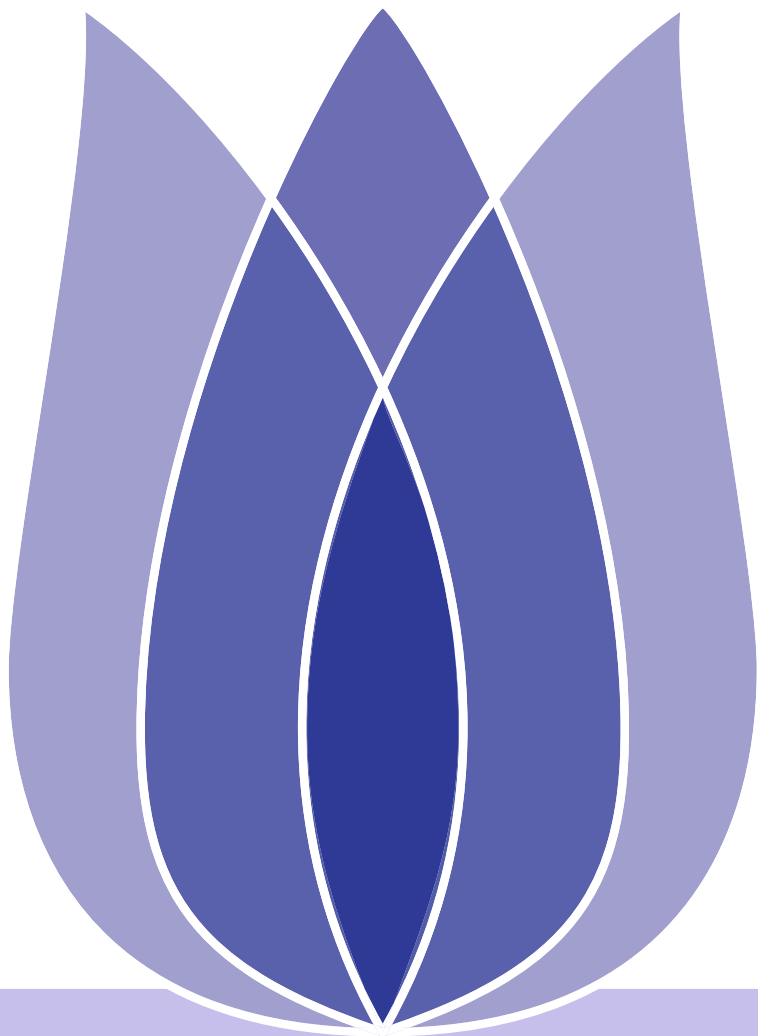


Identifying Customers

Xichen Tang
QUT

January 17, 2020





Directory

Subject Introduce

Naive Bayes

Gaussian naive Bayes

LinearRegression

Catboost

Conclusion

Thanks and Question

Directory



Introduce

Directory
Subject Introduce
Naive Bayes
Gaussian naive Bayes
LinearRegression
Catboost
Conclusion
Thanks and Question

Directory
Subject Introduce
Naive Bayes
Gaussian naive Bayes
LinearRegression
Catboost
Conclusion
Thanks and Question



- [Directory](#)
- [Subject Introduce](#)**
- [Data](#)
- [Avg and Std](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

Subject Introduce



Introduce

- [Directory](#)
- [Subject Introduce](#)
- [Data](#)
- [Avg and Std](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

■ The kaggle subject:Santander Customer Transaction Prediction

In this challenge, we need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.





- Directory
- Subject Introduce
- Data**
- Avg and Std
- Naive Bayes
- Gaussian naive Bayes
- LinearRegression
- Catboost
- Conclusion
- Thanks and Question

■ 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

■ train_data.info

RangeIndex:	200000 entries	0 to 199999
Columns:	202 entries	ID_code to var_199

■ Missing Values

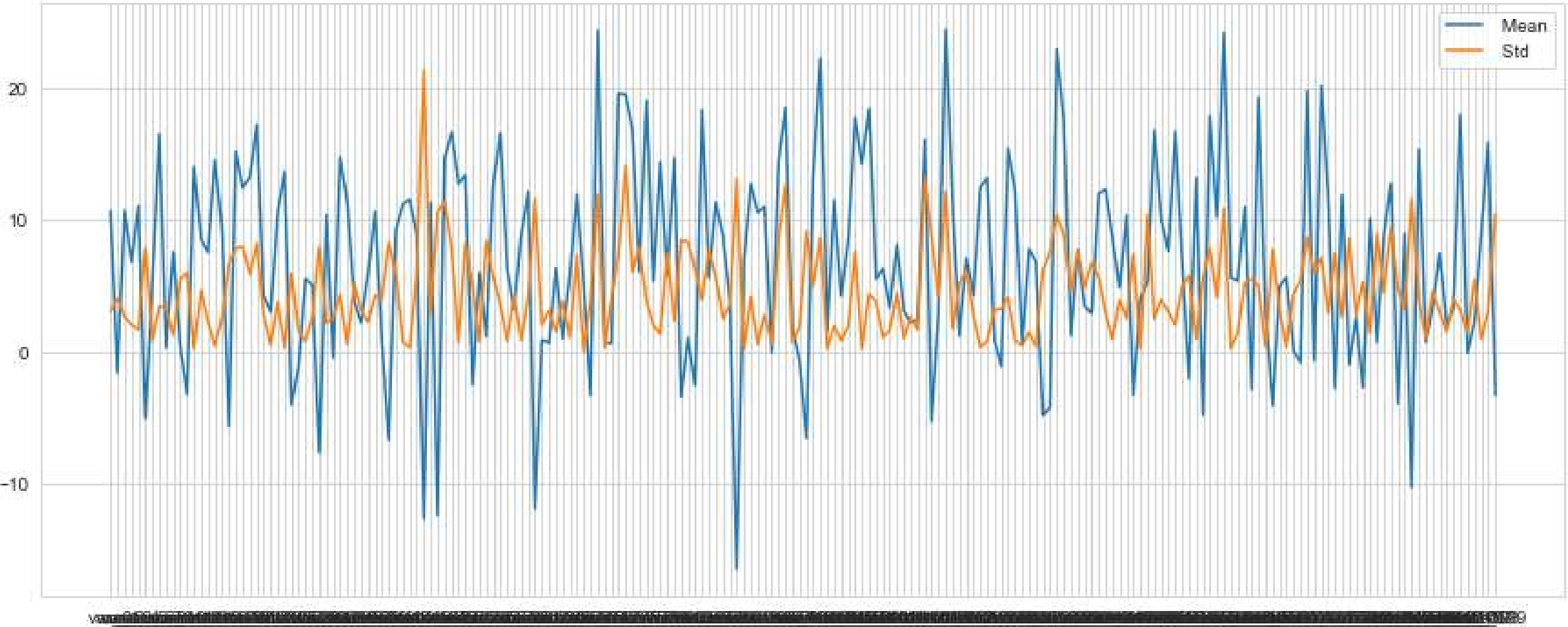
train data missing values? False
test data missing values?False



Avg and Std

- Directory
- Subject Introduce
- Data
- Avg and Std**
- Naive Bayes
- Gaussian naive Bayes
- LinearRegression
- Catboost
- Conclusion
- Thanks and Question

■ by describe()





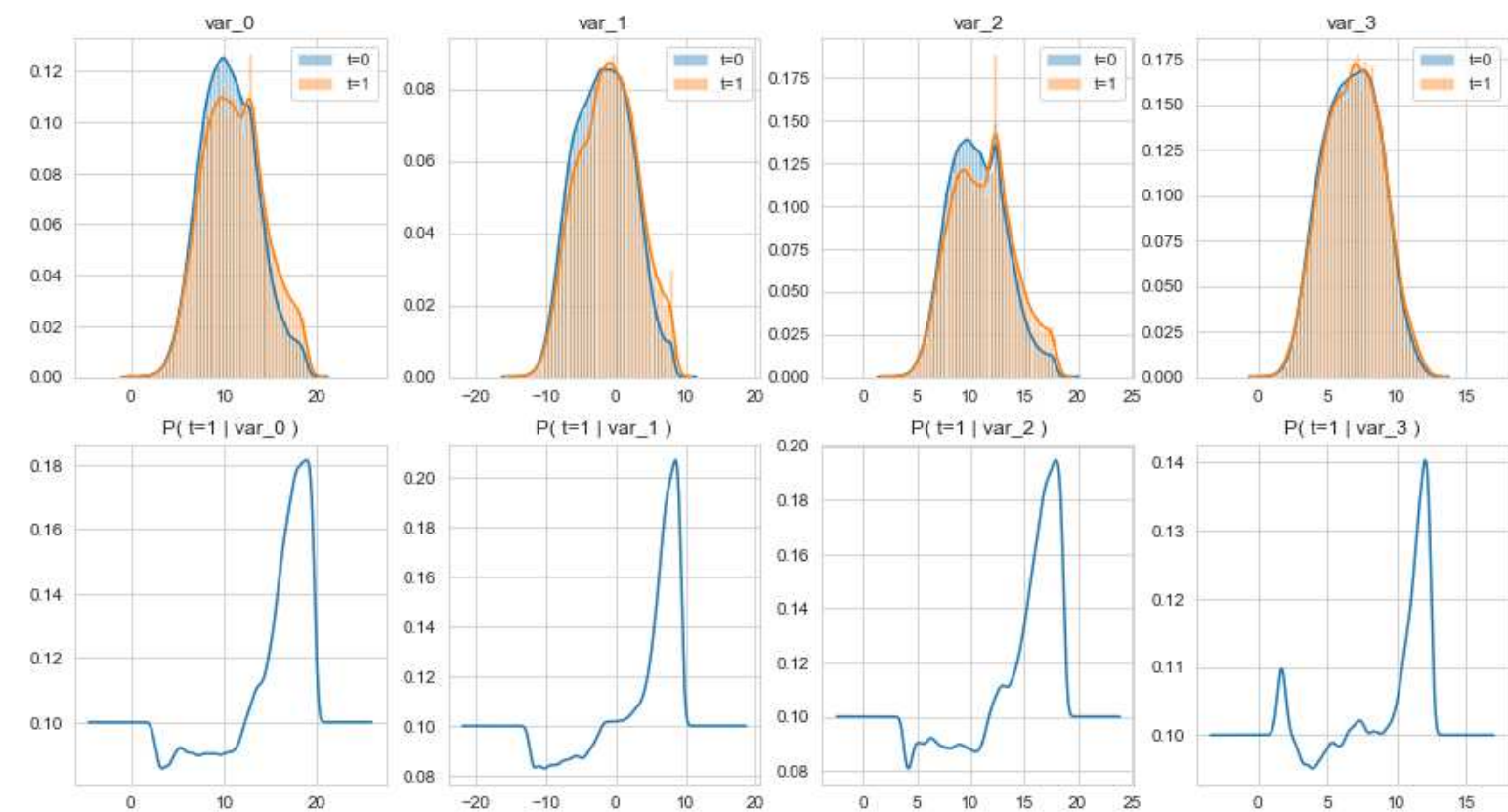
- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)**
- [Statistical Functions](#)
- [AUC](#)
- [Result](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

Naive Bayes



- Directory
- Subject Introduce
- Naive Bayes
- Statistical Functions**
- AUC
- Result
- Gaussian naive Bayes
- LinearRegression
- Catboost
- Conclusion
- Thanks and Question

■ Calculate Prob

$$P(A | B) = \frac{P(AB)}{P(B)}$$


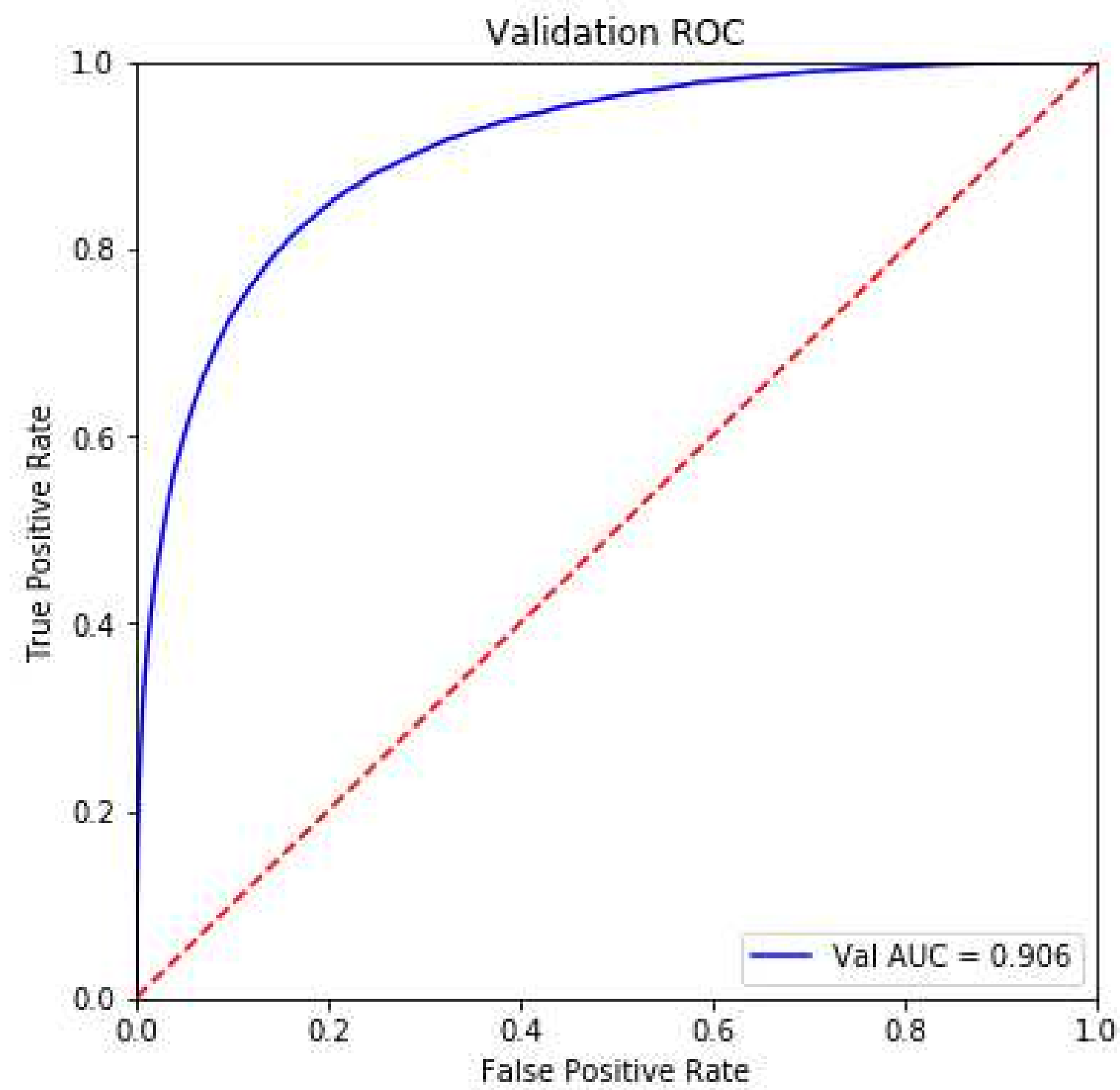
■ Smoothing

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



- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Statistical Functions](#)
- [AUC](#)**
- [Result](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

■ Validation AUC

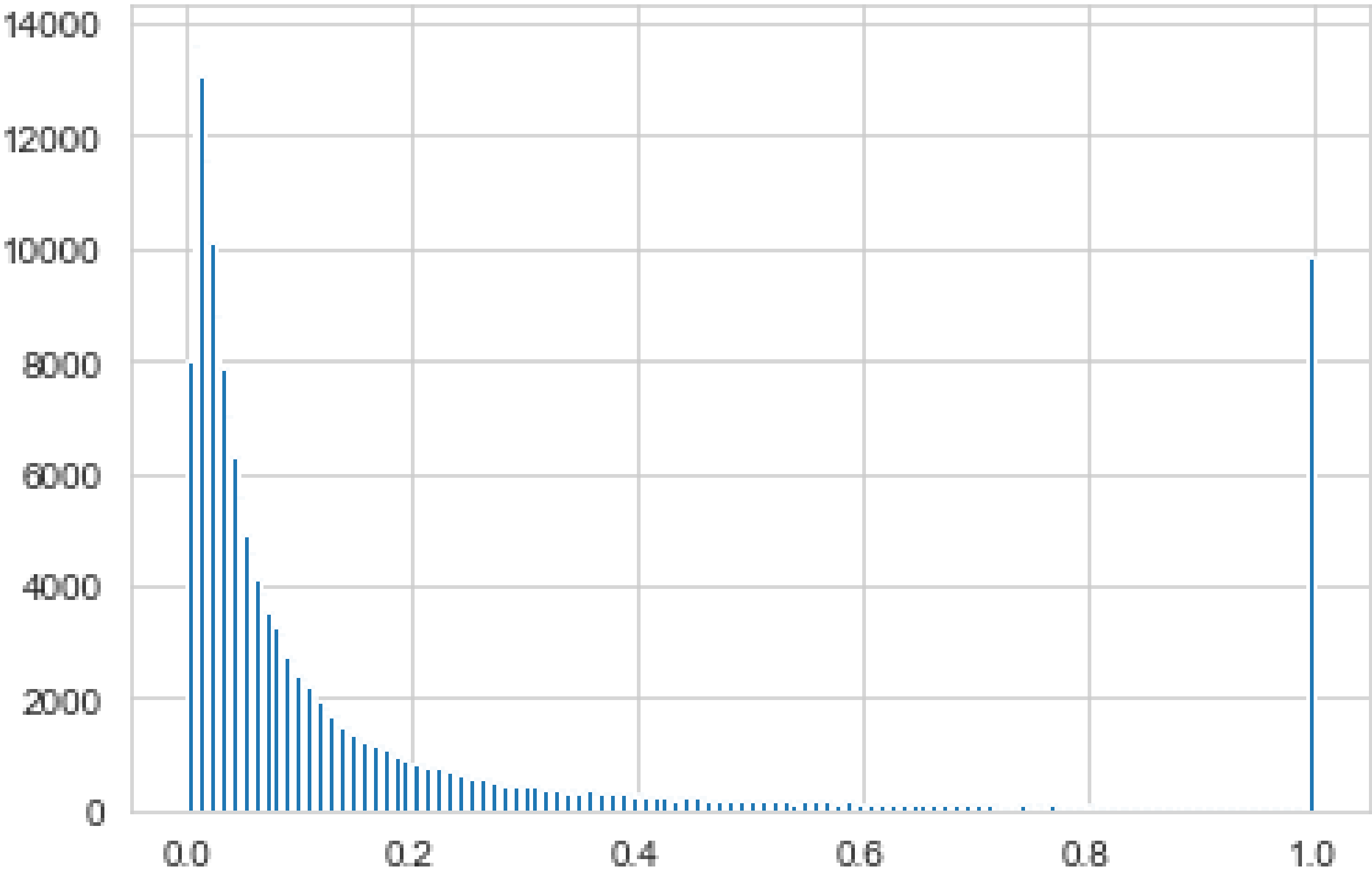


Validation AUC = 0.805571412599524



- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Statistical Functions](#)
- [AUC](#)
- [Result](#)**
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

1. Probability





- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)**
- [Statistical Functions](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

Gaussian naive Bayes



Statistical Functions

- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [Statistical Functions](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

- 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.



- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [process](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

LinearRegression



- Directory
- Subject Introduce
- Naive Bayes
- Gaussian naive Bayes
- LinearRegression
- process
- Catboost
- Conclusion
- Thanks and Question

- 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



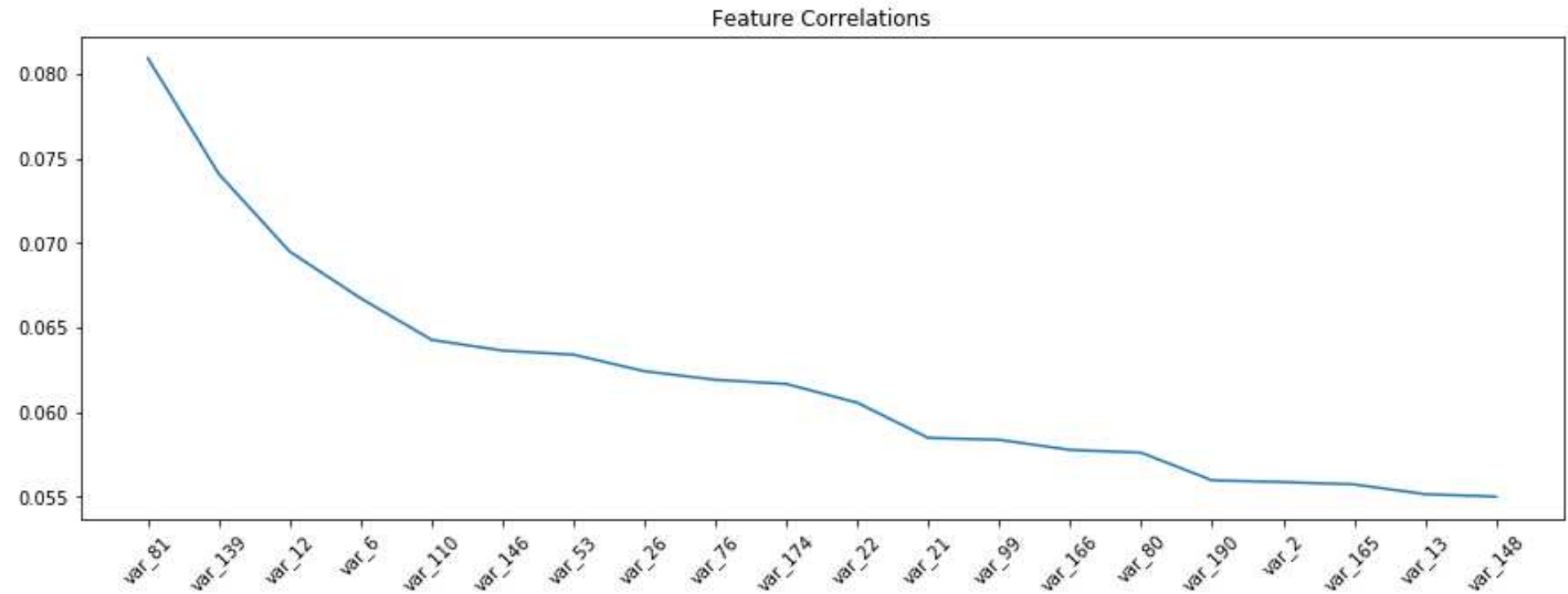
- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)**
 - [process](#)
 - [process](#)
- [Conclusion](#)
- [Thanks and Question](#)

Catboost



- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [process](#)
- [process](#)
- [Conclusion](#)
- [Thanks and Question](#)

■ Feature Correlations



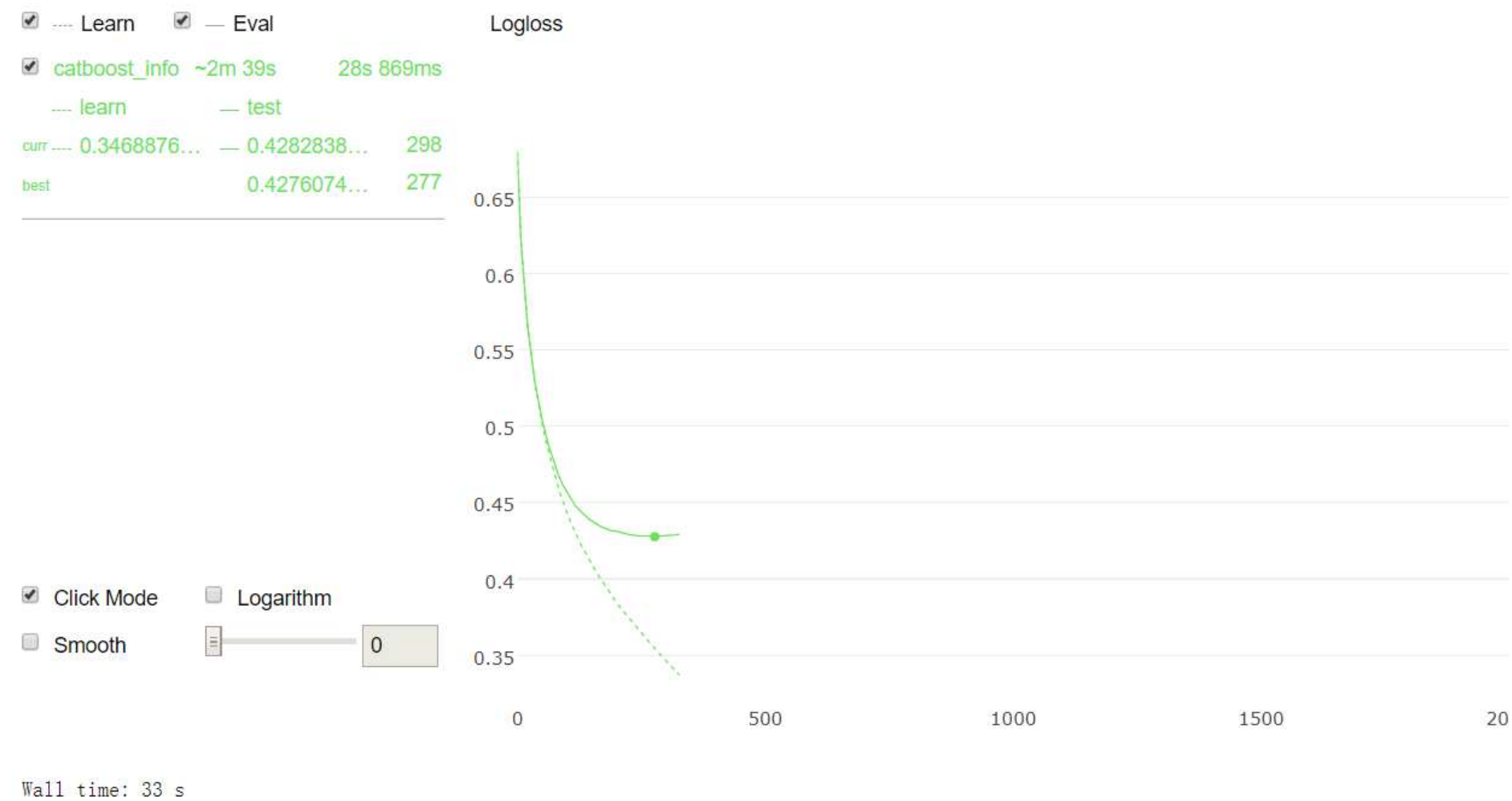
- Get the features Get the top 100 features,merge them and divide the training set and test set



process

- Directory
- Subject Introduce
- Naive Bayes
- Gaussian naive Bayes
- LinearRegression
- Catboost
- process
- process
- Conclusion
- Thanks and Question

- 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



- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)**
- [Thanks and Question](#)

Conclusion



Compare

- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

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



- [Directory](#)
- [Subject Introduce](#)
- [Naive Bayes](#)
- [Gaussian naive Bayes](#)
- [LinearRegression](#)
- [Catboost](#)
- [Conclusion](#)
- [Thanks and Question](#)

Thanks and Question