Bank Loan Case Study

Project Description

This case study tries to find trends that show whether a client has trouble making their payments, which may be used to decide whether to grant the loan, reduce its size, charge riskier applicants a higher interest rate, etc. By doing this, it will be ensured that only borrowers who can repay the loan will be accepted. The objective of this case study is to identify such applications using EDA.

In other words, the organization seeks to understand the characteristics that are reliable predictors of loan default, also known as the driving factors (or driver variables) behind loan default. This information can be used by the business in portfolio management and risk analysis.

Approach

Because of the people's weak or non-existent credit histories, loan providers find it challenging to grant loans to them. Because of this, some customers take advantage of it by defaulting. Our finance company that specializes in providing urban customers with several kinds of loans. To analyse the patterns found in the data, you must employ EDA. By doing this, it will be ensured that only those applicants who can repay the loan would be accepted.

Data cleansing will be our first step, and we'll count how many null values there are in each column. then analysing our data using statistical methods. After cleaning, we'll examine the relationship between our data and visualization. Finally, we will summarize all the conclusions we reached from our data.

Tech – stack used

Microsoft Excel is a Microsoft software tool that uses spreadsheets to arrange numbers and data using formulas and functions. Microsoft Excel is a robust data visualization and analysis program that employs spreadsheets to store, organize, and track data sets using formulae and functions.

Insights

Exploratory Data Analysis

1. **Data Inspection**
    a. To check the shape of our data, we are counting the total number of records in it.
    b. We have total 307511 rows and 123 columns in our data.
2. **Calculating the Percentage of Null Values in Each Column**
    a. To begin our analysis, we must first determine the number of null values in our data. We will calculate this number by dividing the total number of rows by the number of blank rows. This value is shown as a percentage.
    b. We will eliminate columns with null values once we obtain the percentage of null values for each column greater than 35%.

A	B	C	D	E	F	G	H	I
Column2	SK_ID_CU	TARGET	NAME_CONTRA	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REA	CNT_CHILDREN	AMT_INCOMI
Blank Rows count	307511	0	0	0	0	0	0	0
% of Missing value		0.00	0.00	0.00	0.00	0.00	0.00	0.00

Column2	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPL	YEARS_BUILD_AVG	COMMONAREA_	ELEVATORS_AVG	ENTRANCES_AVG
Blank Rows count	660	60965	156061	179943	150007	204488	214865	163891	154828
% of Missing value	0.21	19.83	50.75	58.52	48.78	66.50	69.87	53.30	50.35

3. **Removing Unnecessary Columns**
    a. First, we are eliminating columns with more than 35% of null entries.

b. There are several Flag_document columns, and we are tallying each flag document in comparison to targets 0 and 1. With the use of this information, we can determine that only Flag_document3 is utilized in both targets, as opposed to other documents. Therefore, we are eliminating all but the Flag_document 3 column.

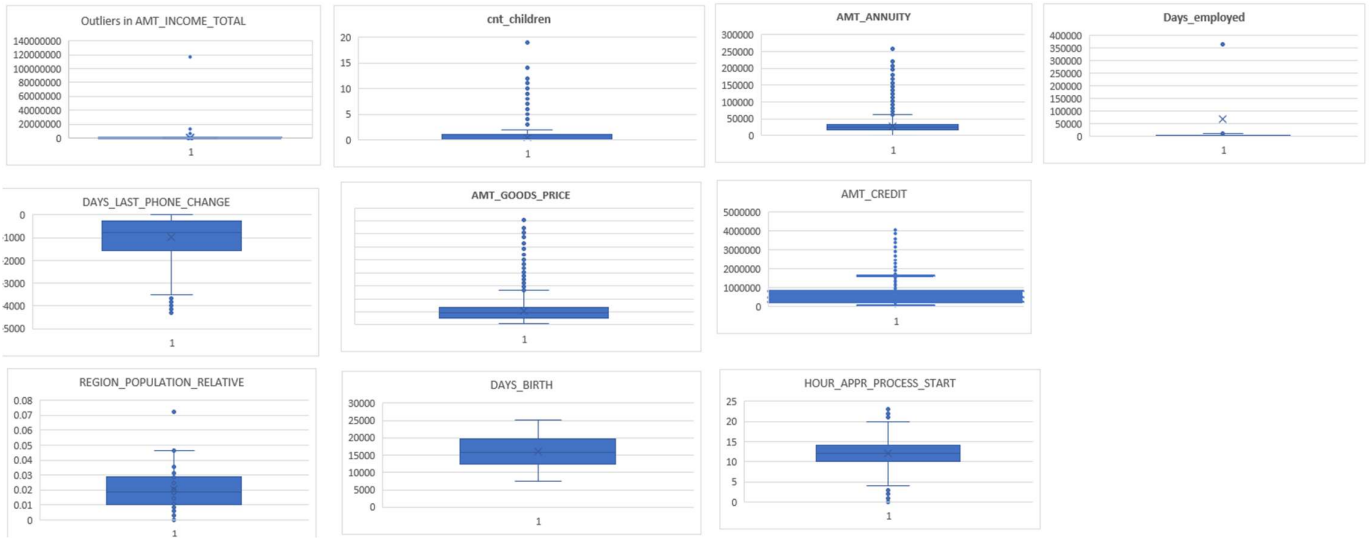
c. Additional columns contain information concerning contact information. We are computing the columns' correlation to the target field. This computation demonstrates that there is no direct correlation between the target field and the contact information. Therefore, since the contact column has no impact on target, we are eliminating it.

4. **Adjusting Values**
   a. For Code_Gender, there are three subcategories: F, M, and XNA. For the variable XNA, there are only 4 records in the data. We are substituting F for XNA because F Code_Gender Category has a significant number of records in the data.
   b. Our data for days columns show that we have negative data. To make the values positive, we'll use the abs () function.
   c. In order to compute Age precisely, we must divide days_birth by 365.
   d. The occupation_type field contains 31.35% null values. Null values are being converted to "Unknown" in order to imput the occupation field.
5. **Statistics regarding Numeric Columns**
   a. In order to offer statistical data regarding numerical columns, we compute the mean, standard deviation, minimum, maximum, and 25%, 75% using the number of records for each column.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
Count	307511	307511	307511	307499	307233
Mean	0.42	168797.92	599026.00	27108.57	538396.21
Std	0.722121384	237123.1463	402490.777	14493.73732	369446.4606
Min	0	25650	45000	1615.5	40500
25%	0	112500	270000	16524	238500
75%	1	202500	808650	34596	679500
Max	19	117000000	4050000	258025.5	4050000

	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
Count	307511	307511	307511	307511	307511
Mean	0.02	16037.00	67724.74	4986.12	2994.20
Std	0.01383128	4363.988632	139443.7518	3522.886321	1509.450419
Min	0.00029	7489	0	0	0
25%	0.010006	12413	933	2010	1720
75%	0.028663	19682	5707	7480	4299
Max	0.072508	25229	365243	24672	7197

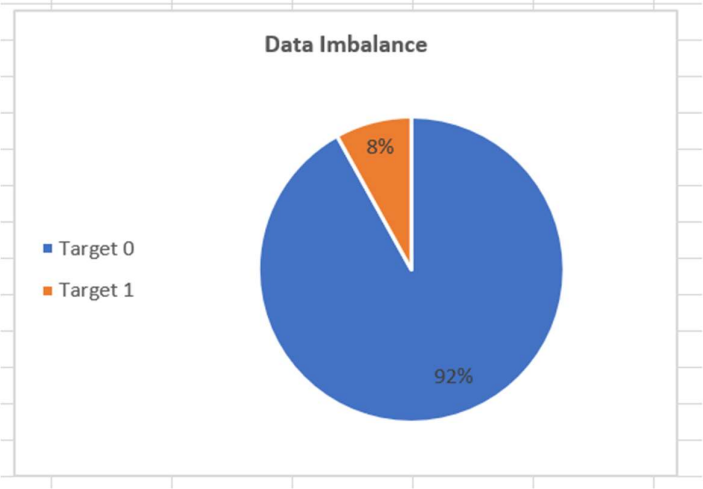
6. **Outliers** - We are using a box plot to look for any outliers in our data.



a. This study shows that days_employed and income_total have more outliers than other columns.
b. The Cnt_children outlier value exceeds 3.

c. The third quartile of the variables amt_annuity, goods_proce, and credit contains outliers.
d. Likewise, there are outliers in the first quartile of Last_phone_changed

7. **Data imbalance**
   a. We are using the Target field to calculate data imbalance. Each category in the Target field is counted, and the result is divided to determine the proportion of data imbalance.
   b. This pie chart illustrates the percentage of unbalanced data.



c. This image demonstrates that the distribution of observations inside the target field is unequal.
d. Unbalanced data can make it difficult for us to accurately forecast outcomes.
e. There are a variety of methods we might employ to balance the data.

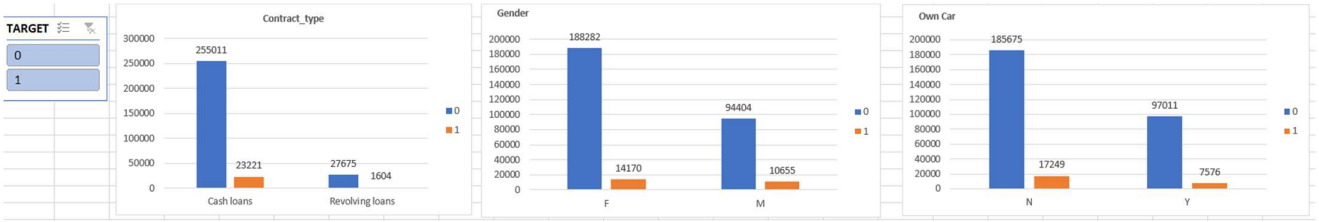
8. **Data Distribution**
   a. Target 0 and Target 1 are the two major groups for our data. By comparing the columns with the target field, we are able to determine the distribution of data for this field.

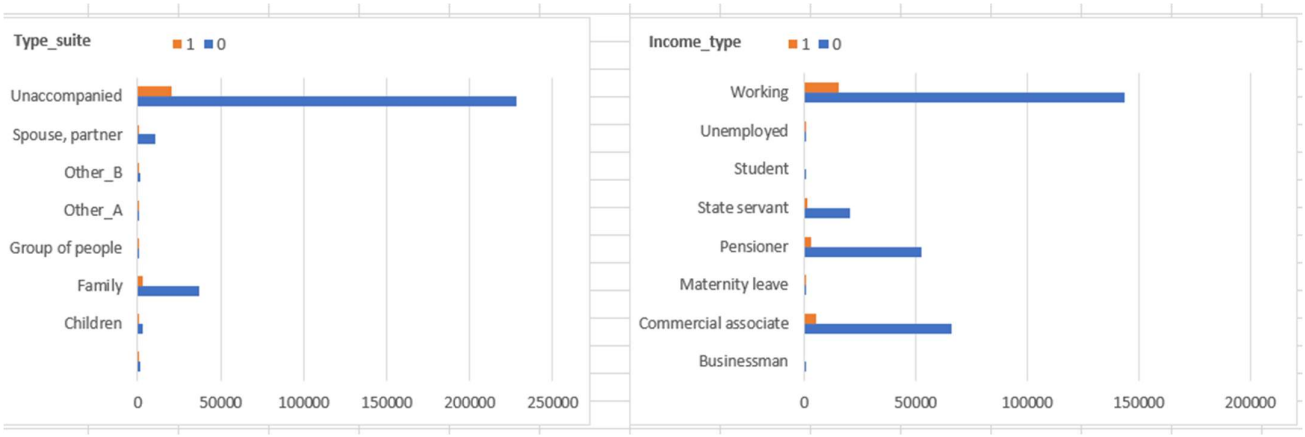
b. Customers who earn between $25,000 and $30,000 per year are more likely to pay back the loan.

c. Working professionals and business partners are two types of consumers who will return the loan with credit between $45,000 and $104,5,000.
d. Married individuals, rather than singles, are the ones who are repaying loans in significant numbers.
e. Customers who are between the ages of 30 and 60 are more likely to repay their loans.

9. Categorical Variables

a. To analyze categorical variables from our data, we are employing target field against all categorical fields.





b. The vast majority of our clients favor cash loans over revolving loans.
c. No matter their family structure, female clients are more likely to repay their loans.
d. Customers with secondary or special education tend to repay loans more frequently.
e. The majority of loan borrowers don't have cars.

10. Correlation

a. To determine how numeric columns are affecting our target field, we are computing the correlation between each numerical variable and the target field.

Target 0	AMT_INC	AMT_CREDIT	AMT_ANNUITY	AMT_GOOD	REGION_P	Age	DAYS_BIR	DAYS_EMI	DAYS_REG	DAYS_ID_	HOUR_AP
AMT_INCOME_TOTAL	1										
AMT_CREDIT	0.342799	1									
AMT_ANNUITY	0.418953	0.771308946	1								
AMT_GOODS_PRICE	0.349462	0.987250457	0.776685779	1							
REGION_POPULATION_RELATIVE	0.167851	0.100603799	0.120988482	0.1038272	1						
Age	0.001673	-0.002034274	-0.002741497	-0.0018017	0.004448	1					
DAYS_BIRTH	0.001744	0.000619604	0.001311382	0.0011837	0.000751	-0.33072	1				
DAYS_EMPLOYED	0.004237	0.003385553	0.002681343	0.0034824	-0.00084	-0.05641	0.025659	1			
DAYS_REGISTRATION	0.00084	0.004395418	0.001889465	0.004629	0.001836	0.055408	0.002538	0.339465	1		
DAYS_ID_PUBLISH	0.00047	0.005464892	0.004982027	0.0057319	0.001969	-0.0094	0.022194	0.414699	0.770174	1	
HOUR_APPR_PROCESS_START	0.076743	0.053618783	0.053588877	0.0627661	0.172814	0.000578	-0.00311	-0.00266	-0.00358	-0.00239	1

Target 1	AMT_INC	AMT_CREDIT	AMT_ANNUITY	AMT_GOOD	REGION_P	Age	DAYS_BIR	DAYS_EMI	DAYS_REG	DAYS_ID_	HOUR_AP
AMT_INCOME_TOTAL	1										
AMT_CREDIT	0.038131	1									
AMT_ANNUITY	0.046421	0.752194735	1								
AMT_GOODS_PRICE	0.037583	0.983102519	0.752699196	1							
REGION_POPULATION_RELATIVE	0.009135	0.069161087	0.07169025	0.0760489	1						
Age	-0.00217	-0.000232372	-0.003168461	-0.0008539	0.003594	1					
DAYS_BIRTH	-0.00261	0.011399456	0.004636069	0.0127598	-0.00493	-0.33357	1				
DAYS_EMPLOYED	0.001159	-0.000986766	0.002981652	-0.0023684	-0.00535	-0.01303	0.007961	1			
DAYS_REGISTRATION	-0.00357	-0.004253795	-0.010053685	-0.0047159	0.005255	0.055696	-0.00224	0.049882	1		
DAYS_ID_PUBLISH	-0.00331	-0.006635685	-0.006240236	-0.006151	0.00116	-0.00969	0.012627	0.060959	0.769051	1	
HOUR_APPR_PROCESS_START	0.013775	0.031781954	0.03123613	0.0443147	0.142744	0.011175	-0.0149	-0.00493	0.002334	0.002196	1

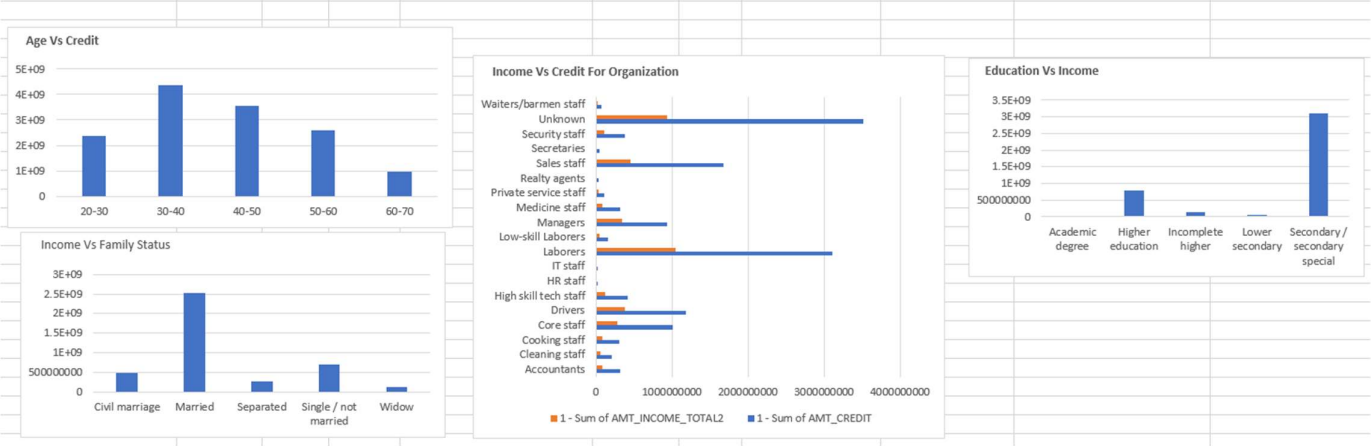
b. Income field is correlated with credit, annuity and goods_price.
c. Credit is highly correlated to annuity and goods price and annuity is highly correlated with goods price.
d. Similarly, days employed is correlated with days_resignation.

11. No Difficulties in Payment

a. To determine how much each group can comfortably repay a loan, we are comparing categorical numbers to credit and income.
b. According to the data below, customers who are married, between the ages of 30 and 50, have a home of their own, and have received special education or a higher education are customers who are not having any trouble repaying their loans.

12. Difficulties in Payment

a. In order to determine which group is most likely to have trouble repaying the loan amount, we have compared credit and income amounts with categorical characteristics including family status, age, and education.
b. We observed the same outcomes as no payment difficulties.



Results

- All of the pre-processing stage's necessary steps have been covered.
- As a result of data cleaning, we now have 46 columns and 307511 rows of final data.
- We have discovered outliers from numerical data using statistics and a box plot.
- Days_employed, credit, annuity, product price, and target 0 and target 1 have strong correlations.
- We can employ a variety of techniques, such as the confusion matrix, resampling, or SMOTE approach, to balance the data that we have in order to foresee acceptable results.

Excel File Drive Link:

https://docs.google.com/spreadsheets/d/1pzCBtSL-wYnRwEFXzE6IOq78paBHLYzr/edit?usp=drive_link&ouid=102714888778675675783&rtpof=true&sd=true