_EMAIL	0%
REGION_RATING_CLIENT	0%
REGION_RATING_CLIENT_W_CITY	0%
WEEKDAY_APPR_PROCESS_START	0%
HOUR_APPR_PROCESS_START	0%
REG_REGION_NOT_LIVE_REGION	0%
REG_REGION_NOT_WORK_REGION	0%
LIVE_REGION_NOT_WORK_REGION	0%
REG_CITY_NOT_LIVE_CITY	0%
REG_CITY_NOT_WORK_CITY	0%
LIVE_CITY_NOT_WORK_CITY	0%
ORGANIZATION_TYPE	0%
DAYS_LAST_PHONE_CHANGE	0%
FLAG_DOCUMENT_2	0%
FLAG_DOCUMENT_3	0%
FLAG_DOCUMENT_4	0%
FLAG_DOCUMENT_5	0%
FLAG_DOCUMENT_6	0%
FLAG_DOCUMENT_7	0%
FLAG_DOCUMENT_8	0%
FLAG_DOCUMENT_9	0%
FLAG_DOCUMENT_10	0%
FLAG_DOCUMENT_11	0%
FLAG_DOCUMENT_12	0%
FLAG_DOCUMENT_13	0%

FLAG_DOCUMENT_14	0%
FLAG_DOCUMENT_15	0%
FLAG_DOCUMENT_16	0%
FLAG_DOCUMENT_17	0%
FLAG_DOCUMENT_18	0%
FLAG_DOCUMENT_19	0%
FLAG_DOCUMENT_20	0%
FLAG_DOCUMENT_21	0%

4.DataTypes

column	DataType
SK_ID_CURR	int64
TARGET	int64
NAME_CONTRACT_TYPE	object
CODE_GENDER	object
FLAG_OWN_CAR	object
FLAG_OWN_REALTY	object
CNT_CHILDREN	int64
AMT_INCOME_TOTAL	float64
AMT_CREDIT	float64
AMT_ANNUITY	float64
AMT_GOODS_PRICE	float64
NAME_TYPE_SUITE	object
NAME_INCOME_TYPE	object
NAME_EDUCATION_TYPE	object
NAME_FAMILY_STATUS	object
NAME_HOUSING_TYPE	object
REGION_POPULATION_RELATIVE	float64
DAYS_BIRTH	int64
DAYS_EMPLOYED	int64
DAYS_REGISTRATION	float64
DAYS_ID_PUBLISH	int64
OWN_CAR_AGE	float64
FLAG_MOBIL	int64
FLAG_EMP_PHONE	int64
FLAG_WORK_PHONE	int64
FLAG_CONT_MOBILE	int64
FLAG_PHONE	int64
FLAG_EMAIL	int64
OCCUPATION_TYPE	object
CNT_FAM_MEMBERS	float64
REGION_RATING_CLIENT	int64
REGION_RATING_CLIENT_W_CITY	int64

WEEKDAY_APPR_PROCESS_START	object
HOUR_APPR_PROCESS_START	int64
REG_REGION_NOT_LIVE_REGION	int64
REG_REGION_NOT_WORK_REGION	int64
LIVE_REGION_NOT_WORK_REGION	int64
REG_CITY_NOT_LIVE_CITY	int64
REG_CITY_NOT_WORK_CITY	int64
LIVE_CITY_NOT_WORK_CITY	int64
ORGANIZATION_TYPE	object
EXT_SOURCE_1	float64
EXT_SOURCE_2	float64
EXT_SOURCE_3	float64
APARTMENTS_AVG	float64
BASEMENTAREA_AVG	float64
YEARS_BEGINEXPLUATATION_AVG	float64
YEARS_BUILD_AVG	float64
COMMONAREA_AVG	float64
ELEVATORS_AVG	float64
ENTRANCES_AVG	float64
FLOORSMAX_AVG	float64
FLOORSMIN_AVG	float64
LANDAREA_AVG	float64
LIVINGAPARTMENTS_AVG	float64
LIVINGAREA_AVG	float64
NONLIVINGAPARTMENTS_AVG	float64
NONLIVINGAREA_AVG	float64
APARTMENTS_MODE	float64
BASEMENTAREA_MODE	float64
YEARS_BEGINEXPLUATATION_MODE	float64
YEARS_BUILD_MODE	float64
COMMONAREA_MODE	float64
ELEVATORS_MODE	float64
ENTRANCES_MODE	float64
FLOORSMAX_MODE	float64
FLOORSMIN_MODE	float64
LANDAREA_MODE	float64
LIVINGAPARTMENTS_MODE	float64
LIVINGAREA_MODE	float64
NONLIVINGAPARTMENTS_MODE	float64
NONLIVINGAREA_MODE	float64
APARTMENTS_MEDI	float64
BASEMENTAREA_MEDI	float64
YEARS_BEGINEXPLUATATION_MEDI	float64
YEARS_BUILD_MEDI	float64
COMMONAREA_MEDI	float64

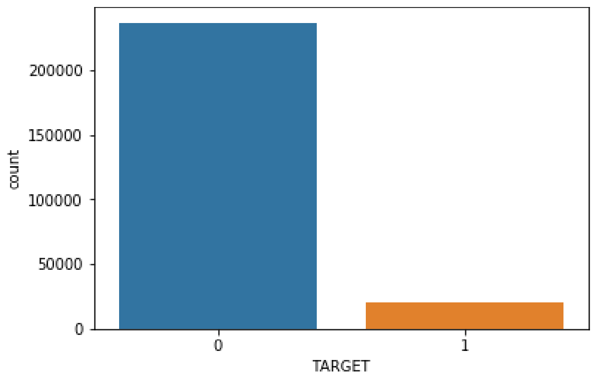
ELEVATORS_MEDI	float64
ENTRANCES_MEDI	float64
FLOORSMAX_MEDI	float64
FLOORSMIN_MEDI	float64
LANDAREA_MEDI	float64
LIVINGAPARTMENTS_MEDI	float64
LIVINGAREA_MEDI	float64
NONLIVINGAPARTMENTS_MEDI	float64
NONLIVINGAREA_MEDI	float64
FONDKAPREMONT_MODE	object
HOUSETYPE_MODE	object
TOTALAREA_MODE	float64
WALLSMATERIAL_MODE	object
EMERGENCYSTATE_MODE	object
OBS_30_CNT_SOCIAL_CIRCLE	float64
DEF_30_CNT_SOCIAL_CIRCLE	float64
OBS_60_CNT_SOCIAL_CIRCLE	float64
DEF_60_CNT_SOCIAL_CIRCLE	float64
DAYS_LAST_PHONE_CHANGE	int64
FLAG_DOCUMENT_2	int64
FLAG_DOCUMENT_3	int64
FLAG_DOCUMENT_4	int64
FLAG_DOCUMENT_5	int64
FLAG_DOCUMENT_6	int64
FLAG_DOCUMENT_7	int64
FLAG_DOCUMENT_8	int64
FLAG_DOCUMENT_9	int64
FLAG_DOCUMENT_10	int64
FLAG_DOCUMENT_11	int64
FLAG_DOCUMENT_12	int64
FLAG_DOCUMENT_13	int64
FLAG_DOCUMENT_14	int64
FLAG_DOCUMENT_15	int64
FLAG_DOCUMENT_16	int64
FLAG_DOCUMENT_17	int64
FLAG_DOCUMENT_18	int64
FLAG_DOCUMENT_19	int64
FLAG_DOCUMENT_20	int64
FLAG_DOCUMENT_21	int64
AMT_REQ_CREDIT_BUREAU_HOUR	float64
AMT_REQ_CREDIT_BUREAU_DAY	float64
AMT_REQ_CREDIT_BUREAU_WEEK	float64
AMT_REQ_CREDIT_BUREAU_MON	float64
AMT_REQ_CREDIT_BUREAU_QRT	float64
AMT_REQ_CREDIT_BUREAU_YEAR	float64

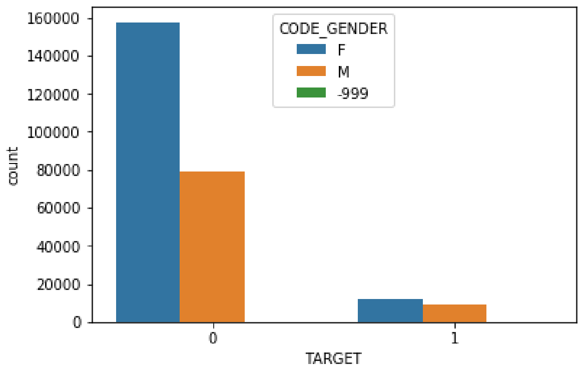
5.Visualization

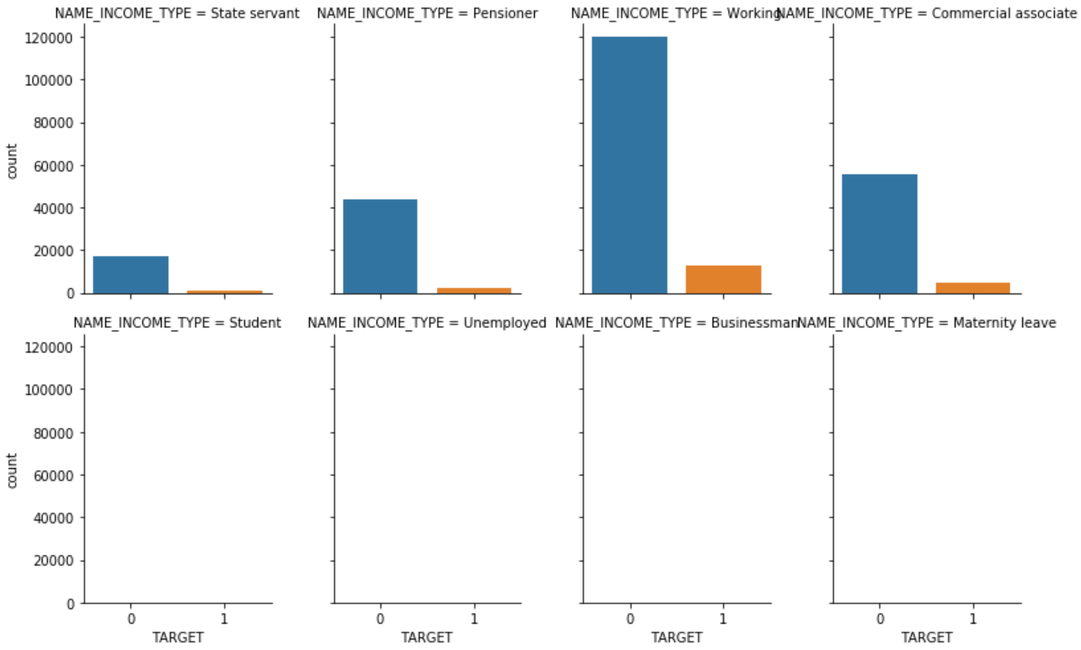
1.Class imbalance cab be seen from below plot.

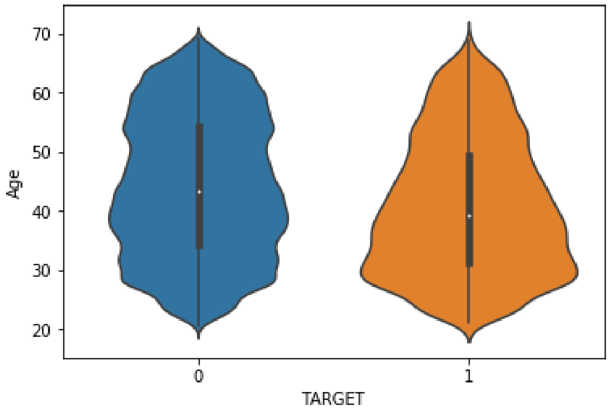
2.Female count is more than male in dataset ,and also as defaulter is higher.

3.4 major income group type are evident from below plot state servant, pensioner ,working and commercial associates. Working group is major contributor to default in loan.





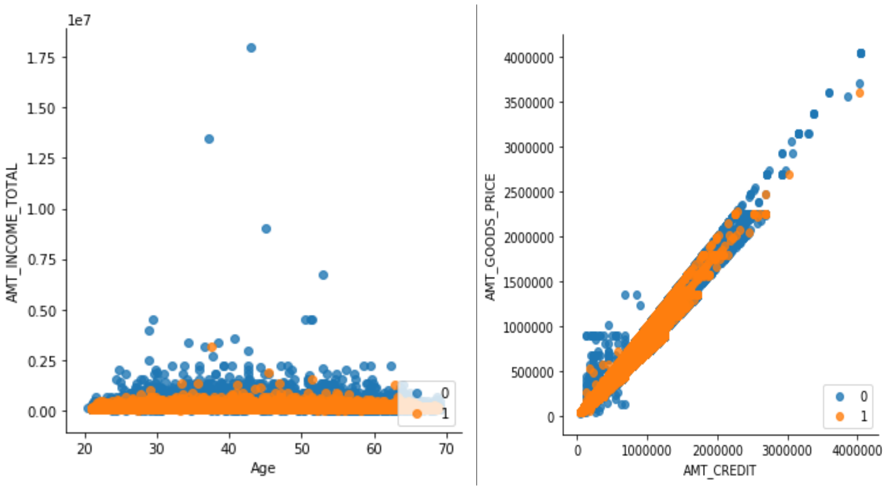




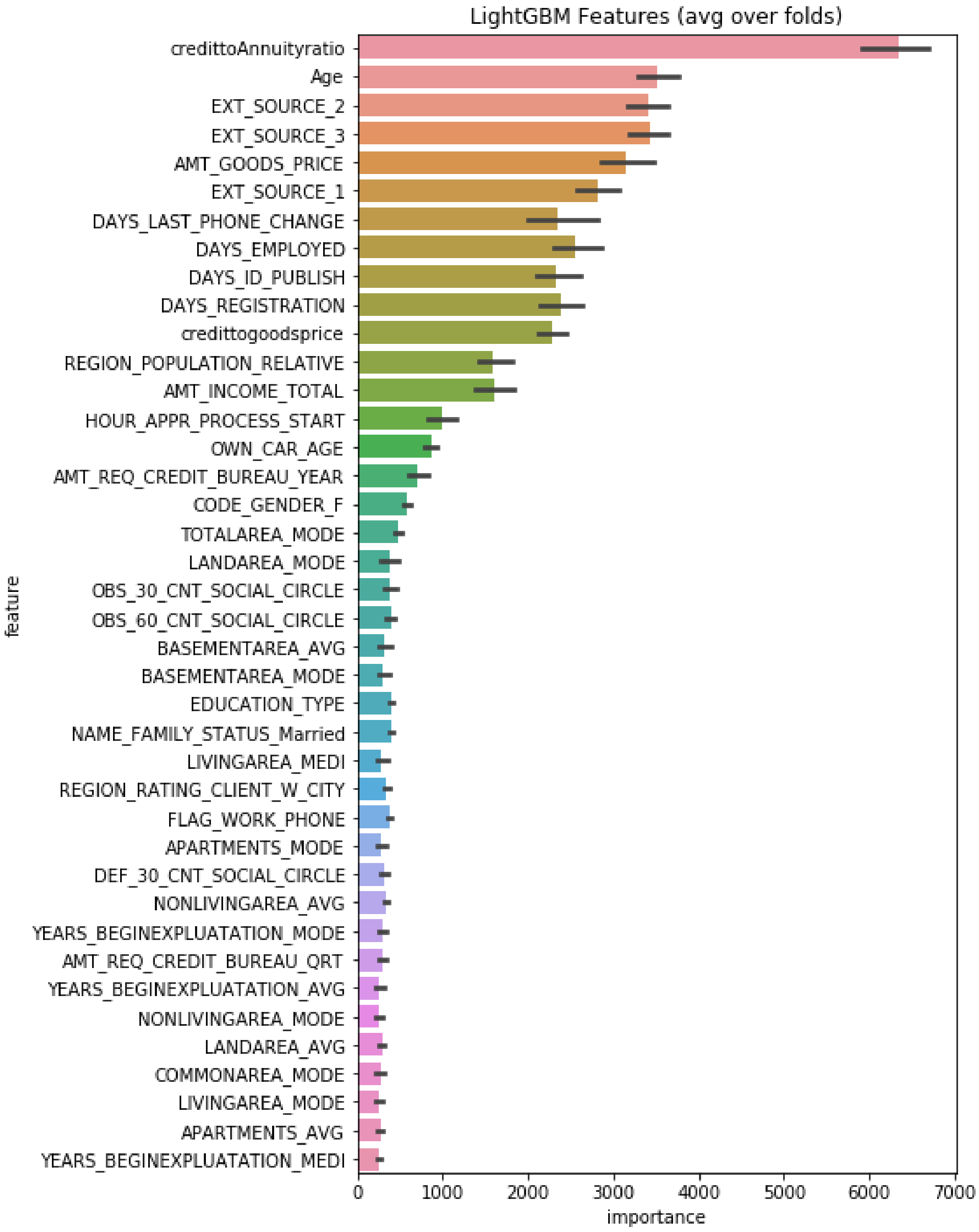
5.Age range is between 20-70 and majority of defaulters fall around 30 Years age group.

6.lower income group(<.25) default more .

7.Amount of credit and amout of good price follows a strong linear trend .Most of defaulters are present below >3000000 credit amount.



6.Feature importances:



6.Correlation:

High correlation can be seen for average and mode ,median columns related to apartment size .