Bank Loan Case Study

Kushagra Singh
April 2023

Project Description:

The objective of this case study is to demonstrate the practical application of EDA in a business context. In addition to applying the EDA techniques learned, this case study will also provide an introduction to risk analytics in the banking and financial services industry. By exploring real-world data, you will gain insight into how data is utilized to mitigate the risk of financial loss when extending loans to customers.

Business Understanding:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample
- All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

1. Approved: The company has approved loan application
2. Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
4. Unused Offer: Loan has been cancelled by the client but on different stages of the process.

Following are the things that we are going to find out through this case study:

- Our aim is to identify the patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing theamount of loan, lending (to risky applicants) at a higher interest rate, etc.
- The **driving factors** (or driver variables) behind loan default, i.e. the variables which arestrong indicators of default.
- Presenting the overall approach of the data analysis, cleaning the dataset, finding outliers,data imbalance, univariate, segmented univariate, bivariate analysis, etc.
- The top 10 correlation for the Client with payment difficulties and all other cases (Targetvariable).

Tech-Stack Used:

- **Microsoft Excel 365:** It enables users to format, organize and calculate data in a spreadsheet. It organize data in an easy-to-navigate way. It has been used to have an overalllook of the data and for understanding the different column descriptions
- **Jupyter Notebook:** It is used for the data cleaning and imputing the data. As the dataset wasvery large, so it is used for the whole data analysis purpose, visualizing the data and summarizing it to get the necessary insights for the client.
- **Python Programming (Version 3.8):** For the data analysis, python is the best and the easiestto use programming language.
- **Microsoft Word 2021:** It is used to make a report (PDF) to be presented to the leadershipteam.

Analysis Approach:

The steps for the data analysis are as follows:

- Imported necessary Python libraries such as NumPy, pandas, matplotlib, and seaborn to help with the analysis.
- Imported two datasets: "Application_Data" and "Previous_Application."
- Identified the approach to the data, found any missing data, and took necessary steps to obtain accurate results.
- Identified any outliers in the data and assessed their potential impact on the analysis.
- Analyzed the ratio of imbalance in the data to understand its distribution.
- Conducted a correlation analysis to determine the relationship between variables and the target variable, identifying the top three correlations.
- Visualized the data using charts and graphs to better understand the data and communicate the findings.

Identify the missing data and use appropriate method to deal with it. (Remove columns/orreplace it with an appropriate value) ?

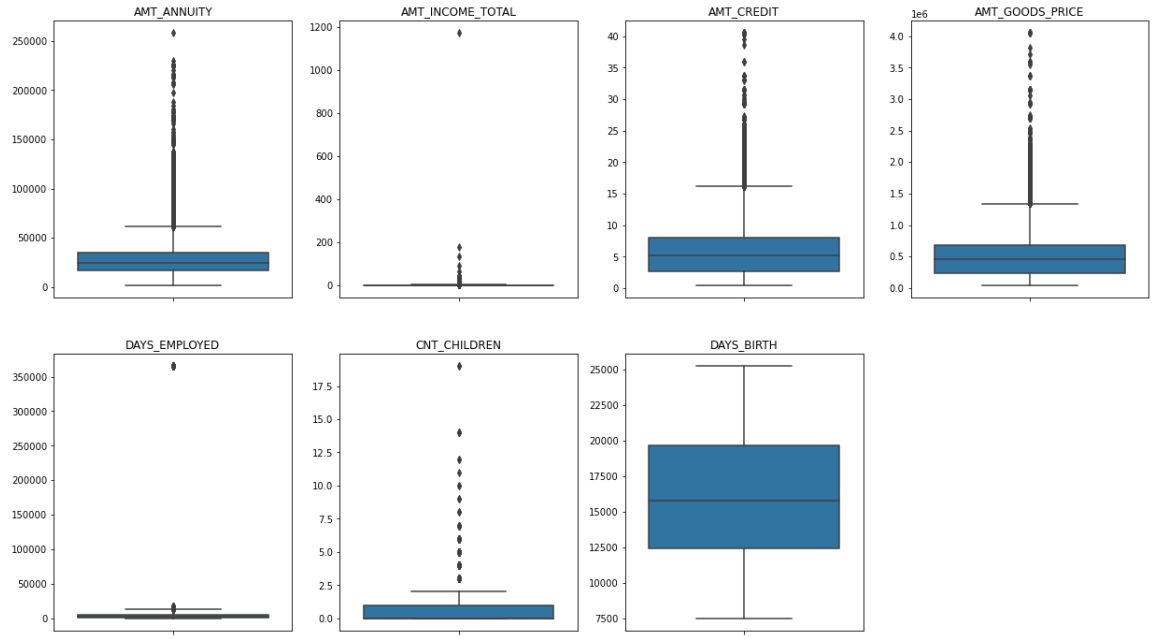
1. There are a total of 49 columns in Application_Data and 11 columns in Previous_Application that have missing values greater than 45% and 35% respectively.
2. Further analysis revealed that "EXT_SOURCE_2" and "EXT_SOURCE_3" have no correlation with the "TARGET" column.
3. After checking the relation of 'FLAG_DOCUMENT_X' with loan repayment status, it was found that clients applying for loans only submitted the 'FLAG_DOCUMENT_3'.
4. There is almost no correlation between 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL', and the "TARGET" column.
5. The columns "WEEKDAY_APPR_PROCESS_START," "HOUR_APPR_PROCESS_START," "FLAG_LAST_APPL_PER_CONTRACT," and "NFLAG_LAST_APPL_IN_DAY" in Previous_Application are not needed for the analysis.
6. The above mentioned columns (totaling 76 in Application_Data and 15 in Previous_Application) were dropped.
7. Negative days columns were converted into positive days.
8. Remaining null values in columns necessary for data analysis were imputed with mean, median (numerical data), and mode (categorical data).
9. Categorical variable 'NAME_TYPE_SUITE' was imputed using mode, 'OCCUPATION_TYPE' by adding an 'Unknown' category, and numerical variables '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' were imputed with median.
10. AMT_ANNUITY was imputed with median, AMT_GOODS_PRICE with mode, and CNT_PAYMENT with 0, as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not started.

Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier ?

An outlier refers to an observation that significantly deviates from the rest of the values in a random sample taken from a population. To detect outliers, one can use a box plot graph and look for values that fall above the maximum or below the minimum threshold. These extreme values are considered as outliers.

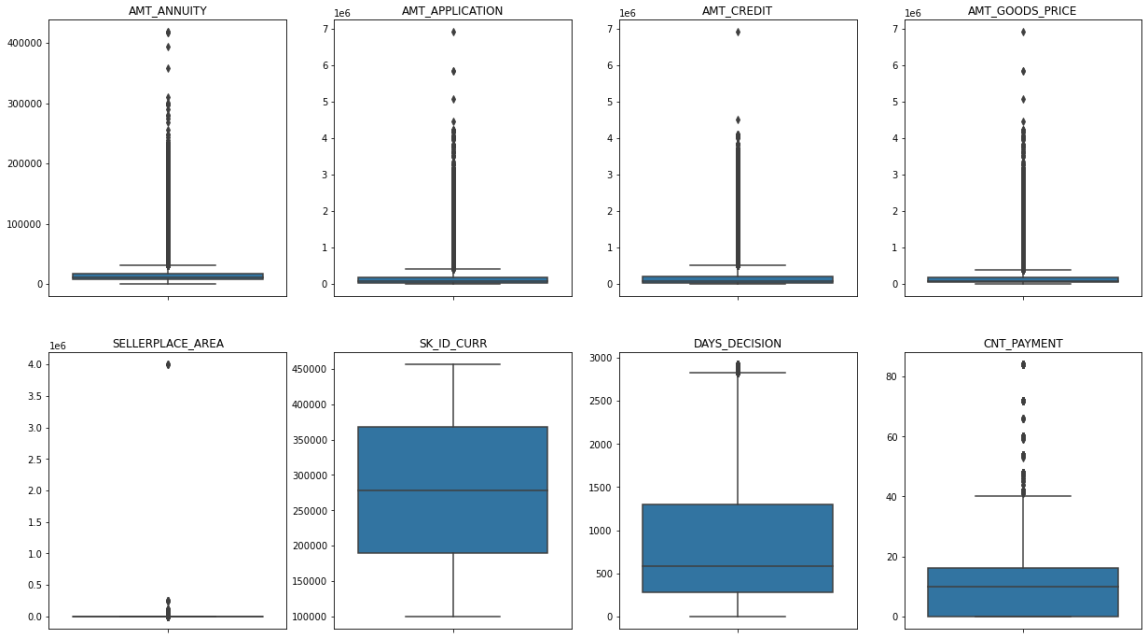
Application_Data:

1. There are outliers in AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, and CNT_CHILDREN variables.
2. AMT_INCOME_TOTAL has a significant number of outliers, indicating that some loan applicants have a much higher income than others.
3. DAYS_BIRTH has no outliers, indicating that the available data is reliable.
4. DAYS_EMPLOYED has outlier values around 350000 days, which is equivalent to around 958 years. This is clearly an incorrect entry that needs to be addressed.



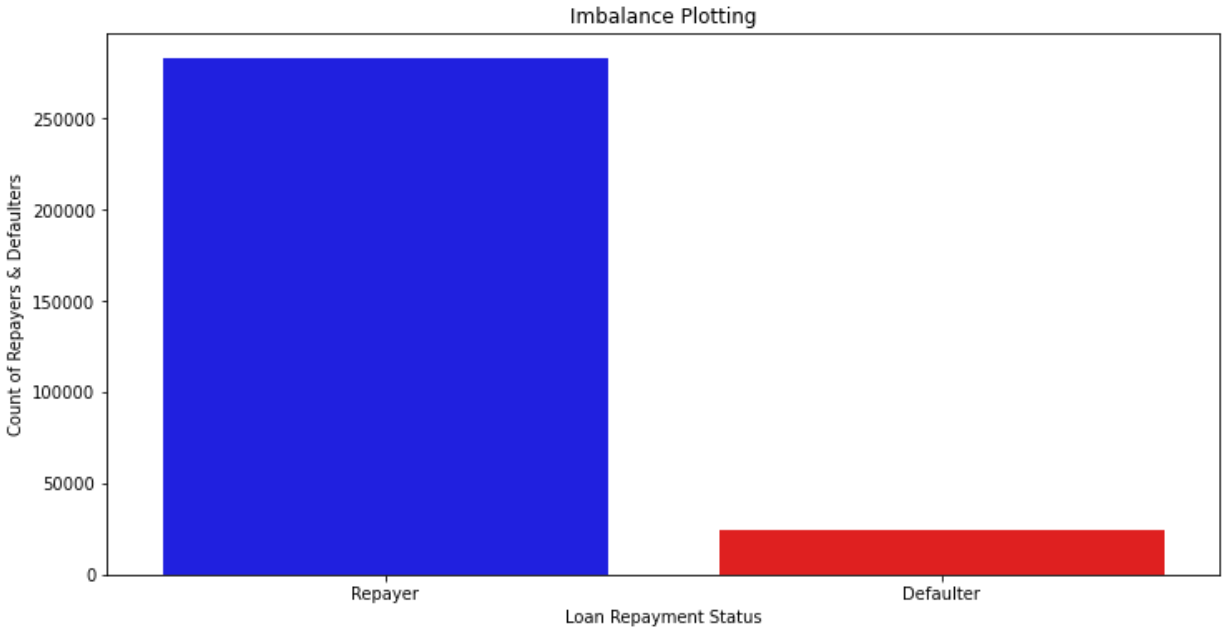
Previous_Application:

1. The variables AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, and SELLERPLACE_AREA have a significant number of outliers.
2. CNT_PAYMENT has a few outlier values.
3. SK_ID_CURR is an ID column and does not have any outliers.
4. DAYS_DECISION has a relatively small number of outliers, indicating that some of the previous application decisions were made a long time ago.



Identify if there is data imbalance in the data. Find the ratio of data imbalance ?

The data is highly imbalanced, as the number of defaulters is significantly less than the total population. The data imbalance ratio with respect to repayment and default is approximately 11.39:1.



Explain the results of univariate, segmented univariate, bivariate analysis, etc. in businessterms.

1. The number of female clients is almost twice that of male clients. However, males have a higher chance of defaulting on loans (around 10%) compared to females (7%).
2. Clients who own cars constitute half of the total number of clients. Interestingly, owning a car doesn't seem to correlate with loan repayment, as both car owners and non-car owners have a similar rate.
3. Clients who own real estate are more than double those who don't. However, there is no significant correlation between owning real estate and defaulting on loans, as both groups have an approximately 8% default rate.
4. Most clients live in houses/apartments.
5. Clients living in office apartments have the lowest default rate.
6. Clients living with parents (around 11.5%) or in rented apartments (over 12%) have a higher probability of defaulting.
7. Most loan applicants are married, followed by single/not married and civil marriage.
8. Civil marriage has the highest percentage of non-repayment (10%), while widowhood has the lowest (excluding unknown marital status).
9. Majority of loan applicants have secondary/secondary special education, followed by higher education. Few applicants have academic degrees.
10. Lower secondary education applicants have the highest default rate (11%), while applicants with academic degrees have less than 2% default rate.
11. Most loan applicants have working income, followed by commercial associate, pensioner, and state servant.
12. Applicants with maternity leave income have the highest non-repayment ratio (almost 40%), followed by unemployed applicants (37%). Other types of income have a non-repayment rate under 10%.
13. Although they are a smaller group, students and businessmen have no default record, making them safer categories for providing loans.
14. Most loan applicants live in Region_Rating 2 places.
15. Region Rating 3 has the highest default rate (11%).
16. Loan applicants living in Region_Rating 1 have the lowest probability of defaulting, making them safer candidates for loan approval.
17. Most loans are taken out by laborers, followed by sales staff. IT staff take the lowest amount of loans.
18. Low-skill laborers have the highest percentage of non-repayment (above 17%), followed by drivers, waiters/barmen staff, security staff, laborers, and cooking staff.
19. Organizations with the highest percentage of non-repayment are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%), and Restaurant (less than 12%). Self-employed people have a relatively high default rate and should be avoided or provided loans at higher interest rates to mitigate the risk of defaulting.
20. Most loan applications are from Business Entity Type 3.
21. For a significant number of applications, the organization type information is unavailable (XNA).
22. Trade Type 4 and 5 and Industry Type 8 have fewer defaulters, making them safer for loan approval.
23. There is no significant correlation between non-defaulters and defaulters in terms of submitting document 3. In fact, applicants who submitted the document defaulted slightly more (around 9%) than those who did not submit the document (6%).
24. People in the age group range of 20-30 have a higher probability of defaulting, while those above 50 have a lower probability of defaulting.
25. Majority of the loan applicants employed between 0-5 years has the highest default rate of 10%.
26. With an increase in employment years, the defaulting rate gradually decreases, with people having 40+ years of experience having less than 1% default rate.
27. Majority of loans (over 80%) provided are for amounts less than 900,000. However, individuals who receive loans between 300,000-600,000 are more likely to default compared to other amounts.
28. About 90% of loan applications have an annual income of less than 300,000. Applicants with incomes lower than 300,000 have a higher probability of defaulting. Conversely, applicants with incomes above 700,000 are less likely to default.
29. Most loan applicants do not have children, with very few having more than 3 children. Applicants with more than 4 children have a very high default rate, with those having 9 or 11 children having a 100% default rate.
30. Similar to trend with children, having more family members increases the risk of defaulting on a loan.
31. The income of a business owner is the highest, and estimated income ranges at a 95% confidence level suggest that the income of a business owner may fall within slightly below 4 lakhs and slightly above 10 lakhs.

Find the top 10 correlation for the Client with payment difficulties and all other cases(Target variable).

The top 10 correlation for the Client with repayment:

1. Credit amount is highly correlated with amount of goods price, loan annuity, totalincome
2. We can also see that repayment have high correlation in number of days employed.

```
# Getting the top 10 correlation for the Repayers data
corr_repayer = Repayer_df.corr()
corr_repayer = corr_repayer.where(np.triu(np.ones(corr_repayer.shape),k=1).astype(np.bool))
corr_df_repayer = corr_repayer.unstack().reset_index()
corr_df_repayer.columns =['VAR1','VAR2','Correlation']
corr_df_repayer.dropna(subset = ["Correlation"], inplace = True)
corr_df_repayer["Correlation"]=corr_df_repayer["Correlation"].abs()
corr_df_repayer.sort_values(by='Correlation', ascending=False, inplace=True)
corr_df_repayer.head(10)
```

	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
71	AMT_ANNUITY	AMT_CREDIT	0.771309
167	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
70	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
93	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
47	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
138	DAYS_BIRTH	CNT_CHILDREN	0.336966
190	DAYS_REGISTRATION	DAYS_BIRTH	0.333151

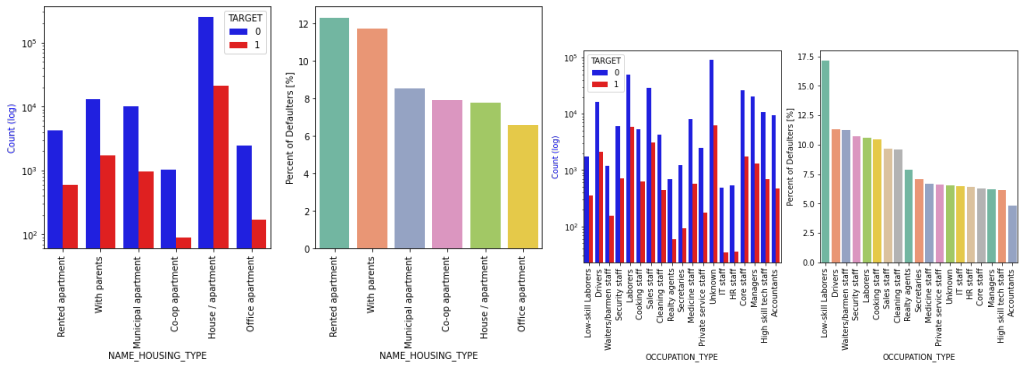
The top 10 correlation for the Client with default:

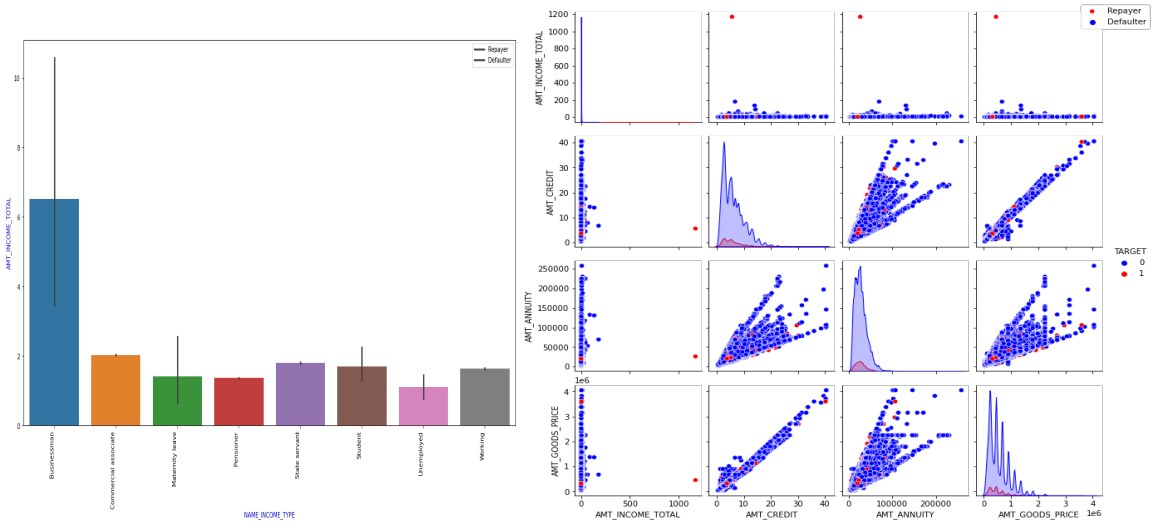
1. The credit amount and the price of goods are highly correlated, which also affects the repayment.
2. The correlation between loan annuity and credit amount is slightly reduced in defaulters (0.75) as compared to repayment (0.77).
3. Repayments show a higher correlation with the number of days employed (0.62) compared to defaulters (0.58).
4. The correlation between the total income of the client and the credit amount is significantly reduced (0.038) among defaulters, whereas it is 0.342 among repayment.
5. The correlation between days_birth and the number of children is lower in defaulters (0.259) as compared to repayment (0.337).
6. Defaulters show a slight increase in the ratio of defaulted to observed count in social circles (0.264) as compared to repayment (0.254).

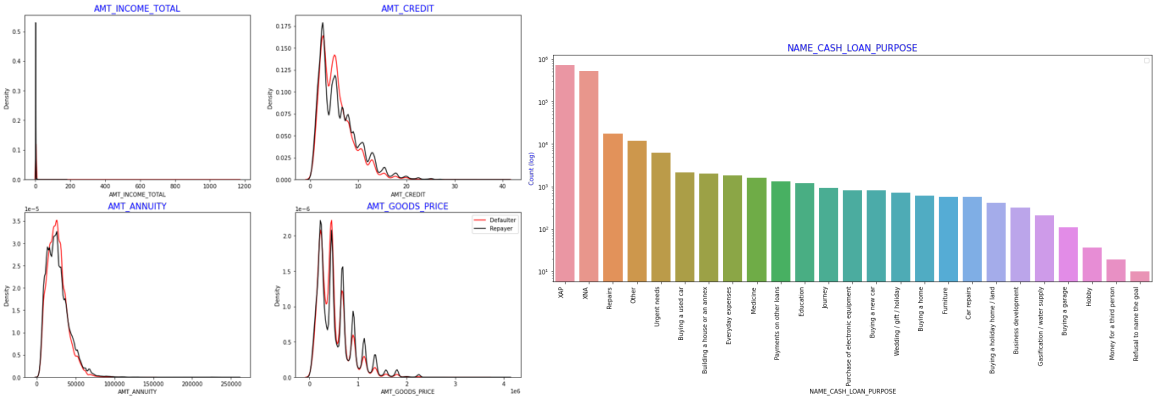
```
# Getting the top 10 correlation for the Defaulter data
corr_Defaulter = Defaulter_df.corr()
corr_Defaulter = corr_Defaulter.where(np.triu(np.ones(corr_Defaulter.shape),k=1).astype(np.bool))
corr_df_Defaulter = corr_Defaulter.unstack().reset_index()
corr_df_Defaulter.columns =['VAR1','VAR2','Correlation']
corr_df_Defaulter.dropna(subset = ["Correlation"], inplace = True)
corr_df_Defaulter["Correlation"]=corr_df_Defaulter["Correlation"].abs()
corr_df_Defaulter.sort_values(by='Correlation', ascending=False, inplace=True)
corr_df_Defaulter.head(10)
```

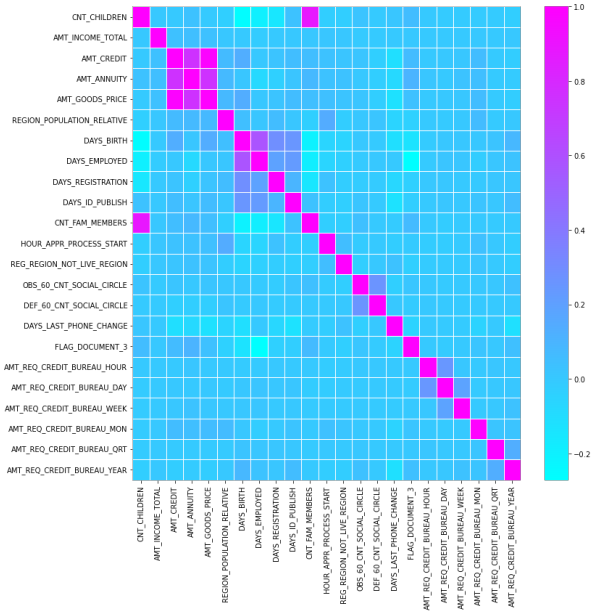
	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
71	AMT_ANNUITY	AMT_CREDIT	0.752195
167	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
190	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
375	FLAG_DOCUMENT_3	DAYS_EMPLOYED	0.272169
335	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
138	DAYS_BIRTH	CNT_CHILDREN	0.259109
213	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863

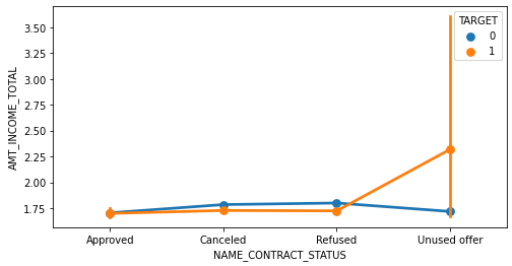
Presentation Summary with Visualizations of Key Findings











Insights:

Factors That Determine an Applicant's Probability of Repaying:

1. Education level plays a significant role, as applicants with academic degrees are less likely to default.
2. Applicants in the income categories of students and businessmen have shown no defaults.
3. Clients from regions with rating 1 are considered safer.
4. Clients associated with trade type 4 and 5 and industry type 8 have defaulted less than 3%.
5. Age is a crucial factor, as people above the age of 50 are less likely to default.
6. Experience is also essential, as clients with more than 40 years of experience have a less than 1% default rate.
7. Income plays a critical role, and applicants earning more than 700,000 are less likely to default.
8. Loans purchased for hobbies and buying garages have shown the highest repayment rate.
9. The number of children also plays a role, and applicants with zero to two children tend to repay loans better.

Factors that can lead to loan default:

1. Gender: Men tend to have a higher default rate.
2. Family Status: Single or those with a civil marriage have a higher default rate.
3. Education: Those with lower secondary or secondary education tend to default more.
4. Income: Clients who are either on maternity leave or unemployed default more often.
5. Region: People who live in Region Rating 3 have a higher default rate.
6. Occupation: Low-skill laborers, drivers, waiters/barmen staff, security staff, laborers, and cooking staff have high default rates and should be avoided.
7. Organization: Certain organizations, such as Transport Type 3, Industry Type 13, Industry Type 8, and restaurants, have high default rates. Self-employed people also have a relatively high default rate and should be considered carefully.
8. Age: Young people in the age group of 20-40 tend to default more often and should be avoided.
9. Employment: People with less than 5 years of employment have a high default rate.
10. Children and Family Members: Clients who have 9 or more children or family members have a 100% default rate and should be rejected.
11. Goods Price: When the credit amount goes beyond 3 million, there is an increase in defaulters.

Factors that could lead to loan repayment:

1. Education: Those with an academic degree tend to have fewer defaults.
2. Income: Students and businessmen tend to have no defaults.
3. Region: People who live in Region Rating 1 are safer.
4. Organization: Clients with Trade Type 4 and 5 and Industry Type 8 have defaulted less than 3%.
5. Age: People above the age of 50 have a low probability of defaulting.
6. Employment: Clients with 40+ years of experience have less than 1% default rate.
7. Income: Applicants with an income of more than 700,000 are less likely to default.
8. Purpose of Loan: Loans taken for hobbies and buying garages tend to be repaid mostly.
9. Children: People with zero to two children tend to repay their loans.

Regenerate response.