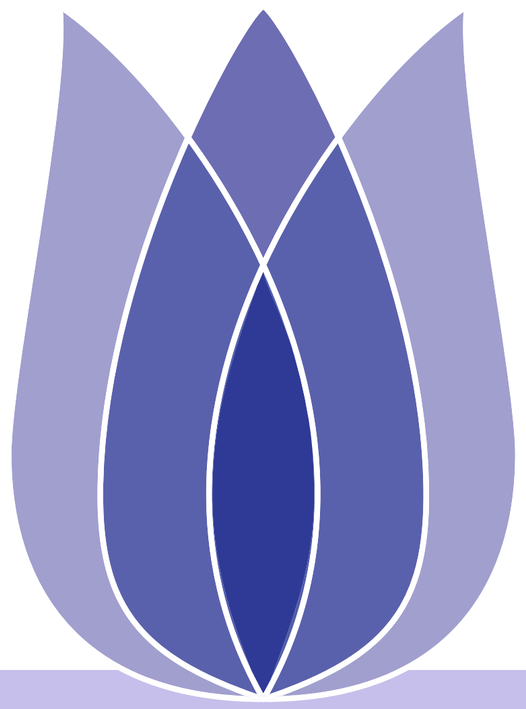
Identifying Customers

Xichen Tang
QUT

January 18, 2020



Directory

Introduce

Directory

Subject Introduce

The data observed

Forecasts

Thanks And Question

Subject Introduce

Introduce

- 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.

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

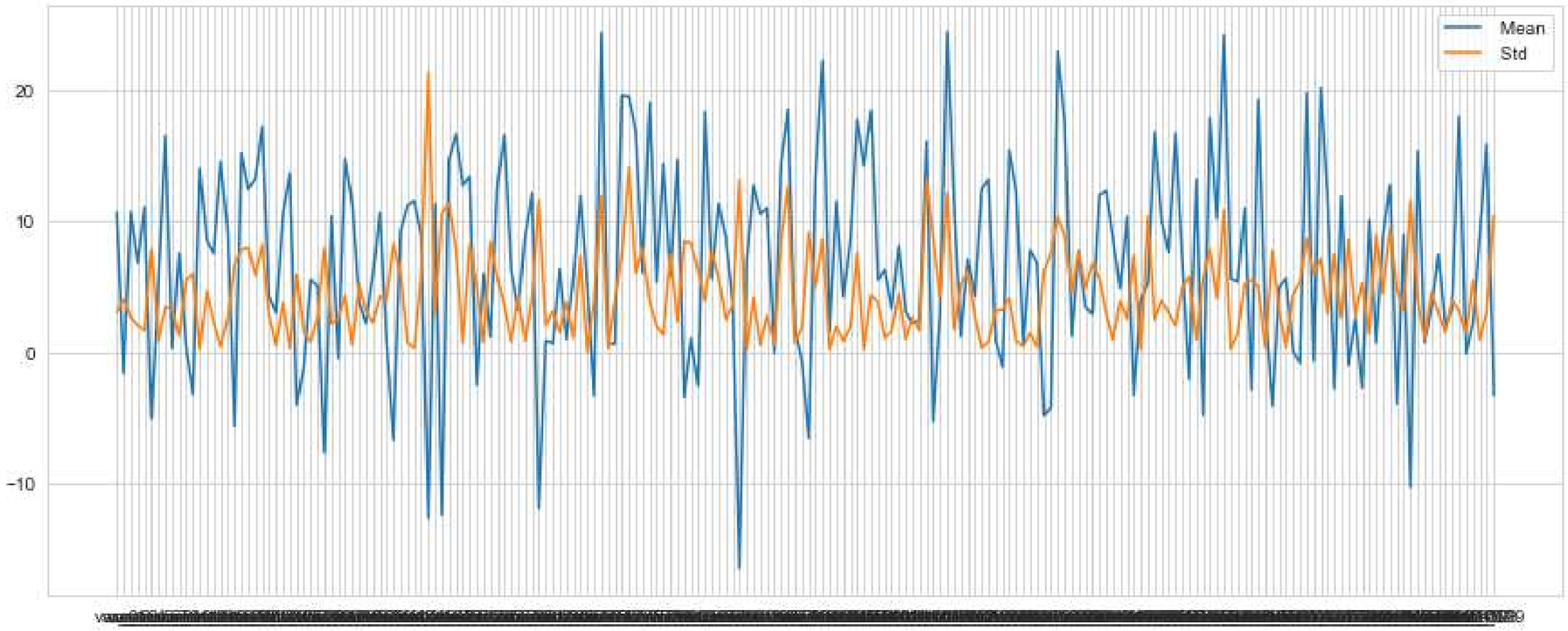
- train_data.info

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

- by describe()

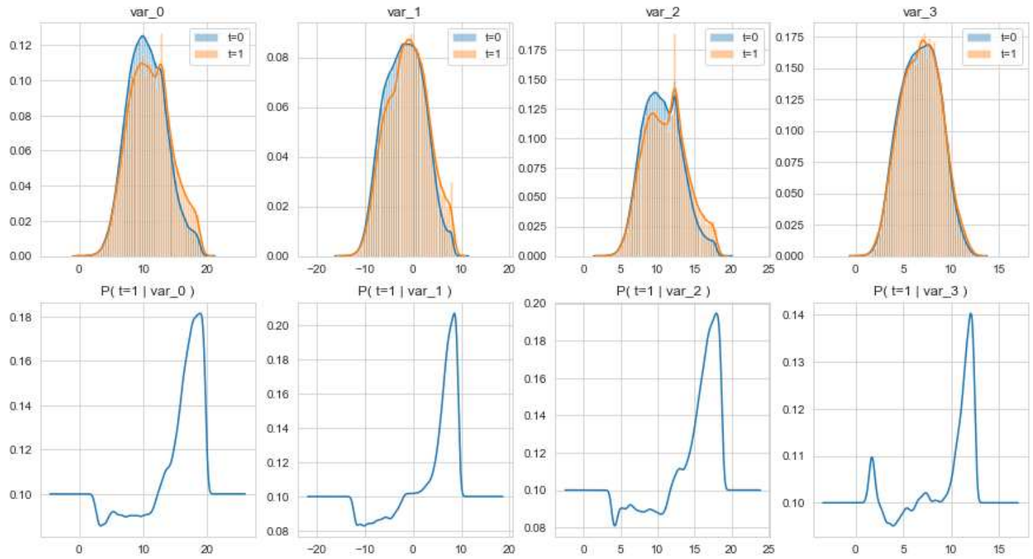


Naive Bayes

Statistical Functions

- Calculate Prob

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

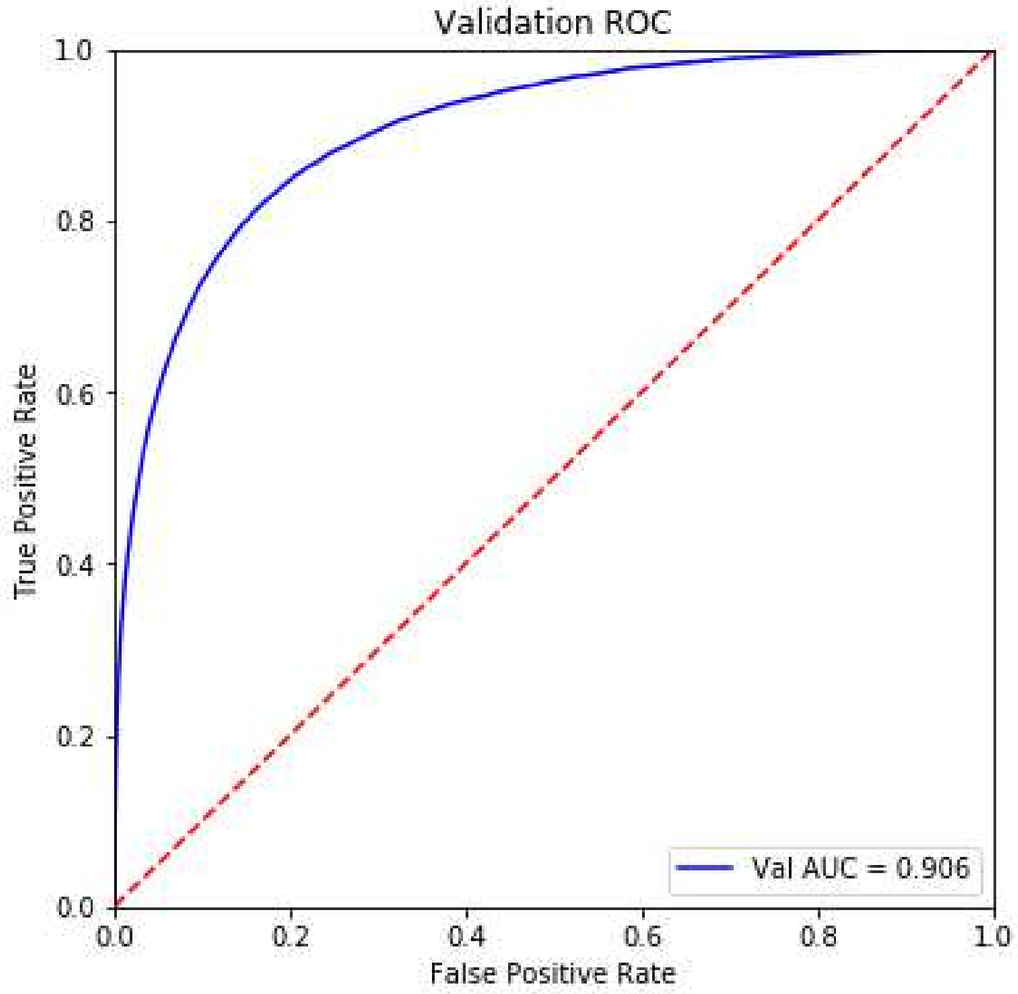


- Smoothing

  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

AUC

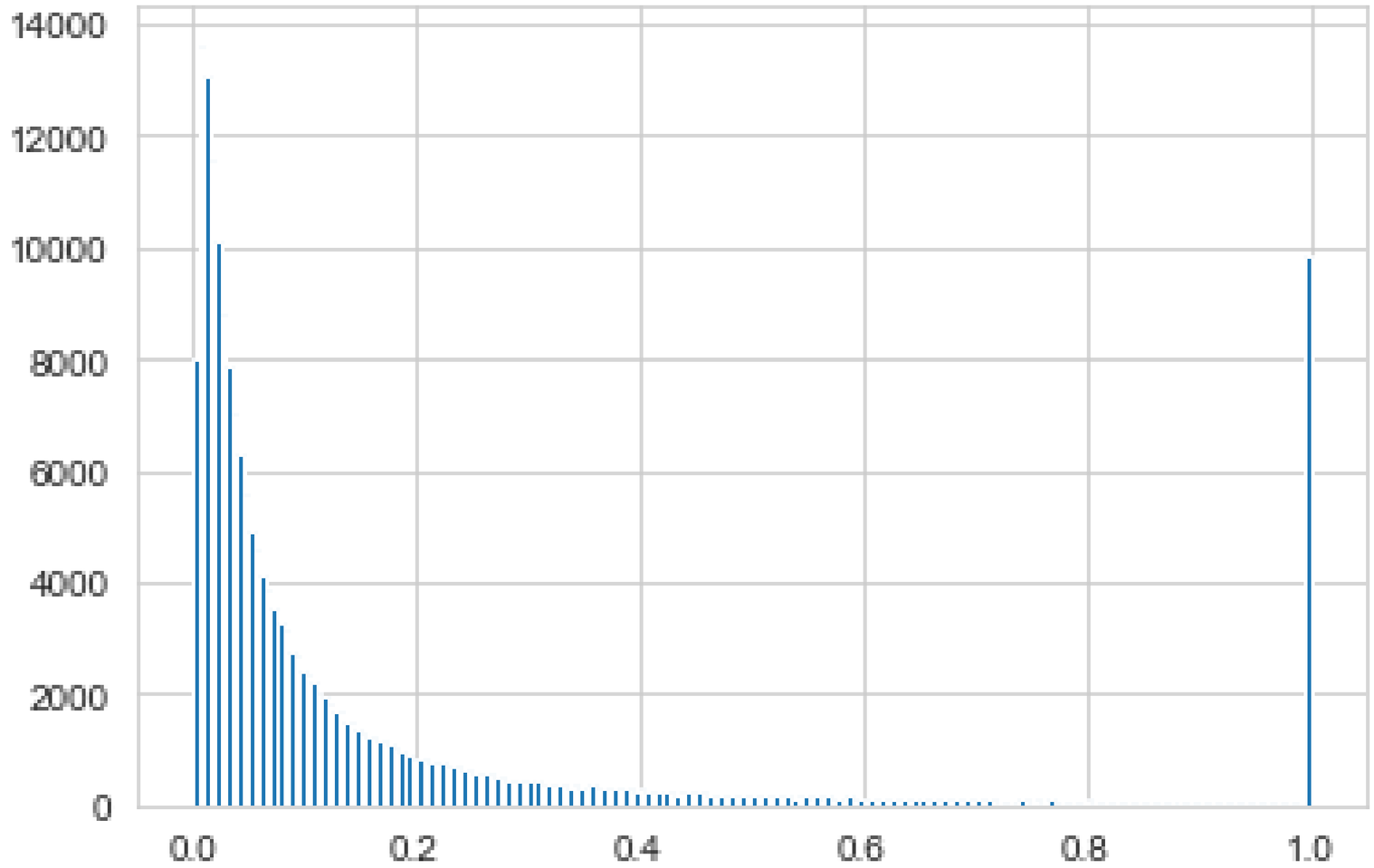
Validation AUC



Validation AUC = 0.8055714125995 24

Result

1. Probability



Gaussian naive Bayes

Statistical Functions

- 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.

LinearRegression

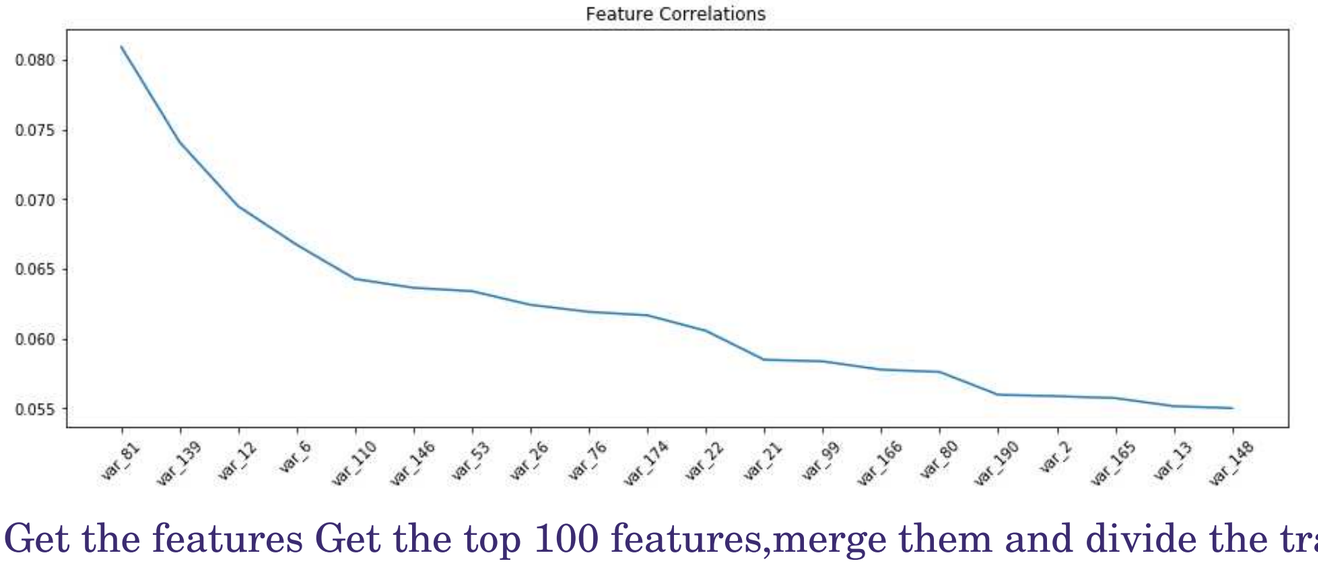
process

- 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

Catboost

process

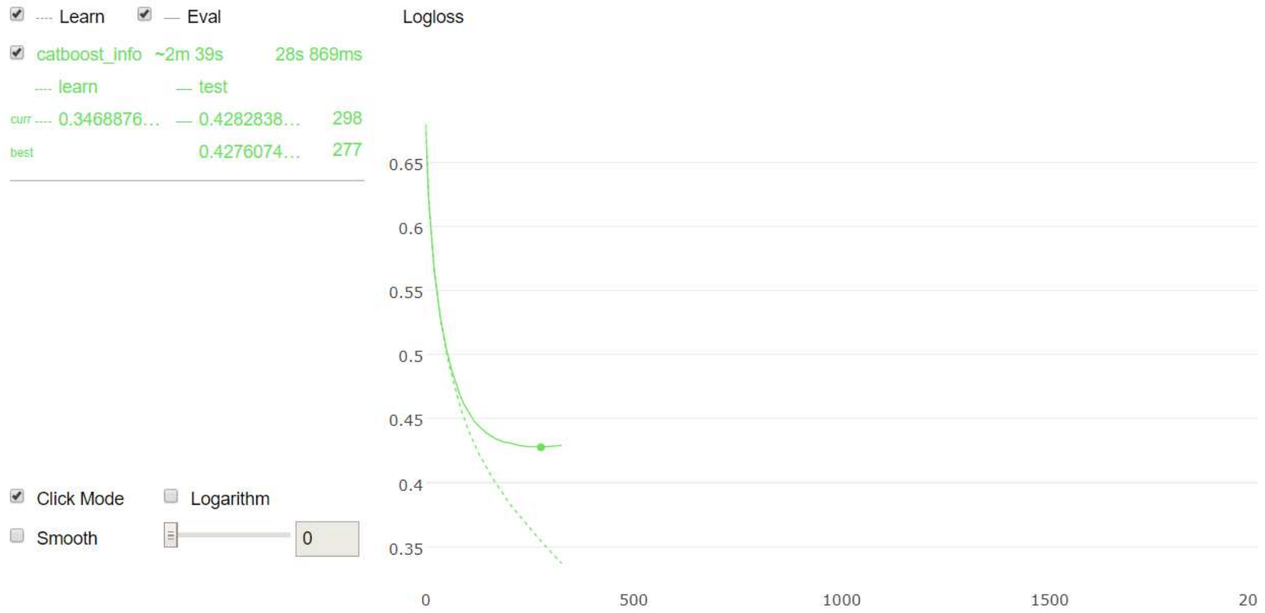
Feature Correlations



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

process

- 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




```
Wall time: 33 s
```

- fit and prediction
  AUC: 0.80399151

Conclusion

Compare

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

Thanks and Question