Big Data Analytics in Internet Finance

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

Zijun (Terrence) Shao

OUTLINE

- 1 Problem Description
- 2 Solution Summary
- 3 Solution Details
- 4 Results and Recommendations
- 4.1 Results
- 4.2 Recommendations

1. Problem Overview

- **1.1 Background and problem:**
- 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 high-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.
- Alibaba Group hopes us to research on this problem. The key objective is two.
- (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 collection to make a better-quality data source.

1. Problem Overview

- 1.2 Current solutions:
- Internet finance is quite a new concept that few companies set foot in this field. Thus, there is still little experience on how to analyze problems.
- Now the company almost relies on analytical methods in traditional finance to search the churn risk. However, they are not sure whether the traditional methods can still apply to Internet finance.
- 1.3 Resources available and the importance of implementing new method
- Ant Financial has been in the Internet finance field for more than 2 years. They have billions of data about customers' information. Thus, with the data, we can get result through statistics analysis and machine learning.
- These new methods are important, because Internet finance is trying to apply new technology to traditional finance. We are not sure whether traditional method is still effective. Thus, we need to check it.
- 1.4 Cost with unsolvedproblem
- With unsolved problem, the company could lose its market shares. For example, if Ant Financial did not identify and predict their churn factor precisely, and markets at wrong targets. It could lost tons of profit.

2. SOLUTION SUMMARY

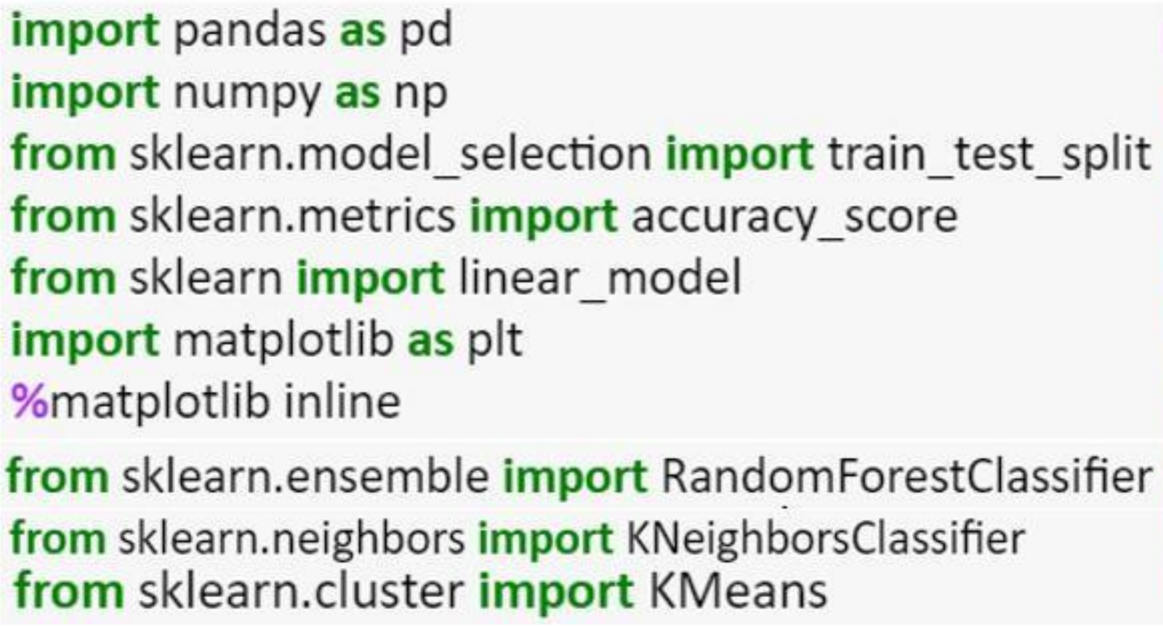
- 2.1 Overview of the solutions:
- (1) data import and analysis
- In this part, we first import the related data. Then, we draw the distribution of different variables, to have a general idea of the total dataset.
- (2) data preprocessing:
- In this part, we define a function to deal with the original dataset. The objective is to handle with two problems: filling in missing value and encoding non-numeric columns to numeric columns
- (3) using different algorithm to make prediction:
- In this part, we use different kinds of model to identify the best model with highest prediction score.
- 2.2 Major results:
- (1) Distribution charts of different variables
- (2) Get the accuracy score of different models

2.3 Potential improvement:

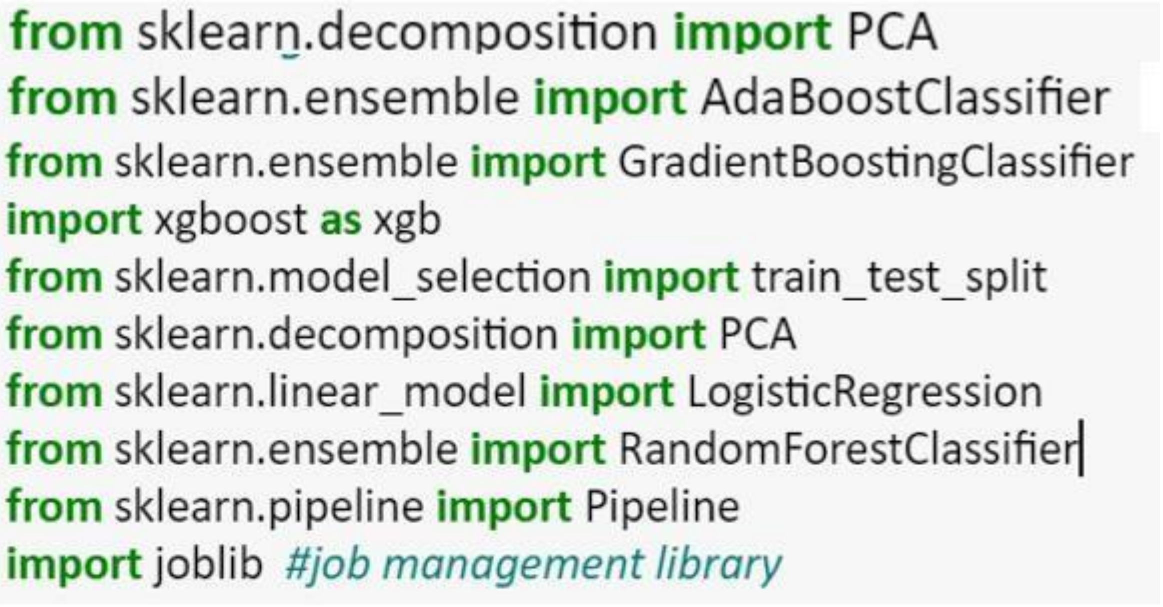
The potential improvement is in the data preprocessing. A better processing with missing value and non-numeric columns can result in a better score in the third part.

2. SOLUTION SUMMARY

- 2.4 Tools used and required
- The related modules are listed at the bottom of the slides. Generally, most of module of machine learning algorithms should be used.
- 2.5 Projected timeline
- The project is estimated to achieved in 1 months. The most important part is to determine and write appropriate function to deal with missing data and encode with non-numeric columns, which takes us most of time.



```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import linear_model
import matplotlib as plt
%matplotlib inline
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.cluster import KMeans
```



```
from sklearn.decomposition import PCA
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
import xgboost as xgb
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline
import joblib  #job management library
```

3. Solution Details

- 3.1 assumption:
- The most important assumption is that the data is all within recent months. Because of Covid-19, the dataset could be quite different as before. If many data from previous year are included, the prediction model cannot work well.
- 3.2 data gathering and import
- two csv files were imported – bankChurn.csv and ExternalData.csv. , represents internal and external data, respectively.
- (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.

3. Solution Details

- 3.3 Data cleanup and normalization:
- A function called “MakeupMissing” was defined whose inputs are df, col, type and method. Df is the dataset containing columns with missing value. Col is the column in the dataset. Type is the type of column which should be continuous or categorical. Depending on the type of column, different approaches would be taken. For continuous columns, if there are values greater than mean + 3*sigma, they will be replaced by mean + 3*sigma. Method refers to the method of filling missing values, which would be specified by the user and is either random or mean.
- 3.4 Data transformation and conversion:
- a function called “Encoder” was defined in which categorical values would be converted into numbers. The principle is that the corresponding number is based on the categorical value appearance rate in thecolumn.
- 3.5 Handle the dataset with defined functions:
- firstly, a variable called indepCols was defined which is the combined dataset excluding the dependent variable “CHURN_CUST_IND” and the ID variable “CUST_ID”. Then, a for loop was used to loop over each column in the indepCols variable. After removing the column with only one variable and eliminating the noise, we applied the MakeupMissing function (if there is missing value in the column) and Encoder function (for categorical variables). Finally, the dataset was double checked to ensure there was no missing values.

3. Solution Details

- 3.5 Machine learning setup:
- For this project, we use supervised machine learning mostly
- Train test split: all independent variables and the dependent variable were split into training set (70%) and testing set (30%).
- 3.6 Machine learning algorithm:
- supervised learning was applied for the problem. Algorithms used include logistic regression, decision trees, random forest, KNN, K means, PCA, boosting, XGB and pipeline.
- ***a. Logistic regression***
- As we need to predict a Boolean value, linear regression was not appropriate because we did not want to get decimals. Logistic regression was performed and accuracy score was examined.
- ***b. Decision trees***
- Both criterion – “gini” and “entropy” were used. Random state was set to be 100. We set maximum depth = 3 and minimum sample leaf = 5 to avoid the problem of overfitting. Then the accuracy score was examined to see how well the model performed.
- ***c. Random forest***
- Random forest can help solve overfitting problem that decision tree has. We used default parameters of the random forest algorithms and checked the accuracy score.

3. Solution Details

- 3.6 Machine learning algorithm(continued):
- ***d. K nearest numbers***
- We used a loop to iterate over the value of k from 1 to 25 and applied KNeighborsClassifier for each k value to find the k with highest accuracy score. For each loop, weights were set to be “uniform” and “distance” to see which method gave better results.
- ***e. K means***
- For K means, we set the number of clusters equal to 5 and random state equal to 100 and then check the local of each cluster.
- ***f. PCA***
- We used Principal Component Analysis to reduce dimensions. We set the number equal to five and checked how much of the variation these new five variables explained.
- ***g. Boosting***
- Both AdaBoost and Gradient Boost were employed. Random state was set to be 100.
- ***h. XGB***
- Gradient Boosting is computationally intensive and XGBoost speeds the process. It is also more logical and uses a more efficient algorithm. We checked the accuracy score to see whether it preformed better than the Gradient Boosting.
- ***i. Pipeline***
- Six pipelines were used, in which we employed three classification methods – Logistic regression, random forest with gini criterion and random forest with entropy criterion – each with and without PCA applied first. We then found the best pipeline and saved the result to file so that we could load it again next time we use it.

3. Solution Details

- 3.7 Additional models and visualization tools:
- Three visualization functions are defined:
- (1) plot_distribution: it is defined to see the distribution of numerical and categorical column. For numerical column, we wanted it to be Normally distributed because it is an assumption for regression analysis.
- (2) NumVarPerf: it is used to show the relationship between attrition rate and different values of corresponding numerical columns.
- (3) CharVarPerf: it is used to show the relationship between attrition rate and different values of corresponding categorical columns.
- 3.8 Input set and predict set:
- There are 65 variables in input set to predict one variable – churn or not. Also, the number of rows is 17,217. Both the number of columns and rows are large enough to make a prediction.
- 3.9 libraries to import
- The related libraries are shown in Slides 5

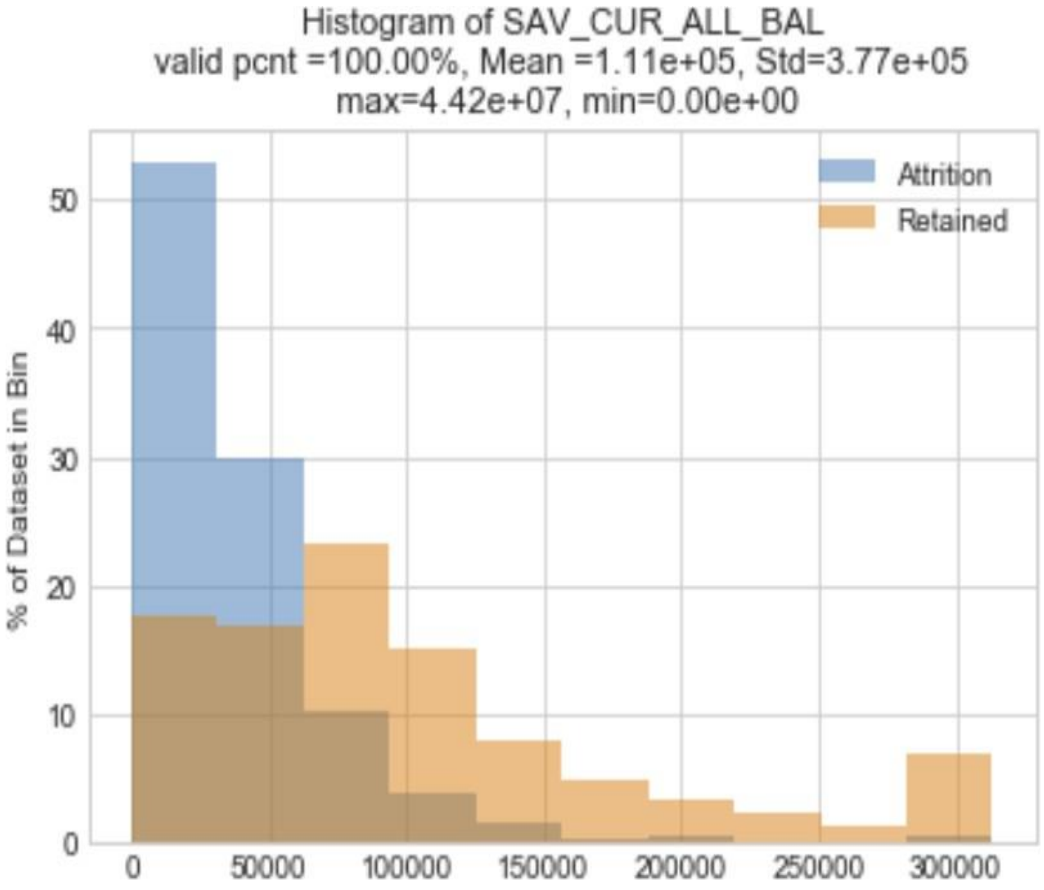
4. 1 Results - Objective(1)

- (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.



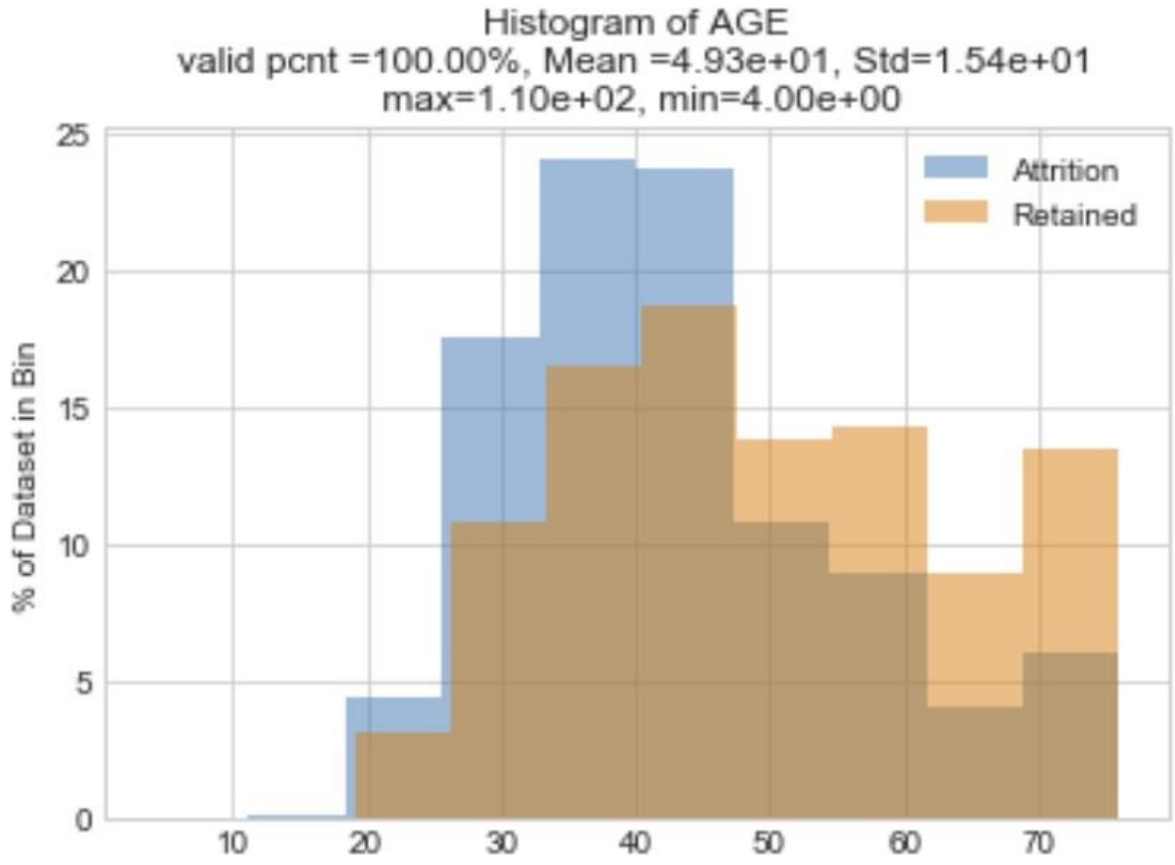
4. 1 Results - Objective (1)

- (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 islow.



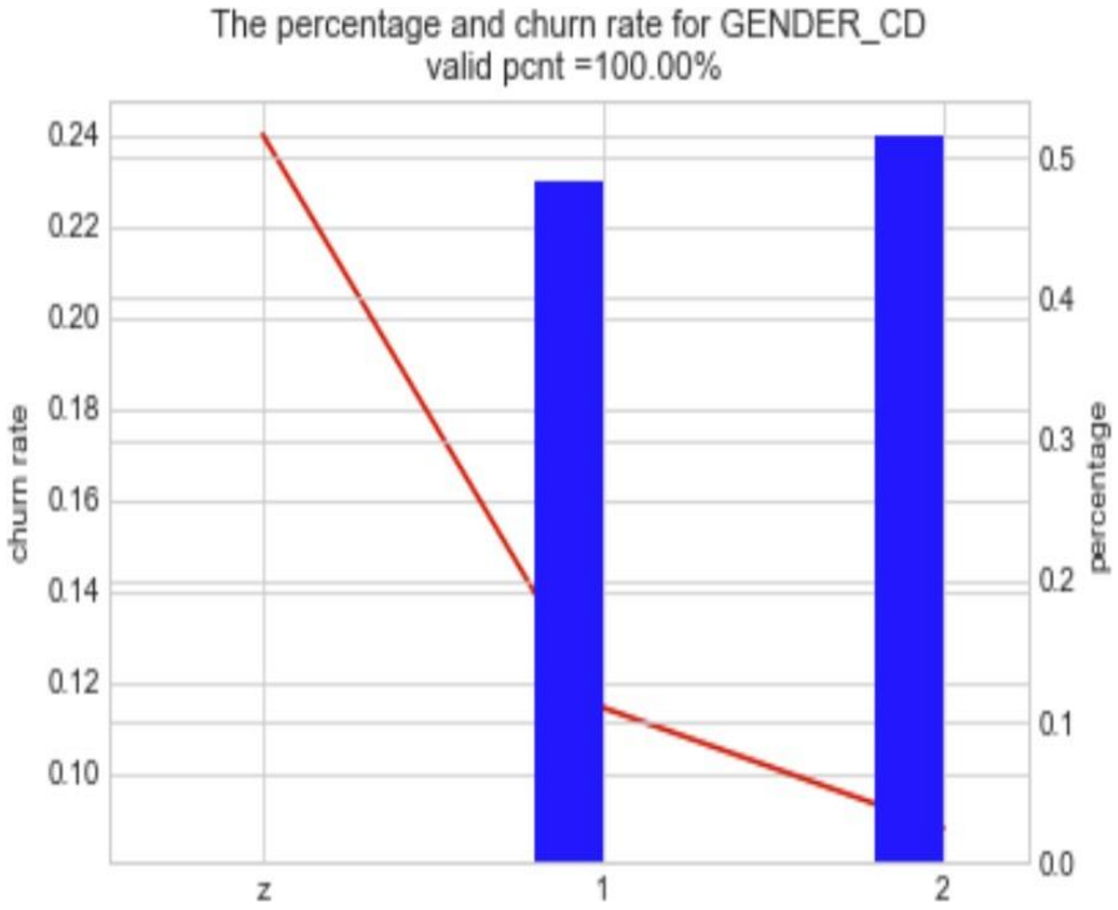
4. 1 Results - Objective (1)

- (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.



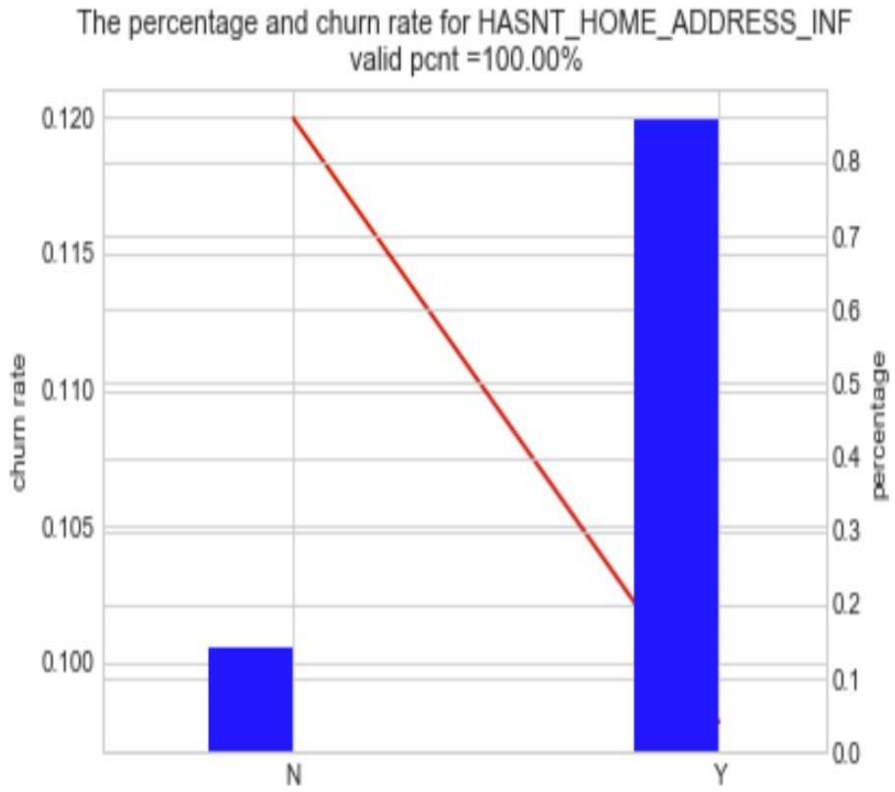
4. 1Results - Objective (1)

- (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.



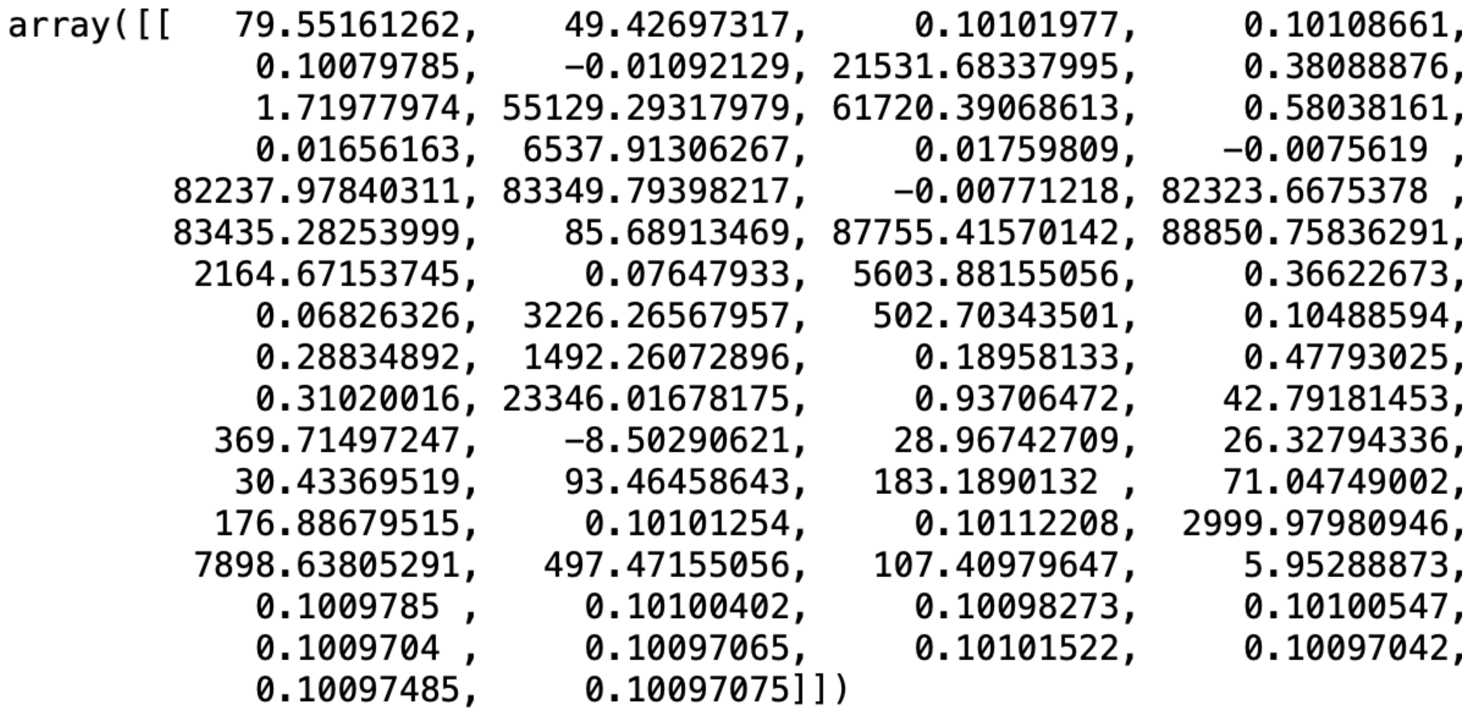
4. 1Results - Objective (2)

	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**

- In general, we use the testing dataset to estimate accuracy scores (without PCA). From the table above, we can find that XGB algorithm provides the best accuracy. Other algorithms like KNN and Random Tree also provide a good prediction.

4. 1 Results - Objective (2)

- **Results of Cluster Techniques (K-Means)**
- We set the number of clusters to 5 and we got five cluster centers. The typical characteristics for the first group clients are shown below:



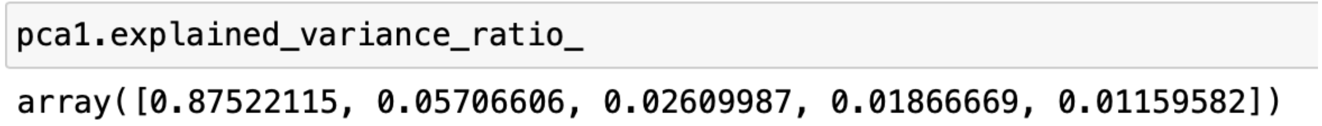
```
array([[   79.55161262,    49.42697317,     0.10101977,     0.10108661,
            0.10079785,    -0.01092129, 21531.68337995,     0.38088876,
            1.71977974, 55129.29317979, 61720.39068613,     0.58038161,
            0.01656163,  6537.91306267,     0.01759809,    -0.0075619 ,
        82237.97840311, 83349.79398217,    -0.00771218, 82323.6675378 ,
        83435.28253999,    85.68913469, 87755.41570142, 88850.75836291,
         2164.67153745,     0.07647933,  5603.88155056,     0.36622673,
            0.06826326,  3226.26567957,   502.70343501,     0.10488594,
            0.28834892,  1492.26072896,     0.18958133,     0.47793025,
            0.31020016, 23346.01678175,     0.93706472,    42.79181453,
          369.71497247,    -8.50290621,    28.96742709,    26.32794336,
           30.43369519,    93.46458643,   183.1890132 ,    71.04749002,
          176.88679515,     0.10101254,     0.10112208,  2999.97980946,
         7898.63805291,   497.47155056,   107.40979647,     5.95288873,
            0.1009785 ,     0.10100402,     0.10098273,     0.10100547,
            0.1009704 ,     0.10097065,     0.10101522,     0.10097042,
            0.10097485,     0.10097075]])
```

- 49 years old
- OPEN_ACC_DUR is 79

- Customers are classified into five different groups; we then might be able to develop different strategies for different groups of clients.

PCA (DIMENSION REDUCTION)

- We reduce variables in the dataset to 5 by using dimension reduction:



```
pca1.explained_variance_ratio_
```

```
array([0.87522115, 0.05706606, 0.02609987, 0.01866669, 0.01159582])
```

PCA / Without PCA	Scores
Logistic Regression withPCA	0.892518
Logistic Regression withoutPCA	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

- However, the PCA techniques did not improve the model
- Algorithm with best accuracy: Random Forest Gini without PCA

4. 2 Recommendation

(1)How to improve currentbusiness:
Based on distribution chart, we can find that currently, the internet financial business is still similar to traditional finance in terms of churn rate.
The possible reason is that internet financial still in its initial stage. Everything is still derived from traditional finance. So experience from traditional finance can still apply into the business.
Another reason is about Covid-19. Because of the pandemic, traditional life has changed a lot. More people tends to stay at home, which incentive more people from traditional finance to turn to internet finance.

4. 2 Recommendation

- (2) Recommendation for the next steps:
- However, this tend cannot keep in the long-run. Because of the totally difference between traditional and internet finance, they will diverge in the future.
- Machine learning is an efficient way to detect the trend and identify the key factors which influence churn rate.
- Based on our project, We would recommend Alibaba to create its own data-driven models by using XGB algorithm or Random Forest Gini without PCA to predict the churn.
- Unsupervised algorithm such as K-Means cluster could be used to group clients and create User Portraits for the groups. Then, the company knows the amount of resource should be allocated to different groups based on their churn rate and potential values.
- If the client is labelled as churn, the company may provide him/her with more promotions or marketing efforts to reduce the churn rate.
- The key point is the dataset should be large and accurate

- XGB and Random Forest Gini performance the best
- Although the number of variables is large, the dimension reduction technique is not necessary for our model training
- Our analysis could be improved by selecting more representative dataset and more refined data processing
- The company would use our recommendations to improve their churn rate analysis and resources allocation

4.3 FURTHER THINKING

- (1) Potential saving:
- Though in the short run, the saving is not obvious. However, in the long run, good data maintenance can help the company to identify key factors to influence churn rate. As a result, it will improve its profit and decrease cost.
- (2) Projected timeline and implementation of the solution.
- In our project, the data is about 70 variables with 17,000 rows. It takes us about one month to build up appropriate preprocessing data function and machine learning model. However, with more data in the future, the model will be more complicated with more time-consuming.
- (3) Summary of ourlearnings
- What we learned is that different project will have different methods to set up functions and method.
- When preprocessing data, we find using what we used in lecture cannot result in a good dataset. As this dataset is quite different from the dataset we used in the lecture.
- Thus, there is not a principle or standard to deal with all projects. Thinking will always be the first thing.

Q&A
THANK YOU!