LOGISTIC REGRESSION

Ratchainant Thammasudjarit, Ph.D.

- Goal

X_1	X_2	X_3	Y
Prepaid	3	Yes	?

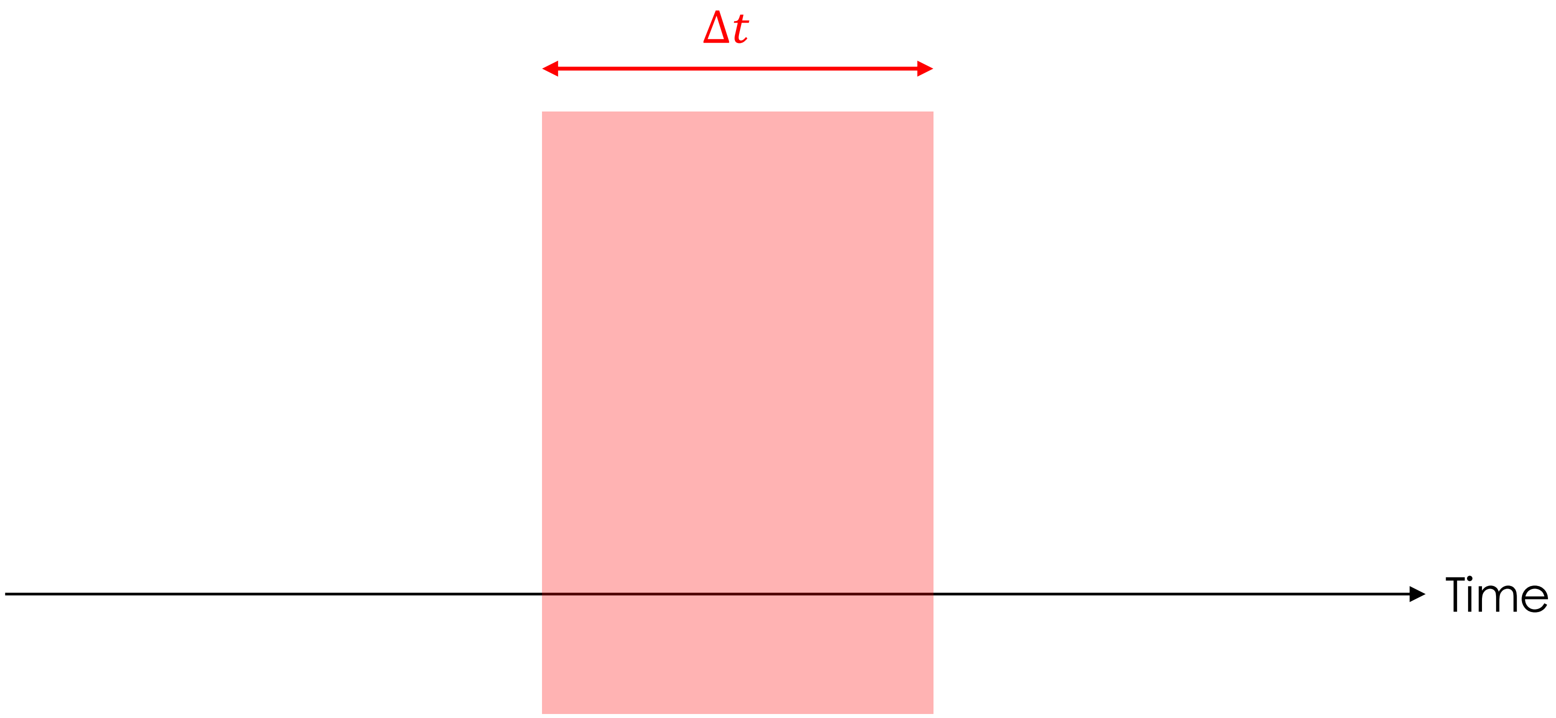
X_1 be customer group either prepaid or postpaid
X_2 be the calling frequency of calling to call center
X_3 be the top-up package (Yes, No)

Scenario

Churn Prediction:

You want to predict how much likely a given customer will churn.

- Data Collection



Inclusion Criteria:

- Any customer who has never churned from our service

Exclusion Criteria

- None

Scenario

Churn Prediction:

Study design is the cross-sectional study

- Application

$P(Churn) = 0.1$

Action

Let's discuss

$P(Churn) = 0.8$

Action

Let's discuss

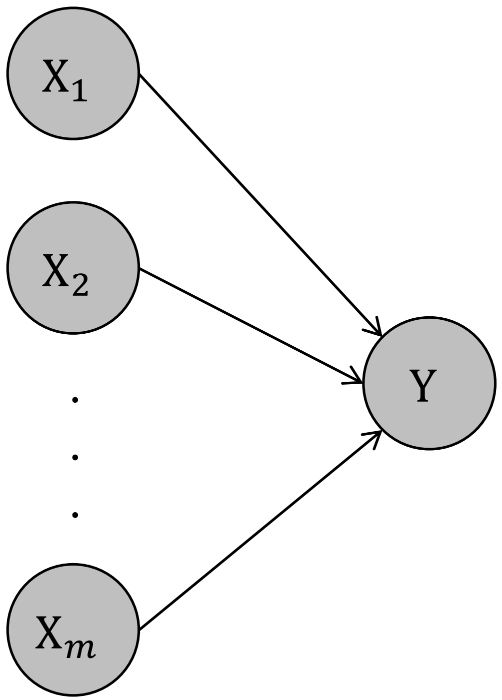
How to estimate $P(Churn)$

Scenario

Churn Prediction:

Make prediction and take actions

◦ Concepts



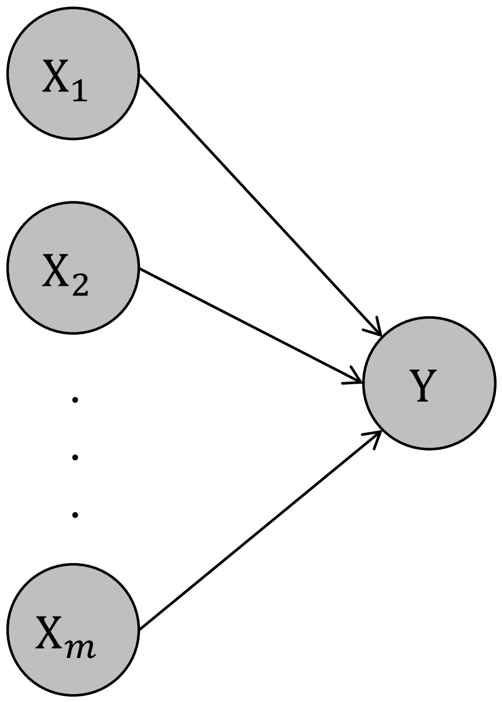
Multivariate Problem

Describing relationship between a set of independent random variables $\mathbf{X}$ and dependent variable Y

Note:

$$\mathbf{X} = \{X_1, X_2, \dots, X_m\}$$

◦ Concepts



Multivariate Problem

For any $X_i, i = 1, 2, \dots, m$ and Y

Their values are defined as

$X_i = \langle x \mid x \in S_i \rangle$ and

$Y = \langle y \mid y \in C \rangle$

where

S be a sample space

C be a set of class labels

Multivariate Problem

For any $\mathrm{X}_i, i = 1, 2, \ldots, m$ and Y

Their values are defined as

$\mathrm{X}_i = \langle x \mid x \in S_i \rangle$ and

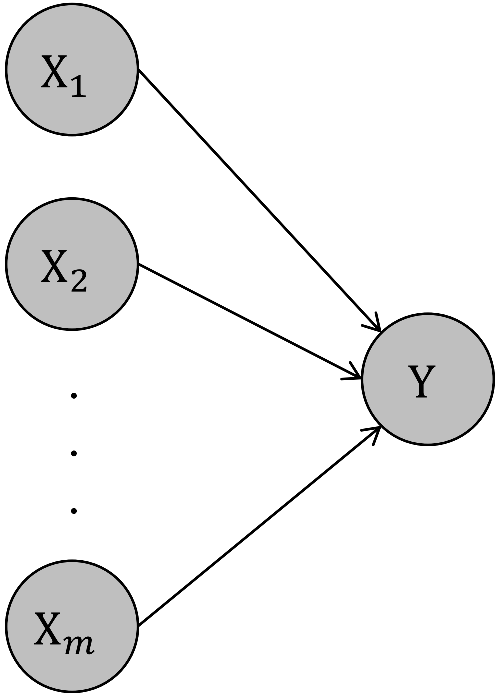
$\mathrm{Y} = \langle y \mid y \in C \rangle$

where

S be a sample space

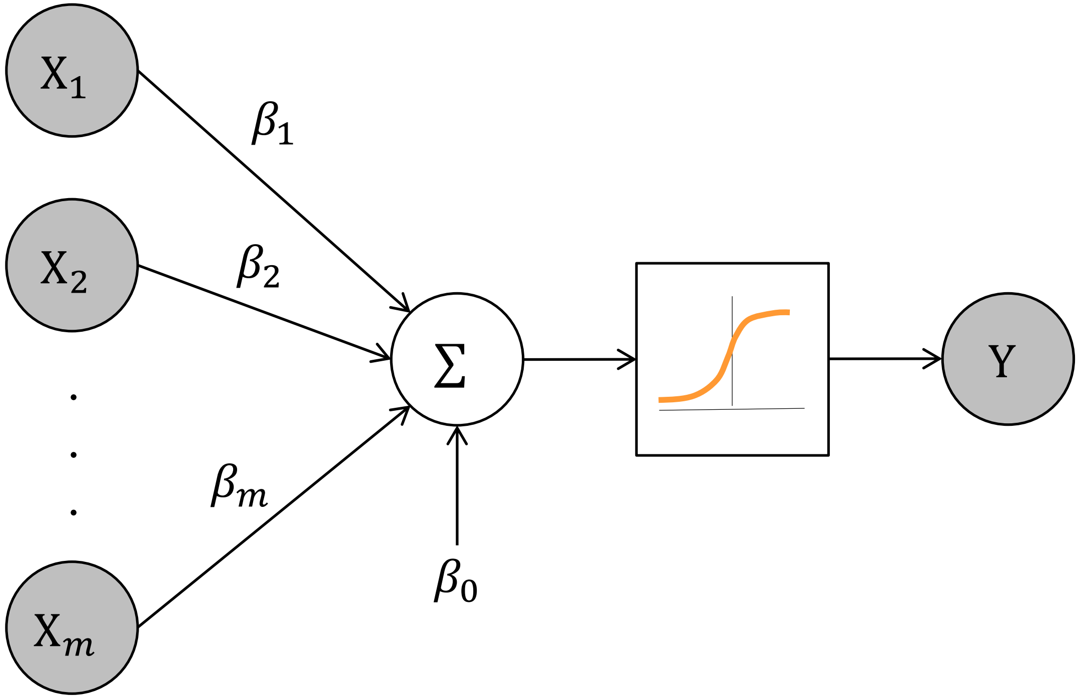
C be a set of class labels

- Example
  - $\mathrm{X}_1 = \langle x_1 \mid x_1 \in \{\mathrm{High}, \mathrm{Low}\} \rangle$
  - $\mathrm{X}_2 = \langle x_2 \mid x_2 \in I^+ \rangle$
  - $\mathrm{X}_m = \langle x_m \mid x_m \in \mathbb{R} \rangle$
  - $\mathrm{Y} = \langle y \mid y \in \{\mathrm{Yes}, \mathrm{No}\} \rangle$



X_1	X_2	…	X_m	Y
High	12	…	-20.5	Yes
Low	10	…	35.2	No
Low	25	…	10.1	Yes
High	30	…	9.5	No

◦ Model



Logistic Regression

Given a dataset $\mathcal{D} = \langle \boldsymbol{X}, \mathbf{Y} \rangle$

The original form of logistic regression is defined as

$$P(y = 1|\mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}$$

- Applying the trained logistic regression to the follow up study

X_1	X_2	X_3	Y
Prepaid	2	No	1
Prepaid	4	No	1
Prepaid	6	No	1
Prepaid	0	Yes	0
Prepaid	2	Yes	0
Prepaid	1	Yes	0
Postpaid	3	No	1
Postpaid	7	Yes	1
Postpaid	5	No	1
Postpaid	1	Yes	0
Postpaid	2	No	0
Postpaid	3	Yes	0

Logistic Regression

$X_1 \in \{\text{prepaid}, \text{postpaid}\}$

$X_2 \in I^+$

$X_3 \in \{Yes, No\}$

12 samples

- Applying the trained logistic regression to the follow up study

X_{pre}	X_{post}	X_2	X_3	Y
1	0	2	0	1
1	0	4	0	1
1	0	6	0	1
1	0	0	1	0
1	0	2	1	0
1	0	1	1	0
0	1	3	0	1
0	1	7	1	1
0	1	5	0	1
0	1	1	1	0
0	1	2	0	0
0	1	3	1	0

Logistic Regression

X_{pre} and X_{post} are dummy variables

We choose either X_{pre} or X_{post} because X_1 is a dichotomous variable

Suppose we choose X_{pre} and discard X_{post}

- Applying the trained logistic regression to the follow up study

$$P(Churn) = \frac{1}{1 + e^{2.61 - 0.33X_{pre} + 0.84X_{topup} - 1.01X_{call}}}$$

where

$\beta_0: -2.61$
$\beta_{pre}: 0.33$
$\beta_{topup}: -0.84$
$\beta_{call}: 1.01$

Logistic Regression

X_{pre} and X_{post} are dummy variables

From machine learning perspective, we choose either X_{pre} or X_{post} because X_1 is a dichotomous variable

However, from interpretation perspective, we may choose both X_{pre} and X_{post}

Prediction

Given the customer who use prepaid. He called to call center 3 times and never use any top-up package

Determine the estimated $P(Churn)$

- Applying the trained logistic regression to the follow up study

$$P(Churn) = \frac{1}{1 + e^{2.61 - 0.33(1) + 0.84(0) - 1.01(3)}}$$

$$= 0.68$$

- To conclude whether this customer will churn or not, in general, we use the cutoff value equal to 0.5 as the decision threshold

$$\hat{y} = \begin{cases} 1 & if\ P(churn) \geq 0.5 \\ 0 & if\ P(churn) < 0.5 \end{cases}$$

- In conclusion, this customer is likely to churn

- Given the model

$$P(Churn) = \frac{1}{1 + e^{2.61 - 0.33X_{pre} + 0.84X_{topup} - 1.01X_{call}}}$$

Risk Ratio

Logistic regression does not only provide prediction results but also describe relationship between each predictor and outcome

Let's consider the Risk Ratio (RR)

- Scenario

<u>Customer A</u>
Group: Prepaid
Top-up: **No**
#Call: 3

<u>Customer A</u>
Group: Prepaid
Top-up: **Yes**
#Call: 3

Risk Ratio

In the follow up study, we measure risk ratio to estimate the treatment effect

Given the scenario, We detected the Customer A is likely to churn

Our call center took action by calling to offer the better top-up package

- Estimate $P(Churn)$ before and after treatment

<u>Before</u>
Group: Prepaid
Top-up: **No**
#Call: 3

$$P_{before}(Churn) = \frac{1}{1 + e^{2.61 - 0.33(1) + 0.84(0) - 1.01(3)}}$$
$$= 0.68$$

<u>After</u>
Group: Prepaid
Top-up: **Yes**
#Call: 3

$$P_{after}(Churn) = \frac{1}{1 + e^{2.61 - 0.33(1) + 0.84(1) - 1.01(3)}}$$
$$= 0.47$$

- Risk Ratio

$$\frac{P_{before}(Churn)}{P_{after}(Churn)} = \frac{68\%\, risk}{47\%\, risk} = 1.44$$

Risk Ratio

Substitute each customer variable into the model to estimate $P(Churn)$

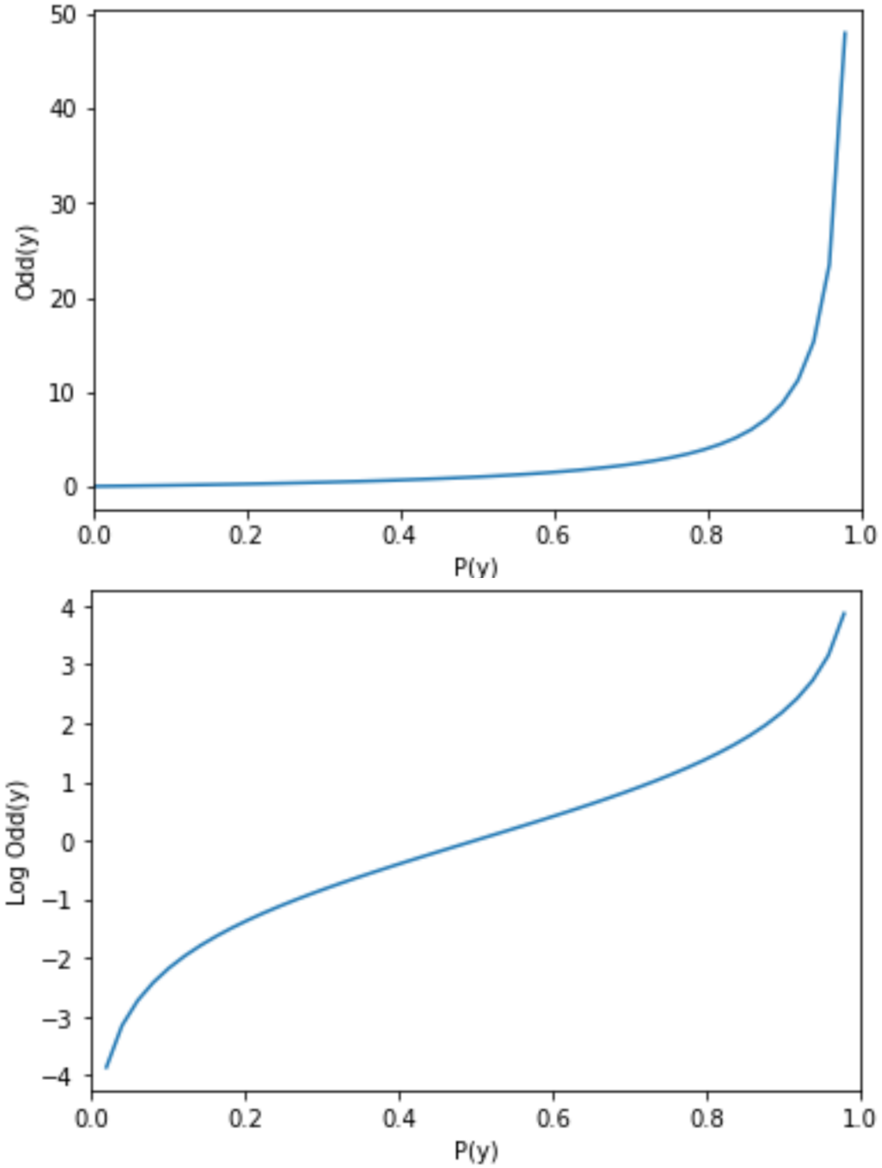
Findings:

Given a prepaid customer who used to call to call center for 3 times

The customer without top-up package has 68% risk to churn, whereas the customer with top-up package has 47% risk to churn over the period of follow-up study

In other words, offer the top-up package to this customer reduce risk to churn about one-third of the existing risk

- Relationship between Odd and Probability



Odd

Odd of any event y is defined as follows

$$Odd(y) = \frac{P(y)}{1 - P(y)}$$

Logit Transformation

An alternative of Logistic regression is the Logit transformation as follows

$$logit\ P(y) = \log Odd(y)$$
$$= \log \frac{P(y)}{1-P(y)}$$

where

$$P(y) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}$$

- Let's do some math

$$\frac{P(y)}{1 - P(y)} = \frac{\dfrac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}}{1 - \dfrac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}}$$

$$= \frac{\dfrac{1}{\cancel{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}}}{\dfrac{e^{-(\beta_0 + \sum \beta_i x_i)}}{\cancel{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}}}$$

$$= \frac{1}{e^{-(\beta_0 + \sum \beta_i x_i)}}$$

- Taking log function for both side

$$\log \frac{P(y)}{1 - P(y)} = \log \frac{1}{e^{-(\beta_0 + \sum \beta_i x_i)}}$$

$$\boxed{\log Odd(y) = \beta_0 + \sum_i \beta_i x_i}$$

Linear form

Logit Transformation

- Original Form

$$P(y) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}$$

- Logit Form

$$\log Odd(y) = \beta_0 + \sum_i \beta_i x_i$$

The main difference between the two formulae is that the expression with the $\boldsymbol{X}$ is more specific

The original formula assumes that the probabilities describe the risk for developing the outcome

The logit form of the logistic model gives an expression for the **log odds of developing the outcome** for an individual with a specific set of $\mathbf{X}$

Logit Transformation

What does it mean if all X are zeros

- Logit Form

$$\log Odd(y) = \beta_0 + \sum_i \beta_i x_i$$

- If all X are zeros

$$\log Odd(y) = \beta_0$$

- One interpretation is that a gives the log odds for a person with zero values for all X but this interpretation has serious limitation
  - There may not be any person in the population of interest with zero values on all the X

- Second interpretation is that β_0 gives the **log of the baseline odds** when all X are ignored or unknown
  - By baseline odds, we mean the odds that would result for a logistic model without any X at all

- What if β_i is varied and others are fixed

- Example

$$\log Odd(y) = \beta_0 + \beta_{pre} X_{pre} + \beta_{topup} X_{topup} + \beta_{call} X_{call}$$

(purple)	Fixed by default
(red)	Fixed by controlled
(green)	Vary

Coefficient

With regard to the odds, we need to consider what happens to the logit when only one of the X varies while keeping the others fixed

- Sample 1: Prepaid, Without top-up, 3 calls

$$\log Odd(y|X_{pre}=0) = \beta_0 + \beta_{pre}(1) + \beta_{topup}(0) + \beta_{call}(3)$$

$$= \beta_0 + \beta_{pre} + 3\beta_{call}$$

1

- Sample 2: Prepaid, With top-up, 3 calls

$$\log Odd(y|X_{pre}=1) = \beta_0 + \beta_{pre}(1) + \beta_{topup}(1) + \beta_{call}(3)$$

$$= \beta_0 + \beta_{pre} + \beta_{topup} + 3\beta_{call}$$

2

- Subtraction 2 and 1

$$\Delta X_{topup} = \beta_{topup}$$

Coefficient

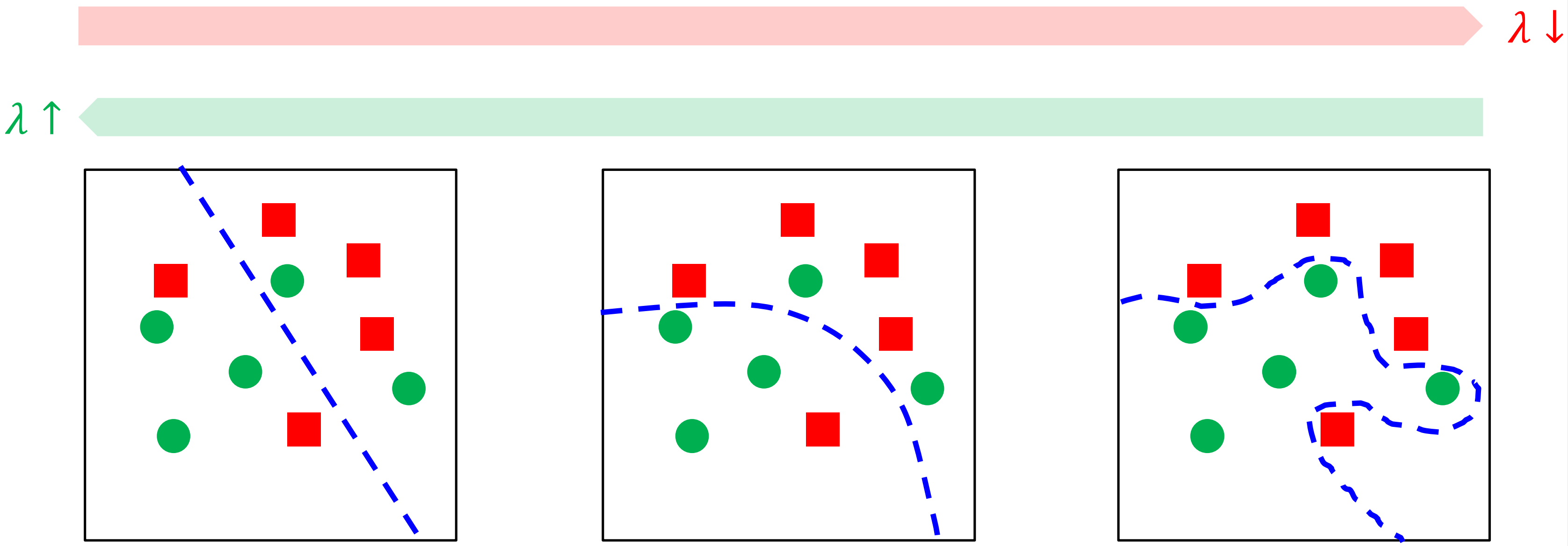
We will demonstrate again with two samples

Conclusions:

For dichotomous variable, the coefficient represents the contribution of such a variable to the log odd

For numeric variable, the coefficient represents the change of log odd when such a variable changes in 1 unit

◦ Bias and Variance



High Bias (Underfitting)

Just right

High Variance (Overfitting)

Regularization

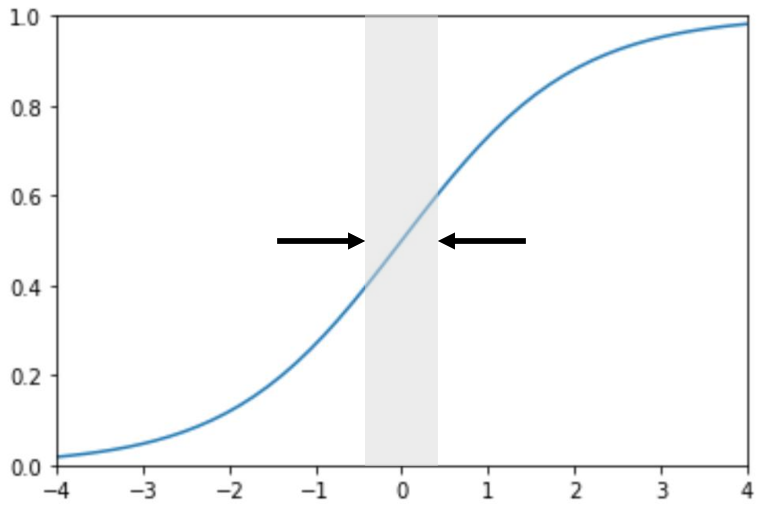
The method to handle overfitting problem

- L1 (Lasso)
- L2 (Ridge)

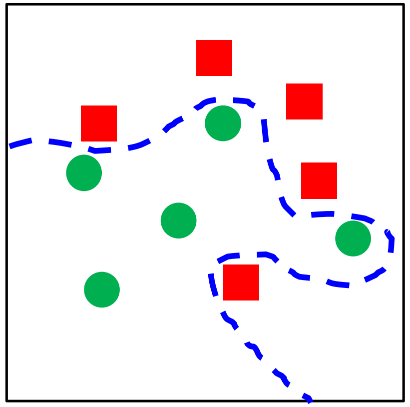
Both regularization technique have parameter λ

- Penalize coefficient adjustment

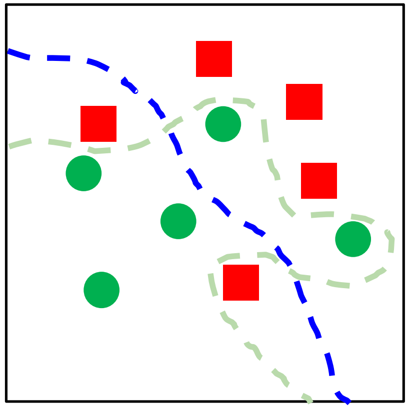
$$\frac{1}{1+e^{-z}}$$



$$z = \beta_0 + \sum_i \beta_i x_i$$



$\lambda \uparrow \quad \boldsymbol{\beta} \downarrow \quad z \downarrow$



Decision boundary is stretched out

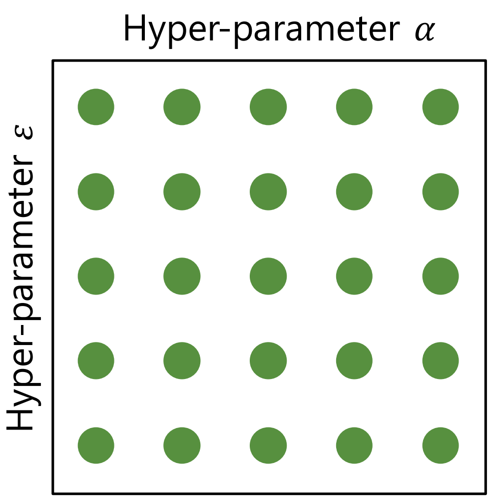
Regularization

How does it work

Model Tuning Strategy

Grid Search

- Limitation
  - Important of some hyperparameters might not be fully address
  - Ex. Grid allows trial on 5 values of α and ε from 25 experiments

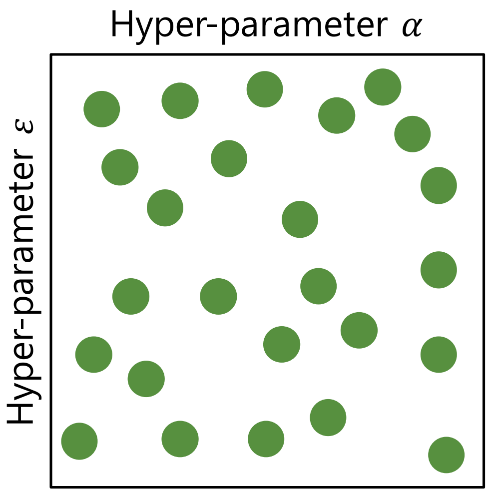


Model Tuning Strategy

Random search

- Random grid allows trial on 25 values of α from 25 experiments

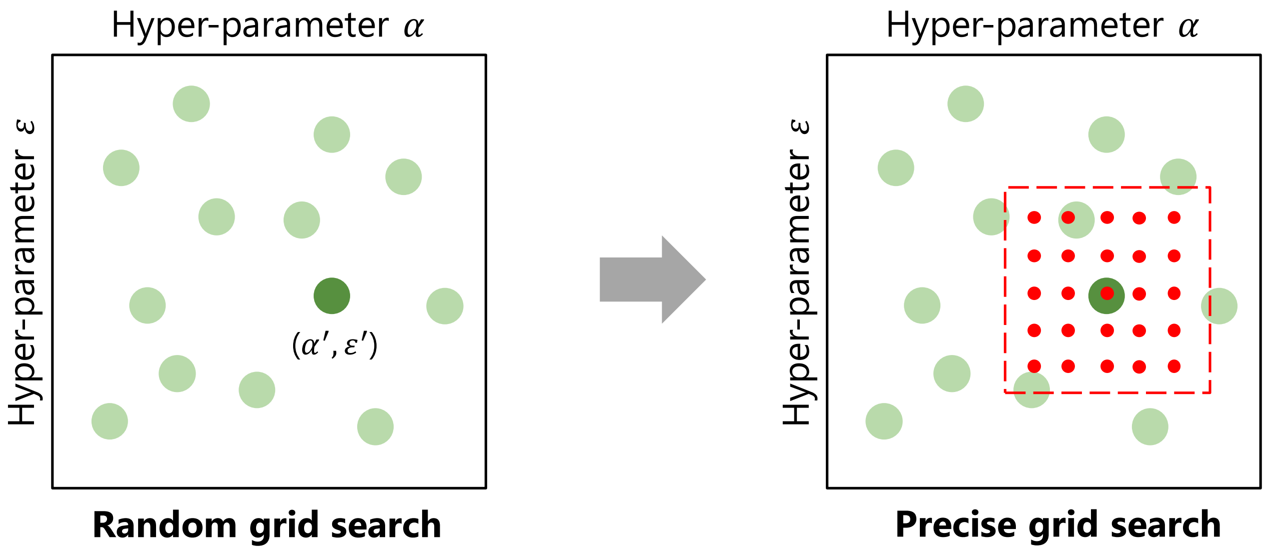
- Random grid gives us more richly to explore sets of possible hyperparameters



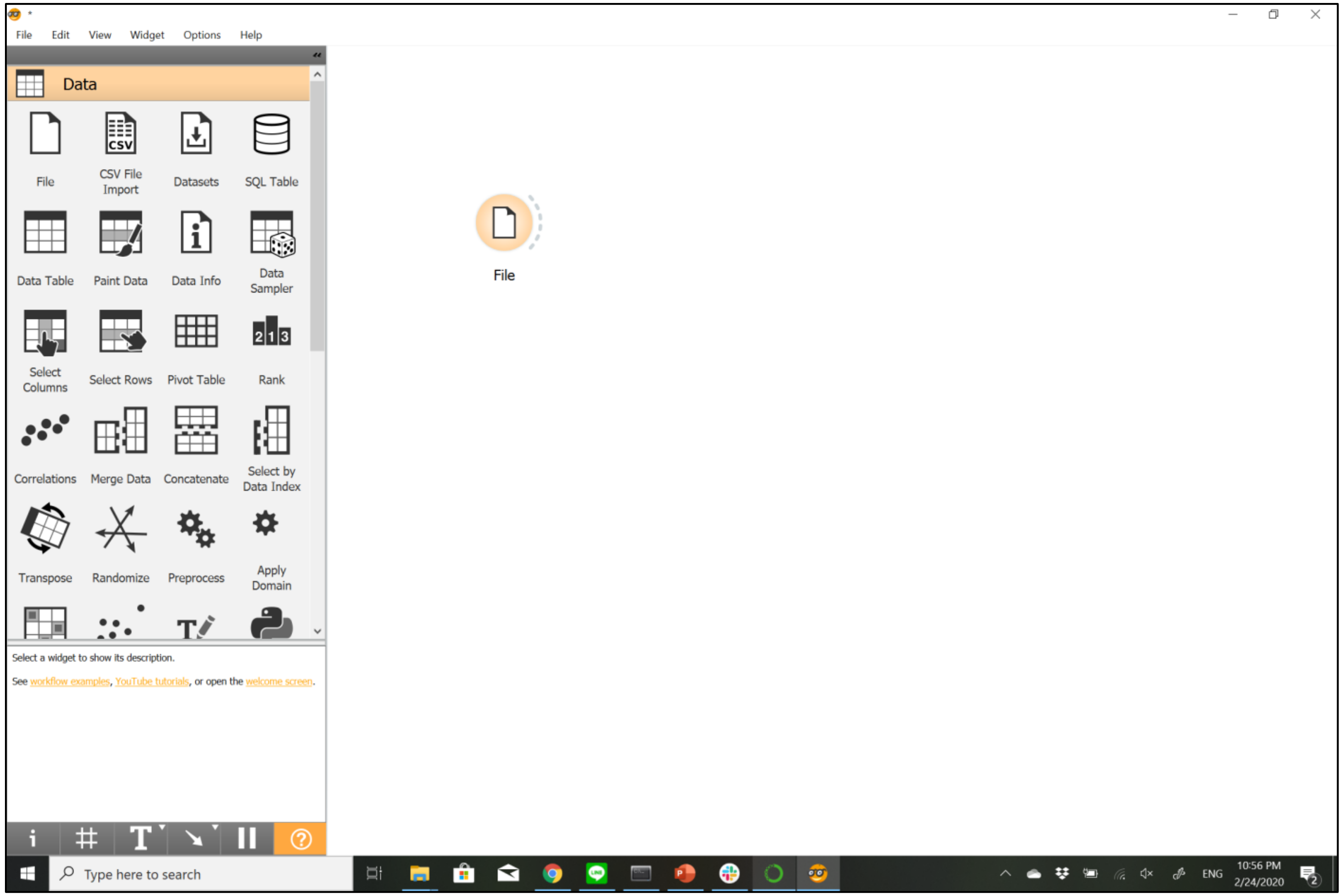
Model Tuning Strategy

Common practice

- Start from randomly coarse
- Randomly fine later



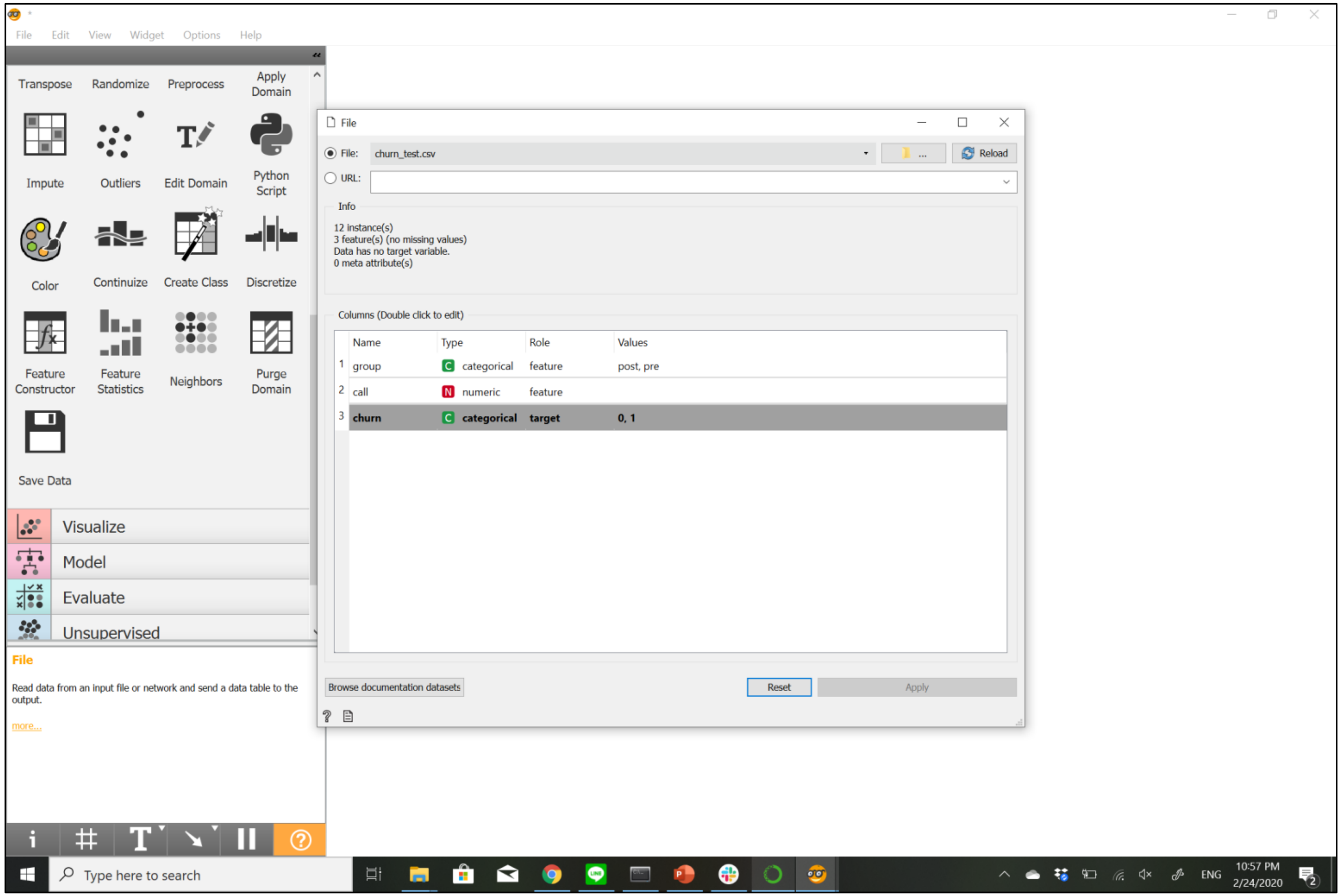
- Import data



Logistic Regression in Orange

Build your model in 10 seconds

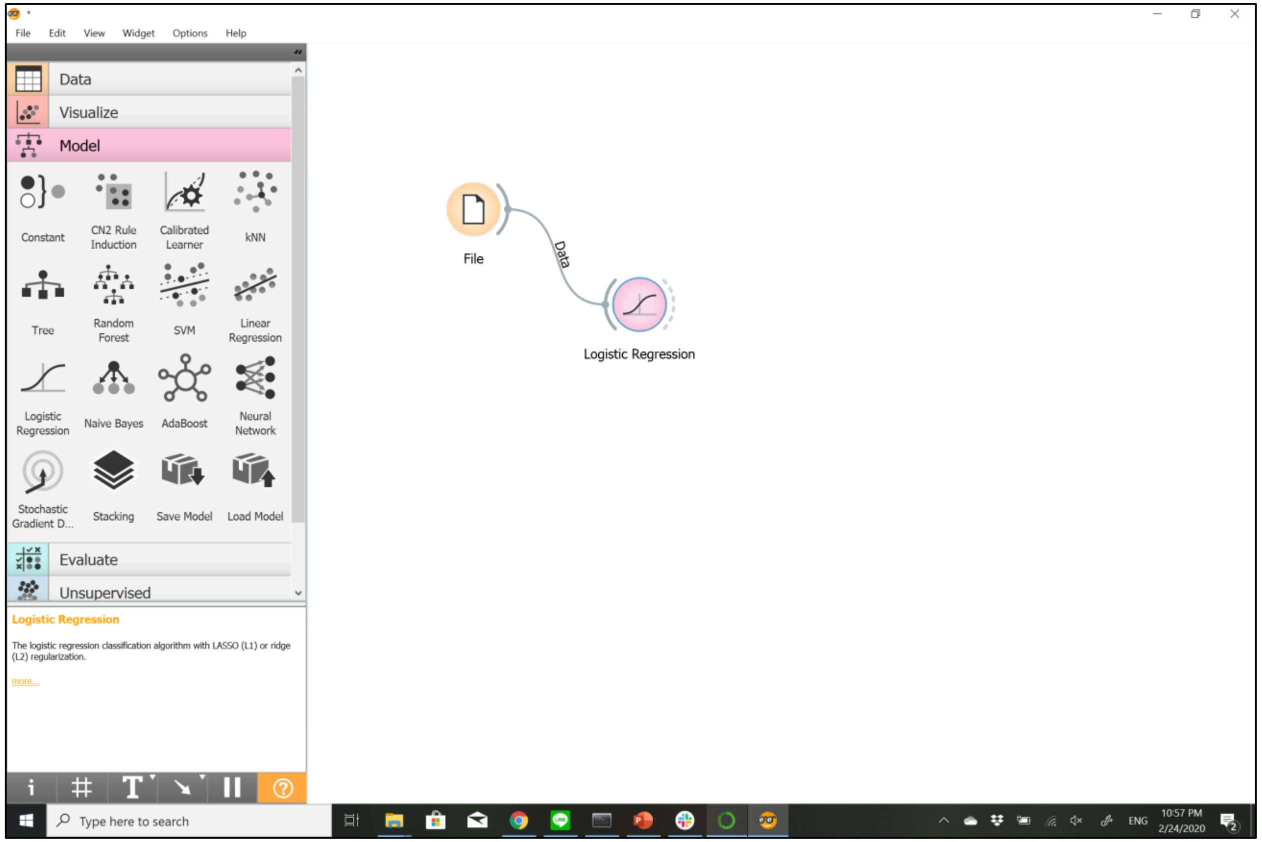
- Identify feature and target



Logistic Regression in Orange

Build your model in 10 seconds

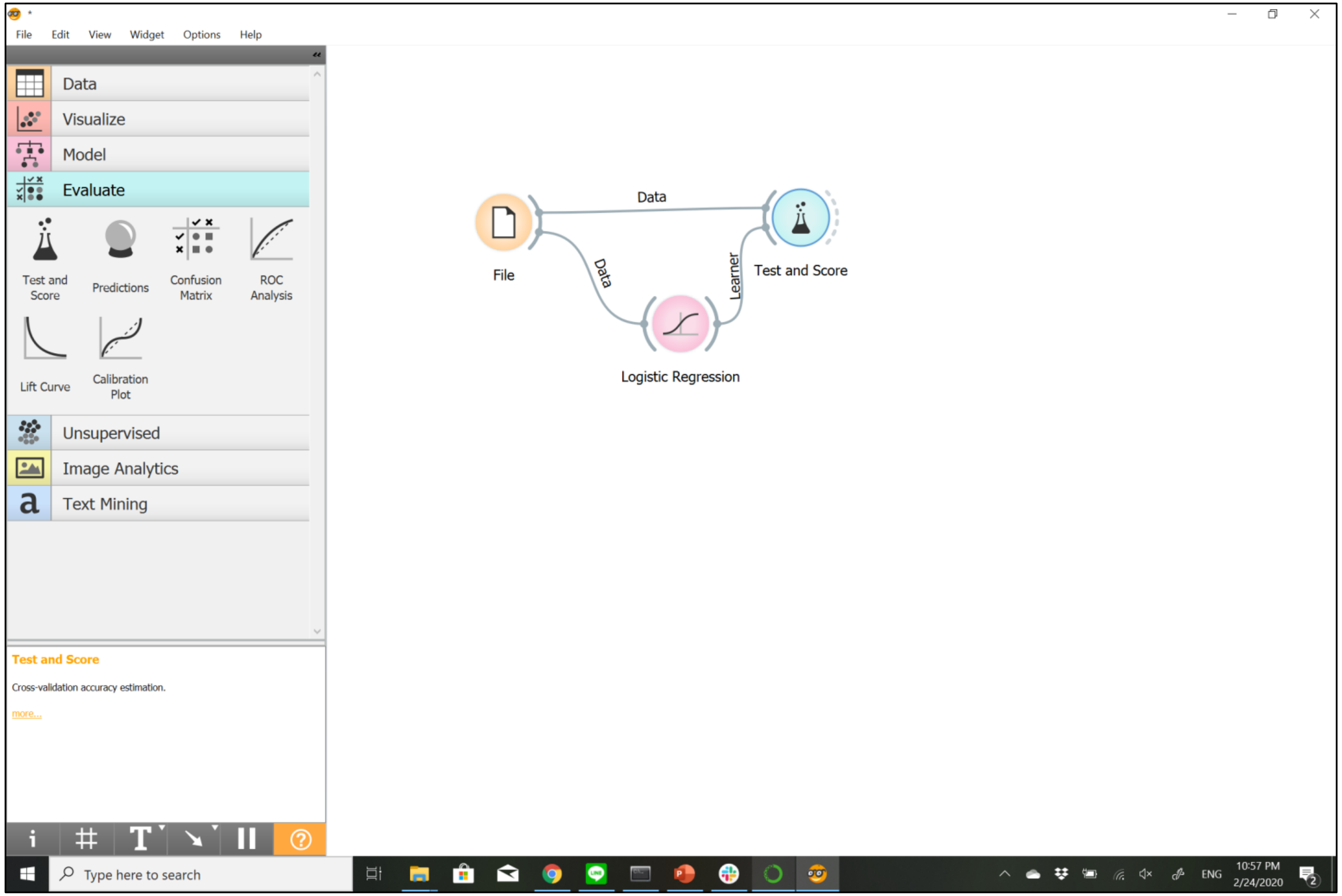
- Add model



Logistic Regression in Orange

Build your model in 10 seconds

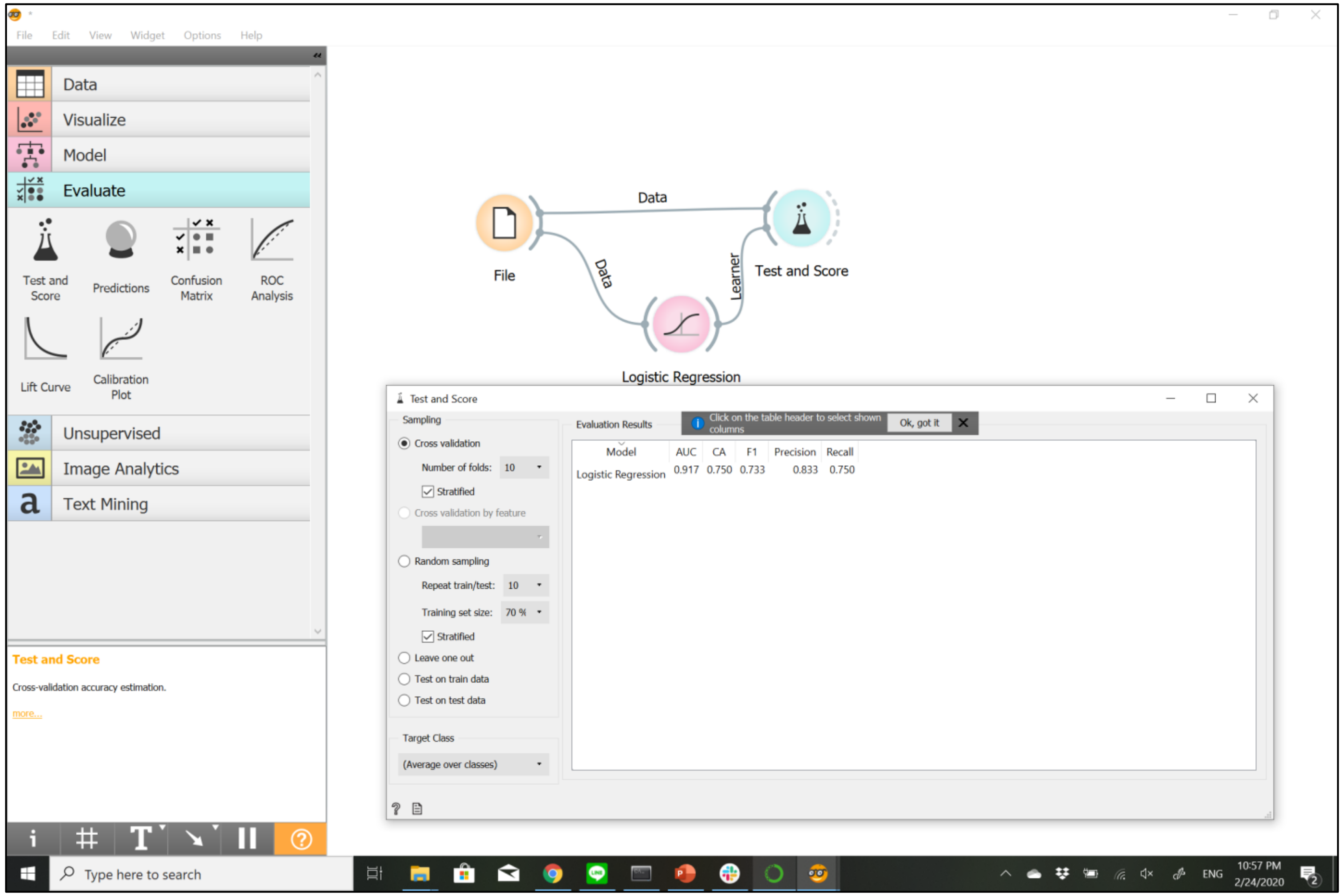
- Evaluate Model



Logistic Regression in Orange

Build your model in 10 seconds

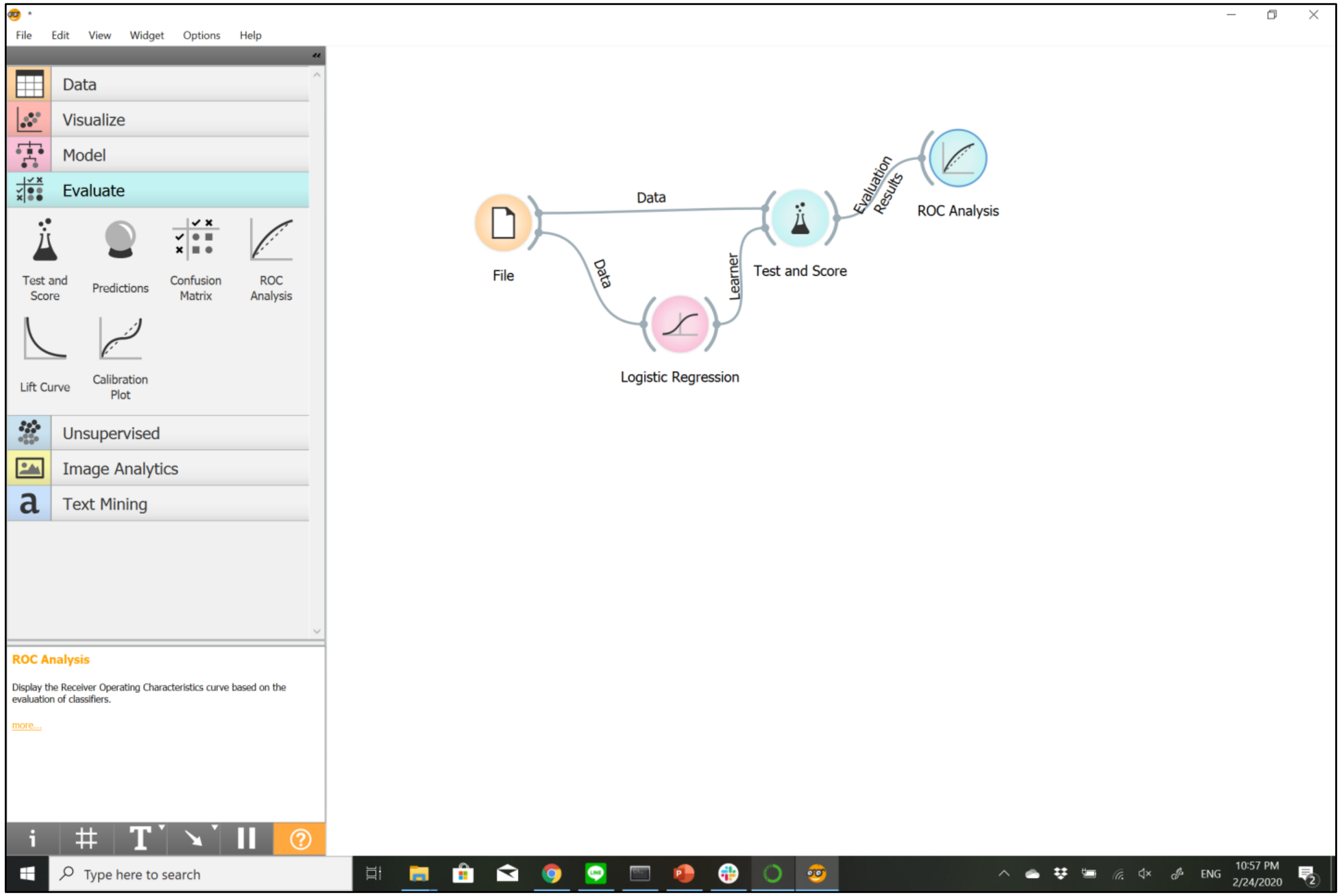
- Evaluate Model



Logistic Regression in Orange

Build your model in 10 seconds

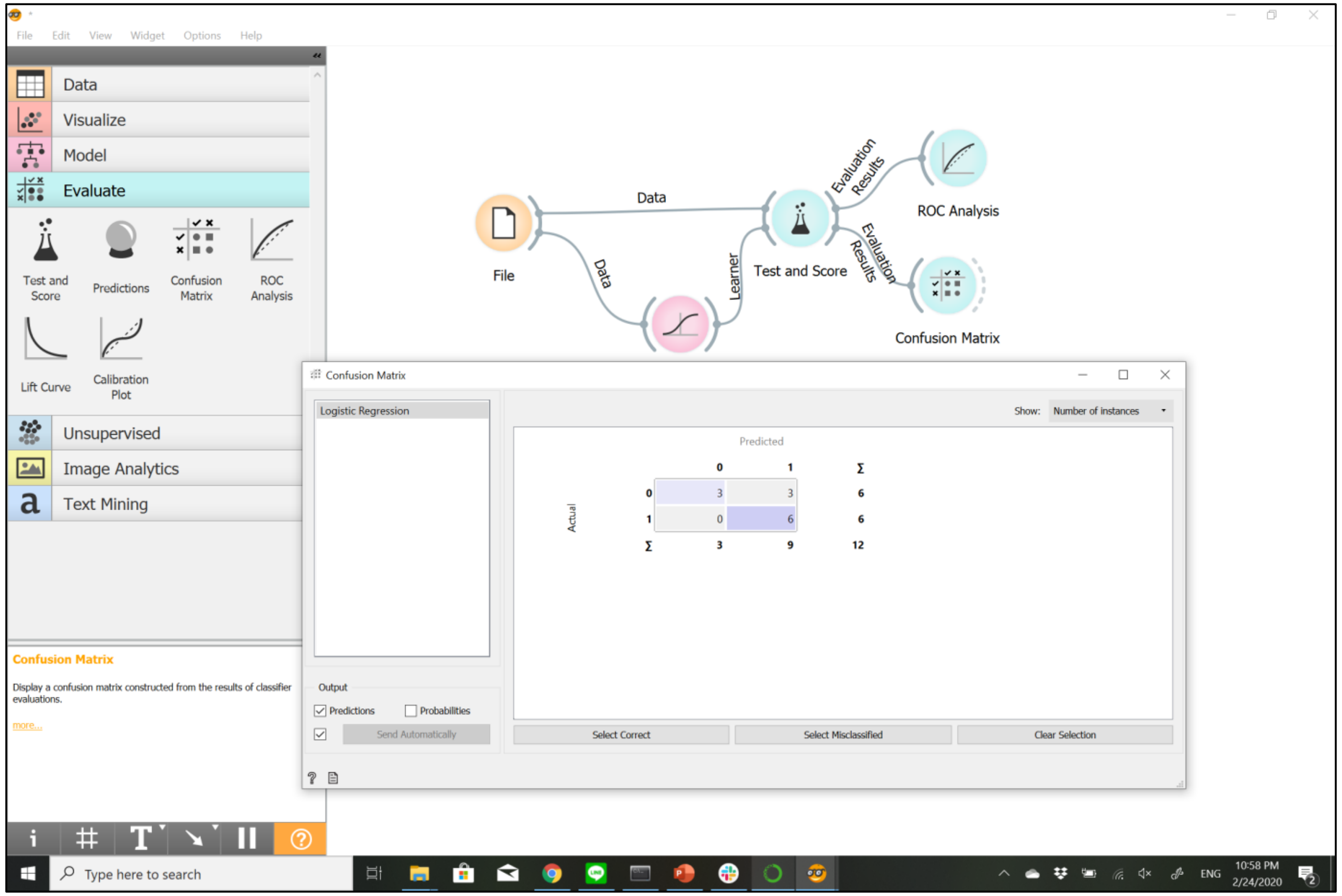
- ROC Plot



Logistic Regression in Orange

Build your model in 10 seconds

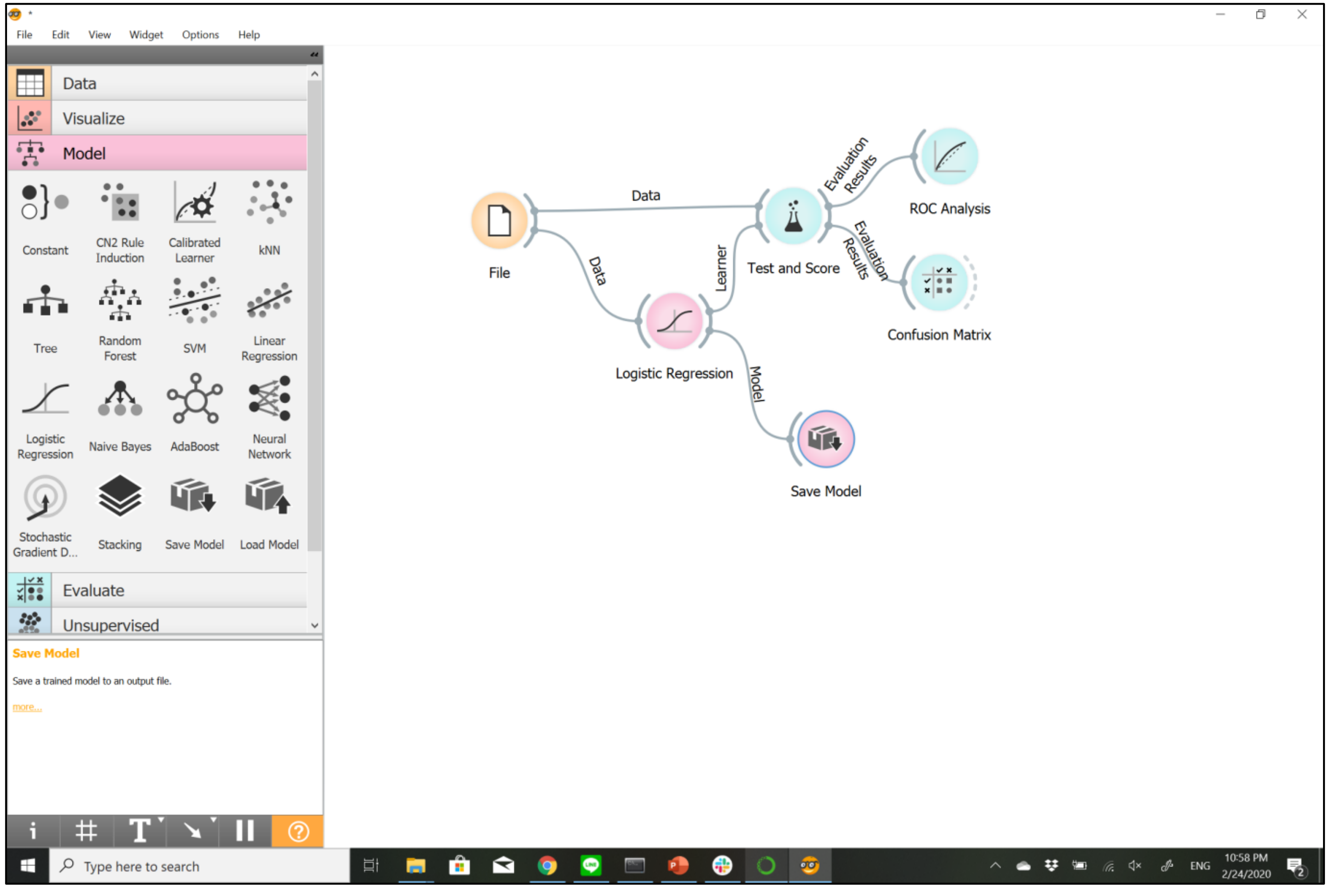
- Confusion Matrix



Logistic Regression in Orange

Build your model in 10 seconds

- Export model



Logistic Regression in Orange

Build your model in 10 seconds

- Build your model using Kaggle dataset

Exercise

Our sample data is very small for demonstration purpose

Now it is the time to work with the dataset churn prediction of telecom in Kaggle Competition