Final Presentation

——2016 Capital One Modeling Competition

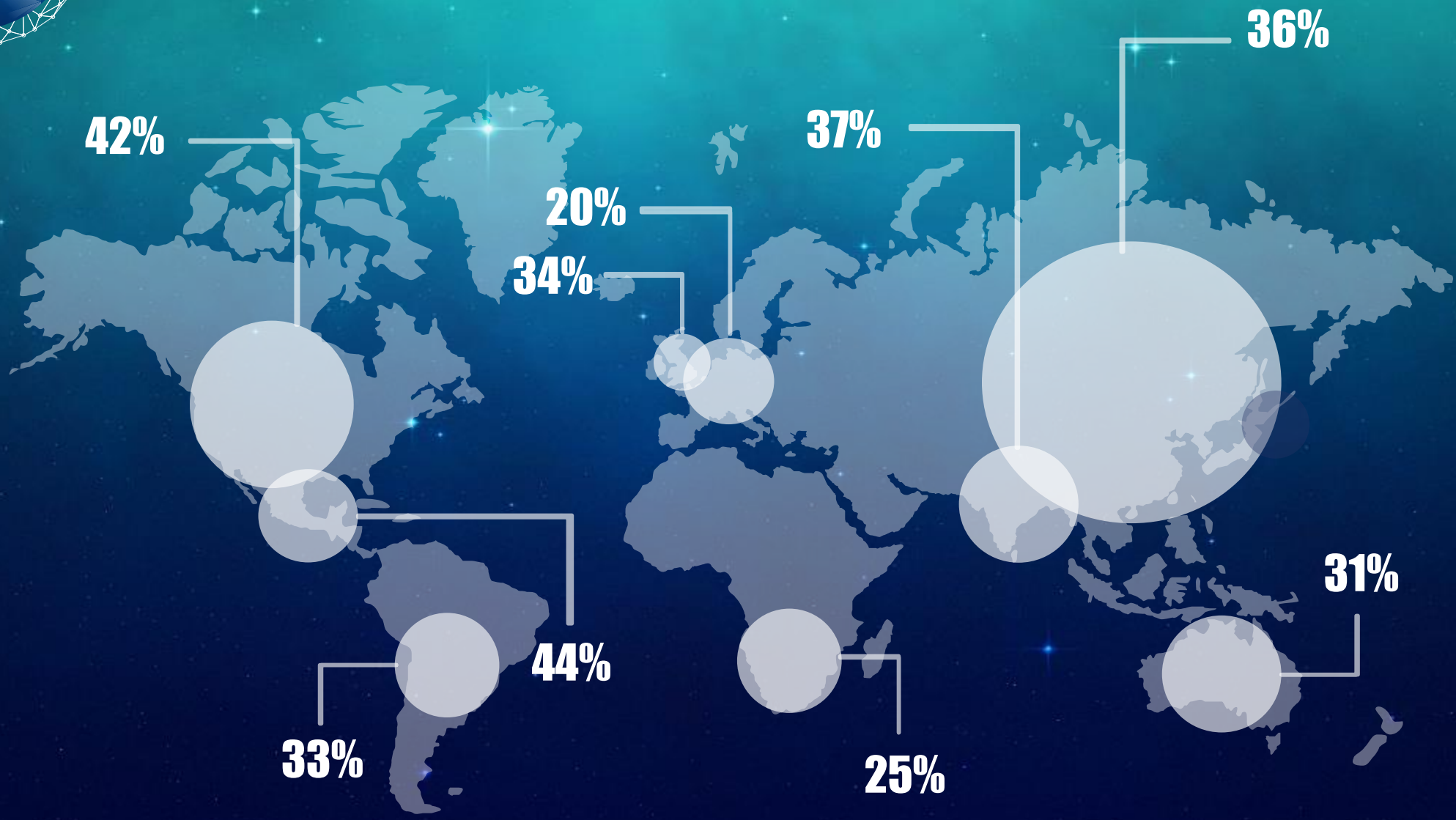
The University of Texas at Austin

Yanyao Shen Weiyan Sun

Chuwen Zhang Mingzhang Yin

0-1 Introduction

Cited from Aite Group and ACI Worldwide
Surveyed over 5000 consumers in 17 countries
Percentage of respondents who have fraud experience



Introduction

Challenge:
Huge amount of data that generated every minute

Result:
1. Slow computation time
2. Fraud pattern is hard to learn

Model

Characteristics:
1. With high AUC result
2. Low model complexity

Introduction

1. Deal with Features
2. Build Suitable Model
3. Avoid Overfitting
4. Tradeoff Due to Big Data
5. Explain The Result
6. Economic Analysis

PREPROCESSING

Feature Preprocessing

Feature	Method	Example
Time variable	Transfer into a float value	Transaction Date Raw value: 2013-07-11 After: 923.5
Categorical variable	One hot encoding	PIN Verification Indicator Raw values: {‘Y’, ‘N’, ‘null’} After： {‘1|0’, ’0|1’, ’0|0’}
Single value variable	Drop	Cryptogram Request raw values: {*2} After: { }
Numerical variable	Keep	Cash Available: Raw value: {2500,500,…}

MODEL

Model

68 features

customer

interaction

terminal

2-2

Model

68 features

customer

Cash Available:
150,2500,etc.

Credit amount:
5000,2000,1750,etc

Distance from home to purchase:
21.63,11.42,22.22, etc

interaction

Transaction date
41315, 41302,etc

Customer Appearance in person:
0/1

terminal

Country code:
840

POS category:
0,3,7,5,etc

2-3

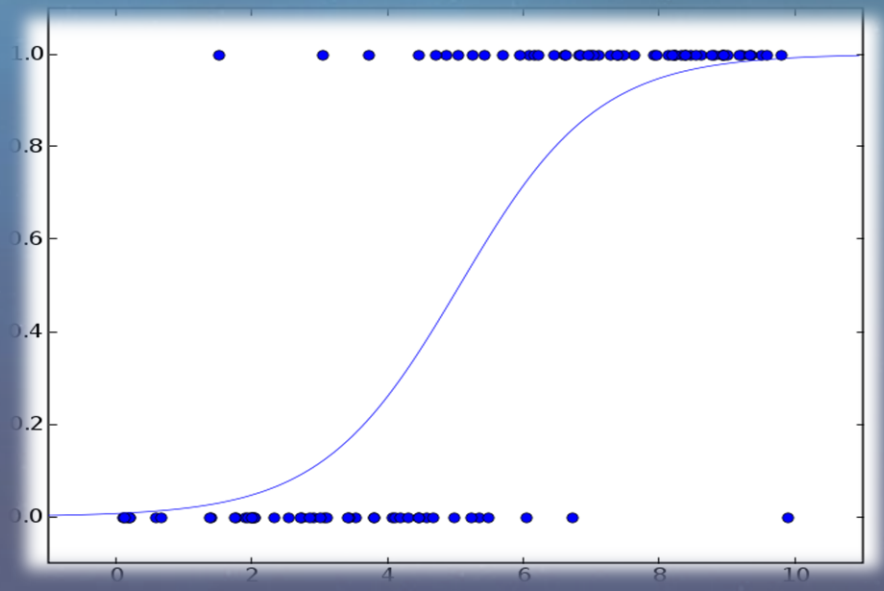
Model – Customer & Terminal End

customer

terminal

Logistics regression

Objective: 0-1 Classification
Estimator: MLE
Method: Gradient Descent



2-4 Model - Interaction End

interaction

Regularized Gradient Boosting

Objective: Classification
Estimator: Tree function
Method: Greedy Boosting

Model Flow Chart

Raw Data

Data Processing

Customer Features

Go to

Customer_1

...

Customer_10

Prediction y_c

Interactions Features

BoostedTrees Interaction

Prediction y_i

Terminal Features

Go to

Terminal_1

...

Terminal_4

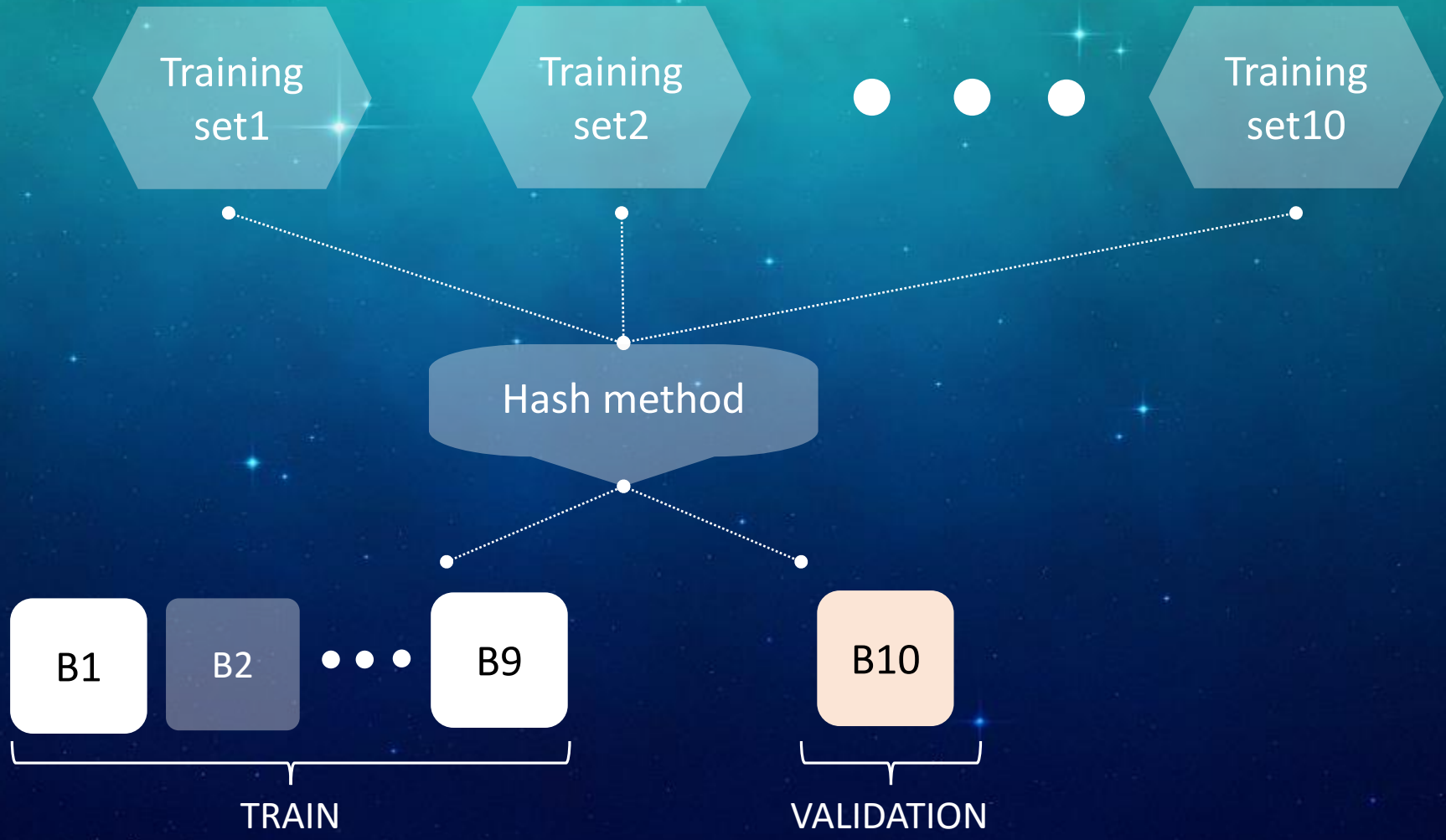
Prediction y_p

Ensemble

Overall prediction

AVOID OVERFITTING

Avoid Overfitting – Data Side



Avoid Overfitting – Model Side

Logistic Regression:

- Ridge regression (L2 norm of coef.)

Boosted Tree:

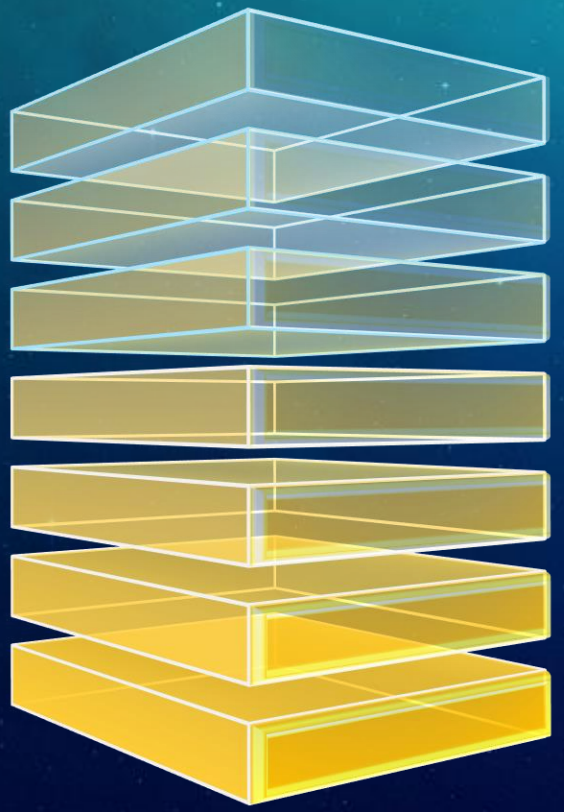
- Number of leaves
- L2 norm of leaf scores

PART FOUR

TRADE OFF

4-1 Trade off for Logistic Regression

Customer part



Transaction Money

Terminal Category

Terminal part

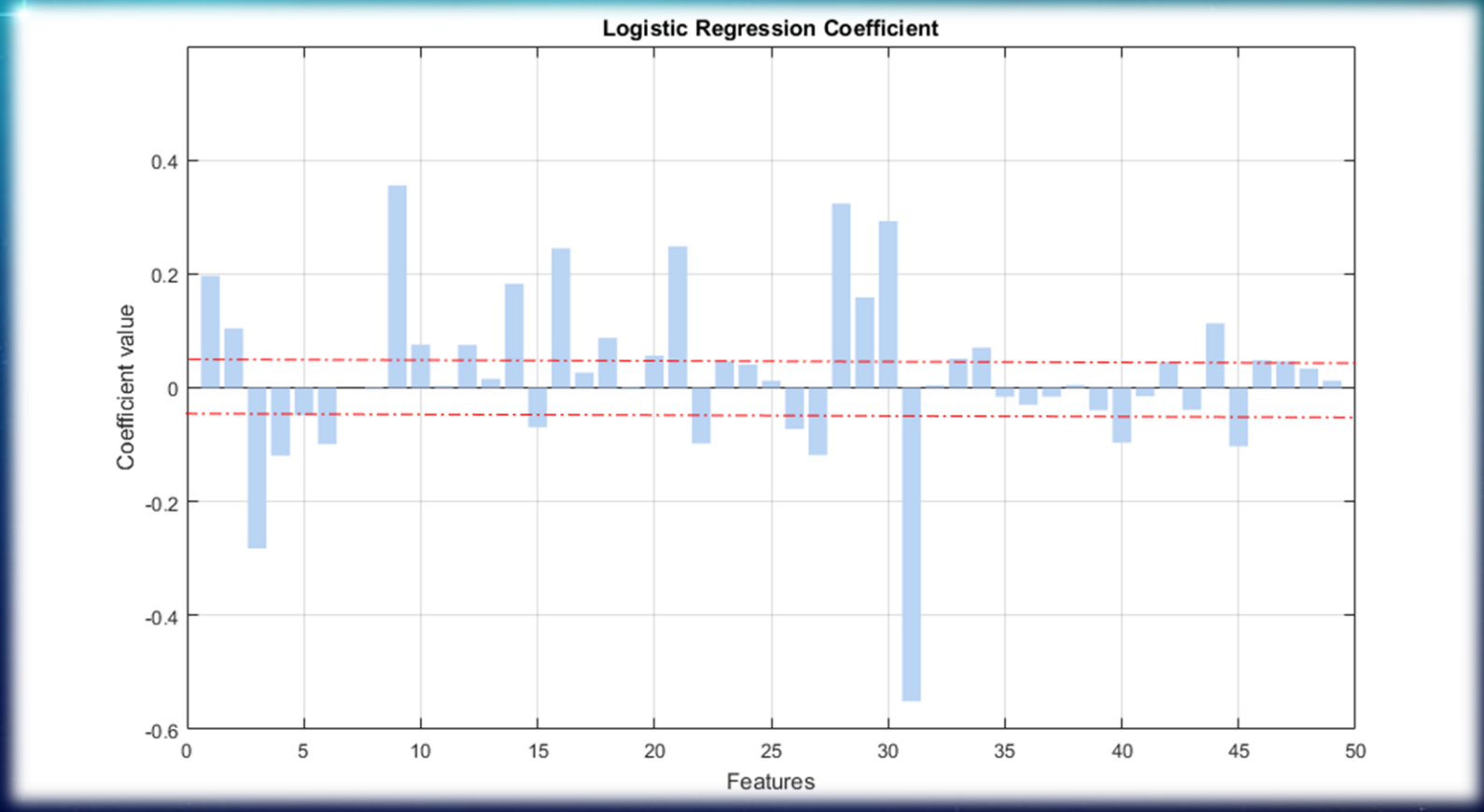
4-2 Interaction end

PART FIVE

Results

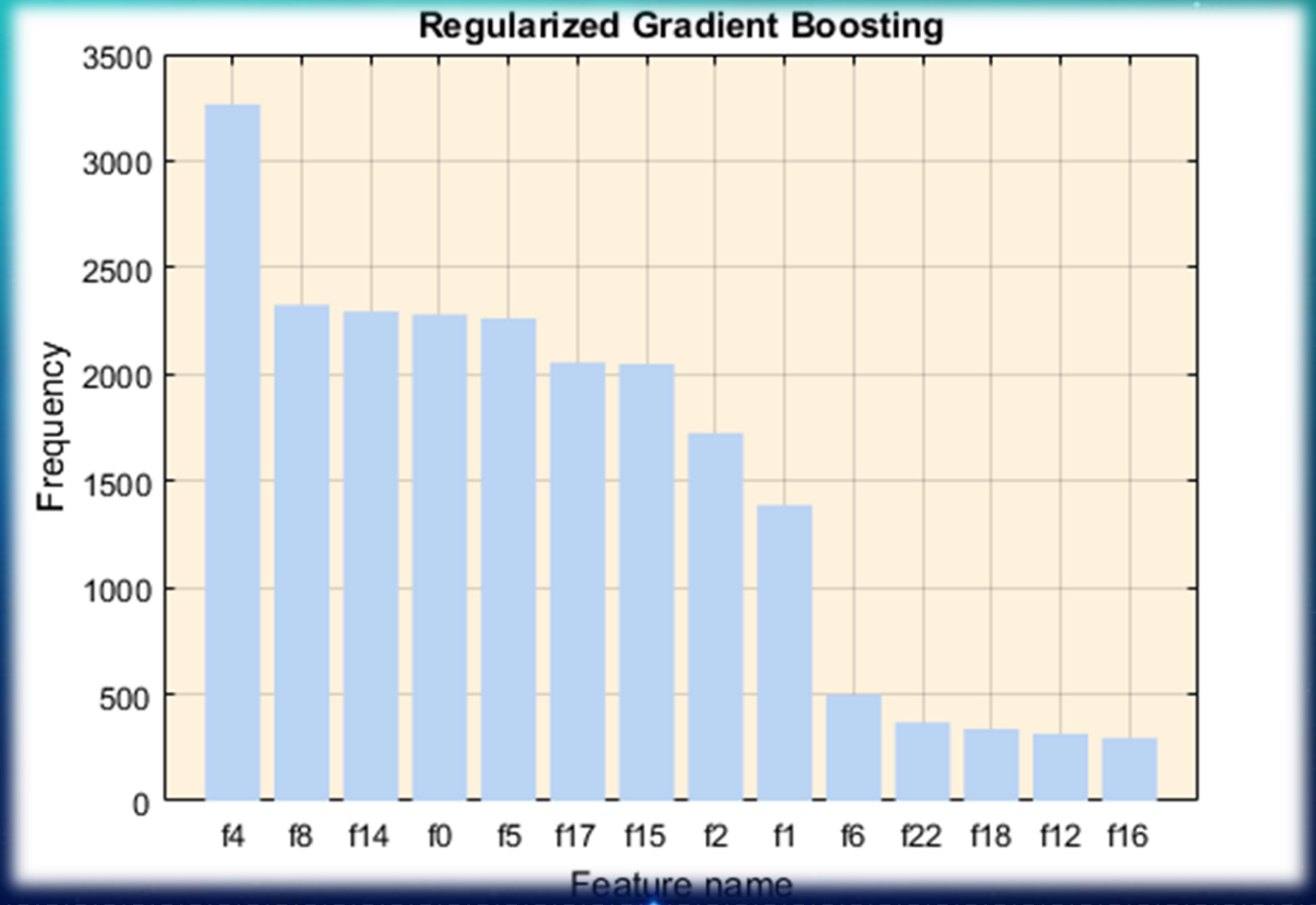
5-1 Logistic Regression

Feature
Address verification code (X14)
Money required (X19)
PIN validation (X50)
Security level (X40)



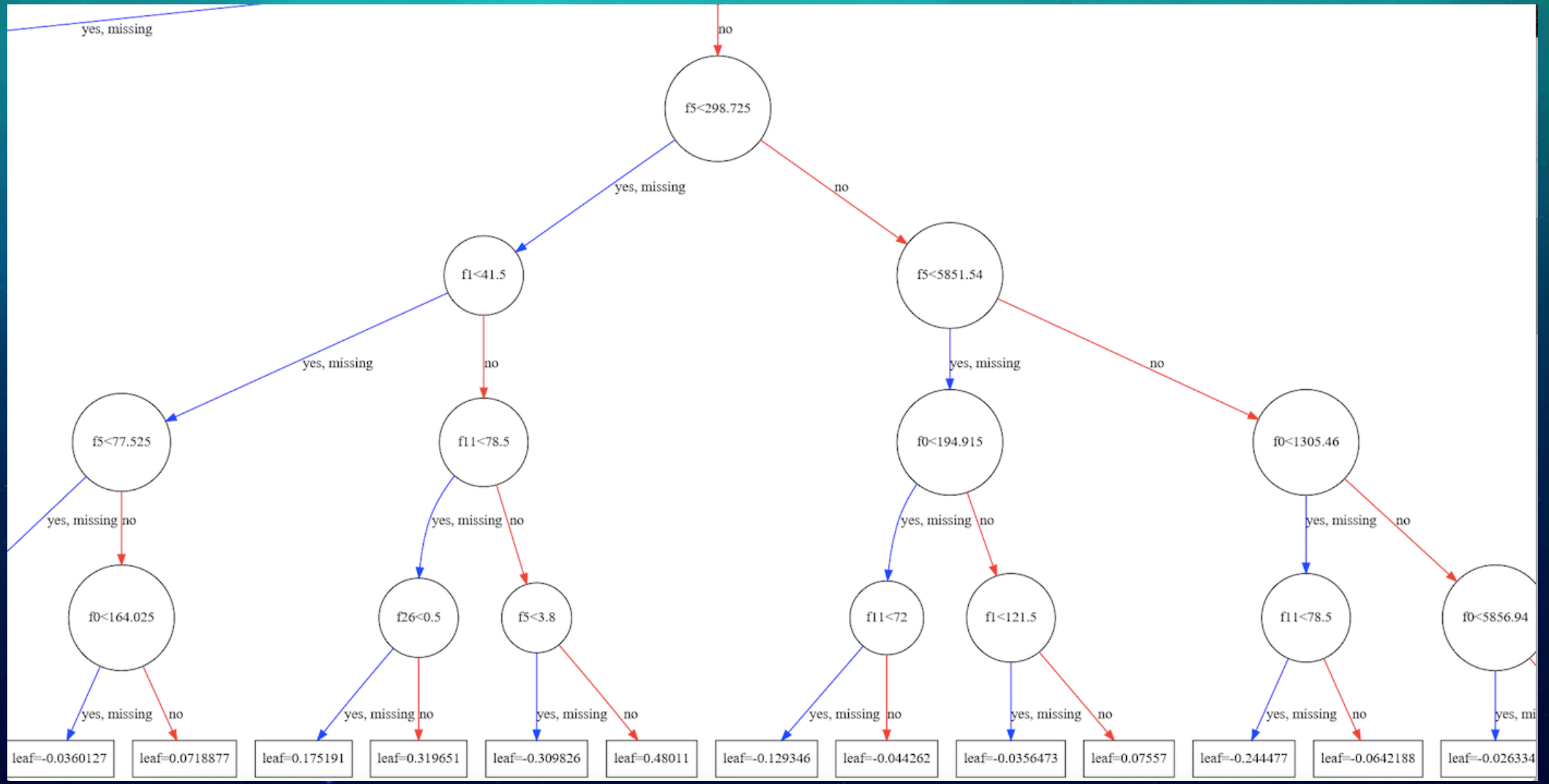
5-2 Regularized Gradient Boosting

Feature name	Meaning
f4	Balance(X8)
f8	Money required(X19)
f14	Authentication Date (X27)
f0	Amount of money (X24)
f5	Available money (X5)
f17	CVV Duration Time(X33)
f15	Authentication Time(X28)



[('max_depth', 6), ('objective', 'binary:logistic'), ('bst:eta', 0.1), ('eval_metric', 'auc'), ('subsample', 0.9), ('colsample_bytree', 0.6), ('early_stopping_rounds',50),('num_round', 500)].

Boosted Tree Results

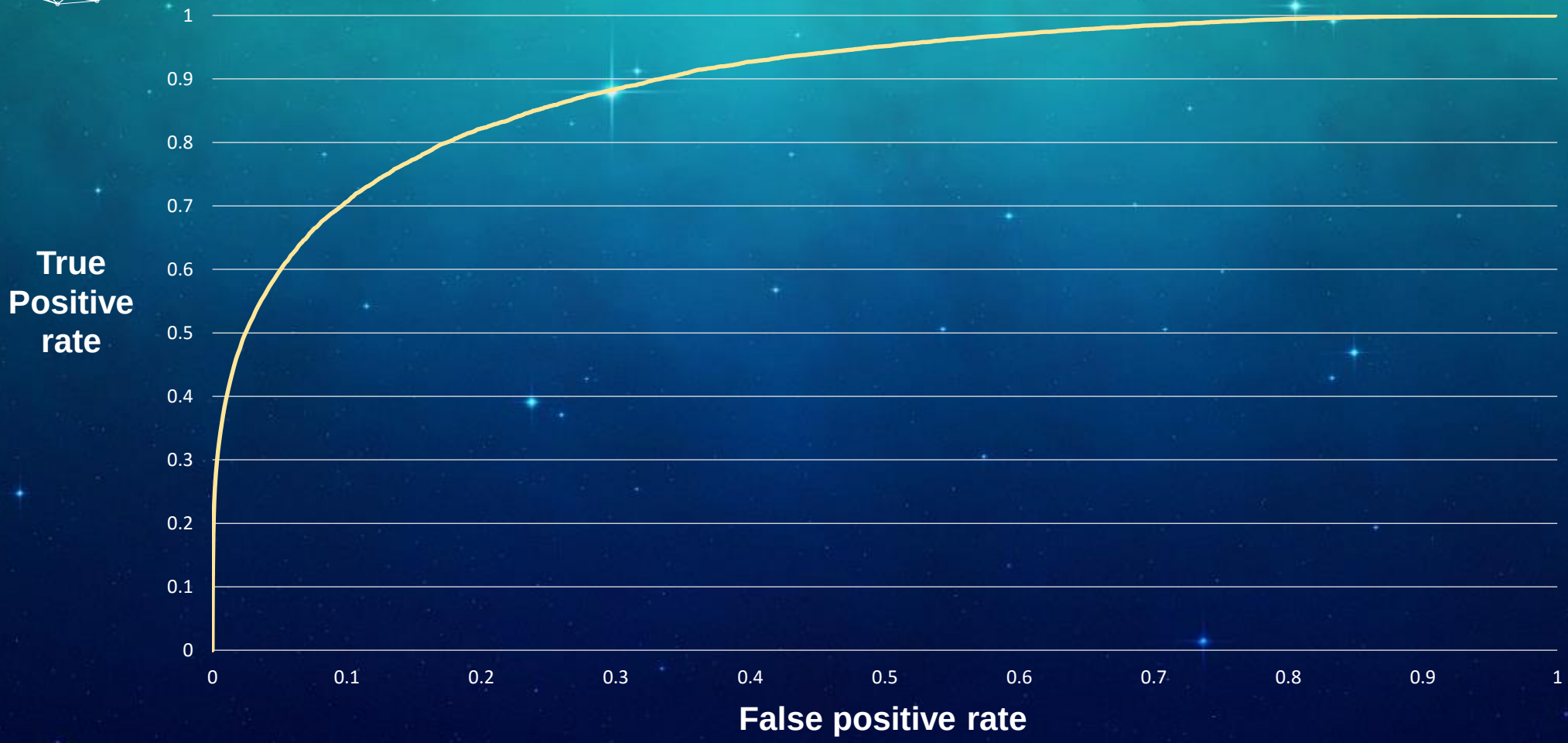


AUC Scores

Data part	Customer	Terminal	Customer	Overall
method	XGBoost	Logistic Reg	Logistic Reg	ensemble
AUC	0.80	0.77	0.89	0.90

5-5

ROC Curve


True Positive rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
False positive rate

PART SIX

Economic Analysis

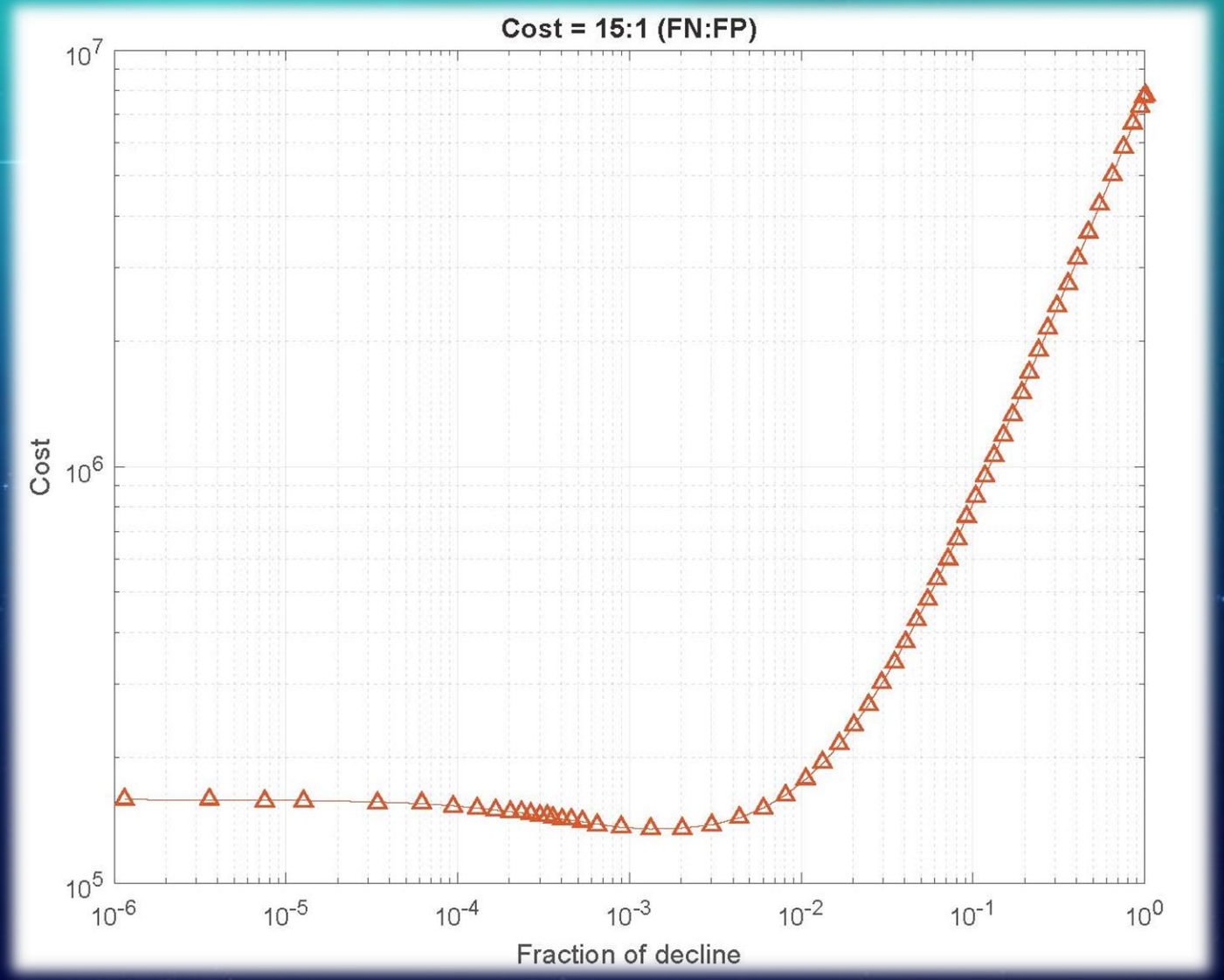
Cost Analysis

- FN harms
- FP is hard to estimate:
  - Cost of reviewing
  - Cost of customer experience
- Total Cost = FN*C + FP (We set C=15)
- Plot the cost versus the fraction of decline.

6-2 Cost Analysis – The Curve

Optimal point (FoD): 0.0013

Threshold: 0.49



Review

1. Deal with Features

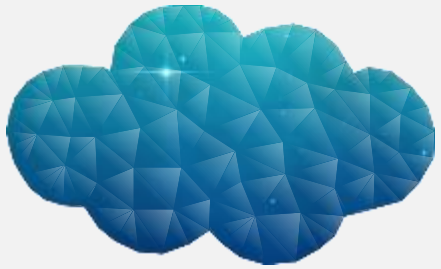
2. Build Suitable Model

3. Avoid Overfitting

4. Tradeoff Due to Big Data

5. Explain The Result

6. Economic Analysis



THANKS FOR YOUR ATTENTION