Implementing Logistic Regression Models in R

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN

www.loonycorn.com

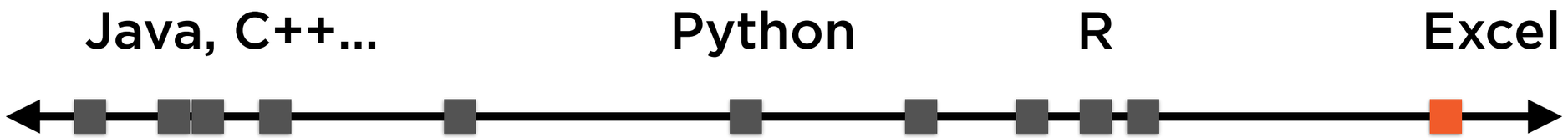
Overview

Set up a logistic regression to predict whether a stock will rise or fall

Solve this logistic regression in R

Extend the logistic regression to include multiple explanatory variables

Ease of Prototyping



Excel is the fastest prototyping tool out there

Robustness and Reuse

No free lunches

“Make the common use-case easy and the difficult use-case possible.”

Regression: Excel, R or Python?

Excel

Create a regression slide for an important presentation

R

Create a regression case study for a seminar

Python

Build trading model that scrapes websites, combines sentiment analysis and regression

Regression: Excel, R or Python?

Excel

Presentations

R

Seminars

Python

Trading models

R for Regression

R

Presentations

R

Seminars

R

Trading models

Use **R for regression**: It makes sense whatever your use-case

Demo

Implement Logistic Regression in R

Logistic Regression in R

Cause

Changes in S&P 500

Effect

Changes in price of Google Stock

Logistic Regression in R

y = Returns on Google stock (GOOG)

x = Returns on S&P 500 (S&P500)

Logistic Regression in R

DATE	GOOG	S&P500
2017-02-01	813.67	2316.10
2017-01-01	796.79	2278.87
2005-01-01	97.71	1181.27

Download prices and we refer to ‘Adjusted Close’

Data Frame: Data in Rows and Columns

Each row represents 1 observation

DATE	OPEN	...	ADJUSTED CLOSE
2016-12-01	772	...	779
2016-11-01	758	...	747
2006-01-01	302	...	309

Each column represents 1 variable (a list or vector)

From File to Data Frame

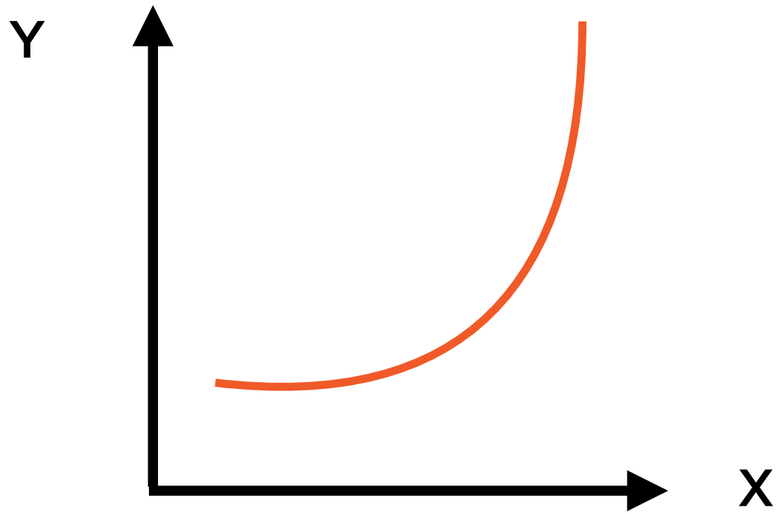
DATE	OPEN	…	ADJUSTED CLOSE
2016-12-01	772	…	779
2016-11-01	758	…	747
2006-01-01	302	…	309

File

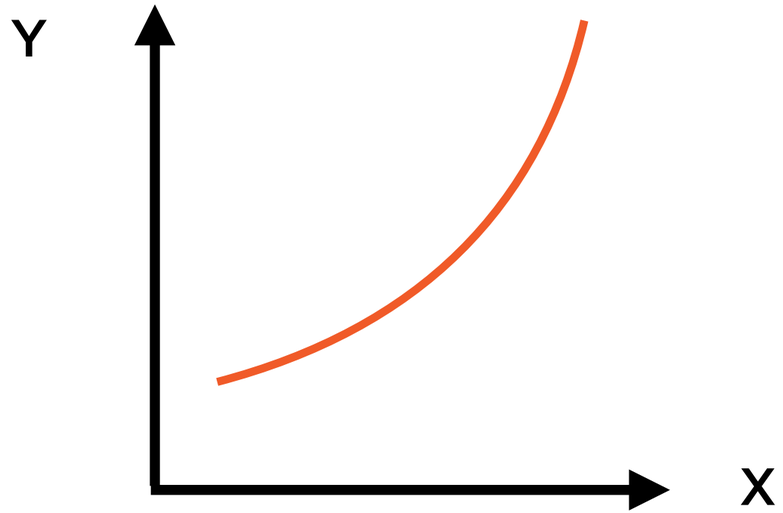
DATE	OPEN	…	ADJUSTED CLOSE
2016-12-01	772	…	779
2016-11-01	758	…	747
2006-01-01	302	…	309

Data Frame

Never Regress Non-Stationary Data



Exponential



Polynomial

Smoothly trending data will lead to poor quality regression models

First Differences

$y'_{12} = \log y_2 - \log y_1$

$x'_{12} = \log x_2 - \log x_1$

Regress y' and x'

Log Differences

$y'_{12} = (y_2 - y_{1)}/y_1$

$x'_{12} = (x_2 - x_{1)}/x_1$

Regress y' and x'

Returns

Take first differences of smooth data converting either to log differences or returns

Negative Indices => Exclude Data

goog

DATE	GOOG. PRICE	NASDAQ. PRICE
2016-12-01	779	5550
2016-11-01	747	5324
2006-01-01	309	1900

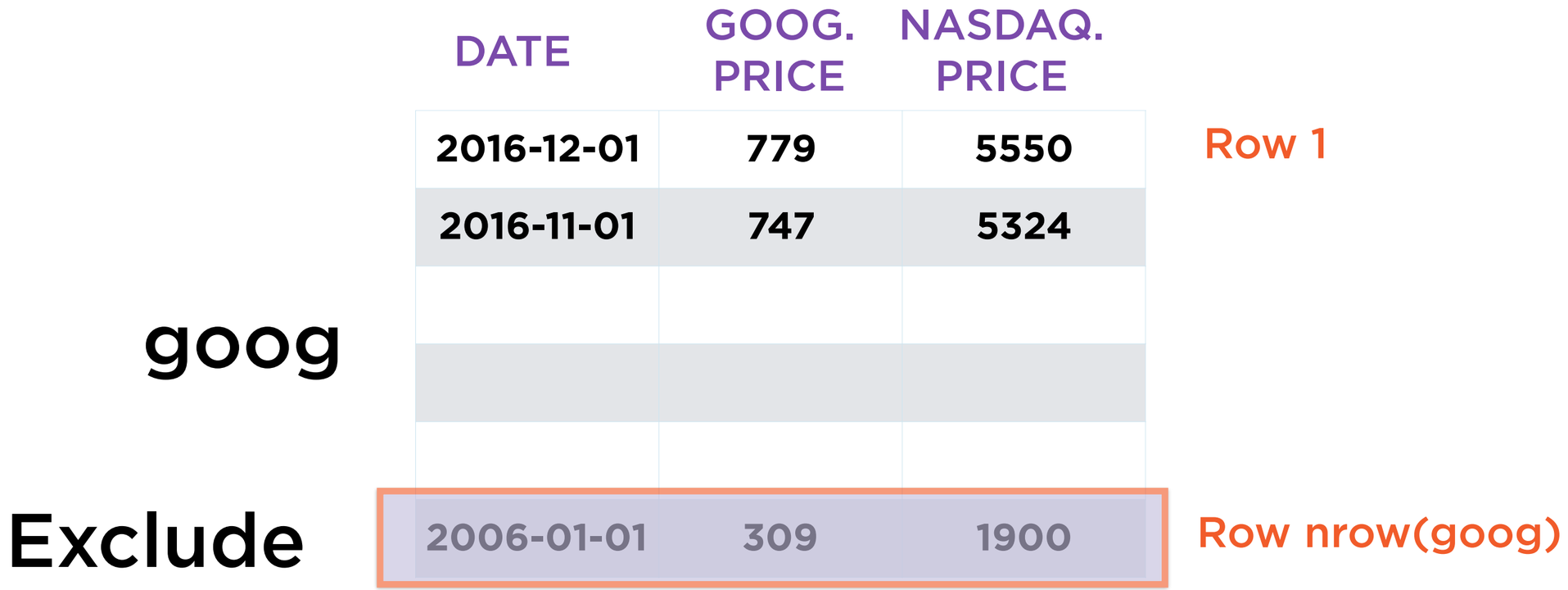
Row 1

Row nrow(goog)

Column 1

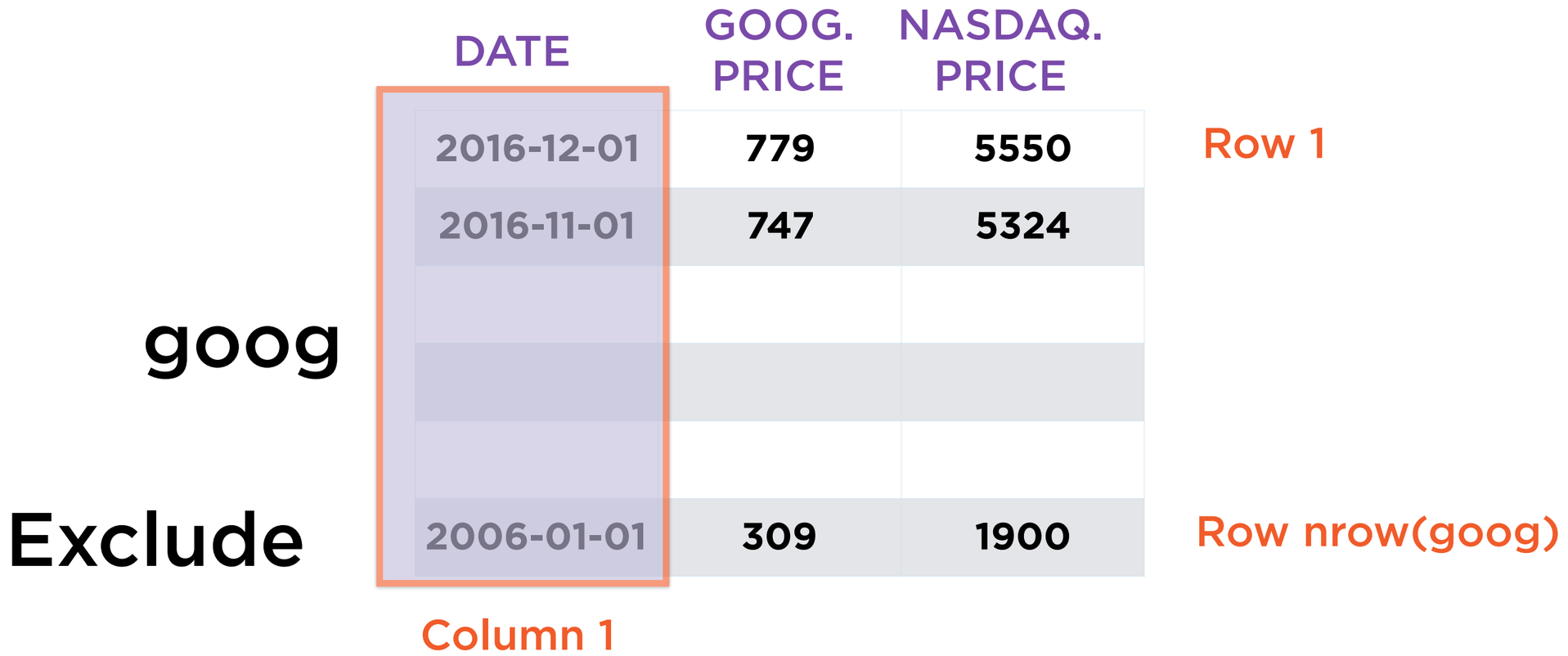
```
goog[-nrow(goog),-1]
```

Negative Indices => Exclude Data



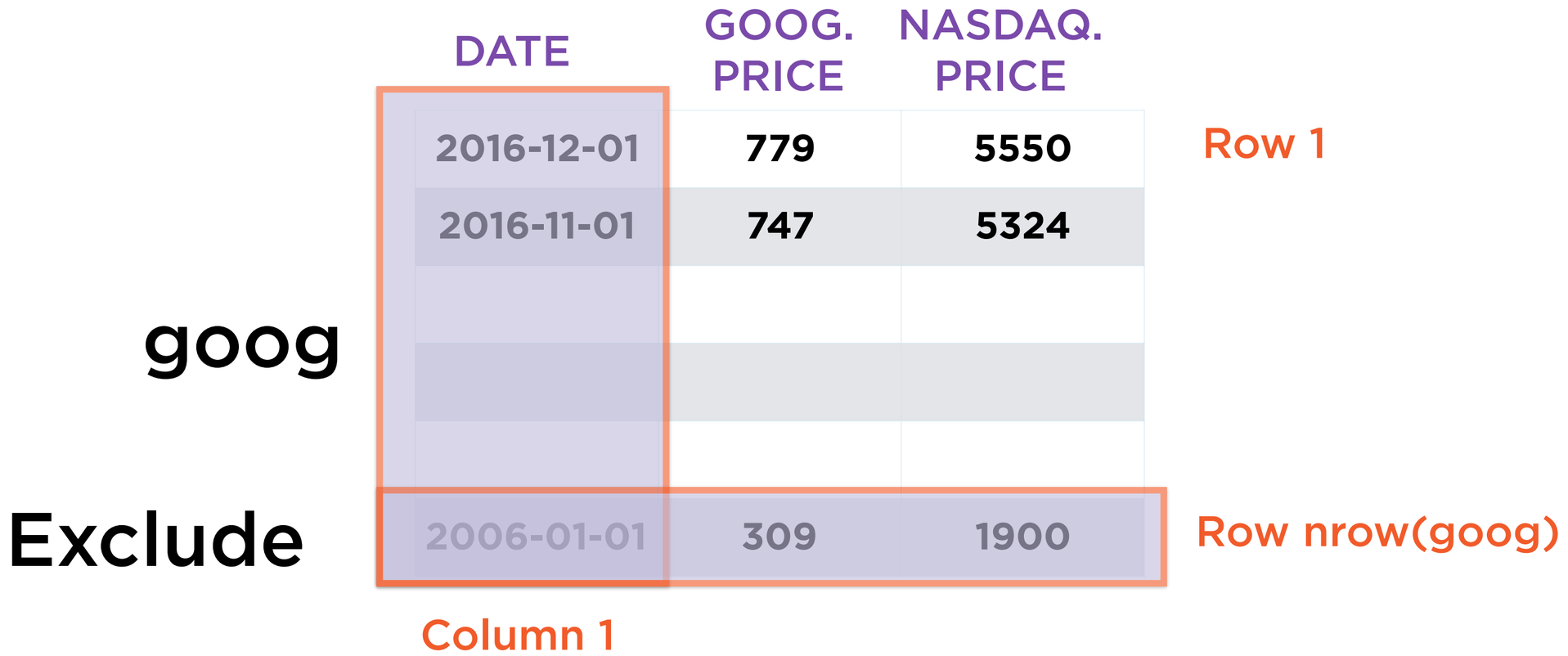

```
goog[-nrow(goog),-1]
```

Negative Indices => Exclude Data




```
goog[-nrow(goog),-1]
```

Negative Indices => Exclude Data




```
goog[-nrow(goog),-1]
```

Element-wise Operations

779	5550

/

747	5324

=

779/747	5550/5324
…	…
…	…

```
goog[-nrow(goog),-1]/
goog[-1,-1]
```

Prices to Returns

779/747	5550/5324
…	…
…	…

-

1	1
1	1
1	1
1	1
1	1

=

779/747 - 1	5550/5324 -1
…	…
…	…

```
goog[-nrow(goog),-1]/
goog[-1,-1] - 1
```

This converts prices to returns

Using Logistic Regression in R

$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

P(y) = Probability of Google going up in the current month i

x = Returns on S&P 500 for current month

Multiple X Variables - Easy

$$p(y_i) = \frac{1}{1 + e^{-(A + B^{GOOG} x^{GOOG}_{i-1} + B^{SP500} x^{SP500}_{i})}}$$

P(y) = Probability of Google going up in the current month i

x^{GOOG} = Returns on GOOG for previous month

x^{SP500} = Returns on S&P 500 for current month

A Much Harder Problem

$$p(y_i) = \frac{1}{1 + e^{-(A + B^{GOOG} x^{GOOG}_{i-1} + B^{SP500} x^{SP500}_{i-1})}}$$

P(y) = Probability of Google going up in the current month i

x^{GOOG}_{i-1} = Returns on GOOG for previous month

x^{SP500}_{i-1} = Returns on S&P 500 for previous month

Very difficult problem to solve - quant hedge funds are very interested in the answer

A Much Harder Problem

$$p(y_i) = \frac{1}{1 + e^{-(A + B^{GOOG} x^{GOOG}_{i-1} + B^{SP500} x^{SP500}_{i-1})}}$$

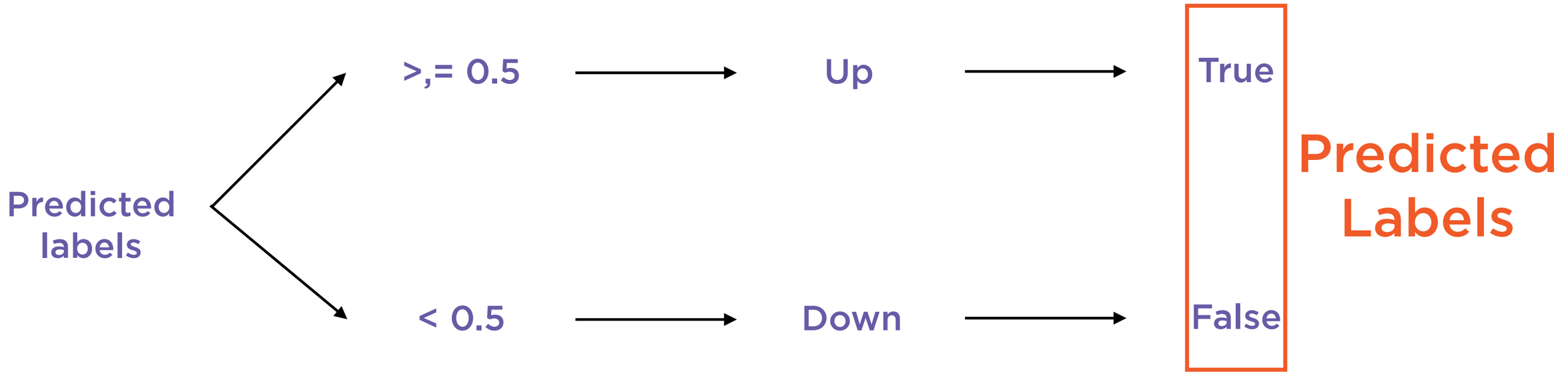
P(y) = Probability of Google going up in the current month i

x^{GOOG}_{i-1} = Returns on GOOG for previous month

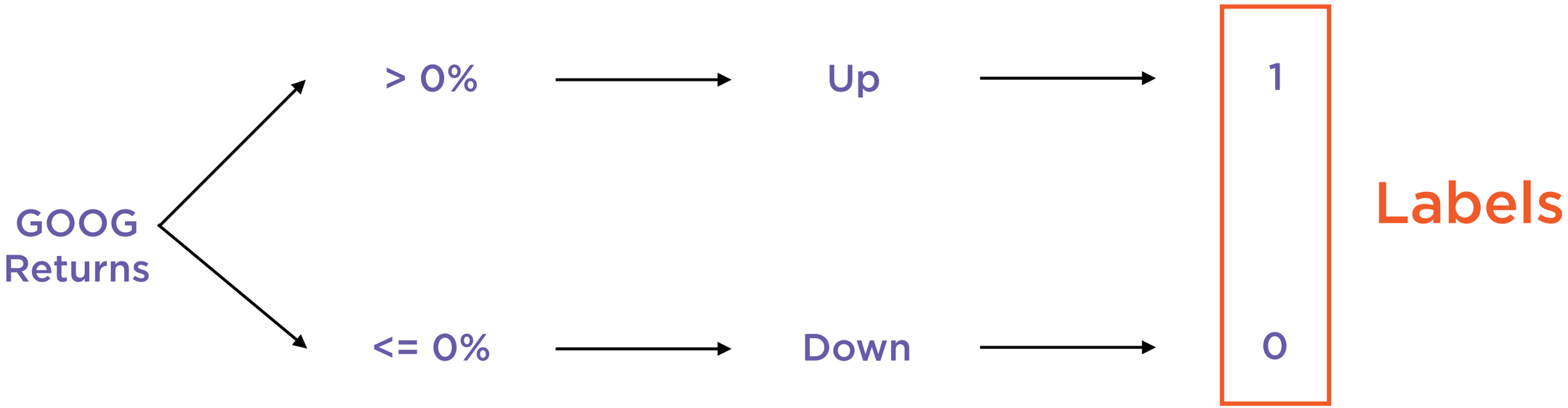
x^{SP500}_{i-1} = Returns on S&P 500 for previous month

Very difficult problem to solve - quant hedge funds are very interested in the answer

Using Logistic Regression in R



Set up the Problem



Label GOOG returns as binary (1,0)

Using Logistic Regression in R

DATE	ACTUAL	PREDICTED
2005-01-01	NA	NA
2005-02-01	0	1
2005-03-01	0	0
2017-01-01	1	1
2017-02-01	1	1

Compare GOOG's actual labels vs. predicted labels

Summary

Logistic regression can be very easily implemented in R