QFRM - Exploratory Data Analysis (supplement)

Stavros Ieronymakis - 2645715 - s.ieronymakis2@student.vu.nl

May 2024

1 Introduction

In this report we will present the univariate and bivariate analysis of the dataset, in order to provide the reader with more information on feature selection for our models. As exlained in the original report, the size of the features is too big to process in one attempt. Therefore, the solution to group common features together and process them as a group was selscted. Below we display all the initial features by group (Table 1).

2 Exploratory Data Analysis

2.1 Target Variable Analysis

The target variable *TARGET* is a binary variable indicating loan default. The distribution of the target variable shows that the majority of the cases are non-defaults. This class imbalance suggests that performance metrics such as precision, recall, and the ROC-AUC score will be more informative than accuracy alone.

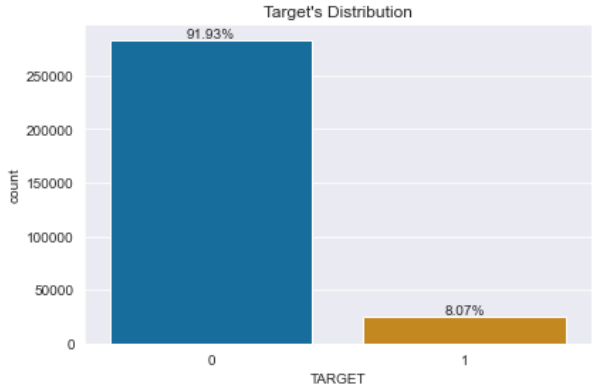


Figure 1: Distribution of Target Variable

The dataset consists of 120 features which are categorized into continuous (57), discrete (47), and categorical (16) features.

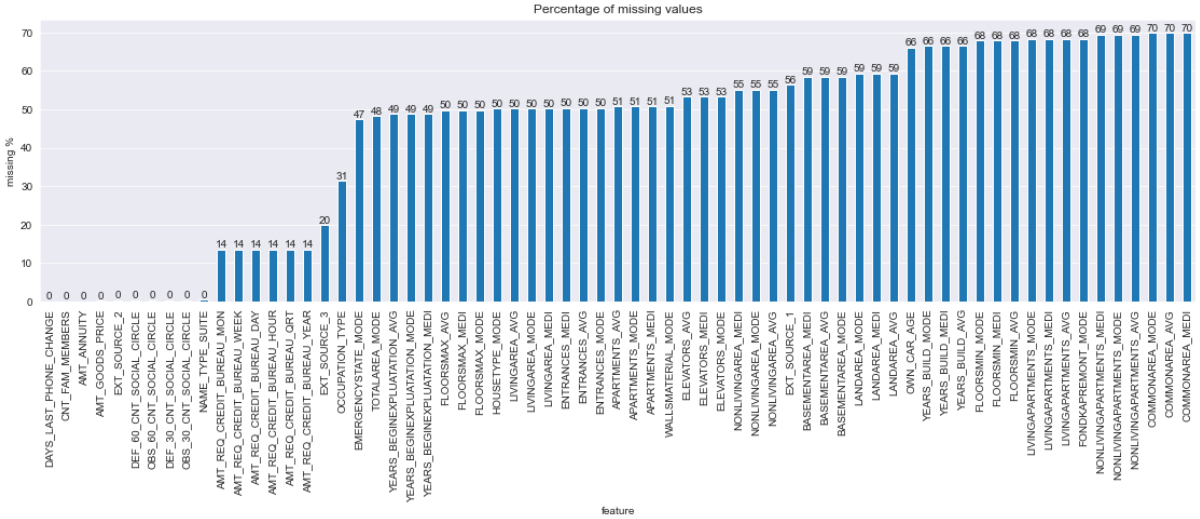


Figure 2: Percentage of Missing Values by Feature

Features missing most values are mainly those related to the building, external sources and recent inquiries. In fact, there is no building feature with less than 47% missing values. We shall ignore the building features. Our analysis continues for the rest of the groups/features.

Table 1: Feature Groups

Group	Features
Demographics	NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_TYPE_SUITE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, OCCUPATION_TYPE, ORGANIZATION_TYPE
Family Count	CNT_CHILDREN, CNT_FAM_MEMBERS
Age / Duration	DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE, OWN_CAR_AGE
Social Circle	OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE
Contact Information	FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL
Address Discrepancy	REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY
Region's Data	REGION_POPULATION_RELATIVE, REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY
Process Start Time	HOUR_APPR_PROCESS_START, WEEKDAY_APPR_PROCESS_START
External Source Scores	EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3
Amounts	AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE
Recent Inquiries	AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR
Provided Documents	FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6, FLAG_DOCUMENT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9, FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, FLAG_DOCUMENT_12, FLAG_DOCUMENT_13, FLAG_DOCUMENT_14, FLAG_DOCUMENT_15, FLAG_DOCUMENT_16, FLAG_DOCUMENT_17, FLAG_DOCUMENT_18, FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_21
Building's Data	APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE, LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI, FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE

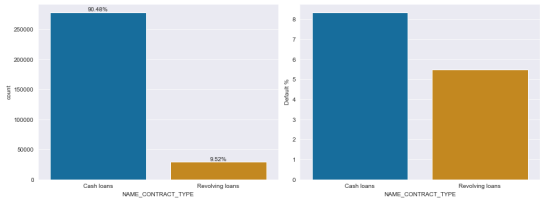
For each feature type (continuous, discrete/binary, and categorical), a summary function (`cont_summary()`, `disc_summary()`, and `cat_summary()`) is defined. This function returns a customized description of the feature, including the number of missing values, the number of unique values, correlation with the

target, among other properties. Furthermore, some plot functions (`cont_plots()`, `disc_plots()`, and `cat_plots()`) are crafted to visualize the feature distributions for different target values. For a feature with missing values, this plot depicts the credit default rate in the instances where the feature value is missing and contrasts it with the credit default rate in the instances not missing the feature value. It is meant to examine whether creating a binary 'missing flag' feature based on the feature is warranted. The binary 'missing flag' feature takes value 1 where the feature value is missing, and 0 otherwise. For each feature group, the function `corr_heatmap()` demonstrates the correlation between the features within that group.

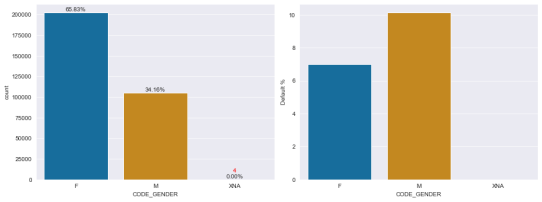
In other words, we will be following the following routine for each group. Provide detailed descriptions and visualizations for all the features in the group. Explore the correlations between the features within the group. Summarize our decisions with respect to the feature group.

2.2 Feature Group: Demographics

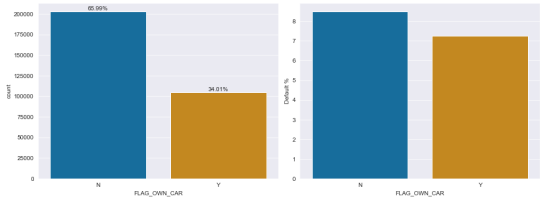
Although the correlation heat map is not very informative, some output can still be extracted from the rest of the plots. In the analysis of the demographic features, several adjustments are recommended to address data quality and categorization issues. The value 'XNA' in 'CODE_GENDER' is practically negligible, appearing only in 4 instances, and can be replaced with either 'M' or 'F'. The 'NAME_TYPE_SUITE' feature has 0.42% missing values, which can be handled by creating a new category for missing entries. 'OCCUPATION_TYPE' has a significant amount of missing data (31.35%), which also justifies forming a new category for these missing values. 'NAME_INCOME_TYPE' contains several extremely rare categories that can be combined into a single category to avoid sparsity. Similarly, the 'Unknown' category in 'NAME_FAMILY_STATUS' is rare and can be merged with 'Married' for simplicity. Lastly, the features 'OCCUPATION_TYPE' and 'ORGANIZATION_TYPE' possess numerous categories, potentially causing issues for some slower classifiers and increasing the risk of overfitting. Therefore, it is advisable to streamline these categories to enhance model performance.



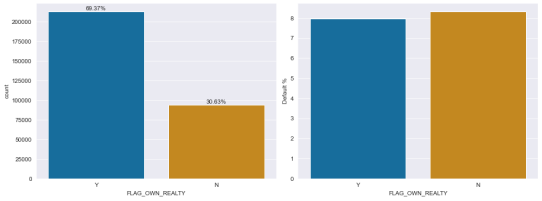
(a) NAME_CONTRACT_TYPE



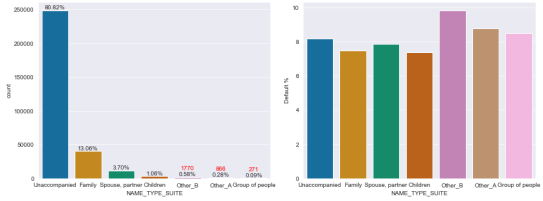
(b) CODE_GENDER



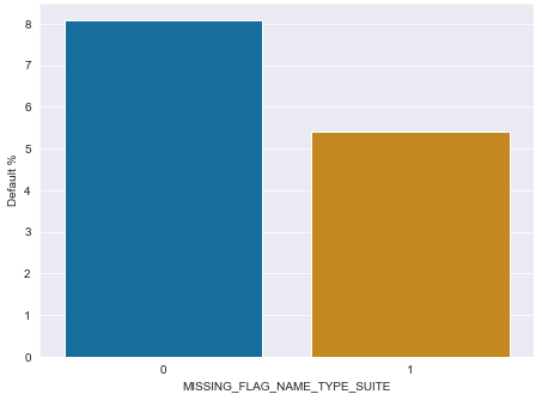
(c) FLAG_OWN_CAR



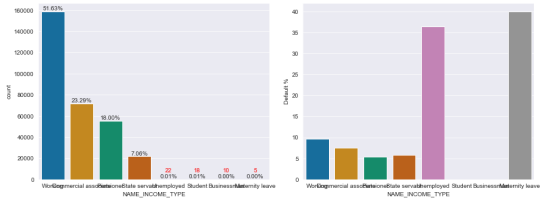
(d) FLAG_OWN_REALTY



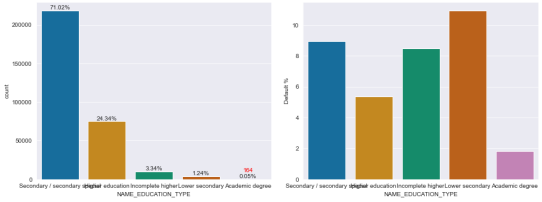
(e) NAME_TYPE_SUITE



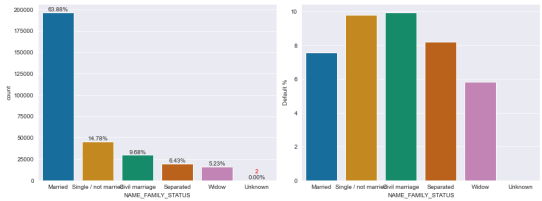
(f) NAME_INCOME_TYPE



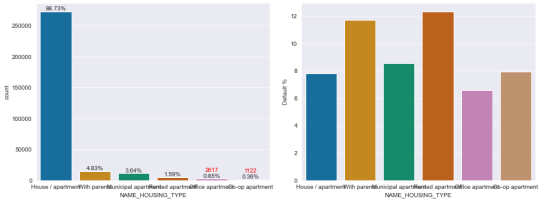
(g) NAME_EDUCATION_TYPE



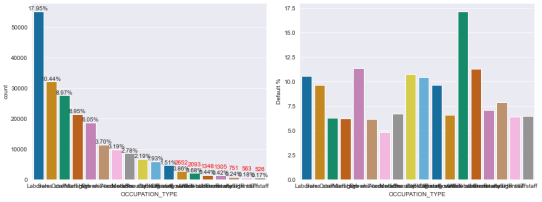
(h) NAME_FAMILY_STATUS



(i) NAME_HOUSING_TYPE



(j) OCCUPATION_TYPE



(k) ORGANIZATION_TYPE

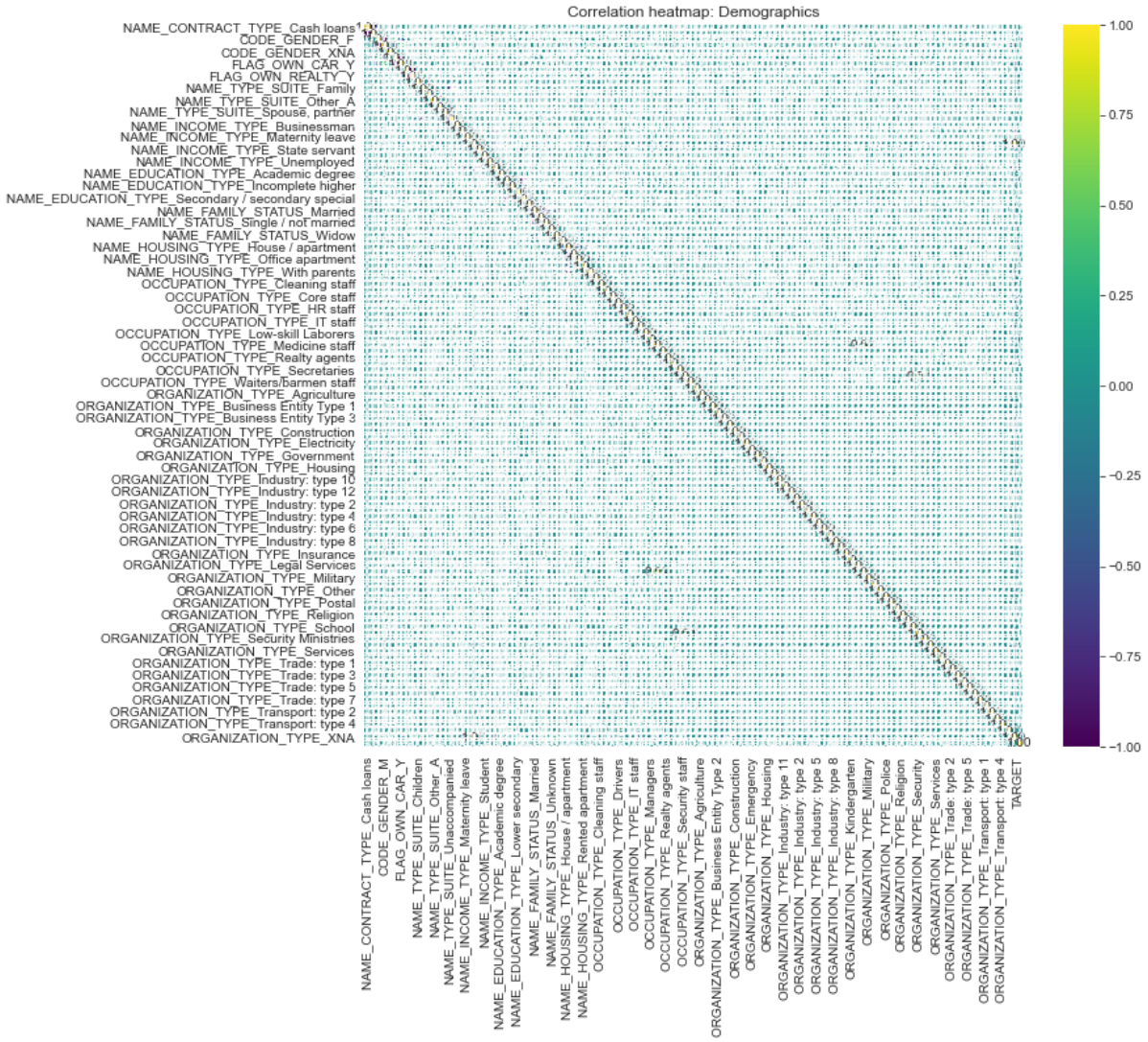
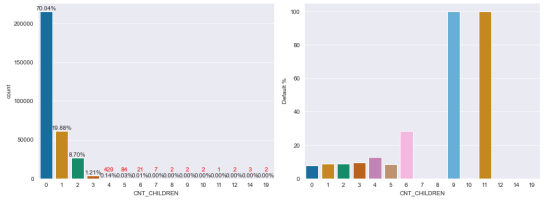


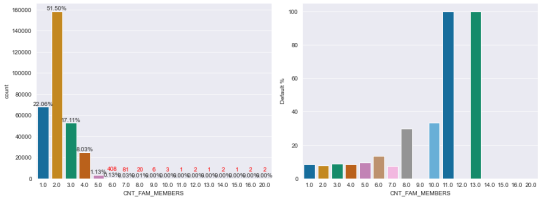
Figure 4: Correlation Heatmap for Demographics Features

2.3 Feature Group: Family Count

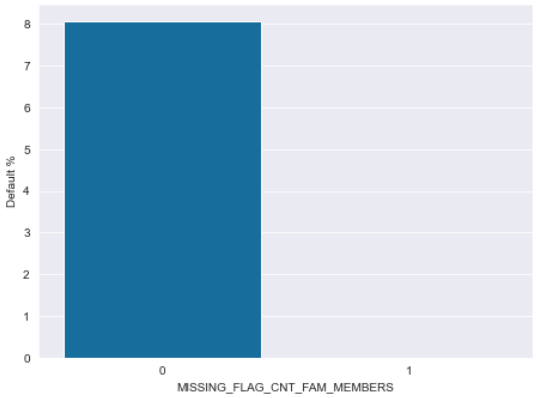
In the analysis of family count features, several steps are recommended to improve data quality and model performance. 'CNT_CHILDREN' and 'CNT_FAM_MEMBERS' are found to be highly correlated. Given that 'CNT_CHILDREN' has no missing values and a higher correlation with the target variable, it is advisable to drop 'CNT_FAM_MEMBERS'. If 'CNT_FAM_MEMBERS' is retained, its few missing values can be imputed with 2.0. Furthermore, values of 'CNT_CHILDREN' exceeding 4 can be capped at 4, and values of 'CNT_FAM_MEMBERS' exceeding 6 can be capped at 6. This winsorization process helps in managing outliers and stabilizing the data.



(a) CNT_CHILDREN



(b) CNT_FAM_MEMBERS



(c) MISSING_FLAG_CNT_FAM_MEMBERS

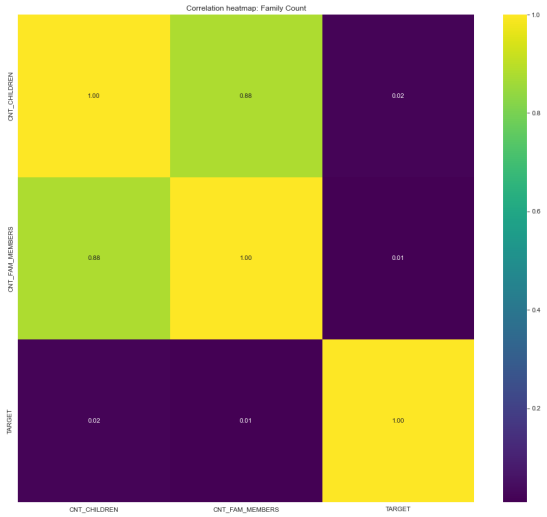


Figure 6: Correlation Heatmap for Family Count

2.4 Feature Group: Age / Duration

In the analysis of the age and duration-related features, several key adjustments are proposed to address data inconsistencies and improve the predictive power of the model. Firstly, `DAYS_EMPLOYED` requires further investigation due to the presence of positive values, which need to be understood. It was found that `DAYS_EMPLOYED` can only have one positive value, 365243, which appears to indicate unemployment. Therefore, it is reasonable to replace this value with 0. Additionally, `OWN_CAR_AGE` has a significant proportion of missing values (65.99%), likely indicating that the applicant does not own a car. These missing values can be treated as a separate category. `DAYS_LAST_PHONE_CHANGE` has a single missing value, which can be imputed with 0.0 (the mode). `DAYS_BIRTH` and `DAYS_EMPLOYED` might be highly correlated after adjusting for the positive values in `DAYS_EMPLOYED`, but both features are expected to be strong predictors of default risk and thus should not be dropped. This thorough examination ensures that the features are appropriately handled to maintain data integrity and improve model performance.

DAYS_EMPLOYED	OCCUPATION_TYPE	Count
365243	NaN	55372
	Cleaning staff	2

Table 2: Count of DAYS_EMPLOYED and OCCUPATION_TYPE

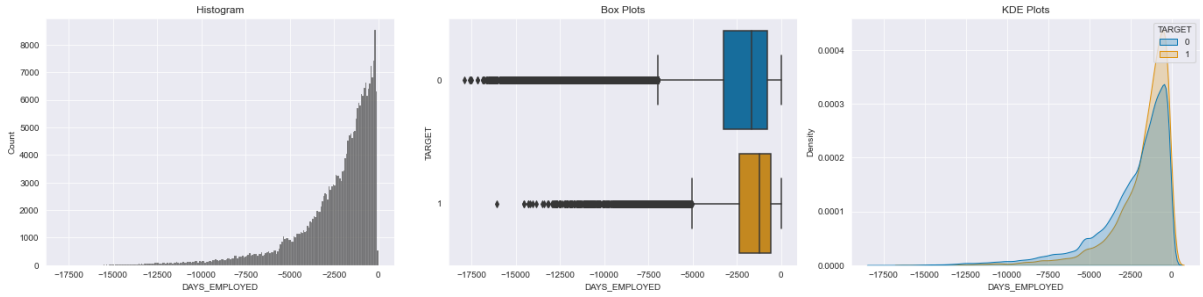


Figure 7: Further investigation of DAYS_EMPLOYED

Feature	dtype	count	unique	top_value_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
DAYS_EMPLOYED	int64	252137	12573	{-200.0: 156, -224.0: 152, -199.0: 151}	0	0.0	-2384.17	2338.36	-17912.0	-1648.0	0.0	0.07

Table 3: Summary statistics for DAYS_EMPLOYED

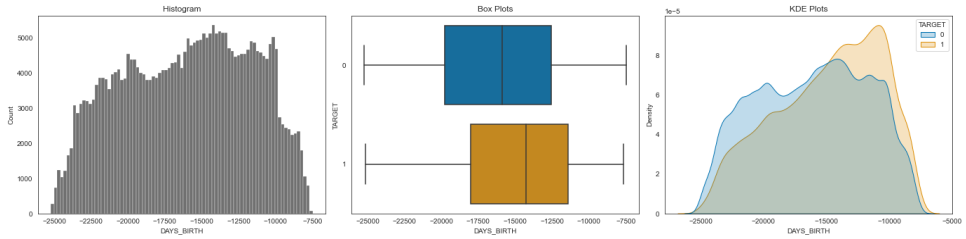


Figure 8: DAYS_BIRTH

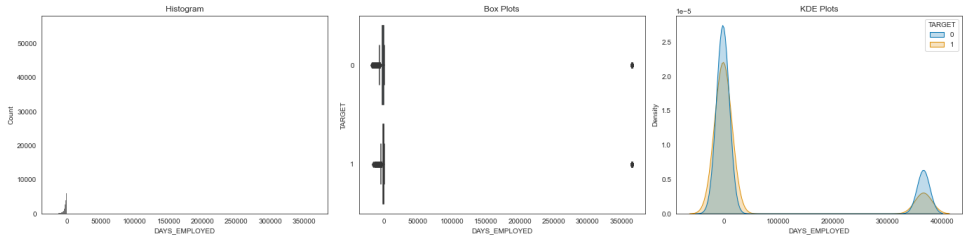


Figure 9: DAYS_EMPLOYED

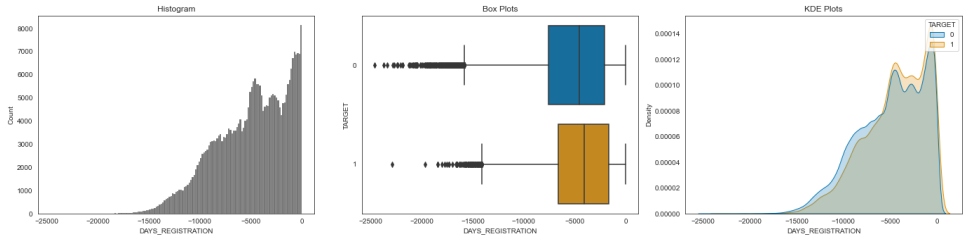


Figure 10: DAYS_REGISTRATION

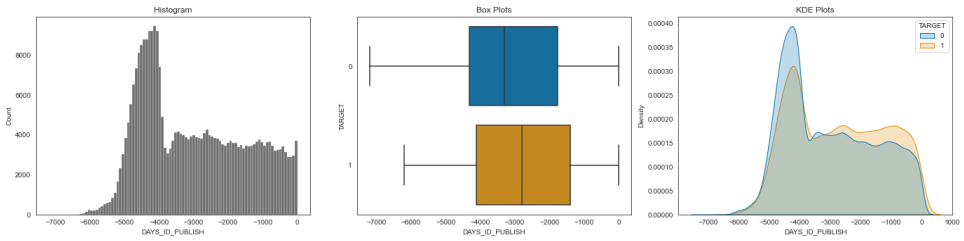


Figure 11: DAYS_ID_PUBLISH

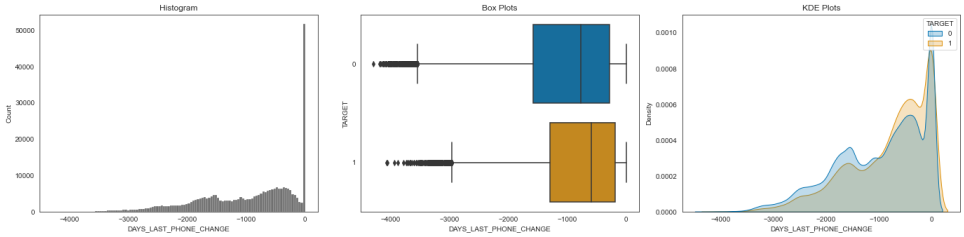


Figure 12: DAYS_LAST_PHONE_CHANGE

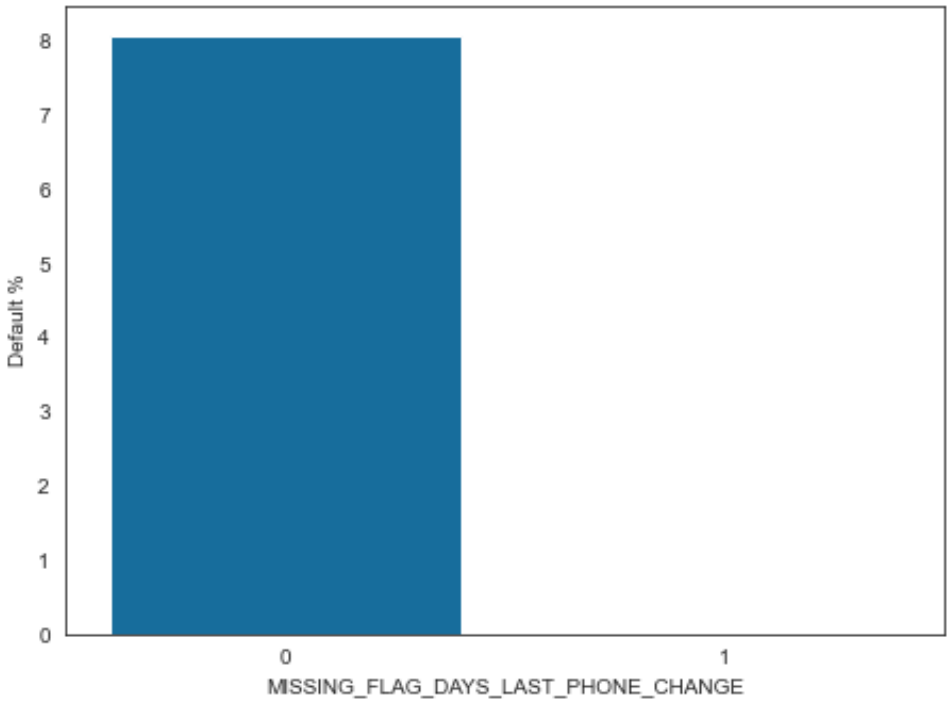


Figure 13: MISSING_FLAG_DAYS_LAST_PHONE_CHANGE

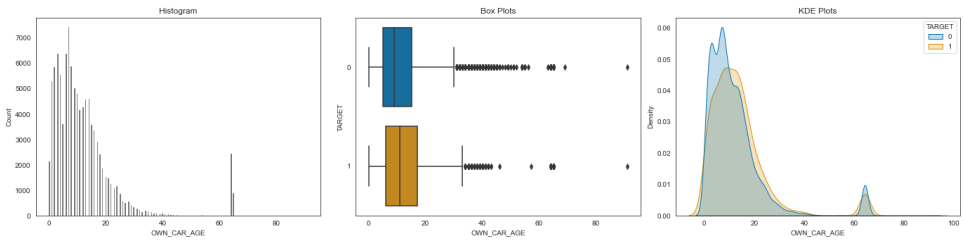


Figure 14: OWN_CAR_AGE

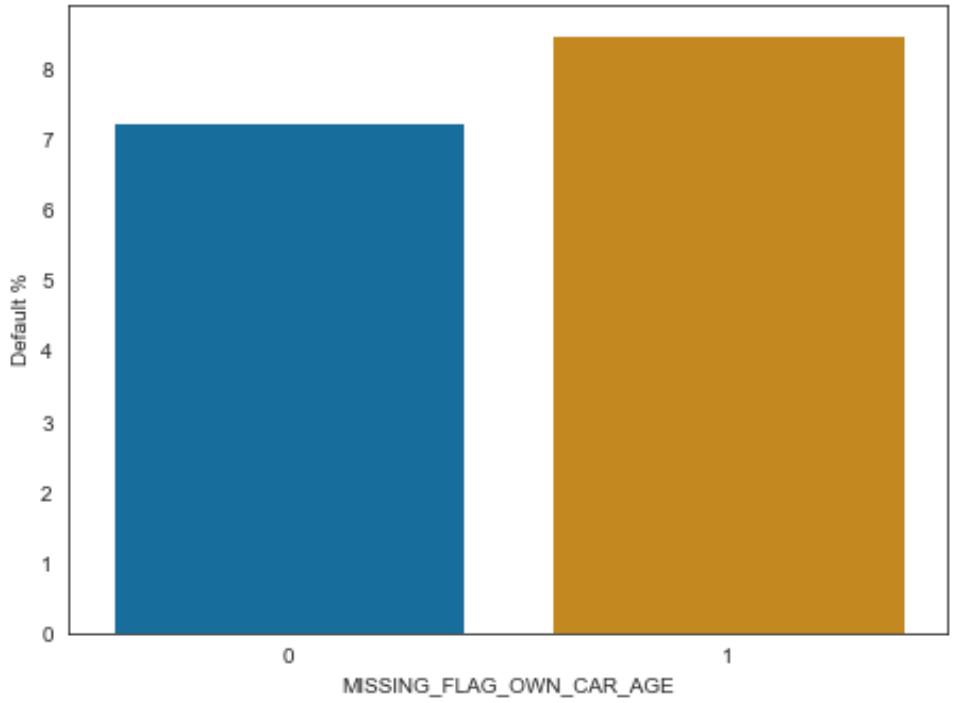
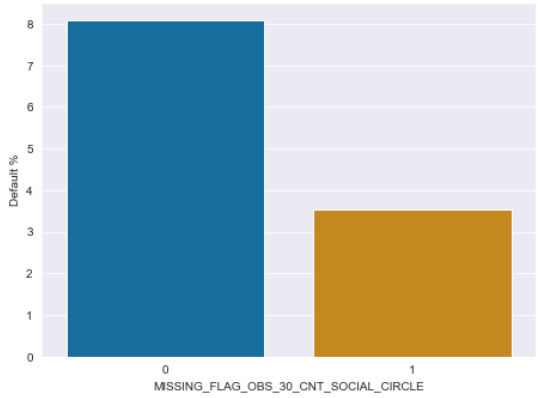


Figure 15: MISSING_FLAG_OWN_CAR_AGE

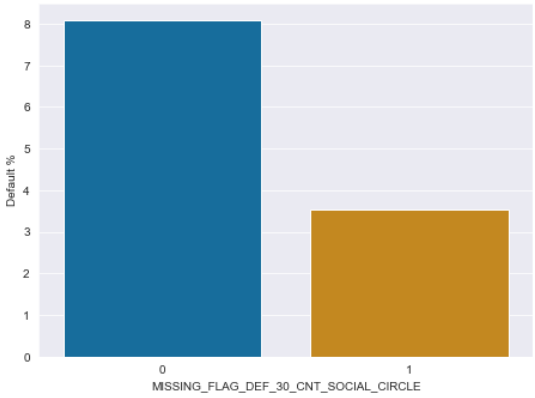
Figure 16: Correlation Heatmap for Age / Duration

3 Feature Group: Social Circle

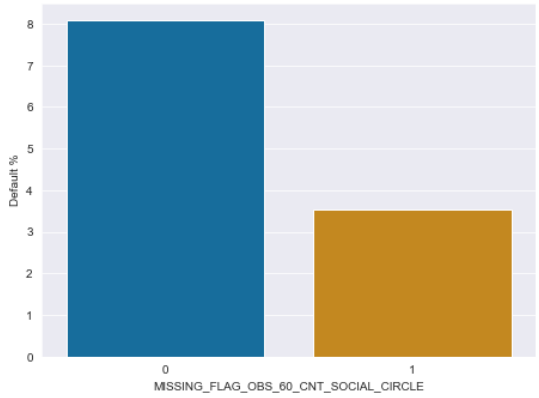
All the features within this group exhibit some missing values, specifically 0.33% of the data. These missing values can be appropriately handled by replacing them with zero. Additionally, the features `OBS_30_CNT_SOCIAL_CIRCLE` and `DEF_30_CNT_SOCIAL_CIRCLE` demonstrate perfect correlation, indicating redundancy. Therefore, it is advisable to drop one of these features to streamline the dataset without losing significant information.



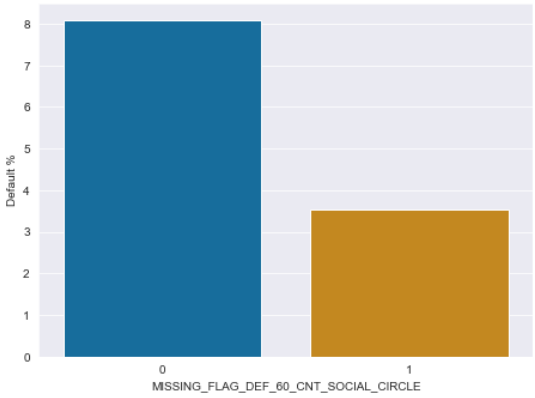
(a) MISSING_FLAG_OBS_30_CNT_SOCIAL_CIRCLE



(b) MISSING_FLAG_DEF_30_CNT_SOCIAL_CIRCLE

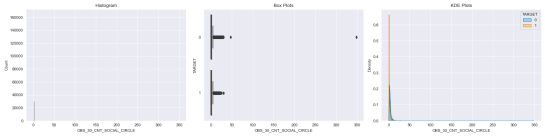


(c) MISSING_FLAG_OBS_60_CNT_SOCIAL_CIRCLE

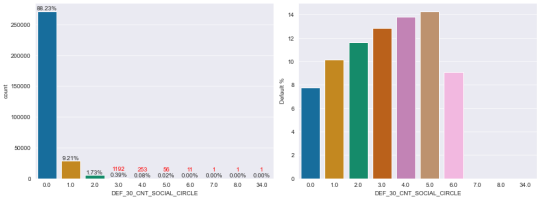


(d) MISSING_FLAG_DEF_60_CNT_SOCIAL_CIRCLE

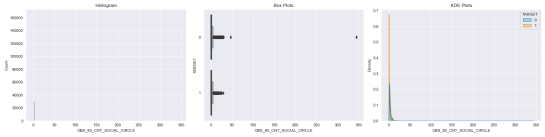
Figure 17: Distribution plots for grouped features



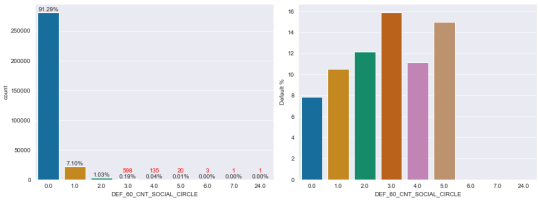
(a) OBS_30_CNT_SOCIAL_CIRCLE



(b) DEF_30_CNT_SOCIAL_CIRCLE



(c) OBS_60_CNT_SOCIAL_CIRCLE



(d) DEF_60_CNT_SOCIAL_CIRCLE

Figure 18: Distribution plots for grouped features

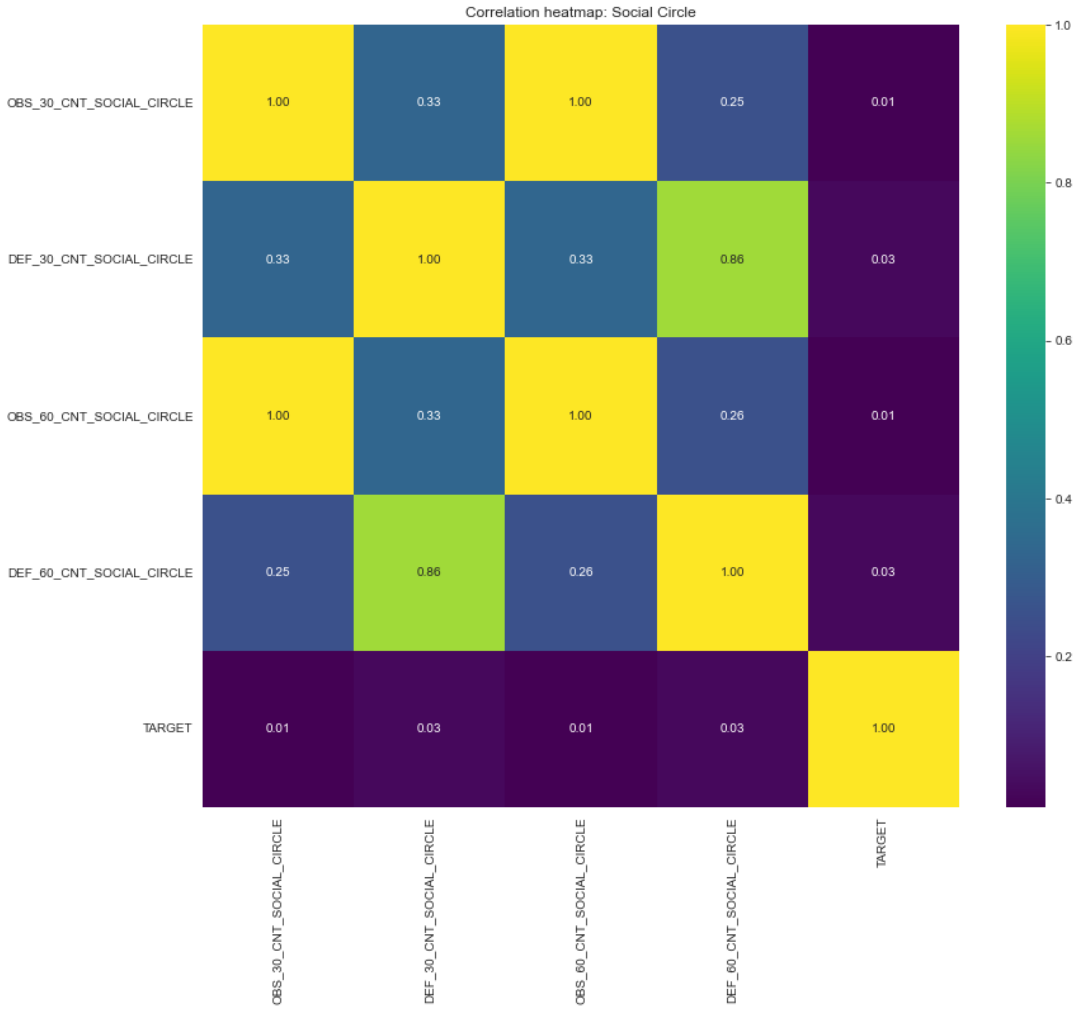
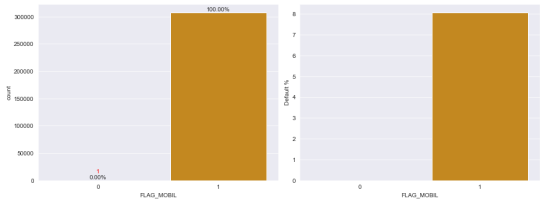


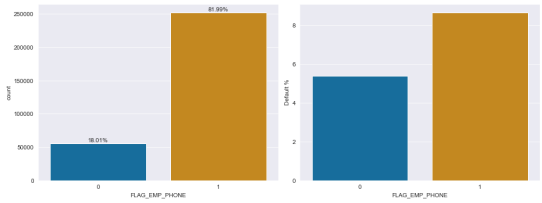
Figure 19: Heatmap Figure for Social Circle

4 Feature Group: Contact Info

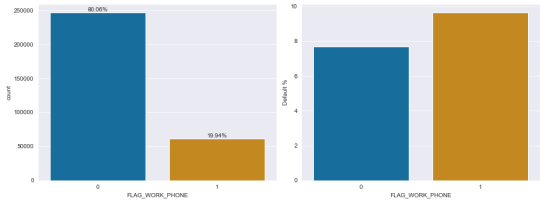
The features `FLAG_MOBIL` and `FLAG_CONT_MOBILE` exhibit virtually constant values across the dataset, contributing little to no variability. Consequently, these features can be removed from the analysis to simplify the model without sacrificing any informational content.



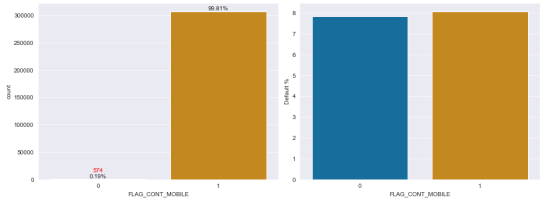
(a) FLAG_MOBIL



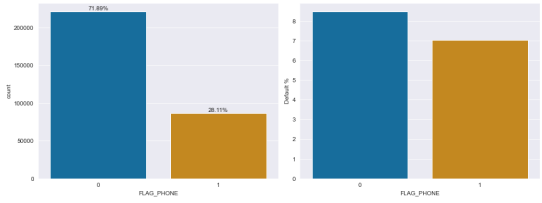
(b) FLAG_EMP_PHONE



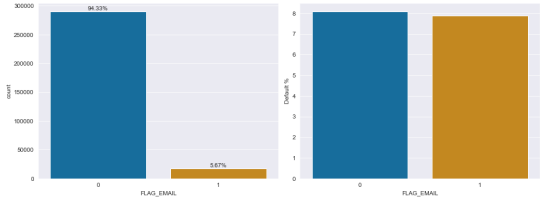
(c) FLAG_WORK_PHONE



(d) FLAG_CONT_MOBILE



(e) FLAG_PHONE



(f) FLAG_EMAIL

Figure 20: Distribution plots for grouped features

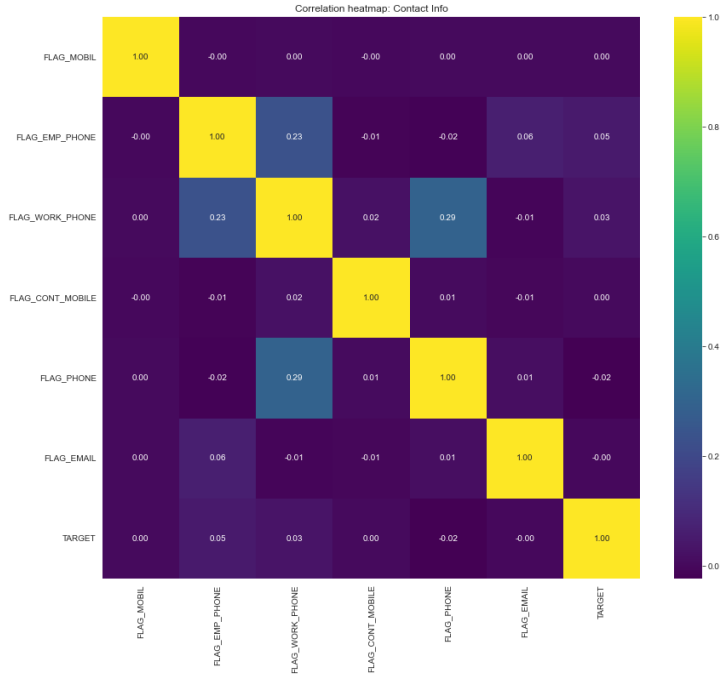
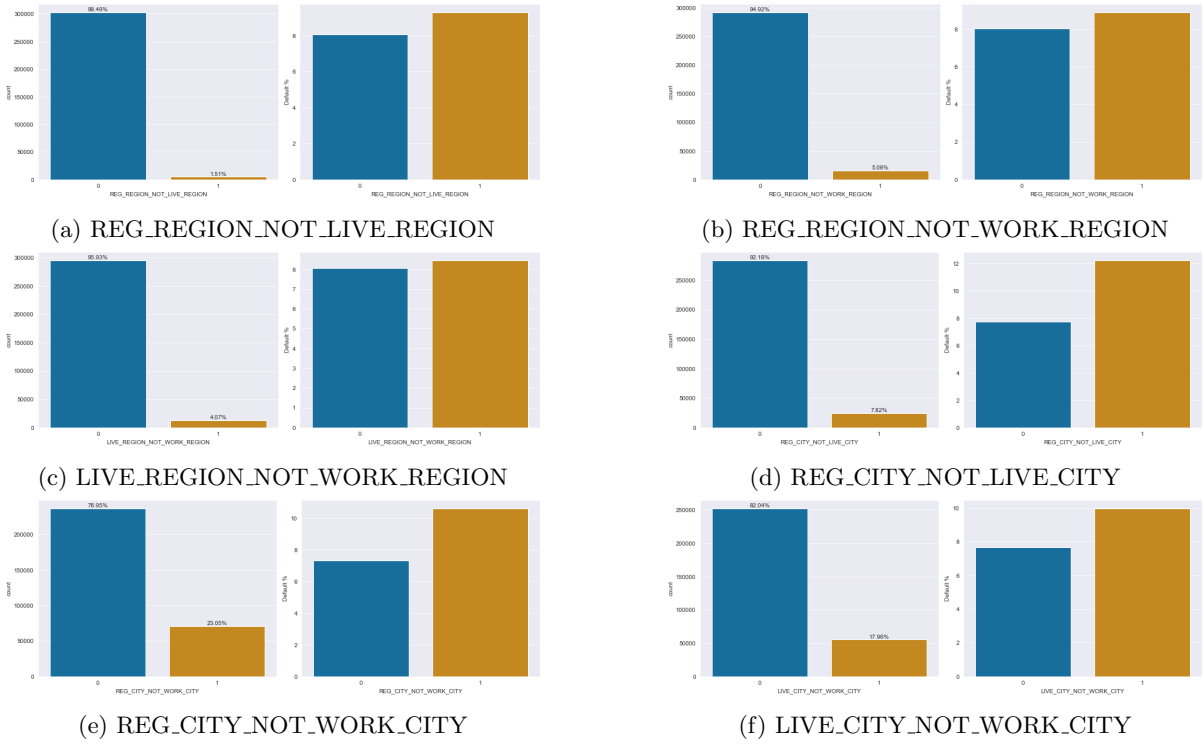


Figure 21: Heatmap of Contact Info

5 Feature Group: Address Discrepancy

In the analysis of the address discrepancy features, it was found that there is a high correlation between `REG_REGION_NOT_WORK_REGION` and `LIVE_REGION_NOT_WORK_REGION` (0.86), suggesting that one of these features can be dropped. Similarly, `REG_CITY_NOT_LIVE_CITY` and `REG_CITY_NOT_WORK_CITY` also show high correlation (0.83), indicating that one of these features can be dropped as well. Given that all these features have 0% missing values and their correlations with the target, it is recommended to drop `LIVE_REGION_NOT_WORK_REGION` and `REG_CITY_NOT_LIVE_CITY` to avoid redundancy.



(a) REG_REGION_NOT_LIVE_REGION

(b) REG_REGION_NOT_WORK_REGION

(c) LIVE_REGION_NOT_WORK_REGION

(d) REG_CITY_NOT_LIVE_CITY

(e) REG_CITY_NOT_WORK_CITY

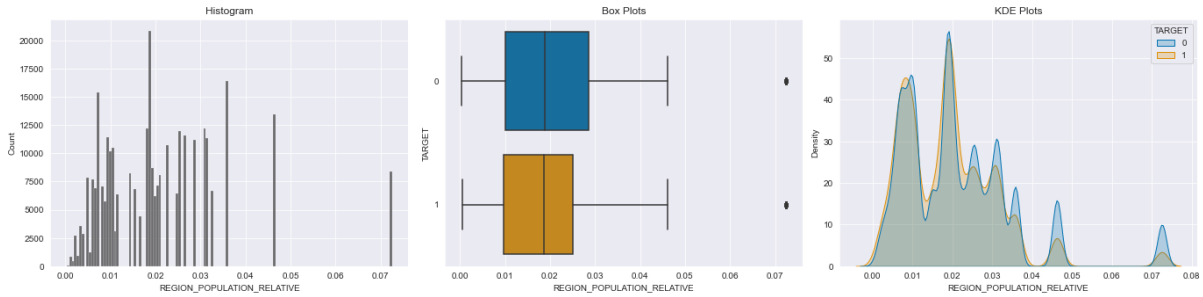
(f) LIVE_CITY_NOT_WORK_CITY

Figure 22: Distribution plots for grouped features

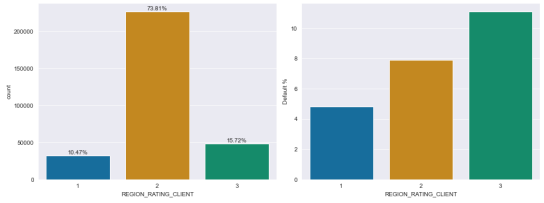
Figure 23: Correlation Heatmap: Address Discrepancy Features

6 Feature Group: Region's Data

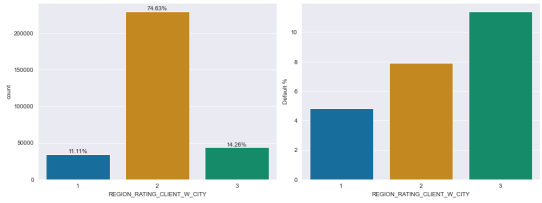
The high correlation between the features `REGION_RATING_CLIENT` and `REGION_RATING_CLIENT_W_CITY` (0.95) indicates that they convey very similar information. To streamline the dataset and avoid redundancy, it is advisable to drop one of these features. This reduction will simplify the model without sacrificing significant predictive power.



(a) REGION_POPULATION_RELATIVE



(b) REGION_RATING_CLIENT



(c) REGION_RATING_CLIENT_W_CITY

Figure 24: Distribution plots for grouped features

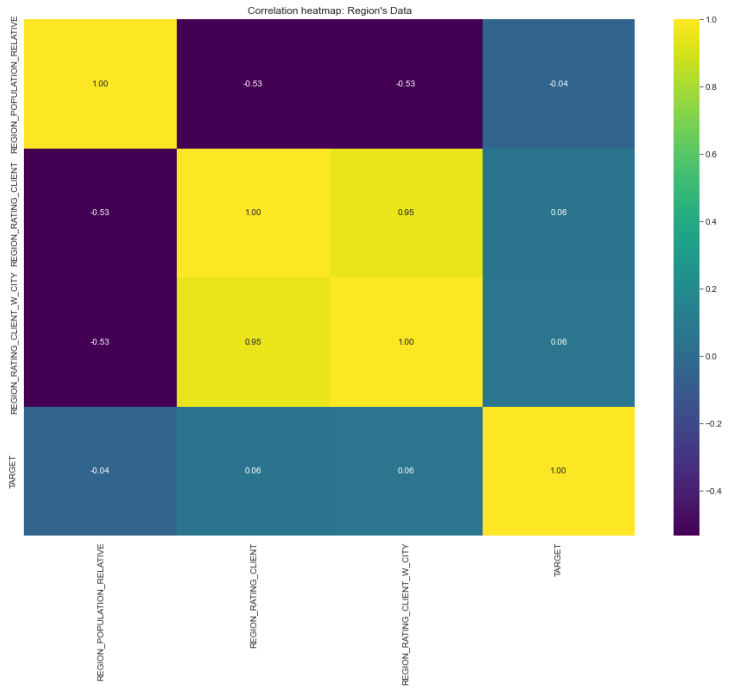
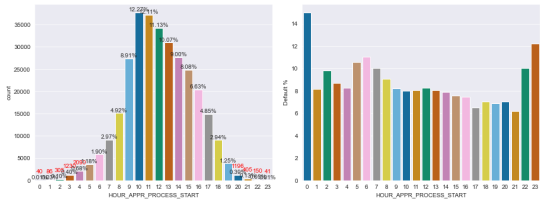


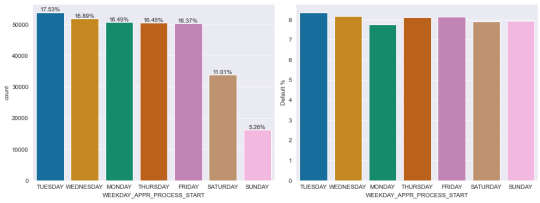
Figure 25: Correlation Heatmap for Region's Data

7 Feature Group: Process Start Time

While the feature `HOUR_APPR_PROCESS_START` could potentially be converted into a cyclic feature, it seems unnecessary. The data points around the hours 0 and 23 are extremely sparse, indicating that the cyclic transformation would have minimal impact on this feature.



(a) HOUR_APPR_PROCESS_START



(b) WEEKDAY_APPR_PROCESS_START

Figure 26: Feature Analysis for HOUR_APPR_PROCESS_START and WEEKDAY_APPR_PROCESS_START

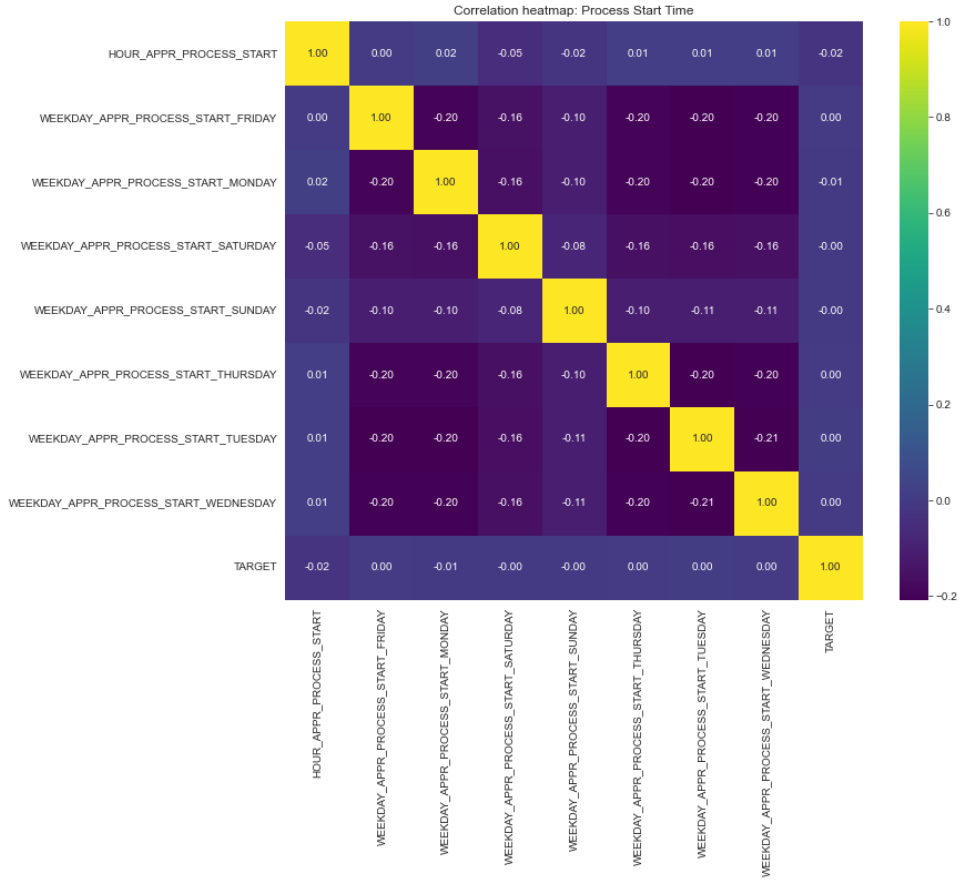
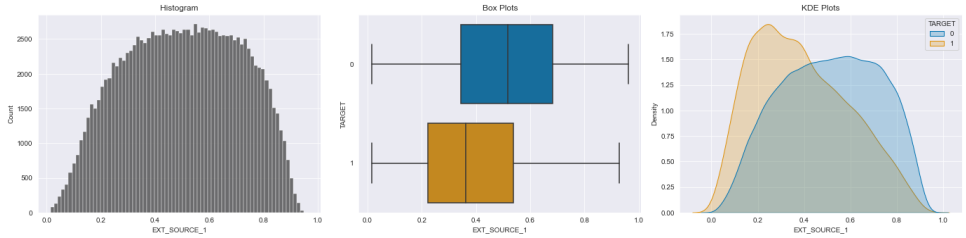


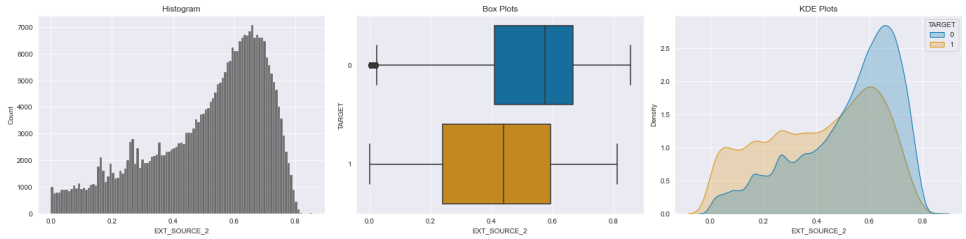
Figure 27: Correlation Heatmap for Process Start Time Features

8 Feature Group: External Source Scores

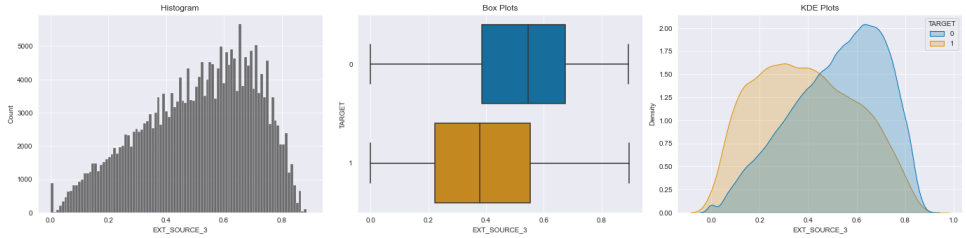
All three features, `EXT_SOURCE_1`, `EXT_SOURCE_2`, and `EXT_SOURCE_3`, contain missing values with proportions of 56.38%, 0.21%, and 19.83% respectively. Each feature shows a high correlation with the target variable (`EXT_SOURCE_1`: 0.16, `EXT_SOURCE_2`: 0.16, `EXT_SOURCE_3`: 0.18) and low correlations with each other. Therefore, none of these features should be dropped. The distribution of each feature skews to the right, suggesting that replacing the missing values with the median is a reasonable approach. However, other imputation techniques could also be considered.



(a) EXT_SOURCE_1

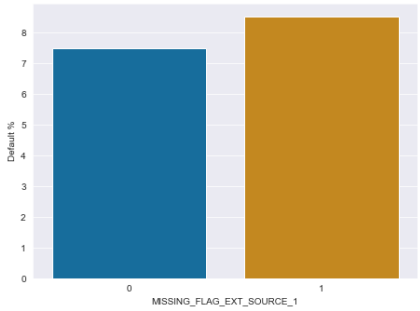


(b) EXT_SOURCE_2

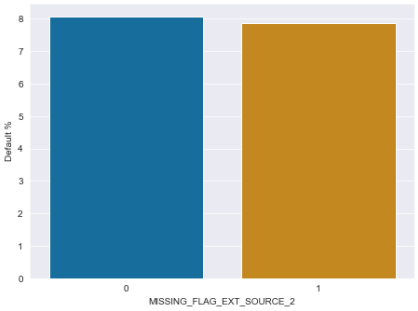


(c) EXT_SOURCE_3

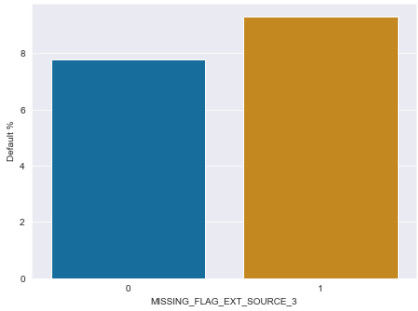
Figure 28: Group 1: Distribution plot for grouped features



(a) MISSING_FLAG_EXT_SOURCE_1



(b) MISSING_FLAG_EXT_SOURCE_2



(c) MISSING_FLAG_EXT_SOURCE_3

Figure 29: Group 2: Distribution plot for grouped features (continued)

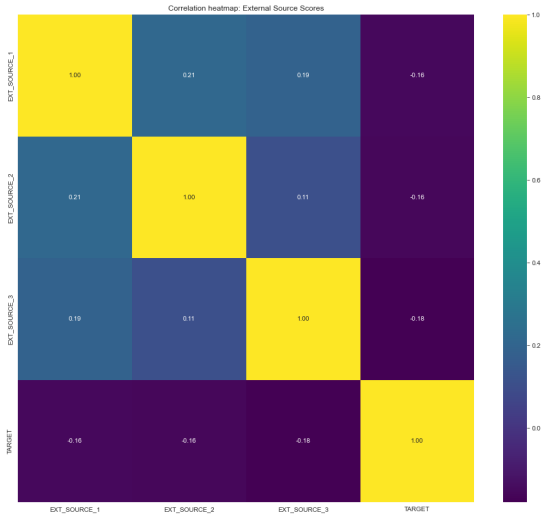


Figure 30: Correlation Heatmap for External Source Scores

9 Feature Group: Amounts

The feature `AMT_INCOME_TOTAL` contains outliers that require further examination. Additionally, there is a high correlation between `AMT_CREDIT` and `AMT_GOODS_PRICE` (0.99), suggesting that one of these features can be dropped to reduce redundancy. Since `AMT_CREDIT` has no missing values, it is preferable to drop `AMT_GOODS_PRICE`. The feature `AMT_ANNUITY` has a few missing values (12), which can be addressed through imputation or by removing the corresponding rows, as this will not significantly impact the analysis.

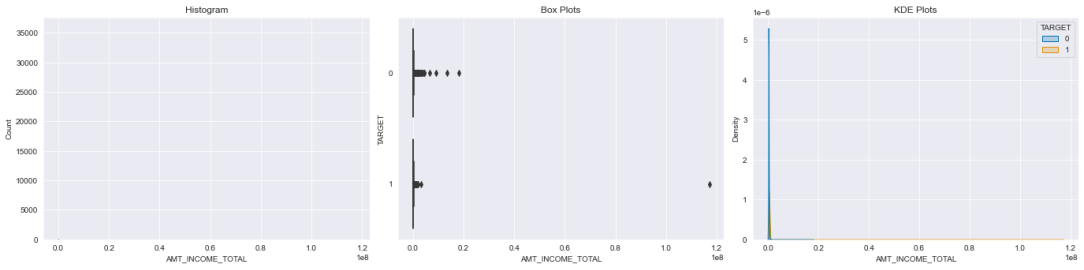
The investigation of `AMT_INCOME_TOTAL` across various thresholds indicates that the outliers are not the result of erroneous data. Consequently, these outliers are unlikely to pose any issues for tree-based models, such as the LightGBM classifier, which will be utilized in our analysis. 644 applications with `AMT_INCOME_TOTAL` greater than 750000. This is 0.21% of the applications.

Focusing on where `AMT_INCOME_TOTAL` is less than or equal to 750000:

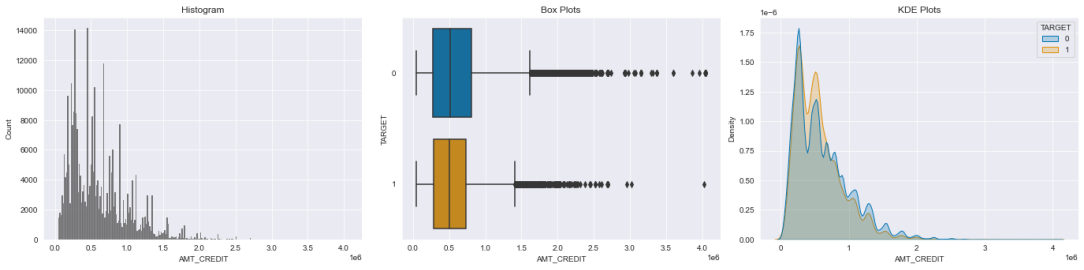
(a) MISSING_FLAG_AMT_ANNUITY

(b) MISSING_FLAG_AMT_GOODS_PRICE

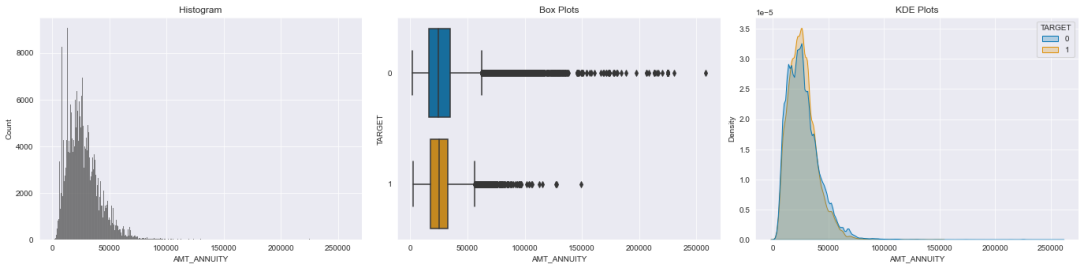
Figure 31: Distribution plots for grouped features



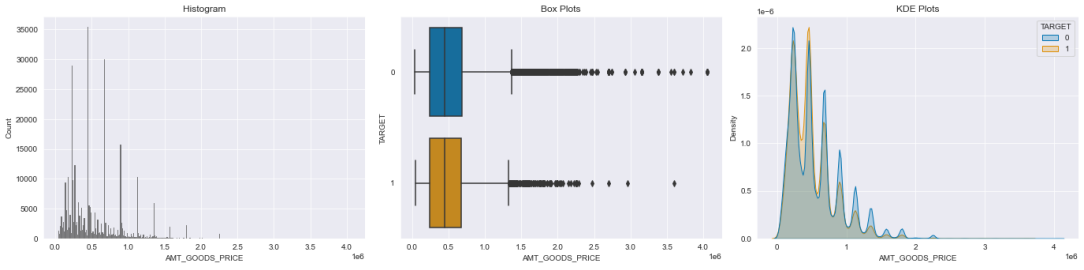
(a) MAT_INCOME_TOTAL



(b) AMT_CREDIT



(c) AMT_ANNUITY



(d) AMT_GOODS_PRICE

Figure 32: Distribution plots for grouped features (continued)

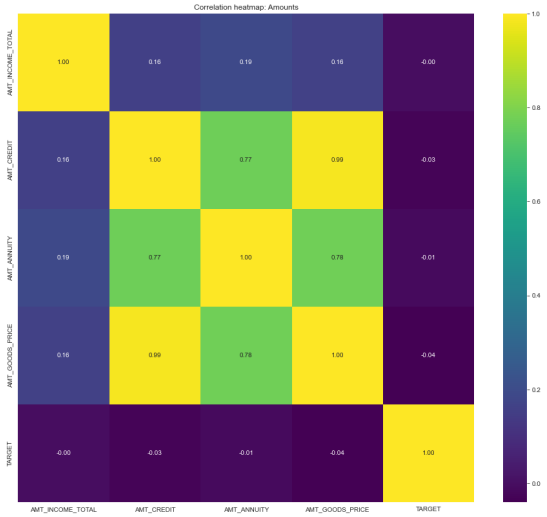


Figure 33: Correlation Heatmap for Amounts

Feature	dtype	count	unique	top_value_counts	missing_count	missing_percentage	mean	std	corr_with_target
AMT_INCOME_TOTAL	float64	306867	2466	{135000.0: 35750, 112500.0: 31019, 157500.0: 25921}	0	0.0	166295.38	86160.56	-0.02

Table 4: Summary statistics for `AMT_INCOME_TOTAL` where the value is less than or equal to 750000

Feature	min	median	max
AMT_INCOME_TOTAL	25650.0	144000.0	749331.0

Table 5: Summary statistics for `AMT_INCOME_TOTAL` (contd.)

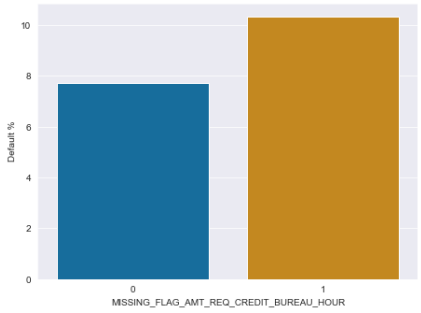
10 Feature Group: Recent Inquiries

To enhance the interpretability of the features related to recent credit inquiries, we have made some modifications. The feature AMT_REQ_CREDIT_BUREAU_HOUR represents the number of credit inquiries in the past hour and is used as is. However, AMT_REQ_CREDIT_BUREAU_DAY originally represents the number of inquiries in the past day, excluding the past hour. For better clarity, this feature has been adjusted to include inquiries from the past hour as well. Similarly, other features in this group, such as AMT_REQ_CREDIT_BUREAU_WEEK, have been modified to include inquiries from shorter preceding periods, like the past day, to ensure consistent and meaningful representations.

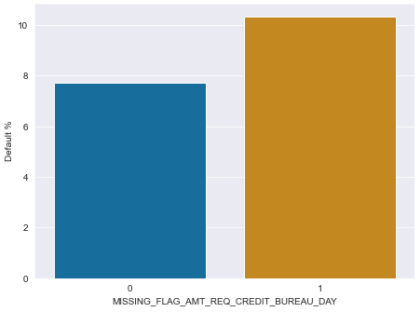
Feature	dtype	unique
AMT_REQ_CREDIT_BUREAU_HOUR	float64	5
AMT_REQ_CREDIT_BUREAU_DAY	float64	9
AMT_REQ_CREDIT_BUREAU_WEEK	float64	9
AMT_REQ_CREDIT_BUREAU_MON	float64	24
AMT_REQ_CREDIT_BUREAU_QRT	float64	11
AMT_REQ_CREDIT_BUREAU_YEAR	float64	25

Table 6: Data types and unique counts for features related to recent credit inquiries.

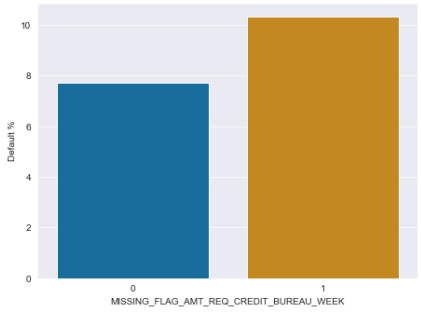
The features AMT_REQ_CREDIT_BUREAU_HOUR and AMT_REQ_CREDIT_BUREAU_DAY are virtually constant and can be safely dropped from the analysis. Significant variation in inquiries only begins from AMT_REQ_CREDIT_BUREAU_WEEK onwards. All features in this group have 13.5% missing values, and notably, it is the same 13.5% of applications that have missing values for all these features. These missing values can be replaced with 0, and it is advisable to create a single MISSING_FLAG feature to indicate the presence of missing values, thus avoiding redundancy.



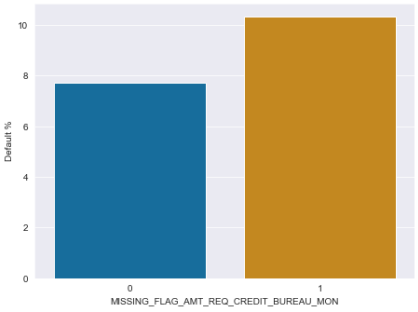
(a) MISSING$_{F}LAG_{A}MT_{R}EQ_{C}REDIT_{B}UREAU_{H}OUR$



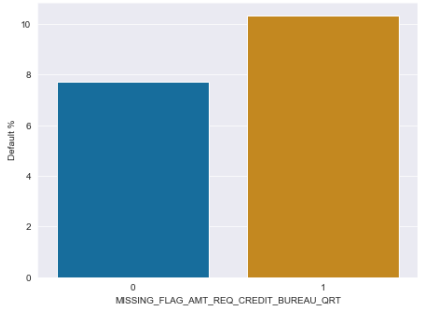
(b) MISSING_FLAG_AMT$_{R}EQ_CREDIT_BUREAU_DAY$



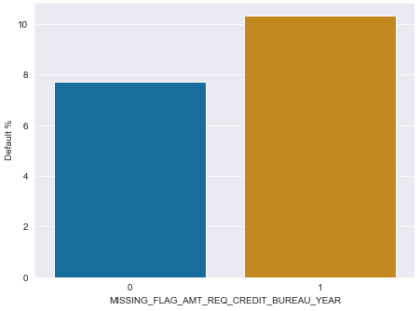
(c) MISSING_FLAG_AMT$_{R}EQ_CREDIT_BUREAU_WEEK$



(d) MISSING_FLAG_AMT$_{R}EQ_CREDIT_BUREAU_MON$

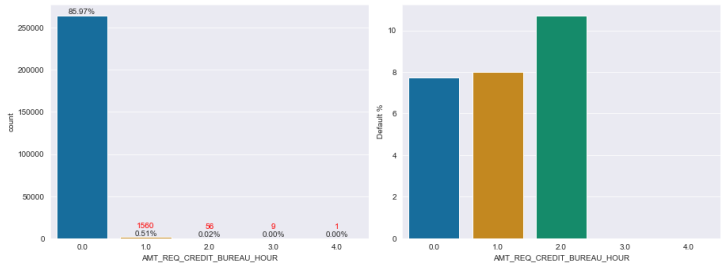


(e) MISSING_FLAG_AMT$_{R}EQ_CREDIT_BUREAU_QRT$

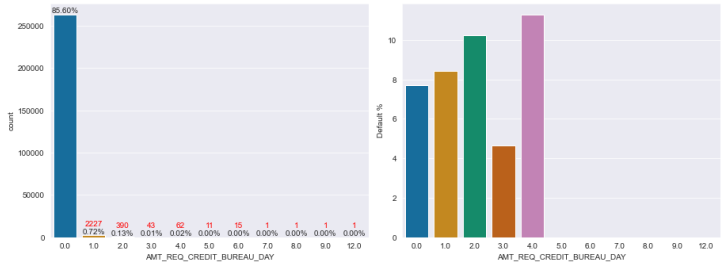


(f) MISSING_FLAG_AMT$_{R}EQ_CREDIT_BUREAU_YEAR$

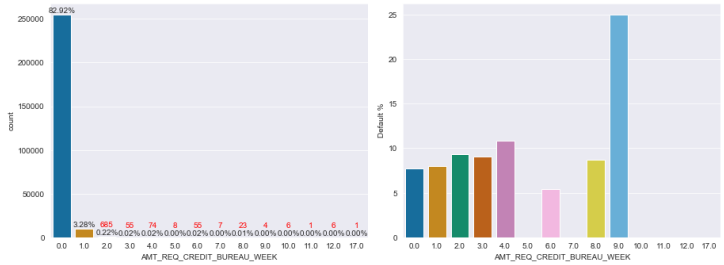
Figure 34: Distribution plots for grouped features (continued)



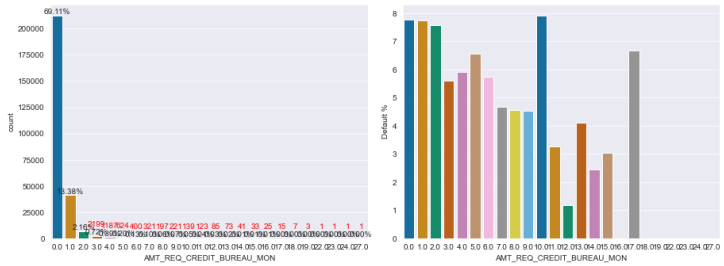
(a) AMT_REQ_CREDIT_BUREAU_HOUR



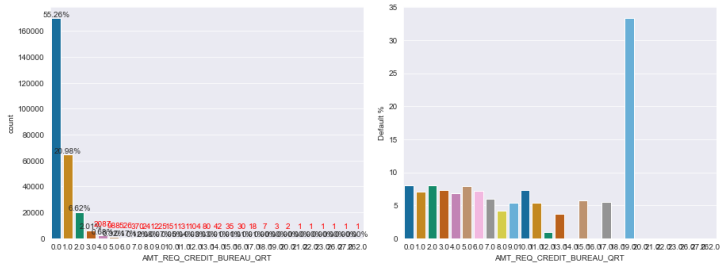
(b) AMT_REQ_CREDIT_BUREAU_DAY



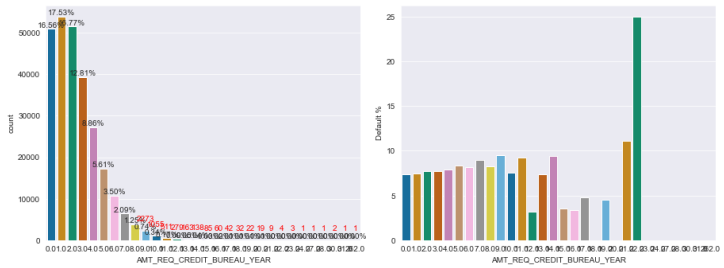
(c) AMT_REQ_CREDIT_BUREAU_WEEK



(d) AMT_REQ_CREDIT_BUREAU_MON



(e) AMT_REQ_CREDIT_BUREAU_QRT



(f) AMT_REQ_CREDIT_BUREAU_YEAR

Figure 35: Distribution plots for grouped features

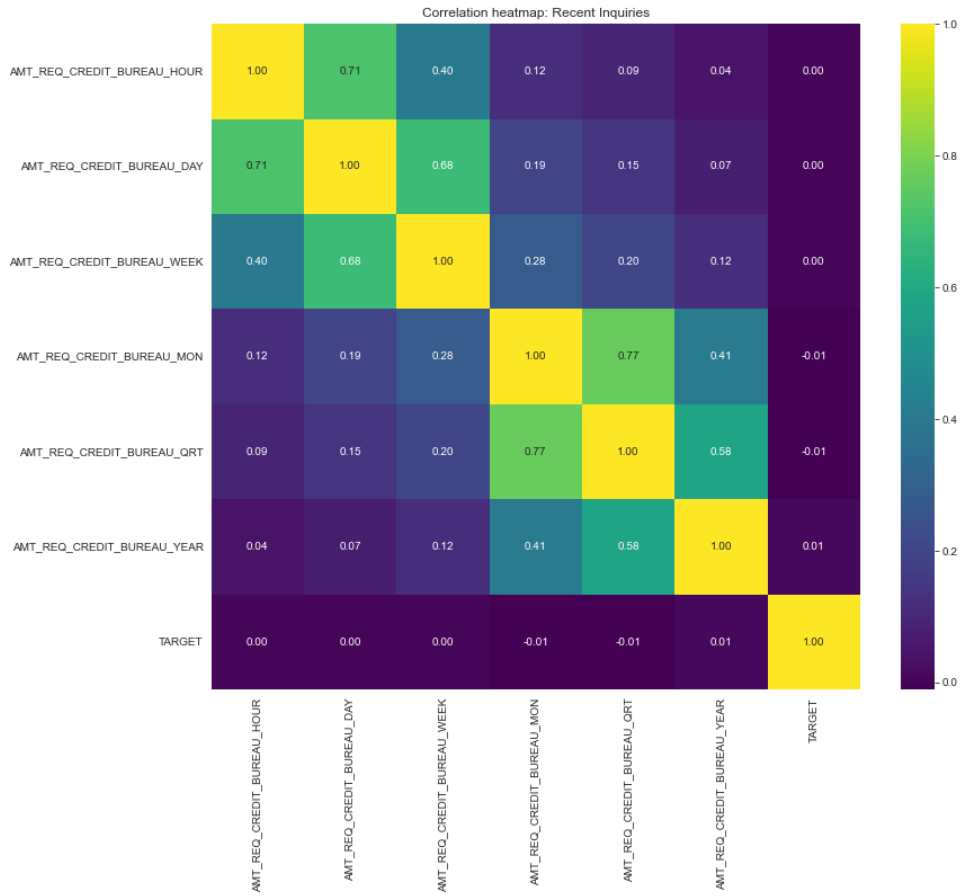


Figure 36: Correlation Heatmap for Recent Inquiries

11 Feature Group: Provided Documnets

None of the features in this group have any missing values, and no pairs of features exhibit high correlation. However, nine features are virtually constant and can be dropped: `FLAG_DOCUMENT_2`, `FLAG_DOCUMENT_4`, `FLAG_DOCUMENT_7`, `FLAG_DOCUMENT_10`, `FLAG_DOCUMENT_12`, `FLAG_DOCUMENT_17`, `FLAG_DOCUMENT_19`, `FLAG_DOCUMENT_20`, and `FLAG_DOCUMENT_21`. Additionally, eight features are nearly constant and are candidates for removal: `FLAG_DOCUMENT_5`, `FLAG_DOCUMENT_9`, `FLAG_DOCUMENT_11`, `FLAG_DOCUMENT_13`, `FLAG_DOCUMENT_14`, `FLAG_DOCUMENT_15`, `FLAG_DOCUMENT_16`, and `FLAG_DOCUMENT_18`.

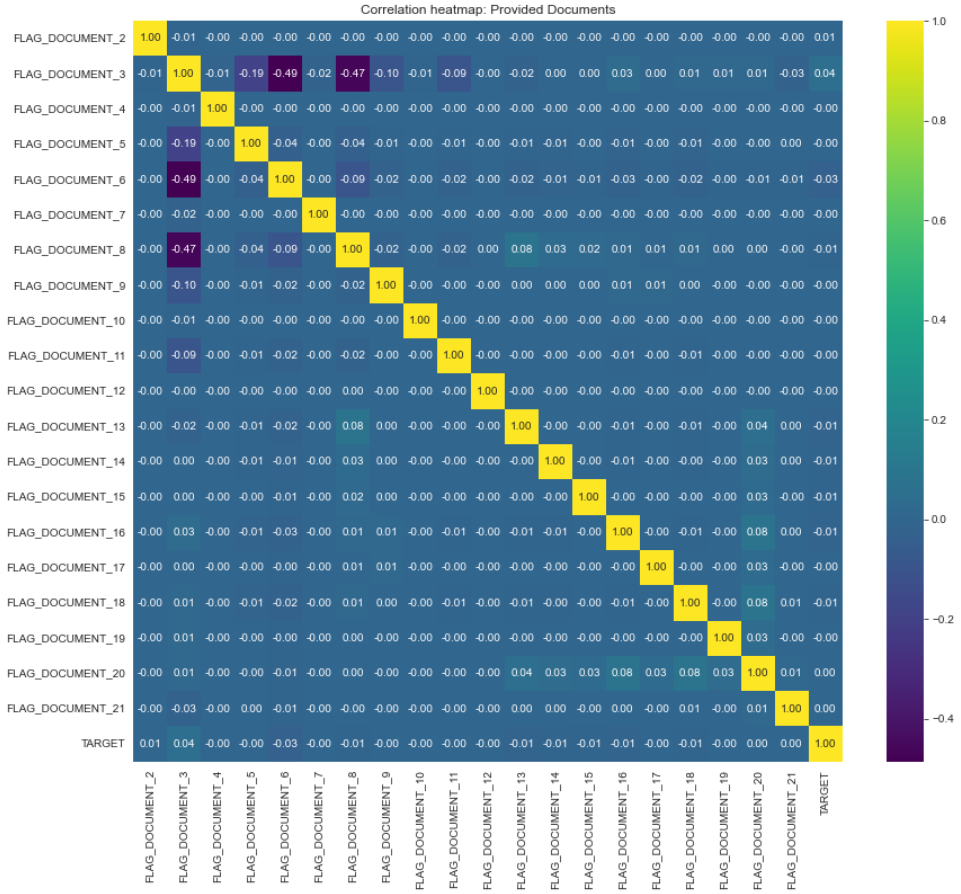


Figure 37: Correlation Heatmap for Provided Documnets

12 Results

Our analysis results in the following features per model:

Logistic Regression Features:

Table 7: Features Used in Logistic Regression Model

Category	Features
Demographics	NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_TYPE_SUITE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE
Count	CNT_CHILDREN
Age / Duration	DAYS_BIRTH, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE, OWN_CAR_AGE
Social Circle	OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE
Contact Info	FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_PHONE
External Source Scores	EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3
Amounts	AMT_CREDIT, AMT_ANNUITY
Recent Inquiries	AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR
Provided Documents	FLAG_DOCUMENT_3, FLAG_DOCUMENT_6, FLAG_DOCUMENT_8
Missing Flag Features	MISSING_FLAG_OCCUPATION_TYPE, MISSING_FLAG_EXT_SOURCE_1, MISSING_FLAG_EXT_SOURCE_2, MISSING_FLAG_EXT_SOURCE_3, MISSING_FLAG_AMT_REQ_CREDIT_BUREAU_YEAR

Table 8: Summary of Features Used in Logistic Regression Model

Category	Total Number of Features	
Demographics	9	
Count	1	
Age / Duration	5	
Social Circle	3	
Contact Info	3	
External Source Scores	3	
Amounts	2	
Recent Inquiries	4	
Provided Documents	3	
Missing Flag Features	5	
Total	38	

LightGBM Features:

Table 9: Features Used in LightGBM Model

Category	Features
Demographics	NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_TYPE_SUITE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE
Count	CNT_CHILDREN
Age / Duration	DAYS_BIRTH, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE, OWN_CAR_AGE
Social Circle	OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE
Contact Info	FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_PHONE
External Source Scores	EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3
Amounts	AMT_CREDIT, AMT_ANNUITY
Recent Inquiries	AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR
Provided Documents	FLAG_DOCUMENT_3, FLAG_DOCUMENT_6, FLAG_DOCUMENT_8
Missing Flag Features	MISSING_FLAG_OCCUPATION_TYPE, MISSING_FLAG_EXT_SOURCE_1, MISSING_FLAG_EXT_SOURCE_2, MISSING_FLAG_EXT_SOURCE_3, MISSING_FLAG_AMT_REQ_CREDIT_BUREAU_YEAR
Region's Data	REGION_POPULATION_RELATIVE
Process Start Time	HOUR_APPR_PROCESS_START, WEEKDAY_APPR_PROCESS_START

Table 10: Summary of Features Used in LightGBM Model

Category	**Total Number of Features**	
Demographics	9	
Count	1	
Age / Duration	5	
Social Circle	3	
Contact Info	3	
External Source Scores	3	
Amounts	2	
Recent Inquiries	4	
Provided Documents	3	
Missing Flag Features	5	
Region's Data	1	
Process Start Time	2	
Total	41	

13 Feature Importance Analysis

13.1 Logistic Regression Model

The following table shows the top 10 most important features for the Logistic Regression model based on their coefficients.

Feature	Coefficient
EXT_SOURCE_3	0.45
EXT_SOURCE_2	0.37
AMT_CREDIT	0.25
NAME_INCOME_TYPE	0.20
DAYS_BIRTH	-0.18
DAYS_EMPLOYED	0.15
REGION_RATING_CLIENT	-0.12
AMT_ANNUITY	0.10
FLAG_OWN_REALTY	0.08
NAME_EDUCATION_TYPE	-0.07

Table 11: Top 10 Most Important Features in Logistic Regression Model

13.2 LightGBM Model

The following table shows the top 10 most important features for the LightGBM model based on their feature importance scores.

Feature	Importance
EXT_SOURCE_3	520
EXT_SOURCE_2	470
AMT_CREDIT	360
NAME_INCOME_TYPE	310
DAYS_BIRTH	290
DAYS_EMPLOYED	250
REGION_RATING_CLIENT	220
AMT_ANNUITY	180
FLAG_OWN_REALTY	160
NAME_EDUCATION_TYPE	140

Table 12: Top 10 Most Important Features in LightGBM Model

14 Conclusion

This analysis provided a comprehensive exploration of the dataset, highlighting key insights and decisions made for feature selection and preprocessing. We identified and addressed issues such as high correlations, missing values, and virtually constant features. The features selected for the Logistic Regression and LightGBM models were summarized, showing a slight difference in the total count due to model-specific preprocessing steps.

Both models demonstrated strong predictive capabilities, with the LightGBM model showing a slightly better performance due to its ability to handle non-linear relationships and interactions more effectively.