Big Data Analytics in Internet Finance

----A Case Study Of Alibaba Group's Customer Churn Rate During COVID-19 Pandemic

Zijun (Terrence) Shao

1. Problem Overview

- Alibaba Group is one of the most famous Internet company, which is founded by Jack Ma.
- Ant Financial, an affiliate of Alibaba, is the biggest Internet financial company in China.
- In October 2020, Ant Financial planned to raise $37 billion in the world's largest IPO, but unfortunately, the large deal was suspended by Chinese government in November due to risk consideration.
- As Internet finance is a new business, which applies high technology and data analysis to traditional finance, Alibaba still does not know how it differentiate from banks and other traditional finance, and whether covid-19 has any influences on their business and risk.
- Thus, Alibaba Group hope us to identify such problems by big data analysis.

2. Data Source And Project Objective

- **Data the company provided includes two files:**
- (1) First is an internal dataset, including required information when applying an account, like age, balance amount, gender, etc.
- The dataset includes 40 columns. The last one is a Boolean value of churn, where 1 represents the customer churns from the company and 0 means not any churn.
- (2)Second is an external dataset, including optional information when applying for accounts, like revenue range, the number of kids they have, etc.
- The internal and external dataset are linked by the first column, guest ID, which is unique and appears in both datasets.

- **Key Objectives:**
- (1) Identify the relationship between key variables and churn rate and make a comparison with traditional finance on churn risk. Thus, we will give our recommendation on management and strategy in internet finance.
- (2) Use different prediction algorithms to identify a model that predict churn rate the best in Internet finance and give our recommendations on data maintenance to make a better-quality data source.

3. Project Solution

- **There are two characteristics of the original datasets**
- (1) So many missing data in the external data
- As the variables in external dataset is optional to offer. Many customers tend not to offer more information to Ant Financials. Thus, how to estimate the missing value is quite important.
- (2) Containing both numeric and un-numeric data
- As there are many non-numeric columns, identifying how to encode the non-numeric to numeric data is one of the most important issue.

- Based on the objectives and dataset characteristics, the solution includes 3 steps.
- **(1) Data import and analysis**

① Import data: two csv files were imported – bankChurn.csv and ExternalData.csv. , represents internal and external data, respectively.

② Data distribution analysis: As we need to identify the relationship between variables and churn; we need to observe the data distribution first.

3. Project Solution(Continued)

- (2) Data preprocessing
  - **Visualize attrition rate**: two functions were defined – NumVarPerf and CharVarPerf
  - **Dealing with missing values**: a function called "MakeupMissing" was defined. The inputs of the function are df, col, type and method.
  - **Encoder**: a function called "Encoder" was defined in which categorical values would be converted into numbers.
  - **Handle the dataset with defined functions**
  - **Train test split**: all independent variables and the dependent variable were split into training set (70%) and testing set (30%).
- (3) Using different machine learning algorithms to determine the best model
- Based on the data characteristics, we make predictions with the models in the table

a. Logistic regression	*b. Decision trees*	*c. Random forest*
d. Knearest numbers	*e. Kmeans*	*f. PCA*
g. Boosting	*h. XGB*	*i. Simple Pipeline*

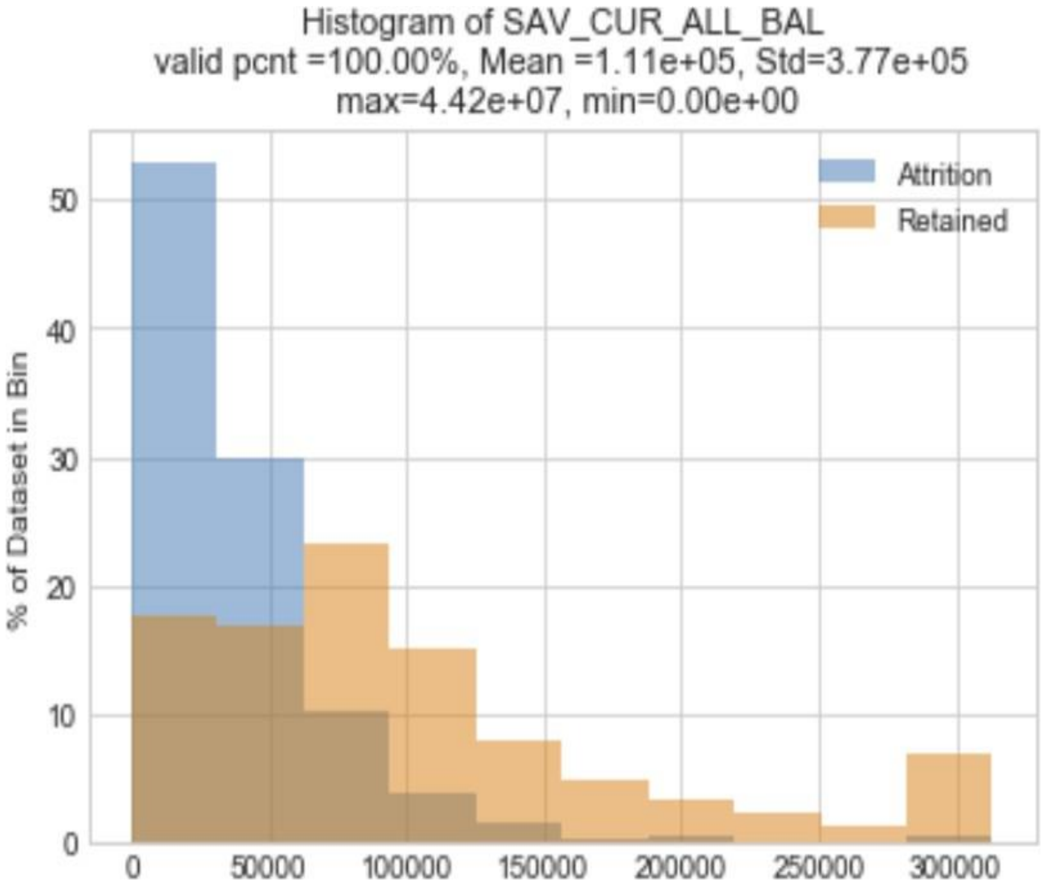
4. Results And Analysis - Objective(1)

- (1) Relationship between churn and savings

The chart on the right is a data distribution between deposit amount and churn.
Blue columns represent churn rate while yellow represents saving amount.

For saving (yellow columns) distribution:
It conflicts with our assumption. As Internet finance is a new business, it is more risky than traditional financial institutions. Customers with more saving tend to not choose internet finance.
However, at 300,000 level, the number of saving accounts is even higher than the sum number of 200,000 and 250,000 level.
This indicates that wealthy people may feel more confident about high-tech finance.

The relationship with churn rate(blue columns):
This is same as traditional finance. People with less deposit in the company tend to churn.



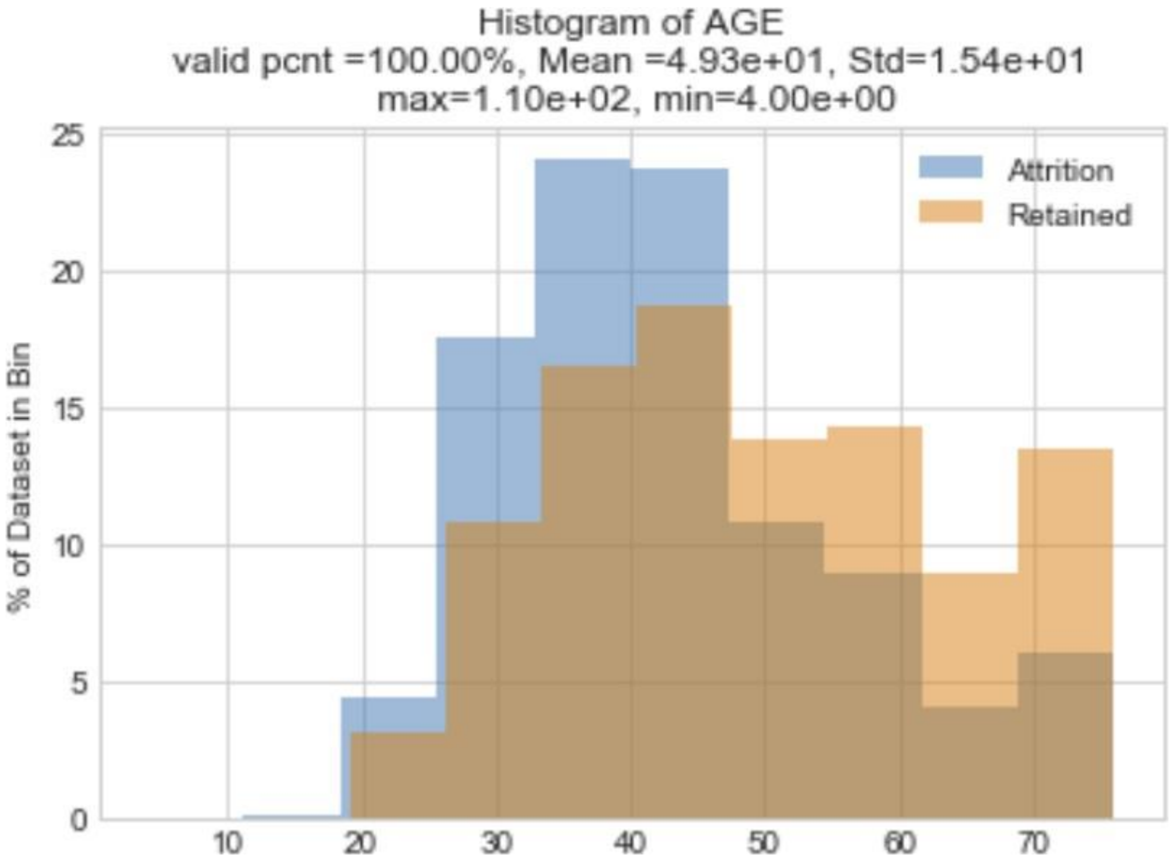
4. Results And Analysis - Objective (1)

- (2) Relationship between churn and ages

The chart on the right is a data distribution between customers' age and churn rate.
Blue columns represent churn rate where yellow represents age.

For age (yellow columns) distribution:
It conflicts with our assumption. High-technology tends to be more popular among young people. However, in this distribution, middle-aged uses internet finance the most.
One of the possible reasons is that during Covid-19 pandemic, most people don't want to go to traditional banks for financial services to keep social distance. Thus, many people turned to internet finance, during this special time.

The relationship with churn rate(blue columns):
This is not the same as traditional finance. Yonge people tend to be more loyal to Internet finance and the burn rate is low.



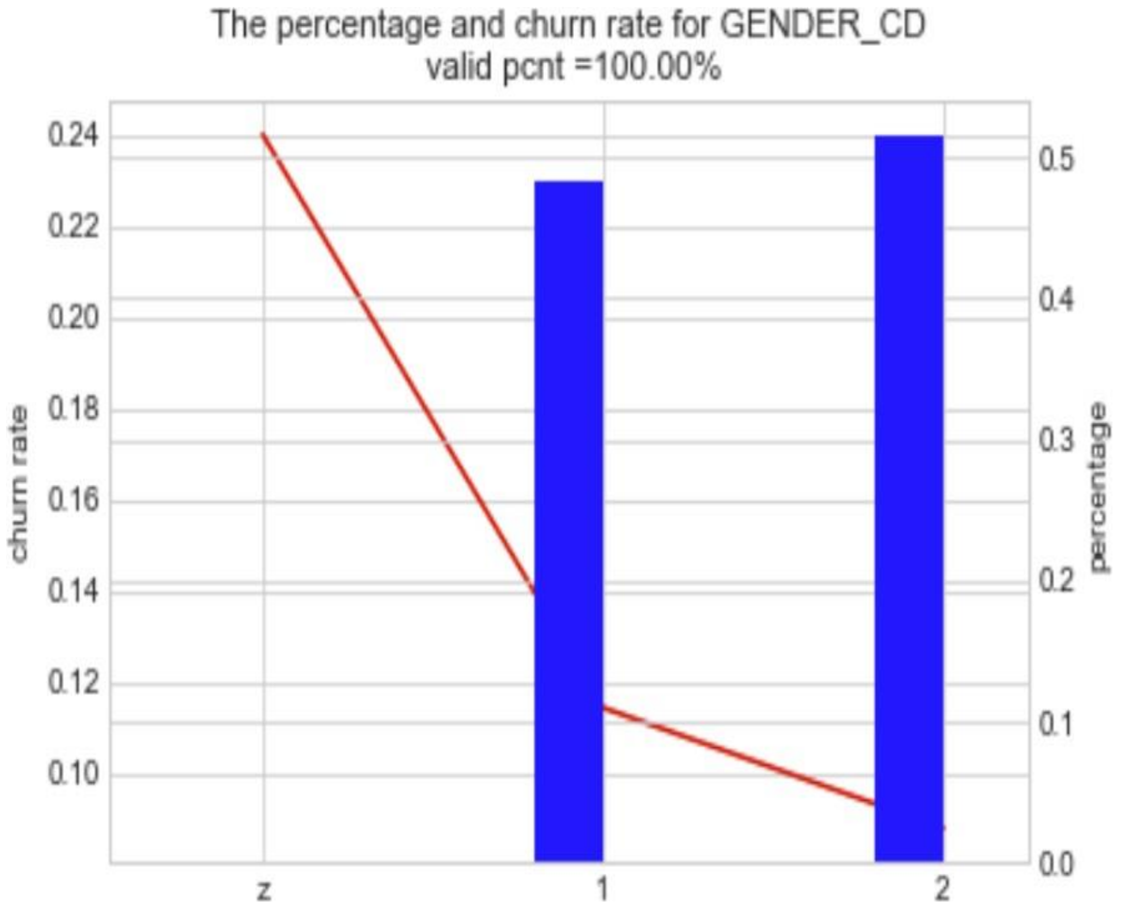
4. Results And Analysis - Objective (1)

- (3) Relationship between churn and gender

The chart on the right is a data distribution between gender and churn. 1, 2 and Z represent men, women and unknown gender.
Red line represents churn rate and blue columns represents gender.

For gender (blue columns) distribution:
It conflicts with our assumption. We assume that men are more likely to choose Internet finance as men are more likely to accept high-tech.
However, the gender is about equally-weight, and women's number is even a little higher.

The relationship with churn rate (red line):
This is the same as traditional finance. People who are reluctant to tell personal information tend to be less loyal to a company, thus with a higher churn rate.



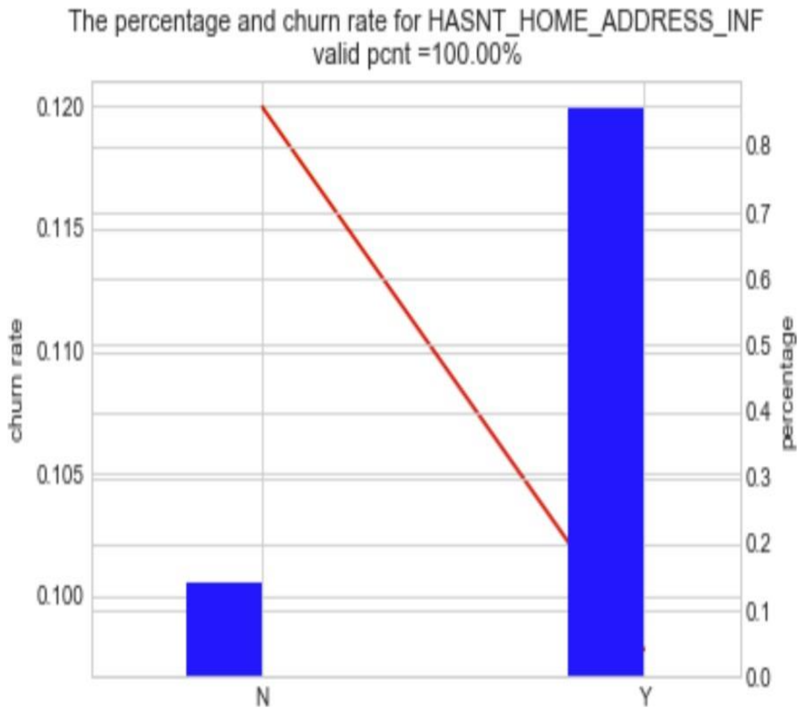
4. Results And Analysis - Objective (1)

- (4) Relationship between churn and contact address

On the right is a data distribution between contact address and churn. Y means company have the customer's address information, while N is not. Red line represents churn rate and blue columns represents whether a address exists.

For address (blue columns) distribution:
It conflicts with our assumption. For optimal information, we thought customers tend not to answer. However, most people offer the information, indicating they are loyal to the company.

The relationship with churn rate (red line):
This is the same as traditional finance. People reluctant to tell the personal information tend to be less loyal to a company, thus with a higher churn rate.



Recommendations:
In general, the customers' characteristics are mainly the same as traditional finance. Thus, traditional finance's management, marketing and other experience can also be useful in Internet finance.

4. Results And Analysis - Objective (2)

Algorithms	Score
Logistic Regression	0.894065
Decision Trees (Gini)	0.898318
Decision Trees (Entropy)	0.897351
Random Tree	0.902184
KNN (uniform)	0.902184 (n = 13)
KNN (weighted)	0.900251 (n = 19)
Boosting	0.899671
XGB	**0.902377**

PCA / Without PCA	Scores
Logistic Regression with PCA	0.892518
Logistic Regression without PCA	0.894065
Random Forest with PCA (Gini)	0.881500
Random Forest without PCA (Gini)	0.901991
Random Forest with PCA (Entropy)	0.884399
Random Forest without PCA (Entropy)	0.901217

In general, the accuracy rate of each model is within the same level, ranging from 89% to 91%。
For a problem to predicting churn rate, 90% score is not bad. This indicates our encoding and other reprocessing ways for the data is effective and could be a reference for further studying.

Specifically, XGB and Random Forest without PCA (Gini) perform the best, with 0.902377 and 0.901991 score.

Recommendations:

- Create data-driven models (XGB or Random Forest)
- Create User Profiles (Cluster Techniques)
- Effectively allocating resources to different client groups
- Keep data collection and analysis