IDENTIFYING CUSTOMERS

XICHEN TANG

Abstract. In this challenge, we asked to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted. The data provided for this competition has the same structure as the real data we have available to solve this problem.

Contents

Date: January 17, 2020.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduce

At Santander our mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals.

Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan?

1.1. **Data.**

The data mainly includes the customer's label number, whether there will be specific transactions in the future, 1 and 0 represent yes or no respectively, and then various relevant information provided by the bank, without specific name.

: ID˙code
customer ID

: target
Whether it will trade in the future, 0 means no, 1 means yes

: var˙0
Relevant information provided by the bank

: var˙1
Relevant information provided by the bank

: ...
Relevant information provided by the bank

: var˙199
Relevant information provided by the bank

1.2. **Train Data and Test Data.**

- train˙data

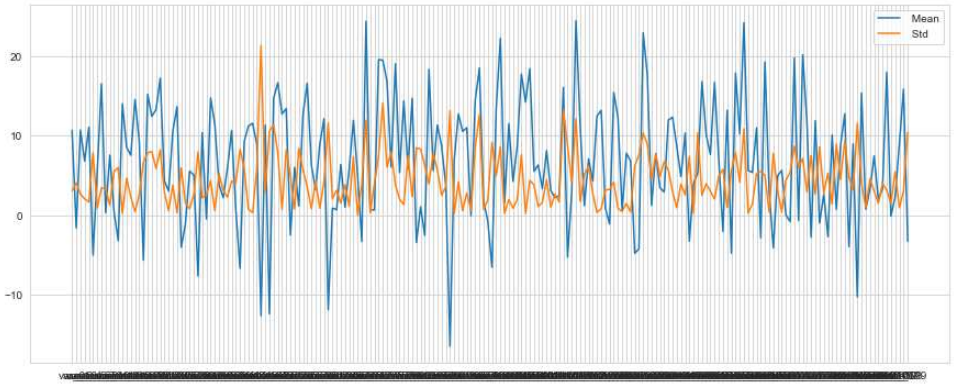
ID˙code	target	var˙0	var˙1	...	var˙198	var˙199
train˙0	0	8.9255	-6.7863	...	12.7803	-1.0914
train˙1	0	11.5006	-4.1473	...	18.3560	1.9518

- test.csv

ID˙code	var˙0	var˙1	...	var˙198	var˙199
test˙0	8.9255	-6.7863	...	12.7803	-1.0914
test˙1	11.5006	-4.1473	...	18.3560	1.9518

- Missing Values
  train data missing values? False
  test data missing values?False
  There are no null values in our dataset. This is good thing else we need to handle the missing values.

- Calculate Avg and Std by describe()


Mean
Std
20
10
0
−10

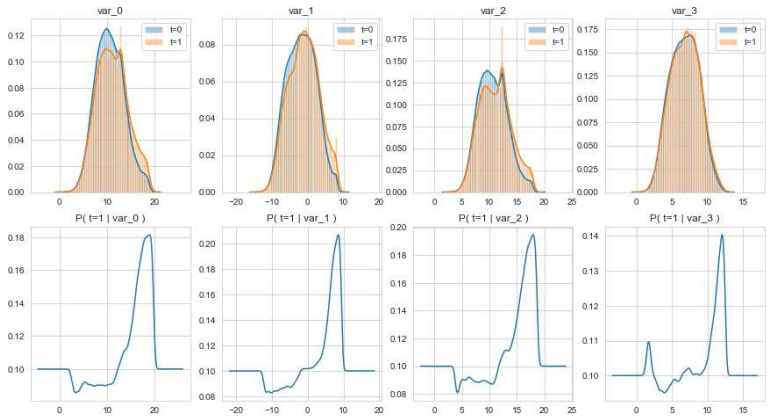
2. Method

We use different methods to predict, and finally through comparison, we get the best value

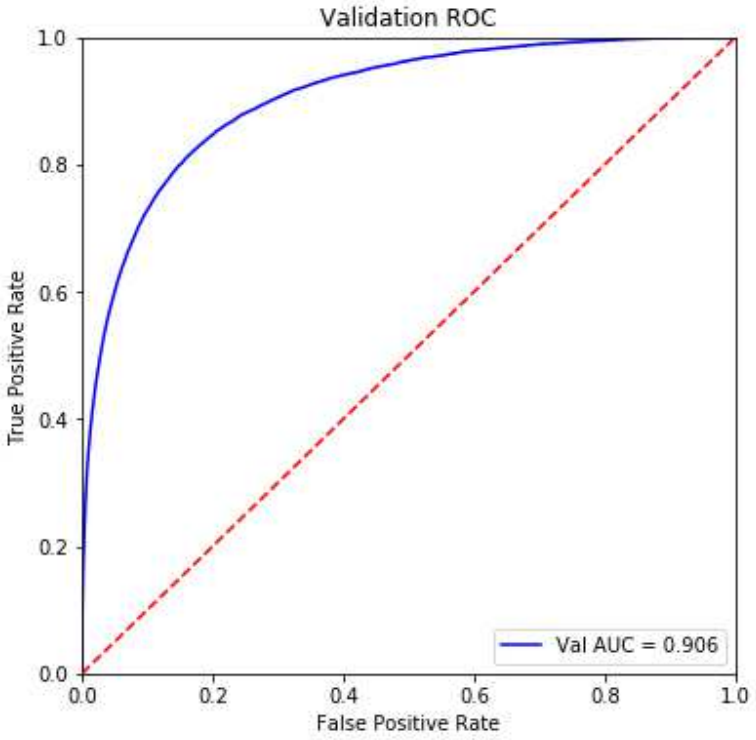
- Naive Bayes
- Gaussian naive Bayes
- LinearRegression
- Catboost

2.1. **Naive Bayes.**

- Calculate Prob
  P(A—B)=$\frac{P(AB)}{P(B)}$

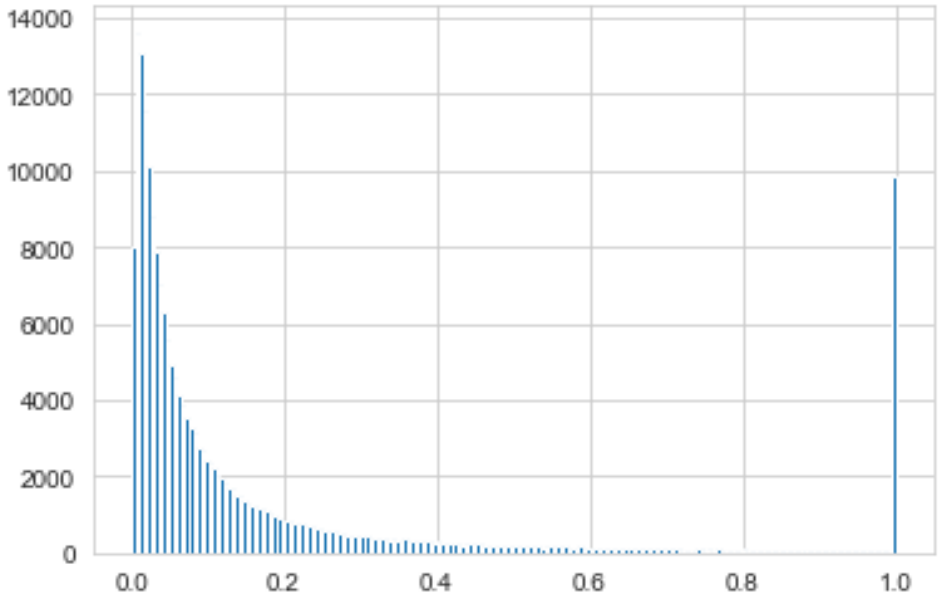


- Smoothing
  If the probability value to be estimated is 0, the calculation result of posterior probability will be affected. The solution to this problem is to use smoothing
- Validation AUC



Validation AUC = 0.905571412599524

- Probability



2.2. **Gaussian naive Bayes.**

- Calculation of prior probability

use Counter() maybe more convenient

- Avg and Std
- Calculate likelihood

  Using probability density function of Gaussian distribution to calculate likelihood and then multiply to get likelihood We can get Raw data, trend data, periodic data, random variables
- Training model and get prediction

  The probabilities of each label are multiplied by the likelihood and then normalized to get the prob of each label.
- AUC

  Validation AUC is 0.8051607443604657.

2.3. **LinearRegression.**

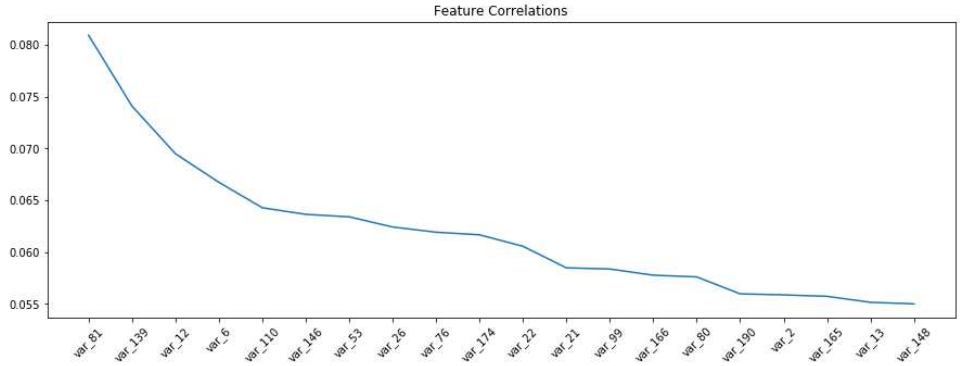
- Merge test/train datasets
- Add more features

  Normalize the data,Standardization of normal distribution,then Square the value, cubic the value,Cumulative normal percentile,Normalize the data,again.Do linear regression,Write submission file

  AUC: 0.8025517936065763

2.4. **Catboost.**

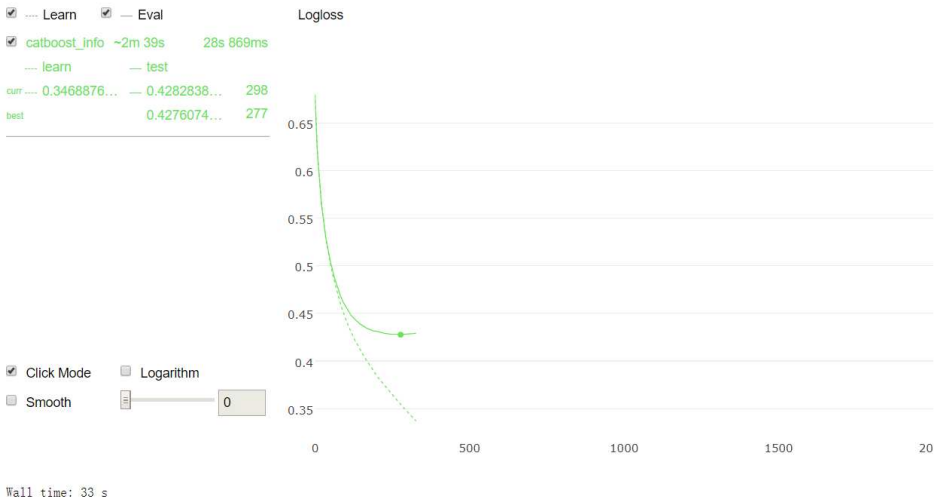
- Feature Correlations



- Get the features Get the top 100 features,merge them and divide the training set and test set
- process data

  In catboost, you don't have to worry about this at all. You just need to tell the algorithm which features belong to category features, and it will help you deal with them automatically

  Finally, we feed the data to the algorithm and train it.



- fit and prediction
  AUC: 0.80399151

3. Conclusion

- Naive Bayes
  AUC: 0.9055714
- Gaussian naive Bayes
  AUC: 0.8051607
- LinearRegression
  AUC: 0.8025517
- Catboost
  AUC: 0.8039915

List of Todos

(A. 1) QUT,, QUT, Xichen Tang 710065, China
Email address, A. 1: xichen@tulip.academy