CLASSIFICATION ANALYSIS ON A BANK MARKETING CAMPAIGN

Group 6

Shutong Fan

Xinyu Yang

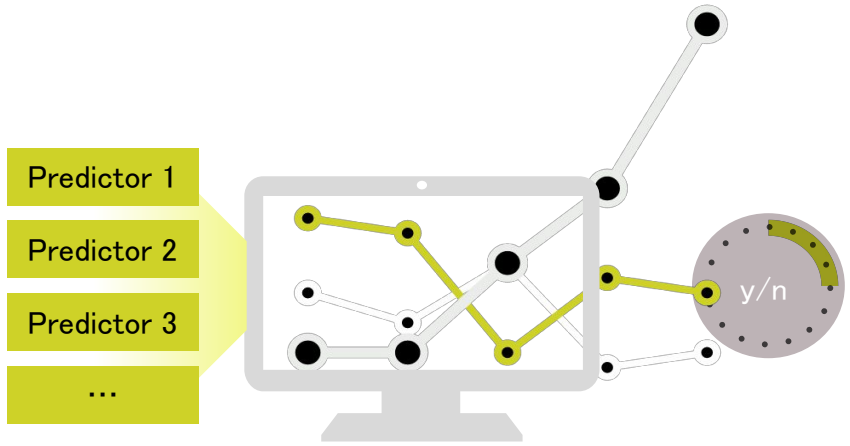
Xueting Deng

Yingnan He

Content

Introduction

- Explore the relationships between response variable and predictors, then train the classification models



- Help the bank predict whether a client will subscribe the term deposit in a telemarketing campaign

Original Dataset
20 Predictors, 1 Binary Response

I. Predictors

San Zhang

1.Age
2.Type of job
3.Marital status
4.Education level
5.Credit in default (Y/N)
6.Housing loan (Y/N)
7.Personal loan? (Y/n)

8.Contact communication type (cellular/telephone)
9.Last contact month of year (march-december)
10.Last contact day of week (monday-friday)
11.Duration
12.Number of contacts performed in this campaign
13.Number of days passed by last contacted from previous campaign
14.Number of contacts performed before this campaign
15.Outcome of the previous marketing campaign (S/F/nonexistent)

16.Employment variation rate
17.Consumer price index
18.Consumer confidence index
19.EURIBOR 3-month rate
20.Number of employees

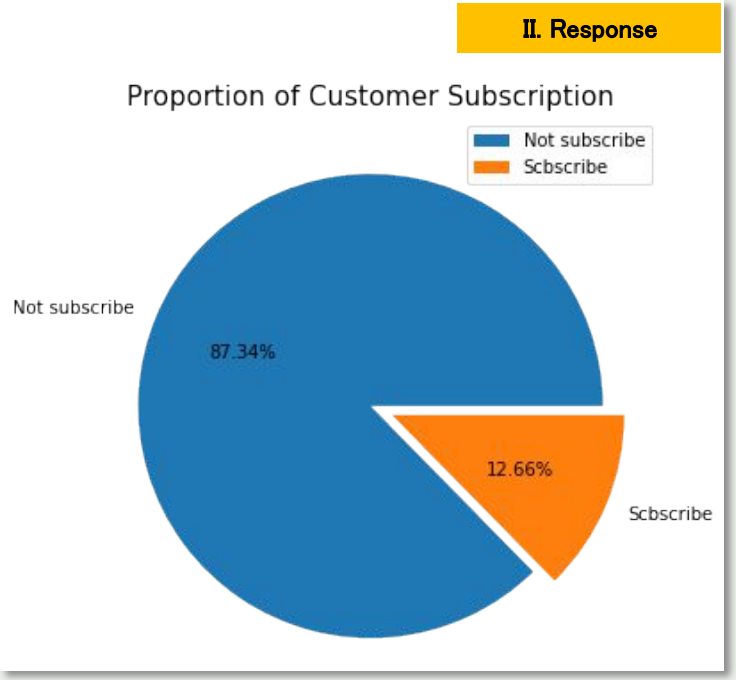
Numeric Categorical

I. Personal Information Of Bank Client

II. Customer Relationship & Contact Info

III. Social & Economic Context Attributes

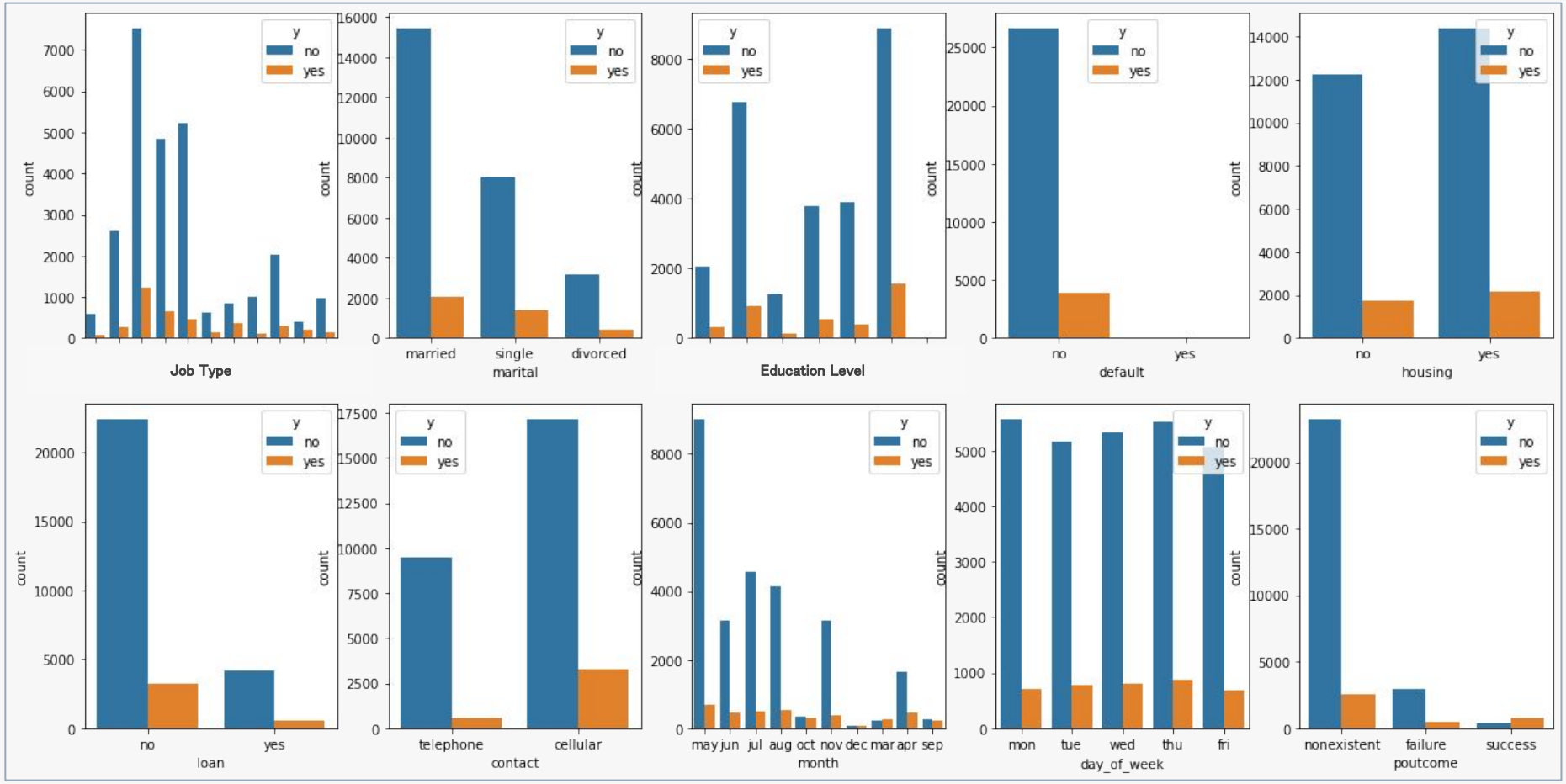
II. Response



Visualization and Feature Selection

Feature Selection of Categorical Variables

Chi-Square Testing



Variable	H0*	P-value
Job	Reject	2.0e-150
Marital	Reject	1.5e-12
Edu	Reject	1.4e-22
Default	Hold	0.83
Housing	Hold	0.08
Loan	Hold	0.38
Contact	Reject	4.90e-139
Month	Reject	0
Day of W	Reject	3.7e-5
P-outcome	Reject	0

**H0: predictor and response variable are independent.*

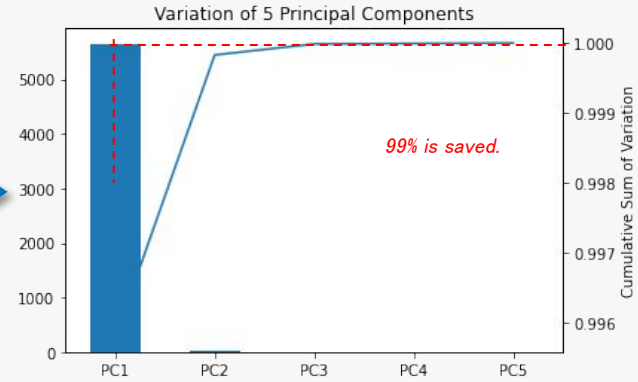
Feature Selection of Numeric Variables

Principle Component Analysis on social and economic context attributes

Look for a linear subspace

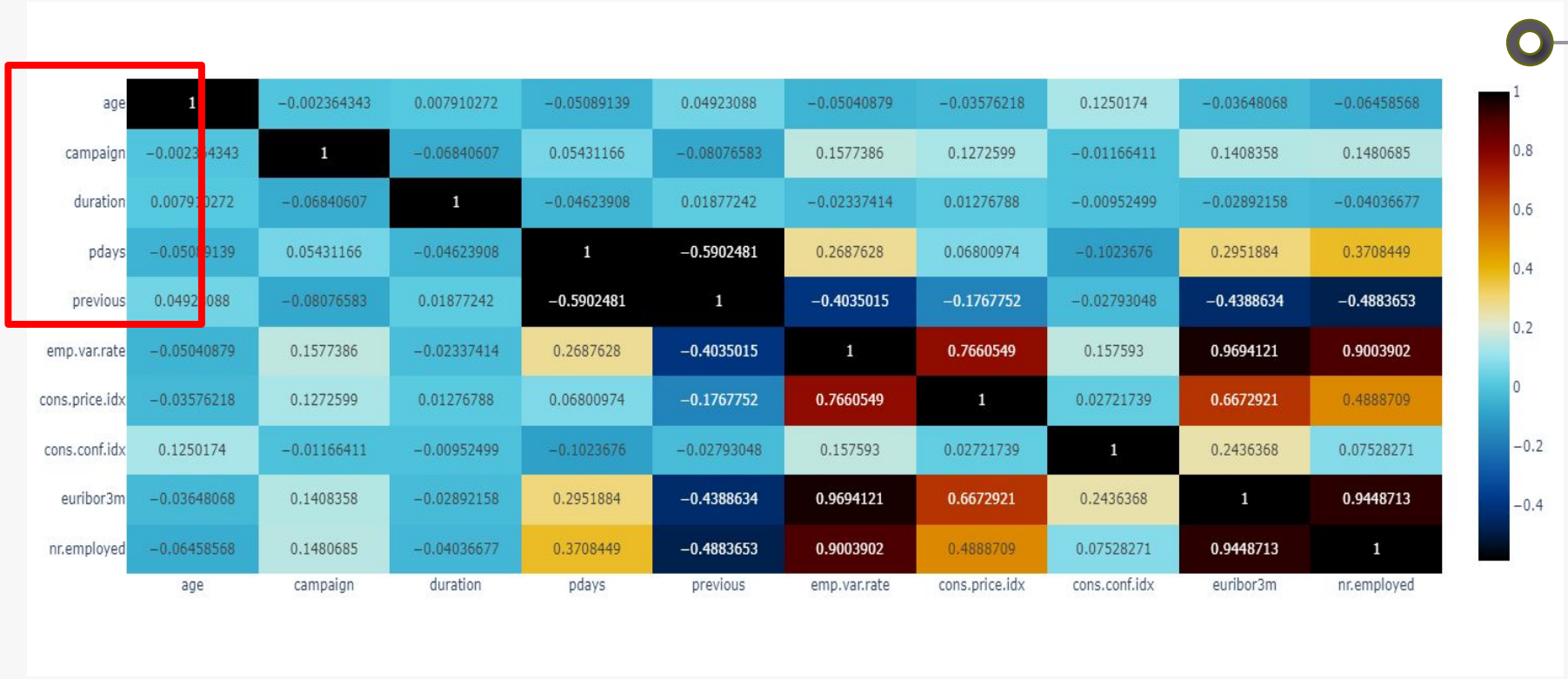
Redundancy information

– Apply PCA



Feature Selection of Numeric Variables (cont.)

1. Point Biserial Testing on other numerical variables; 2. Convert to Categorical Variables



1. Apply point biserial test

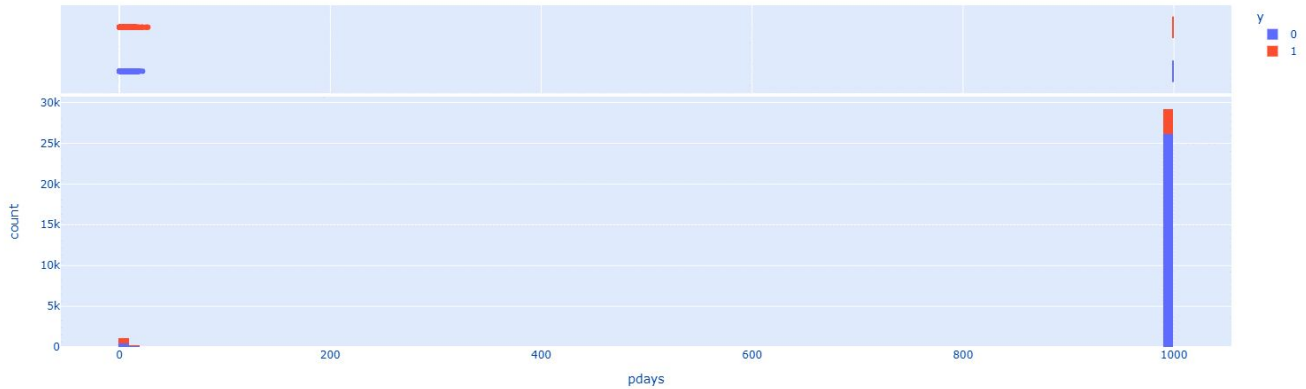
Variable	PB Corr.	P-value
Age	0.05	1.69e-17
Campaign	−0.07	1.08e-33
Duration	0.40	0
P-days	−0.33	0
Previous	0.23	0

2. Convert to a new categorical variables

Feature Selection of Numeric Variables (cont.)

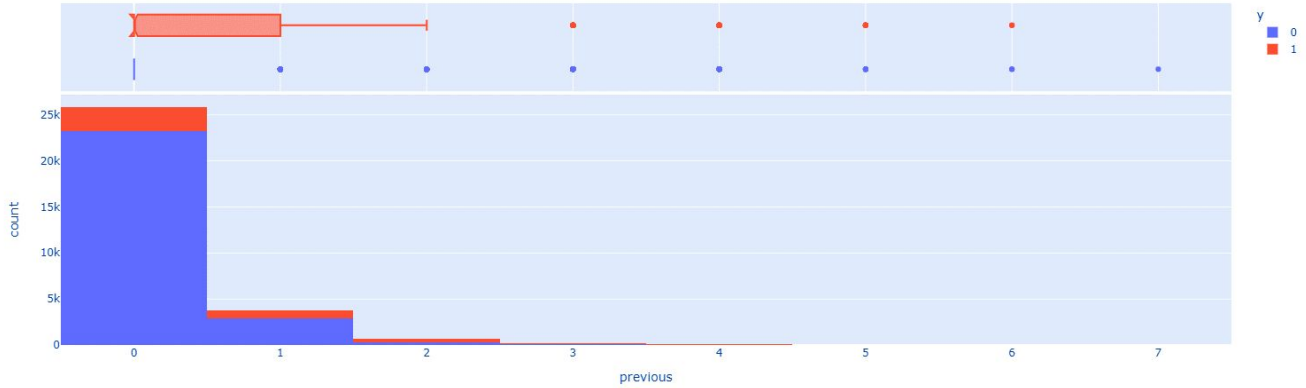
Situation of Previous Contacts

Distribution of **P-Days** in 2 Groups



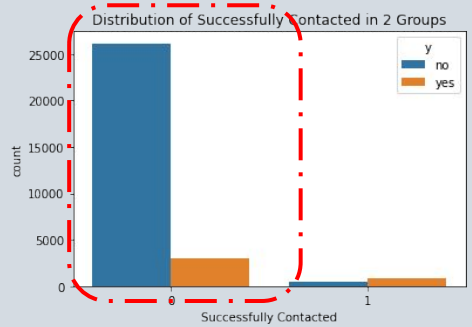
** Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted).*

Distribution of **Number of Previous Phone Calls** in 2 Groups



** Number of contacts performed before this campaign and for this client*

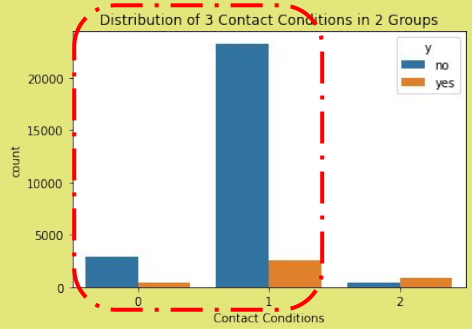
- *Main distribution of P-days are of 999 days, which means **most clients were new clients**.*



1: Successfully contacted previously

0: Contacted but didn't pick up or no previous phone calls at all

- *Among these new clients, they still had non-zero contacts records. These clients might be **on the contact list, but they are not likely to pick up the phone call**.*



2: Successfully contacted previously

1: No previous phone calls at all

0: Contacted but didn't pick up

General Review of Feature Selection Process

Original Feature

1.Age
2.Type Of Job
3.Marital Status
4.Education Level
5.Credit In Default (Y/N)
6.Housing Loan (Y/N)
7.Personal Loan? (Y/N)

8.Contact Communication Type
9.Last Contact Month Of Year
10.Day Of Week
11.Duration (Benchmark)
12.Number Of Contacts In The Campaign
13.Number Of Days Passed
14.Number Of Contacts Performed
15.Outcome Of The Previous Campaign

16.Employment Variation Rate
17.Consumer Price Index
18.Consumer Confidence Index
19.EURIBOR 3-month Rate
20.Number Of Employees

Numeric Categorical

Feature Selection

Statistical Testing

Convert To One Categorical Variable

PCA To Find Linear Subspace

Drop Unnecessary Variables

Feature Selection Result and Evaluation

1.Age
2.Type Of Job
3.Marital Status
4.Education Level

Removed by Statistical Testing

5.Contact Communication Type
6.Last Contact Month Of Year
7.Day Of Week
8.Duration
9.Number Of Contacts In The Campaign
10.Previous Contact Condition

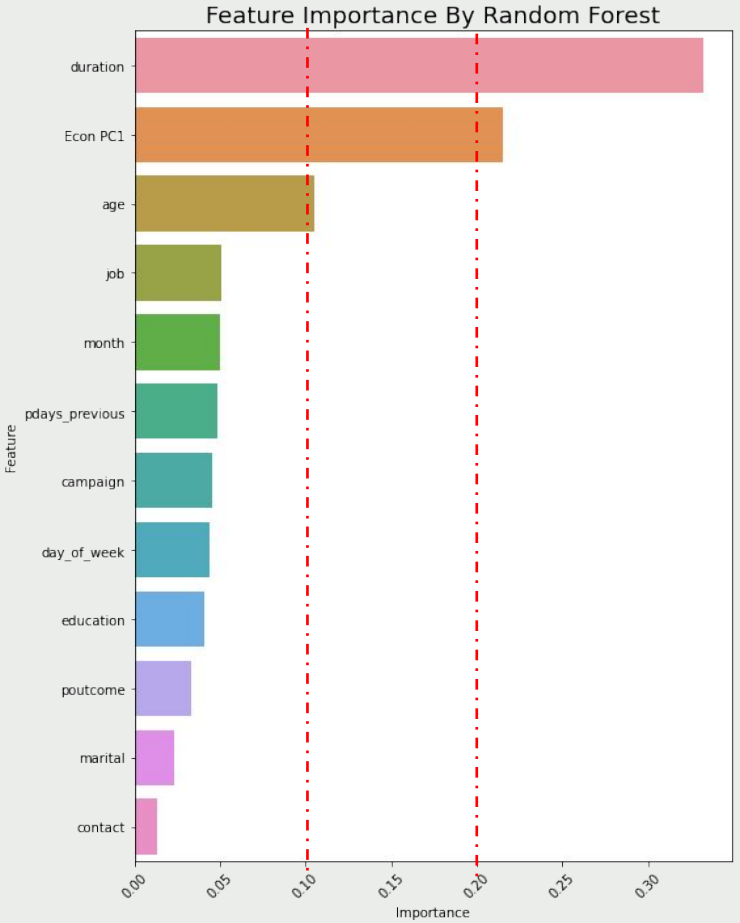
Combined with others

11.Outcome Of The Previous Campaign

12.Econ PC1

Combined with others by PCA

Numeric Categorical



Modeling

Logistic Regression

Logistic Regression

Tolerance *0.0000005*

Max Iteration *50000*

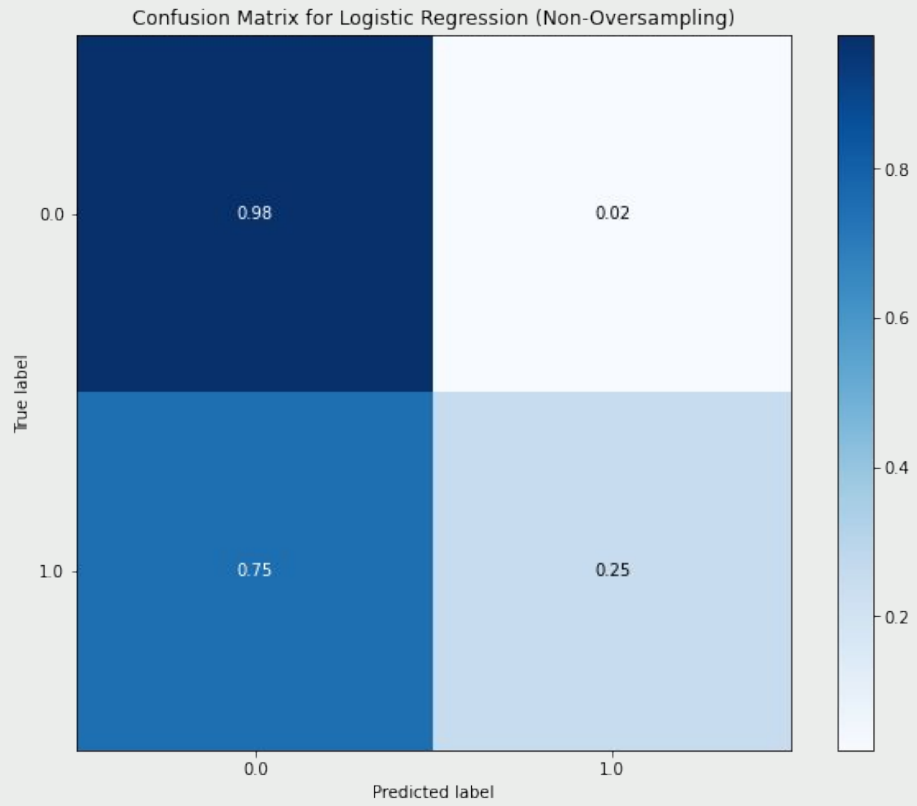
Learning Rate *0.000001*

Penalize *0.05*

Timing *03:04*

Training Score	
F1 Score	0.367
Precision	0.705
Recall	0.248
Validation Score	
F1 Score	0.369
Precision	0.702
Recall	0.251

Non-Oversampling



Logistic Regression

Tolerance *0.0000005*

Max Iteration *50000*

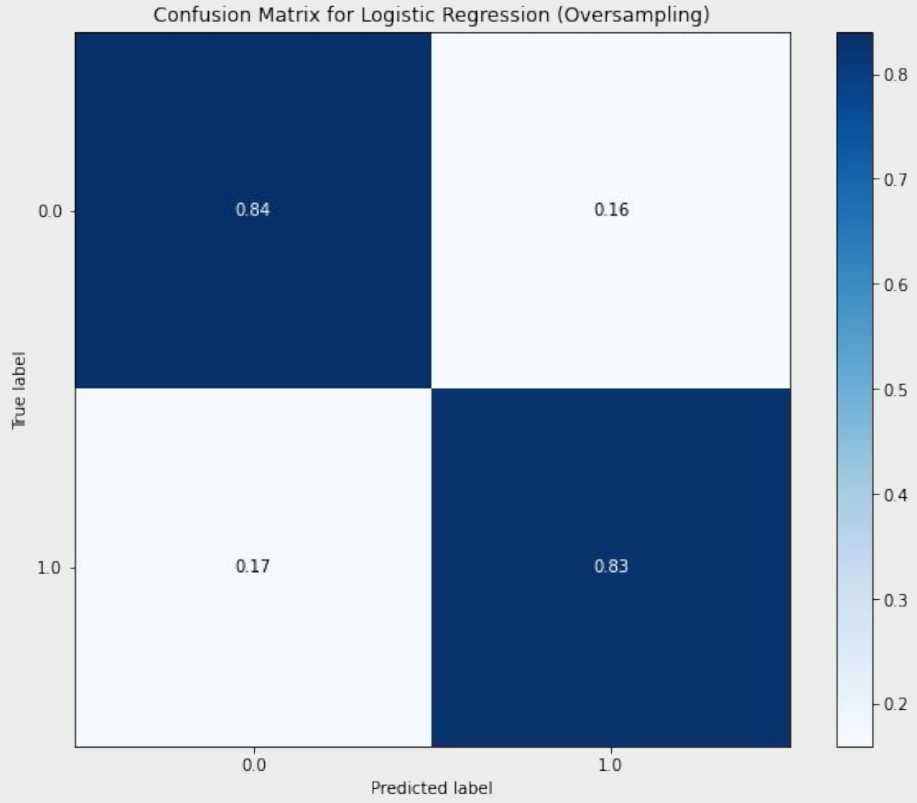
Learning Rate *0.000001*

Penalize *0.05*

Timing *07:38*

Training Score	
F1 Score	0.824
Precision	0.833
Recall	0.815
Validation Score	
F1 Score	0.566
Precision	0.431
Recall	0.825

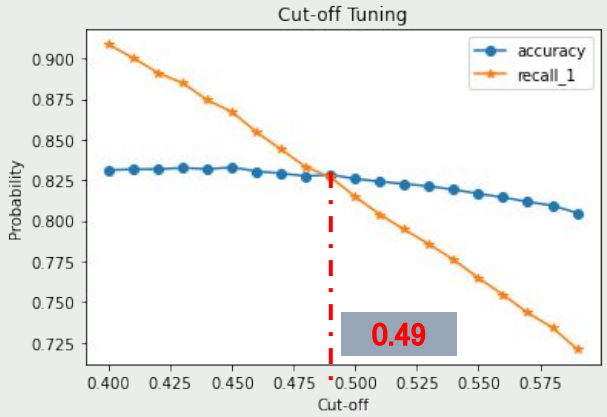
Oversampling

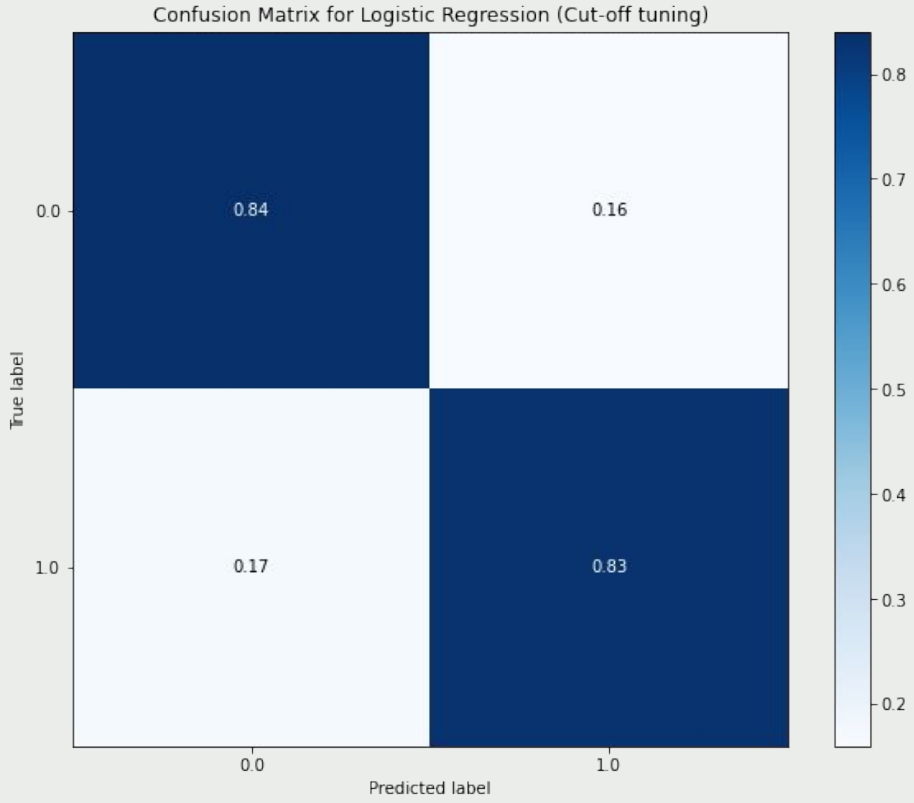


Logistic Regression

Cutoff = 0.49

Training Score	
F1 Score	0.828
Precision	0.829
Recall	0.827
Validation Score	
F1 Score	0.564
Precision	0.425
Recall	0.838



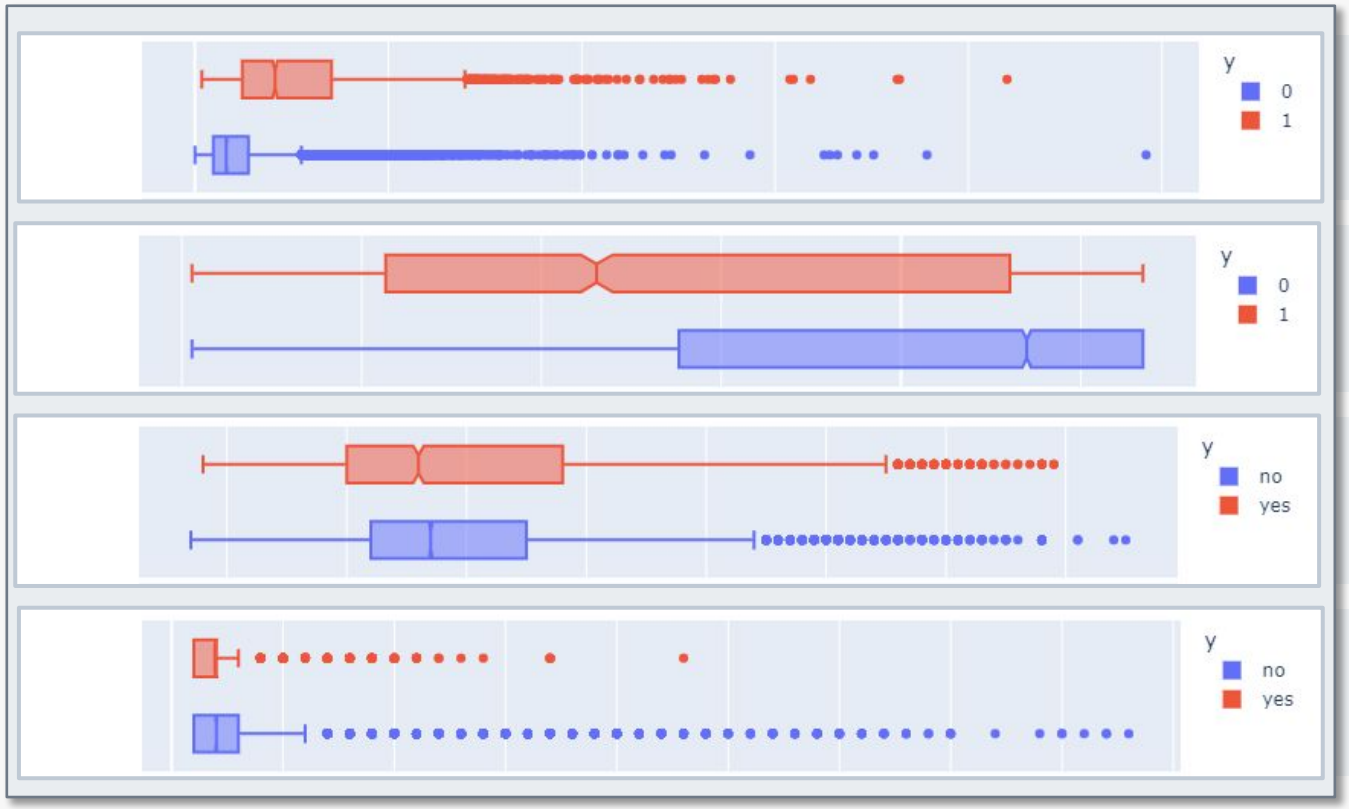


Modeling

Naive Bayes

Naive Bayes Binning Numerical Features

4 Origin Numerical Features Distribution



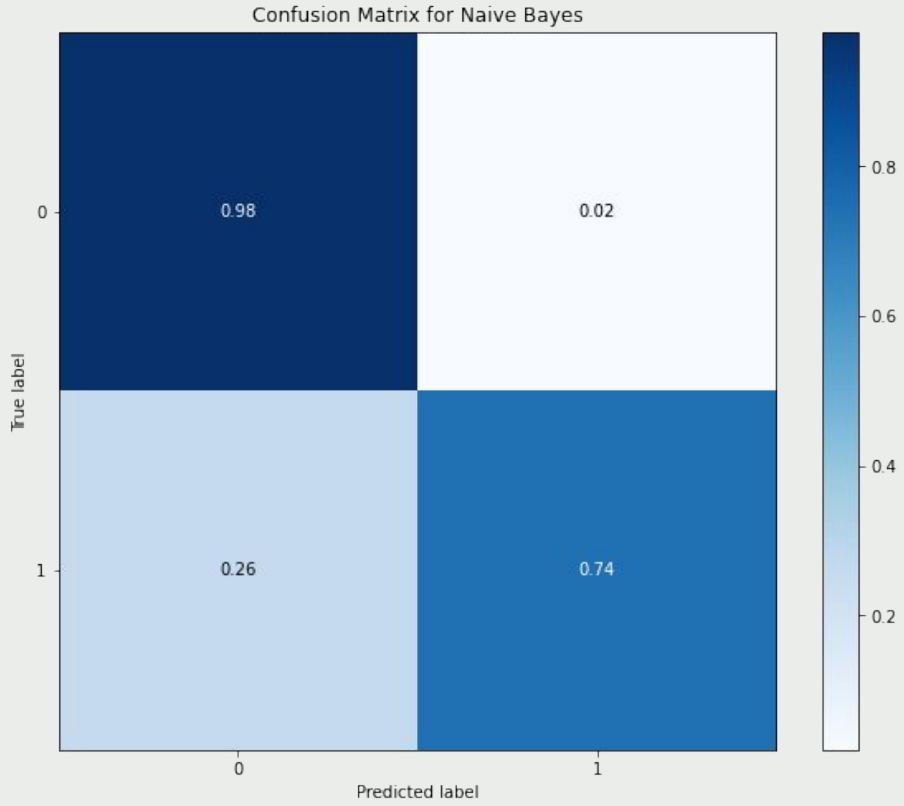
Best combination with lowest error

Duration	in range (5,11)	**Duration** bin = 5
PC1	in range (5,11)	**PC1** bin = 10
Age	in range (2,4)	**Age** bin = 2
Campaign	in range (2,4)	**Campaign** bin = 2

Naive Bayes Classifier

Training Score	
F1 Score	0.82
Precision	0.93
Recall	0.73
Validation Score	
F1 Score	0.82
Precision	0.92
Recall	0.74

No need to Oversampling



Modeling

SVM

SVM (hinge loss)

Max Iteration *1000*

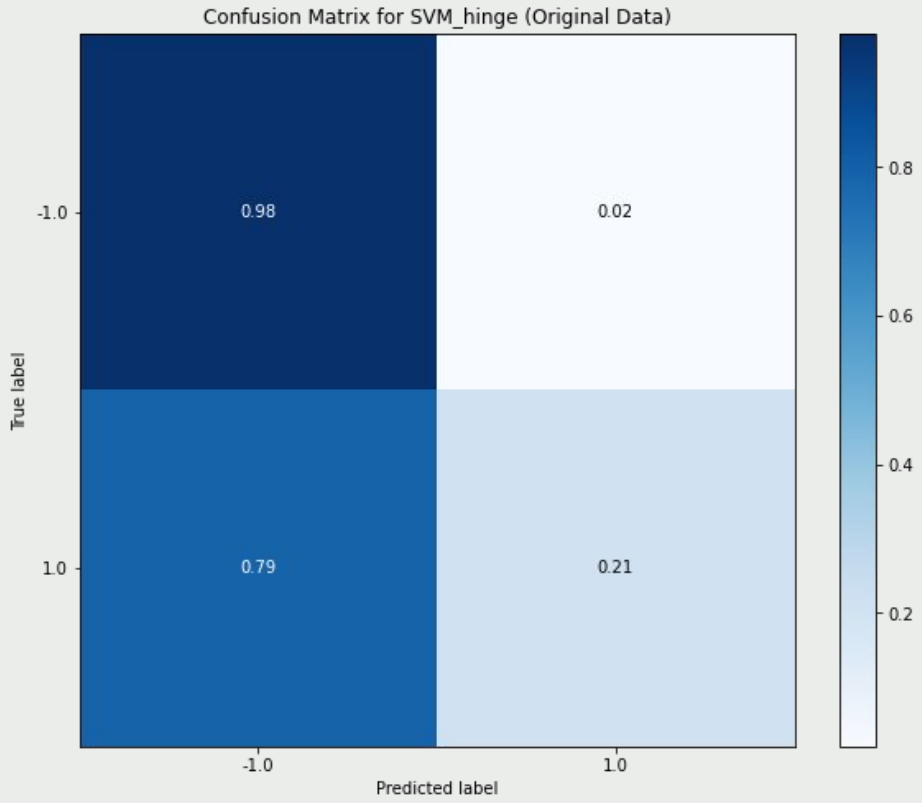
Learning Rate *0.0001*

Penalize *0.001*

Timing *03:11*

Training Score	
F1 Score	0.31
Precision	0.65
Recall	0.20
Validation Score	
F1 Score	0.32
Precision	0.66
Recall	0.21

Non-undersampling



SVM (hinge loss)

Max Iteration *1000*

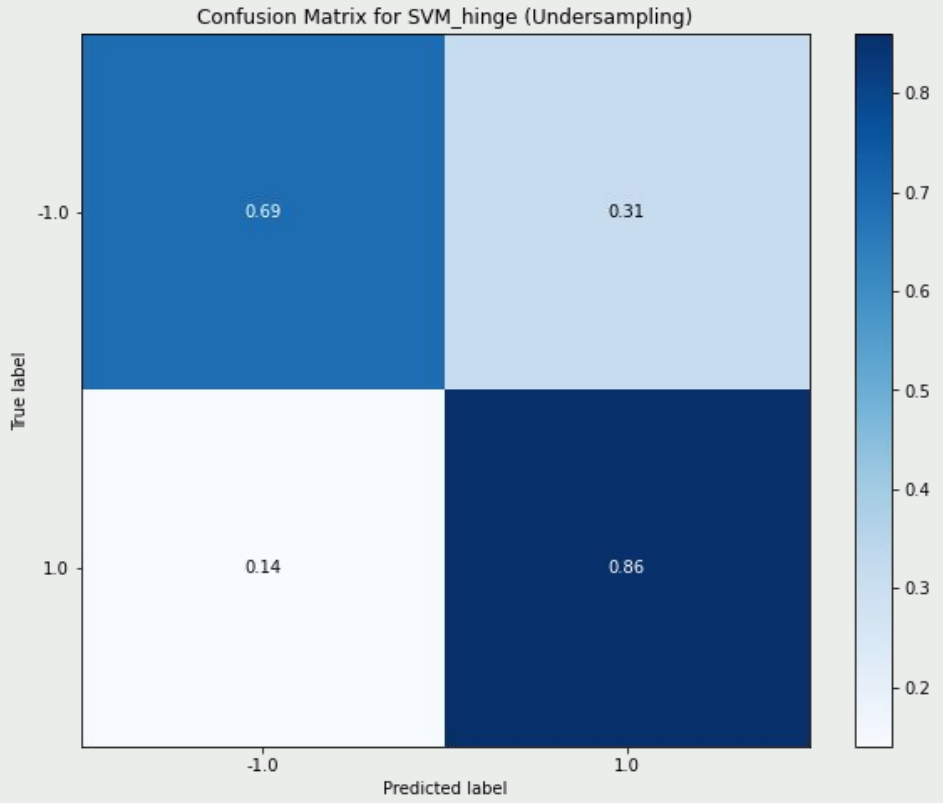
Learning Rate *0.0001*

Penalize *0.001*

Timing *05:42*

Training Score	
F1 Score	0.42
Precision	0.28
Recall	0.87
Validation Score	
F1 Score	0.46
Precision	0.28
Recall	0.86

Undersampling



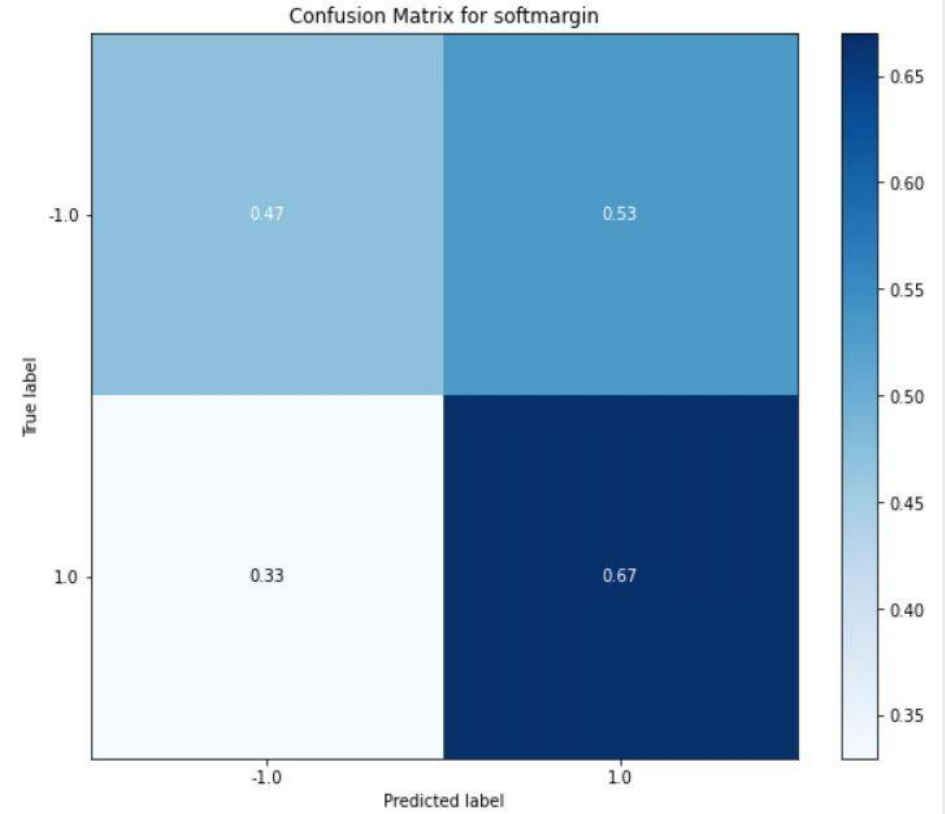
SVM (scipy.optimize)

C = 20

Timing 4:21:00

Training Score	
F1 Score	0.59
Precision	0.52
Recall	0.68
Validation Score	
F1 Score	0.61
Precision	0.56
Recall	0.67

Undersampling



SVM (SMO Algorithm)

C = 1

tolerance = 0.00005

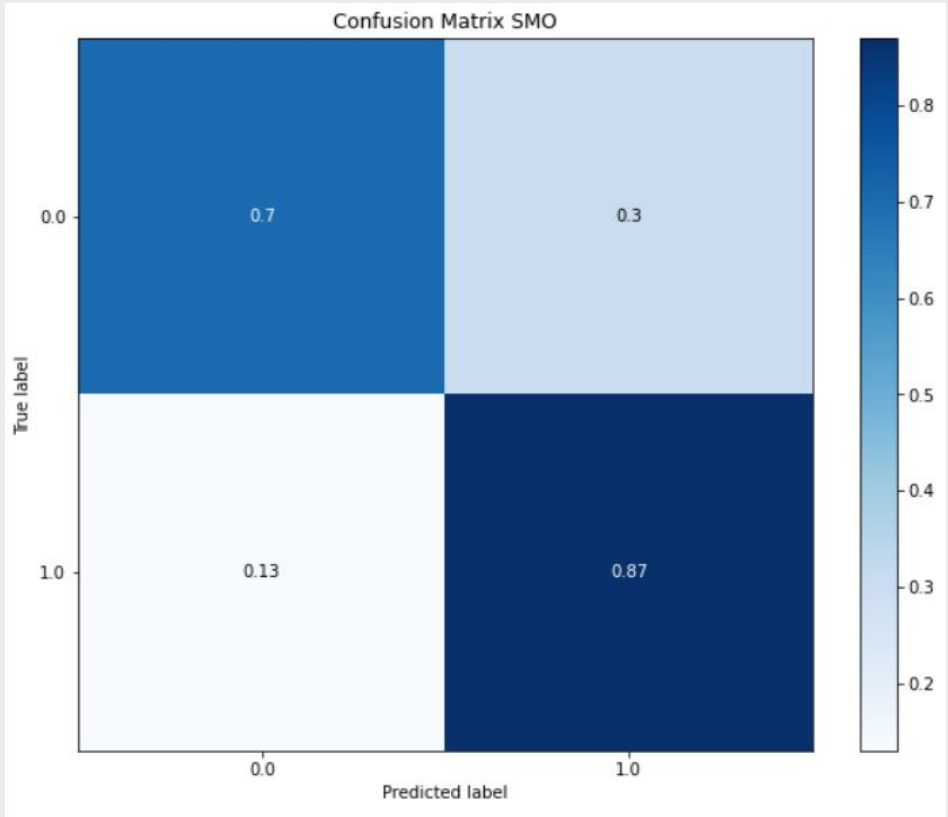
Max Iteration = 100

Kernel Gamma = 5

Timing 1:19:00

Training Score	
F1 Score	0.44
Precision	0.29
Recall	0.94
Validation Score	
F1 Score	0.44
Precision	0.29
Recall	0.87

Undersampling



Modeling

ANN

Neural Network

1. Cost function(self-definition)

Definition:

$$Loss = -\frac{1}{output\ size}\sum_{i}^{output\ size} y_i log(sigmoid\hat{y_i}) + (1 - y_i)log(sigmoid(1 - \hat{y_i}))$$

[1e-10, 1] [1e-10, 1]

Regularization:

$$Cost\ function = Loss + \frac{\lambda}{2m} * \sum \|w\|^2$$

2. Cost function(sigmoid cross-entropy)

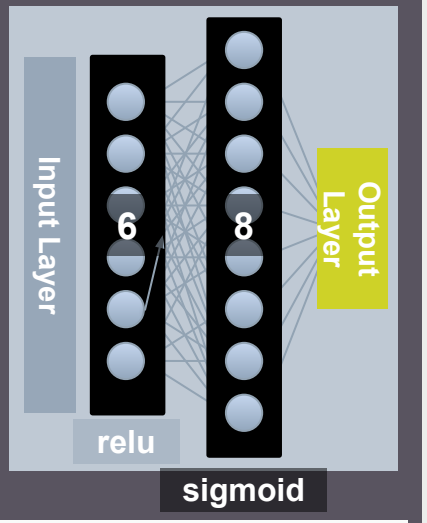
Neural Network

Solver Adam

Learning Rate 0.0001

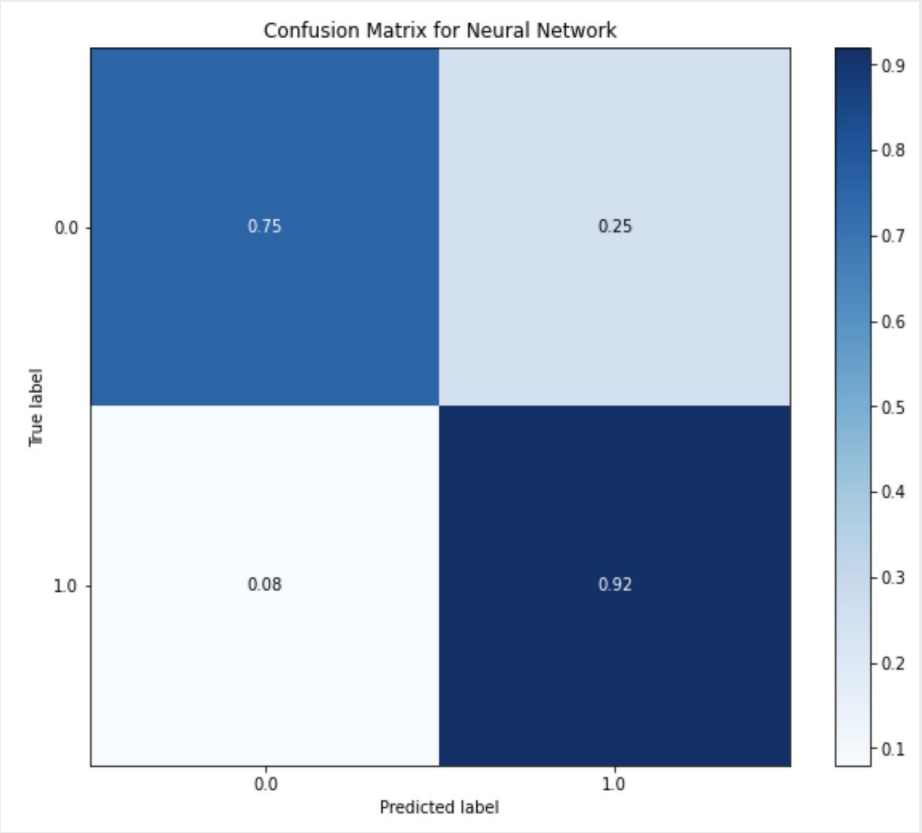
Penalize 0.01

Timing 25:00



Training Score	
F1 Score	0.824
Precision	0.833
Recall	0.815
Validation Score	
F1 Score	0.5
Precision	0.34
Recall	0.92

Cost function (self-definition)



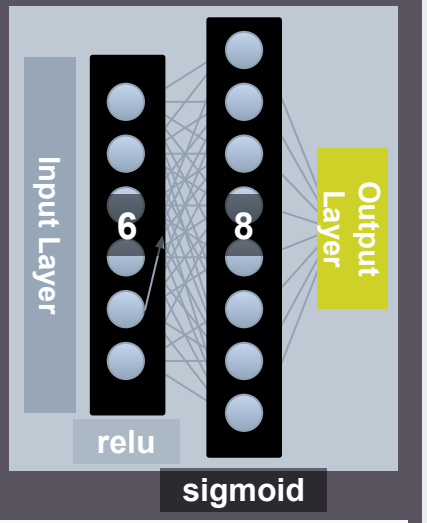
Neural Network

Solver Adam

Learning Rate 0.0001

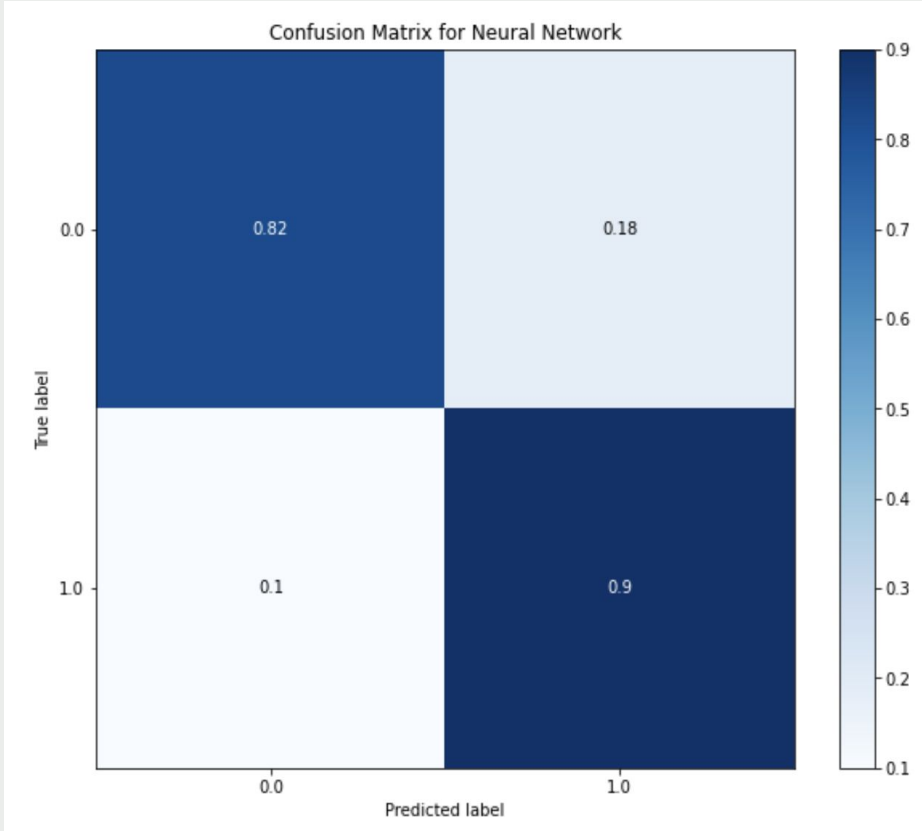
Penalize 0.01

Timing 20:00

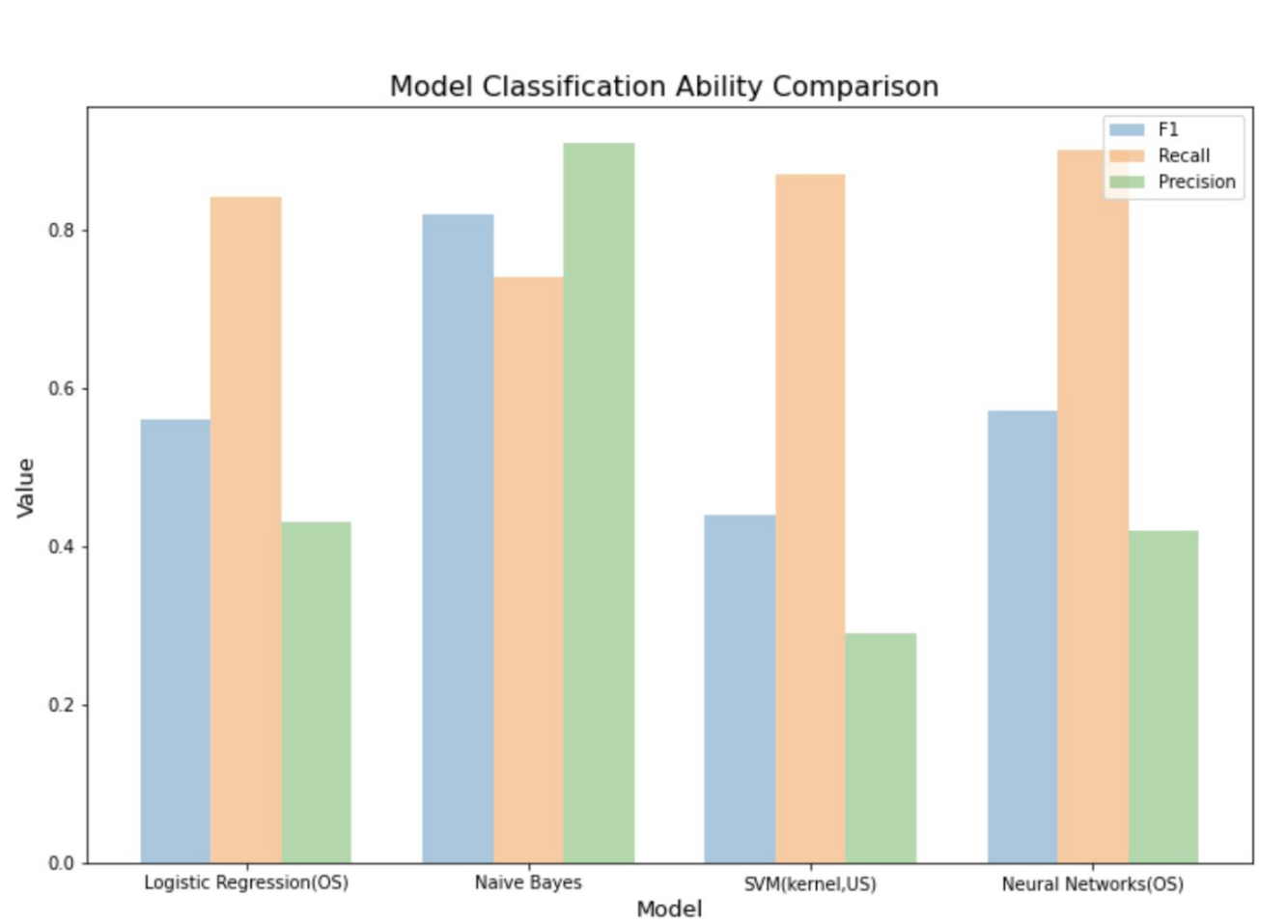


Training Score	
F1 Score	0.855
Precision	0.828
Recall	0.883
Validation Score	
F1 Score	0.571
Precision	0.417
Recall	0.899

Cost function (sigmoid_crossentropy)



Model Comparison



- Expense is not expensive & Don't allow lose new clients(quantity): **Recall**
- Expense is expensive(quality): **Precision**
- Generally: **F1 score**

Improvement

- Data Encode
- Predictors Selection
- Parameters Tuning
- Time and Capacity

Thank You

Q&A

Group 6