The Knowledge Thinking Process

Individual Project: ABC Retail Bank Analysis

Master in Business Analytics and Big Data
Section 2
Ting-Lun, Fan

SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

What is the business problem of the company and how can it be addressed?

BUSINESS UNDERSTANDING

Bank can earn revenue from issuing credit card by…

Revenue from issuing credit card

Dominance income
- Interest income
- Annual payment
- Transaction fee
- Fine
- Commission fee from cooperated shops
- Additional income

Recessive income
- Account Expansion
- Asset Accumulation
- Private & public business synergy

BUSINESS OBJECT& PROBLEM

Object

- Increase revenue
- Reduce risk and speed up admission phase
- Develop a credit risk score to allow the bank to grant credit cards automatically.

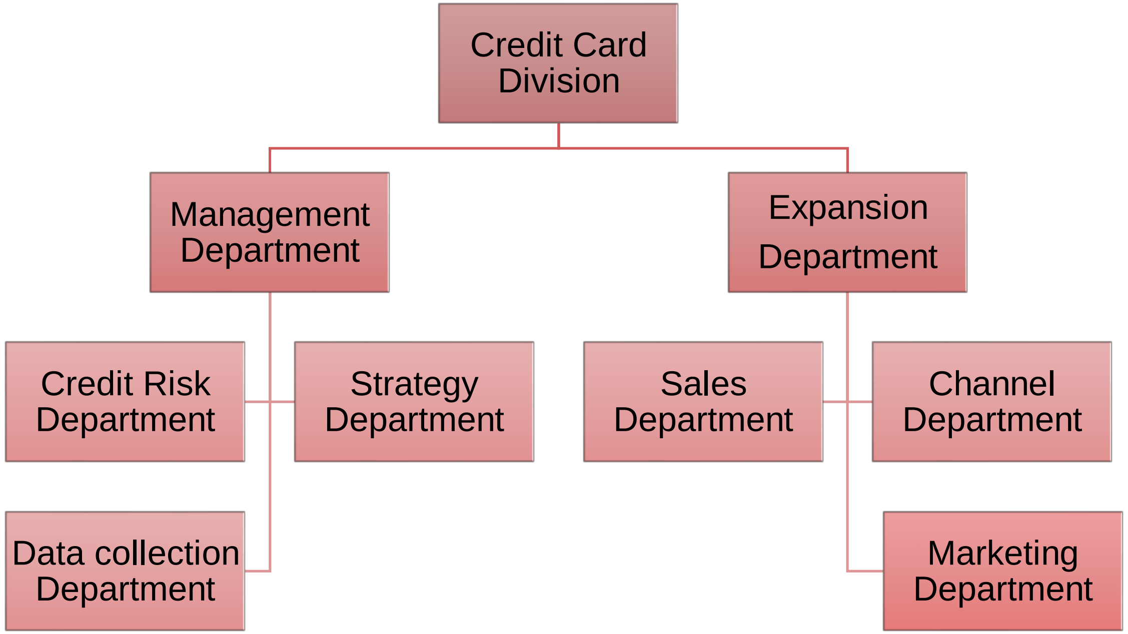
Problem

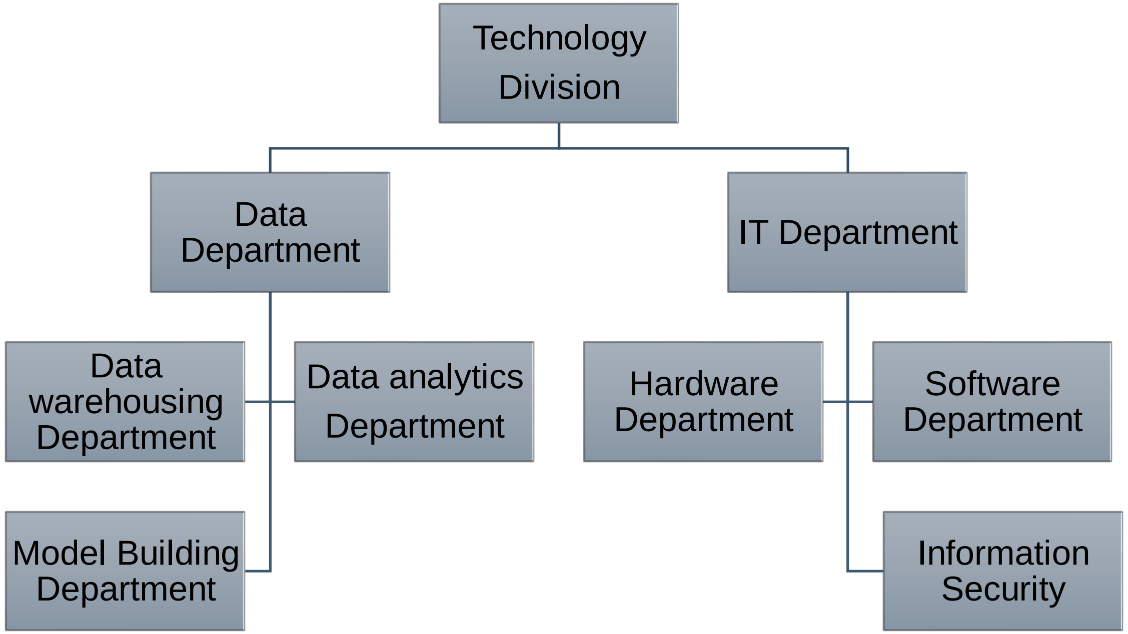
- Incomplete and Inconsistent dataset (missing value and unknown)
- Customer review/feedback could be an important attribute to the data analysis. This would be easier to understand the reason behind that. But this is not in the dataset.

Credit card revenue takes a huge part of the whole company. By cleaning and analyzing the dataset, ABC Bank can create a credit card prediction system to analyze the probability of paying back credits and fees. This system will allow ABC Bank to grant the credit cards to customer automatically. Not only speed up the process but also reduce the total cost, create better earning capability.

What Bank's areas should participate?

The cooperation between "Credit Card Division" and "Technology Division" is necessary in order to create better synergy and achieve the task.



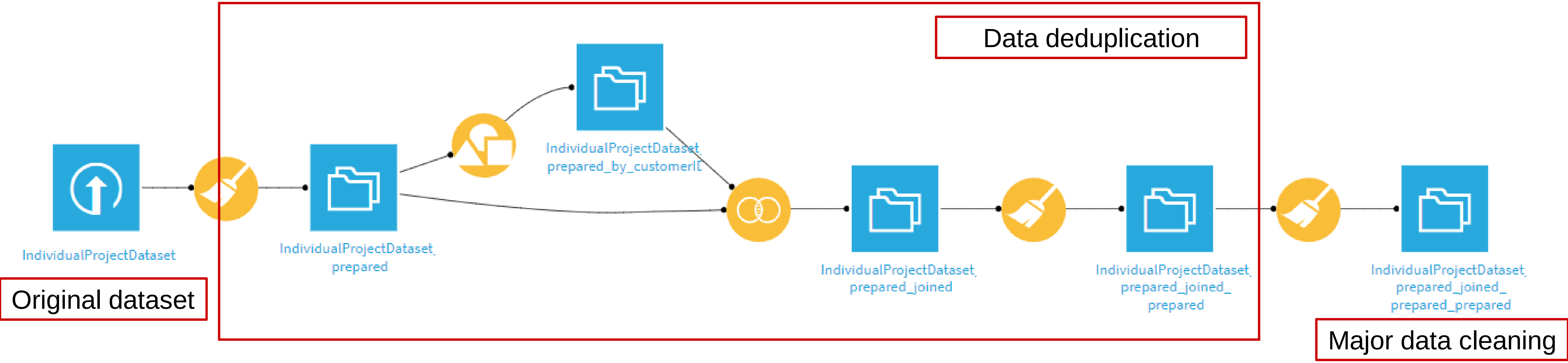


1. Credit Card Division needs to transfer their data correctly to Data department and keep the communication channel active in case there is any possible problem exist.
2. After received data, Data Department needs to start processing the data, including cleaning, understanding and analyzing.

3. After analyzing the data, Data Department need to start building the model to ensure the high accuracy rate of customer credit prediction.
4. After finishing the model, model will be sent it back to Credit Card Division to improve and detect the potential error. Active communication and cooperation is necessary. Then it can be implemented to test its performance.

How will the data be cleaned?

OVERVIEW FLOW



STEP 1 DEDUPLICATION(CUSTOMERID)

- **Create new dataset (the same as original one)**
- **Group dataset with only CustomerID and its count**
- **Join count of CustomerID with original dataset (new column "count")**
- **Add script: Remove rows where 2 ≦ count ≦ 15**
- **Remove those row (left only unique CustomerID)**

- **Original observation: 522939**
- **Duplications: 22077 (4.2% of total observations)**
- **By deleting those duplications in the beginning to ensure the consistency and avoid bias**
- **522939-22077=500862 (These are the unique observations)**

How will the data be cleaned?

STEP 2 OTHER DATA CLEANING PROCESS

Sex	1. Replace	• **Male => 1 Female => 0**
Status	1. Replace	• **Single => 1 Married => 2 Unknown => 3 Widower => 4 Divorced => 5**
Age	1. Clear “NA” value 2. Delete outliers outside 1.5 IQR (18~78) 3. Bin with customer range 4. Fill empty cells with “Mode” <u>1</u>	• **Customer bin range [0:17] => 0, [18:35] => 1, [36:53] => 2 [54:71] => 3, [72:89] => 4**
External Score	1. Clear “NA” 2. Fill with “Median” <u>649</u> 3. Normalized	• **Min-Max normalization (Build new column) Use formula: ((numval("externalScore")-1)/(995-1))*(1-0)-0**

indSimin indXlist indCreditBureau indInternet indBadDebt	1. Change “Meaning”	• **Change to Boolean**
Salary	1. Clear “Unknown” 2. Replace 3. Fill with “Mode” <u>3</u>	• **Replace None => 0, <650 => 1, [650,1000) => 2, [1000,1300) => 3, [1300,1500) => 4, [1500,2000) => 5, [2000,3000) => 6, [3000,5000) => 7, [5000,8000) => 8 <8000 => 9**
numLoans	1. Clear 2. Fill with “Mode” <u>1</u>	• **Clear if not a valid integer**
numMortgages	1. Clear 2. Fill with “Mode” <u>0</u>	• **Clear if not a valid integer**

How will the data be cleaned?

STEP 2 OTHER DATA CLEANING PROCESS

Channel	1. Replace	• **External Agent => 1** **Branch => 2** **Call Center => 3** **Recovery => 4** **App => 5** **Online => 6** **Unknown => 7**
inBadlocation	1. Change "Meaning"	• **Change to Boolean**
Previous	1. Replace	• **Normal => 1** **Restructuring => 2** **Refinancing => 3** **Default => 4** **Unpaid => 5**
SumExternalDefault	1. Clear "NA" 2. Fill empty with "Median" 0 3. Normalized	• **Z-score normalization (Build new column)** **Use formula:** **((numval("sumExternalDefault")-505.5)/11343)**
Target	1. Change "Meaning" 2. Replace	• **Change to Boolean** • **Replace 0(paid),1(unpaid) to 0(unpaid),1(paid)**

To run the correlation with right positive and negative relationship

STEP 3 CHECK&EXPLANATION

Data Cleaning Explanations

Consistency: The dataset has both numeric and categorical variables, integrate the whole dataset into single format (numeric) will ensure the consistency of data and easier to interpret and analyze with other model such as correlation.

Outlier: To lower the effect of extreme values and negative bias, by using normal method (±1.5IQR) to detect and delete outliers.

Normalization: By changing the dataset into same scale without losing its characteristics and distort the relative value, each feature is equally important. Min-Max normalization is used for "ExternalScore" (outliers not obvious) while Z-score normalization is used for "sumExternalDefault" (obvious outliers) based on their different characteristics.

Missing Value: Filling "Mode" for categorical variables and "Median" for continuous numeric variables. This can decrease the amount of bias in the dataset and ensure the data quality.

Exploratory analysis & Insight 1 (Final Dataset + Correlation Analysis + Heat Map)

ANALYSIS 1 **FINAL DATASET**

Column	Catagorical	Importance	Voice of market
Sex	• 61.6% Female • 38.4% Male		In retail bank, *sex* shouldn't be a criteria to analyze if female or male has higher tendency not paying the debt back. Even tough through correlation analysis we can see that Male has higher correlation with *externalscore* and *salary. Sex* should be treated fairly.
Status	• 86.3% Single • 13.5% Married • 0.2% Widower • 0.1% Divorced		Unknown data has been removed after data cleaning. Single stands the highest percentage. It is possible to see that most of their customers are single. ABC Bank can be more targeted to this group. And create suitable strategy to attract them.
Age	• 48.9% [18~35] • 36.5% [36~53] • 12.9% [54~71] • 1.7% [72~89]		It is obvious to see that most of their customer belongs to young age within 18~35 years old. The strategy of ABC Bank should be more tended to young age. Perhaps through social media and internet.
Normalized_ ExternalScore	• Range within [0,1] • Mean 0.59113 • Median 0.64688 • StdDev 0.26623 • Mode 0.65191		This show moderate positive correlation with target and strong negative correlation with *indBadDebt*. This can be interpreted that the higher external score has higher chance to pay credit fee and lower chance to be classified as sub-standard or lower-quality risk. This also show slightly negative correlation with other negative columns such as *indCreditBureau* and *indBadLocation.*
indXlist	• 94.8% Yes • 5.2% No		*indXlist* has strong negative correlation with *indInternet* (-0.43), the higher rate customer has published debt tends to have lower rate to search information online. This could be interpreted that ABC Bank should increase the use rate of their online service.
Salary	• 49.6% [1000,1300] • 12% [1500,2000] • 11.2% [1300,1500] • 10.1% [650,1000] • 17.1% the rest catagories		Salary tends to have moderate positive correlation with number of loans and mortgages, and also have moderate positive correlation with target, it means that customer who has higher salary tends to pay the fee on time.

	age	*indSimin*	*indXlist*	*CreditBur*
age	1			
indSimin	0.183556	1		
indXlist	0.028517	-0.07511	1	
indCreditB	0.023638	-0.00597	0.024195	
indInterne	-0.04045	0.133173	-0.4373	-0.0093
indBadDeb	-0.04233	-0.02589	0.024907	0.16501
salary	0.156086	0.061341	-0.05492	0.04173
numLoans	0.058138	0.050442	-0.02974	-0.073
numMortg	0.024886	0.01411	-0.02489	-0.0017
indBadLoc	0.012414	0.013402	0.01905	-0.0140
target	0.07947	0.080952	-0.06499	-0.0506
normalized	0.29251	0.106435	-0.05095	-0.2157
normalized	0.008434	-0.00142	0.004412	0.0619

Exploratory analysis & Insight 2 (Final Dataset + Correlation Analysis + Heat Map)

ANALYSIS 1 | **FINAL DATASET**

Column	Catagorical	Importance	Voice of market
numLoans **numMortgages**	• 66.4% has 1 loans • 18.4% has 0 loans • 97.4% has 0 mortgages		The *number of loans and mortgages* show positive correlation with *salary*. Besides *salary*, they both don't have too obvious correlation with other attributes. This can be seen as a pure observation, don't need to change strategy base on these two attributes.
Channel	• 52% External Agent • 16.5% Branch • 15.3% Call Center • 9.4% Recovery • 3.9% App • 2.4% Online • 0.6% Unknown		*Channel* is very important for ABC Bank since they want to sign agreement with shopping mall and having new channel to sale and marketing their credit card. It can be seen that from previous analysis the main customer group of ABC Bank is young people. Perhaps it is better for ABC Bank to improve their App and Online channel to attract more youngers. Also, data shows that 52% of channel are from External Agent. ABC Bank could try to switch their priority to App and Online, but more precise and advanced analytics need to be done in order to make sure this is right.
Previous	• 57.6% Normal • 20.1% Restructuring • 15.4% Refinancing • 3.5% Default • 3.4% Unpaid		57.6% of customer belongs to normal from last year, only 7% of total customer have negative problem. The delinquency rate on credit card of all commercial banks in 2019 is 2.56% in Q2 and average 2.5% in 2018. The delinquency rate of ABC bank is a bit higher. This could be their priority to solve this problem.
Normalized_ **sumExternalDefault**	• Min -0.044565 • Max 530.79 • Mean -0.0000072517 • Median -0.044565 • StdDev 1.0000 • Mode -0.044565		The data range of this attribute is very big. It also shows low correlation with other attributes. It needs further analysis to understand the reason behind.
Target	• 66.5% Paid • 33.5% Unpaid		*Target* has a moderate correlation with *normalized_externalScore* and *salary*. The higher external score tends to pay the fee. This observation makes sense.

Appendix: Correlation analysis & Heat map

	age	indSimin	indXlist	lCreditBure	indInternet	ndBadDebt	salary	numLoans	mMortgag	lBadLocati	target	zed_extern	l_sumExternalDefault
age	1												
indSimin	0.183556	1											
indXlist	0.028517	-0.07511	1										
indCreditB	0.023638	-0.00597	0.024195	1									
indInterne	-0.04045	0.133173	-0.4373	-0.00937	1								
indBadDeb	-0.04233	-0.02589	0.024907	0.165011	0.001319	1							
salary	0.156086	0.061341	-0.05492	0.041735	0.043082	-0.04559	1						
numLoans	0.058138	0.050442	-0.02974	-0.0736	0.031244	-0.0495	0.262742	1					
numMortg	0.024886	0.01411	-0.02489	-0.00171	0.015787	-0.02545	0.351623	0.086624	1				
indBadLoc	0.012414	0.013402	0.01905	-0.01402	-0.00571	0.057177	-0.11192	0.000113	-0.03161	1			
target	0.07947	0.080952	-0.06499	-0.05062	0.067161	-0.15968	0.202644	0.091797	0.071533	-0.0831	1		
normalized	0.29251	0.106435	-0.05095	-0.21575	0.019758	-0.49131	0.175658	0.052225	0.076113	-0.0418	0.230545	1	
normalized	0.008434	-0.00142	0.004412	0.06195	-0.0039	0.067615	0.011636	-0.00983	0.004188	0.002515	-0.01004	-0.06696	1

Appendix 2 Reference material

https://fred.stlouisfed.org/series/DRCCLACBS

https://www.zhihu.com/question/20387919

https://www.slideshare.net/phannithrupp/guideline-for-interpreting-correlation-coefficient

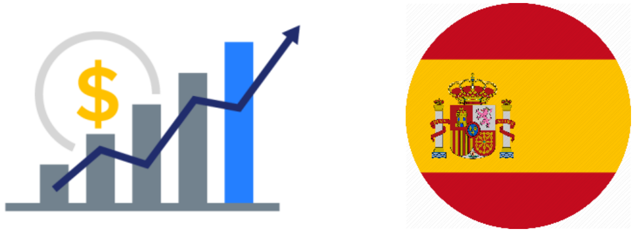
https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

https://medium.com/@claudehung1016/%E8%B3%87%E6%96%99%E5%89%8D%E8%99%95%E7%90%86%E5%AD%B8%E7%%BF%92%E7%AD%86%E8%A8%98-outlier-%E6%AA%A2%E6%9F%A5%E5%8F%8A-%E8%99%95%E7%90%86-98c6bc1821eb

Appendix 3 Company situation

1 Retail bank specialized in financial services to the residential sector

2 Increase Credit Card Revenue Shares in Spain Area



3 Sign Agreement with Shopping Center though different channels

4 Develop Credit Risk Score to Grant Cards Automatically