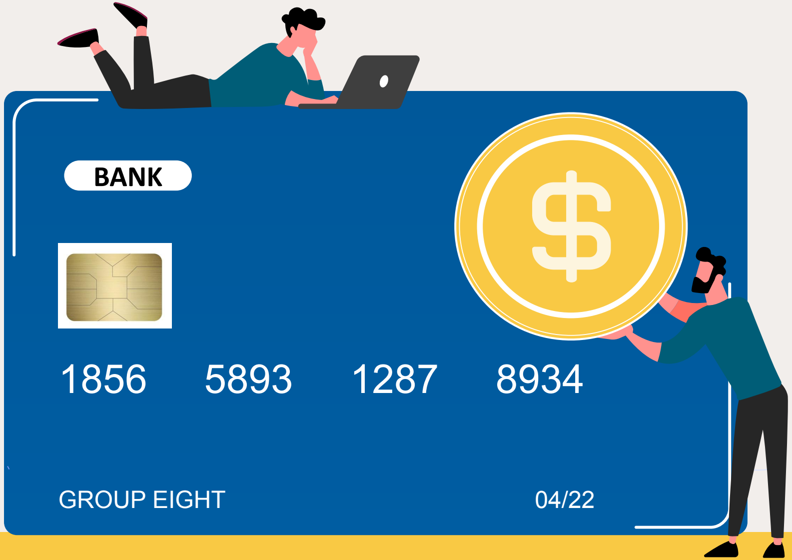
Bank Fraud Detection



Group Members: Jianan Liu, Mengyi Lyu, Weiyi Ni, Shiyuan Ru, Zhenwei Wei, Zhuojing Xie

Columbia University
in the City of New York

Scenario

Small-business lending fraud is a big deal!

- In 2019, a survey on lenders across all types reflects that 33% of SBA (small-business loan) lenders estimated that 1% of their loans were fraudulent.
- Findings reflect that, since the beginning of 2020, smaller banks, credit unions and digital lenders have been hit harder and harder by SBA fraud.

Our Plan

1. Explore and analyze data
2. Structure a 3NF schema
3. Implement the ETL process
4. Establish a comprehensive DBMS
5. Design customer interaction plans

Goals

1. Help clients better understand the driving factors behind loan defaults.
2. Quantify the default risk according to each borrower's financial profile
3. Reduce the SMB fraud for clients

Original Data

Data Structure

122 variables

307511 observations

Data type include float, int, object

Data Wrangling

Delete meaningless variable

Drop high correlation variable

```
df[['AMT_CREDIT','AMT_GOODS_PRICE']].corr()
```

	AMT_CREDIT	AMT_GOODS_PRICE
AMT_CREDIT	1.000000	0.986968
AMT_GOODS_PRICE	0.986968	1.000000

```
drop_list = df[["REGION_POPULATION_RELATIVE","AMT_GOODS_PRICE"]]
df.drop(labels=drop_list, axis=1, inplace=True)
```

Data Source

Our original data was downloaded from Kaggle
https://www.kaggle.com/datasets/mishra5001/credit-card

Data Sample

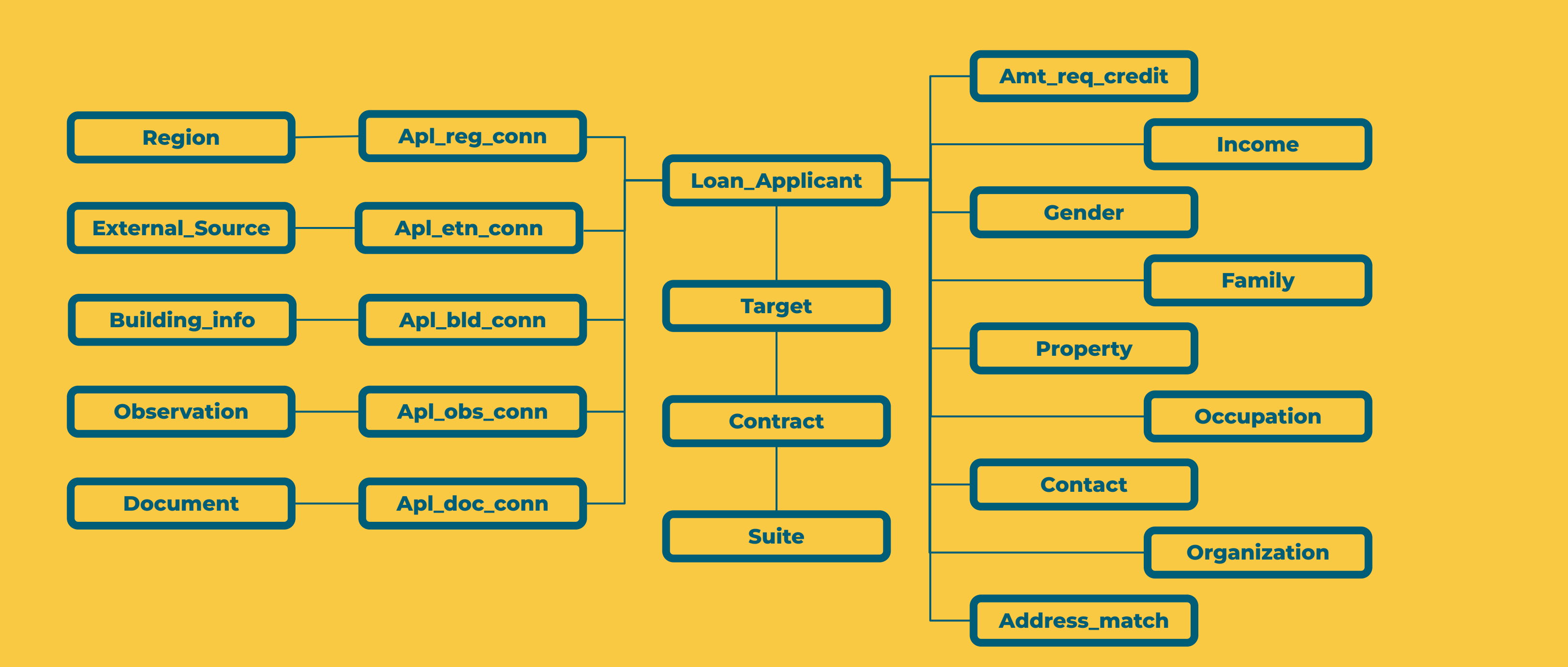
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
0	100002	1	Cash loans	M	N	Y	0
1	100003	0	Cash loans	F	N	N	0
2	100004	0	Revolving loans	M	Y	Y	0
3	100006	0	Cash loans	F	N	Y	0
4	100007	0	Cash loans	M	N	Y	0

5 rows × 122 columns

AMT_ANNUITY	...	FLAG_DOCUMENT_18	FLAG_DOCUMENT_19	FLAG_DOCUMENT_20	FLAG_DOCUMENT_21	AMT_REQ_CREDIT
24700.5	...	0	0	0	0	
35698.5	...	0	0	0	0	
6750.0	...	0	0	0	0	
29686.5	...	0	0	0	0	
21865.5	...	0	0	0	0	

Normalization Plan

- 1NF: Eliminate repeating groups in individual tables, such as building information and document.
- 2NF: Create separate tables for sets of values that apply to multiple records, dividing into 18 tables.
- 3NF: Eliminate fields that do not depend on the key, such as table region and external_source.



ETL Process

Extract

- **Dataset:** credit card fraud data
- Create database & engine
- Create 18 Tables aligning with normalization plan

Transform

- **Drop duplicates**
- **Impute missing value** with mean, medium, or mode
- **Drop some columns** based on the total number of null values
    - Threshold of 30%
- Use **sk_id_cur** as PK in the main dataframe **add incrementing integers** as PK in some subset data frame
- Merge into the main dataframe by using **merge function**

Load

- **Uniform the letter case** of all columns
- **Modify data type** and make sure all of them are correct
- Load the data into postgreSQL

Interaction Plan

Primary Goals:
Reduce Fraud Probability, Increase the Number of Users and Optimize Financial Product

User Side:
- Analyzing user characteristics
- Customer Insight

Product Side:
- Analyzing loan types and loan status
- Mitigate Risks

Which income type has the highest number of frauds? What is the average salary for each income type?

```
CREATE VIEW income_fraud AS
SELECT income.income_type, ROUND(AVG(income.annual_income_total),2)as average_income, target.target_name,
count(loan_applicant.sk_id_curr)as count,
round( 100 * COUNT ( * ) * 1.0 / SUM ( COUNT ( * ) ) OVER ( ), 2 ) || '%' as percentage
FROM income, target, loan_applicant
WHERE  income.income_id=loan_applicant.income_id and loan_applicant.sk_id_curr=target.sk_id_curr and target_name='1'
GROUP BY income_type, target_name
ORDER BY count DESC;
```

	income_type character varying (50)	average_income numeric	target_name character varying (50)	count bigint	percentage text
1	Working	163676.85	1	15224	61.33%
2	Commercial associate	188217.32	1	5360	21.59%
3	Pensioner	135556.94	1	2982	12.01%
4	State servant	164713.35	1	1249	5.03%
5	Unemployed	72000.00	1	8	0.03%
6	Maternity leave	58500.00	1	2	0.01%

What is the number of frauds and non-frauds in different loan types?

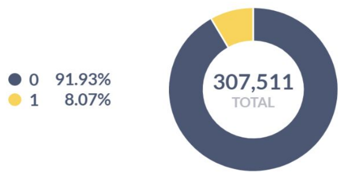
```
CREATE VIEW loan_fraud AS
SELECT contract.contract_name, target.target_name, count(contract.sk_id_curr)as count,
round( 100 * COUNT ( * ) * 1.0 / SUM ( COUNT ( * ) ) OVER ( ), 2 ) || '%' as percentage
FROM contract, target
WHERE contract.sk_id_curr=target.sk_id_curr
GROUP BY contract_name, target_name;
```

	contract_name character varying (50)	target_name character varying (50)	count bigint	percentage text
1	Cash loans	0	255011	82.93%
2	Cash loans	1	23221	7.55%
3	Revolving loans	0	27675	9.00%
4	Revolving loans	1	1604	0.52%

Demo

Credit Card Fraud

Number of late payments



Credit amount of loans and late payment

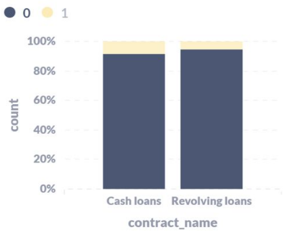
total	lower_loan_credit_amount	median_loan_credit_amount	higher_loan_credit_amount	target_name
282,686	21,634	74,940	186,112	0
24,825	1,429	6,697	16,699	1

Credit amount and late payment status

Borrowers with relatively large loans were more likely to be late on their payments.

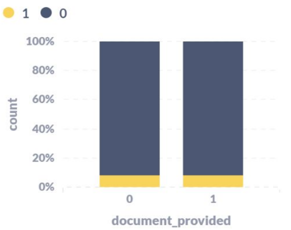
67.3% for late payment of large loan / 65.8% for on-time payment of large loan

Loan types and late payment



The probability of overdue payment for cash loan (8.3%) is higher than that for revolving loans (5.5%).

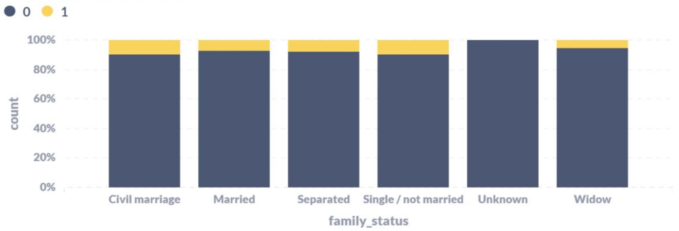
Loan procedure-document and late payment



There is no relationship between provision of relevant loan documents and late payment.

8.2% for late payment of document provided / 8.2% for late payment of document not provided.

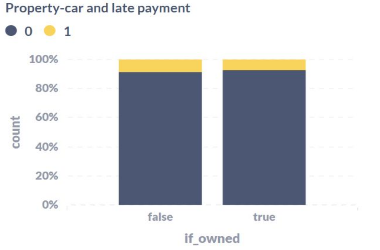
Family status and late payment



Borrowers in civil marriage (9.9%) and Single/Not married (9.8%) were more likely to be late on payments than those in other status, and widow was the least likely to be late on loans (5.8%).

Demo

Laborers(27.8%), especially low-skilled workers(17.2%), drivers(11.3%), waiters/bartenders(11.3%), security personnel(10.7%), cooks(10.4%), cleaners(9.6%), salespeople(9.6%), loan applicants in these categories are more likely to have a late payment.

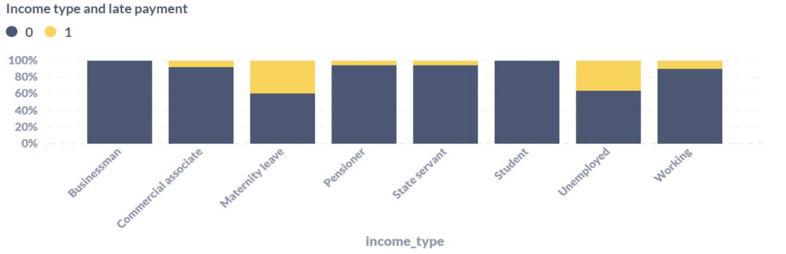


Borrowers who do not own cars or realty are more likely to fall behind on payments.

8.3% of borrowers who do not own a realty are late on their payments / 8.0% of borrowers who own realty are late on their payments.

8.5% of borrowers who do not own a car are late on their payments /7.2% of borrowers who own a car/cars are late on their payments

Borrowers who provided contact information(8.2%) were more likely to be late on their payments than those who did not(7.9%) .



Borrowers whose income type is maternity leave (40%)and unemployed(36.4%) are more likely to make late payments, while businessmen and students almost do not.

Conclusion

Our project aims to reduce credit fraud risk by establishing a comprehensive customer database management system, which allows the financial institution to identify and understand small business lending more effectively.

- <u>For the purpose of characterizing loan applicators precisely</u>, we collected large size real financial loan data to build our relational database and developed this relational database to the third normalization form, with 23 different factors (tables).
- <u>For the purpose of understanding small business lending better</u>, it is important to reveal the driving factors behind each loan default, and we achieved this by implementing a detailed customer interaction plan. In this plan, we designed 12 analytical procedures that derive valuable insights from both the product side and the customer side to the C-level officer.

Successfully carrying out our customer interaction plan will help the institution not only reduce the loan default risk of borrowers but also expand its user population.

Thanks!