STATS 415

DATA MINING PROJECT

Group Member

Hao Xu
Shengqian Jin
Xinye Xu (Presenter)

Contents

- Goals and Issues
- Data Exploration
- Classification
  - Logistic Regression
  - Decision Tree
  - Support Vector Machine
- Conclusion

❑ Goals and Issues

Goal

Propose data mining approaches to predict the success of telemarketing calls for selling bank long-term deposits.

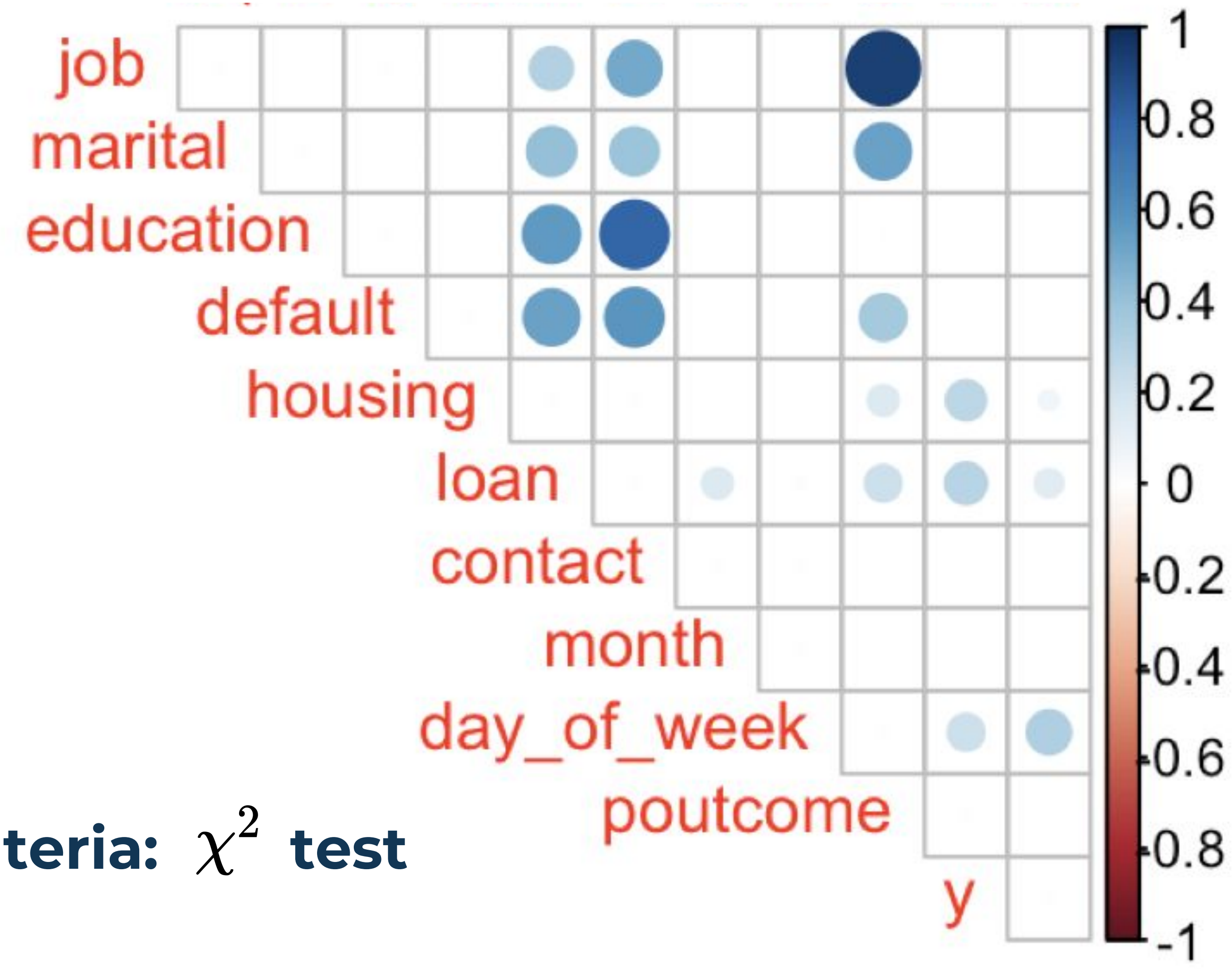
Data Description

- **20** most relevant features out of **150** selected by S. Moro's[1, 2] research from 2008 to 2013.
- **7** variables about client data, **8** about contact information with customers, and **5** about social and economic.
- **41,188** valid observations.

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
[2] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011

❑ Data Exploration

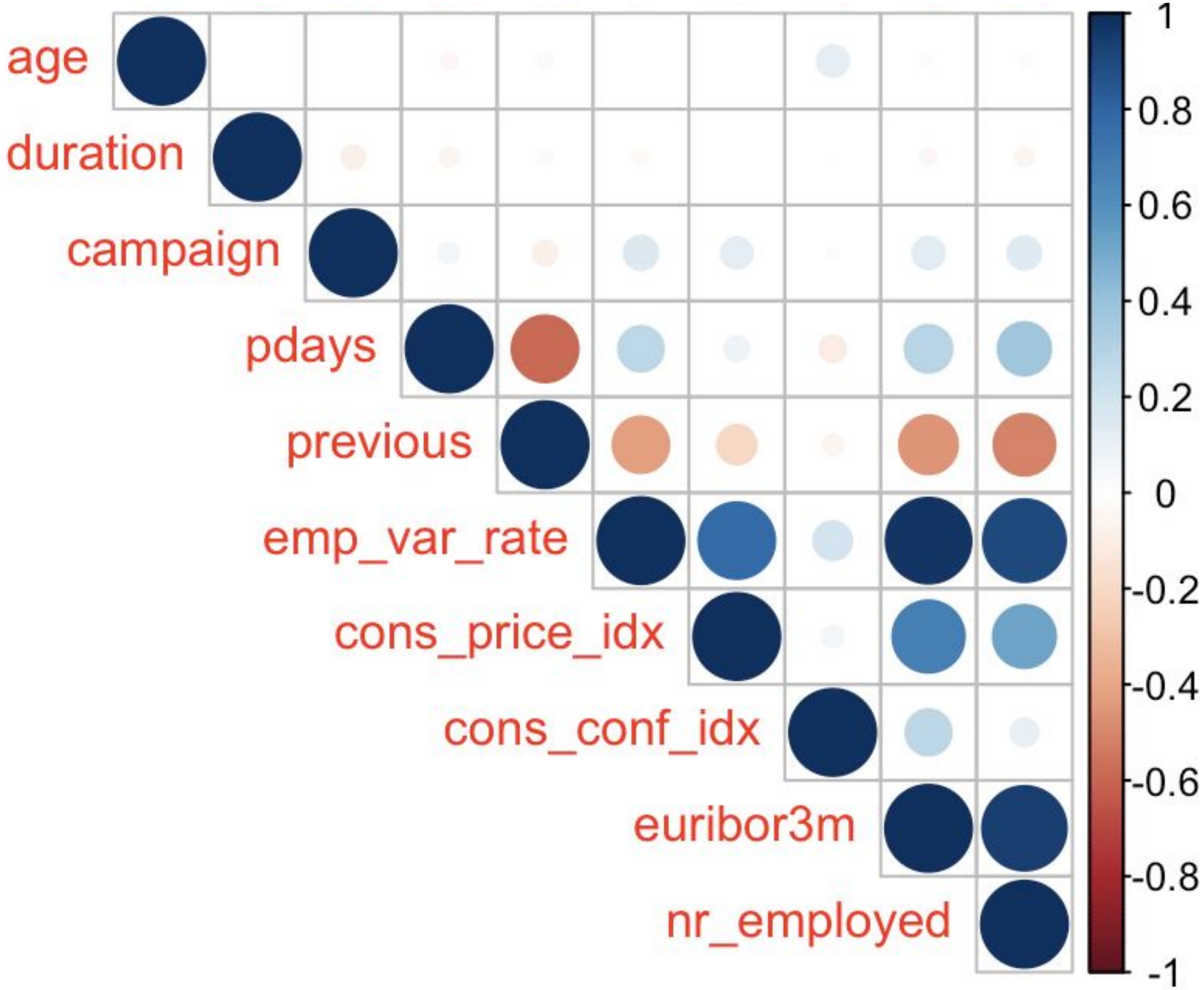
- **Correlation (Categorical)**



Criteria: χ^2 **test**

❑ Data Exploration

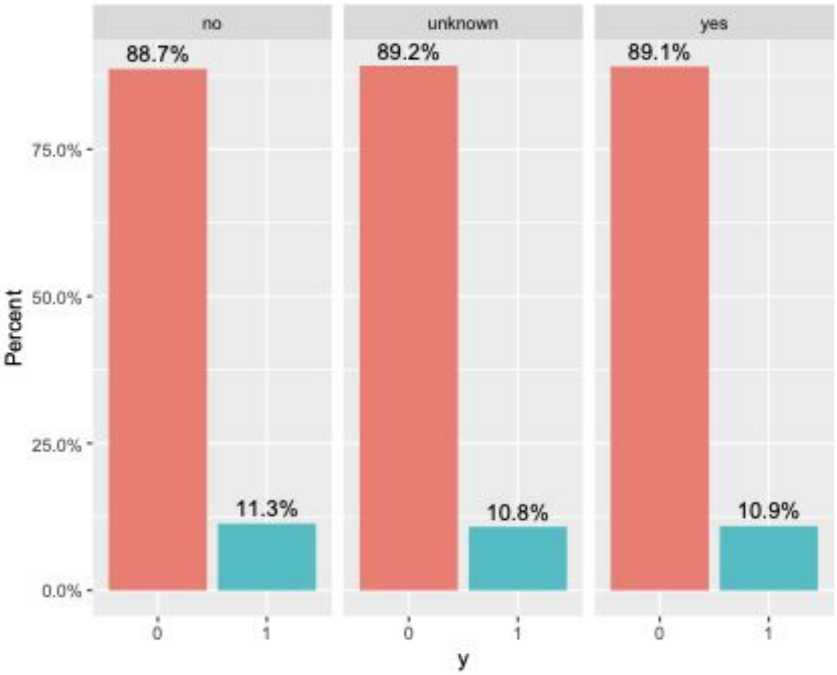
- **Correlation (Continuous)**



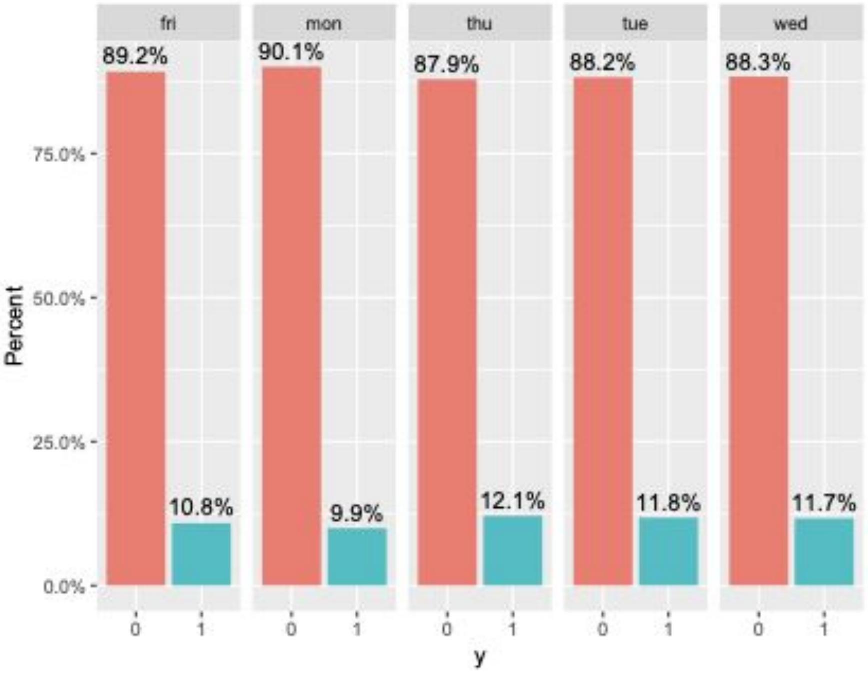
❑ Data Exploration

- **Several Histograms**

Some predictors might be insignificant.



Loan

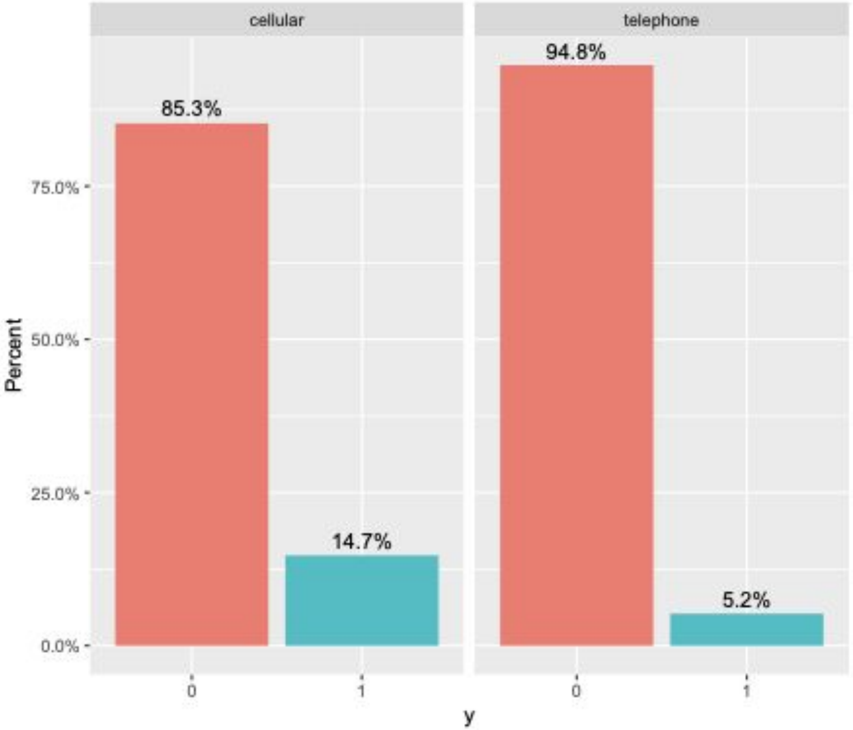


Day_of_week

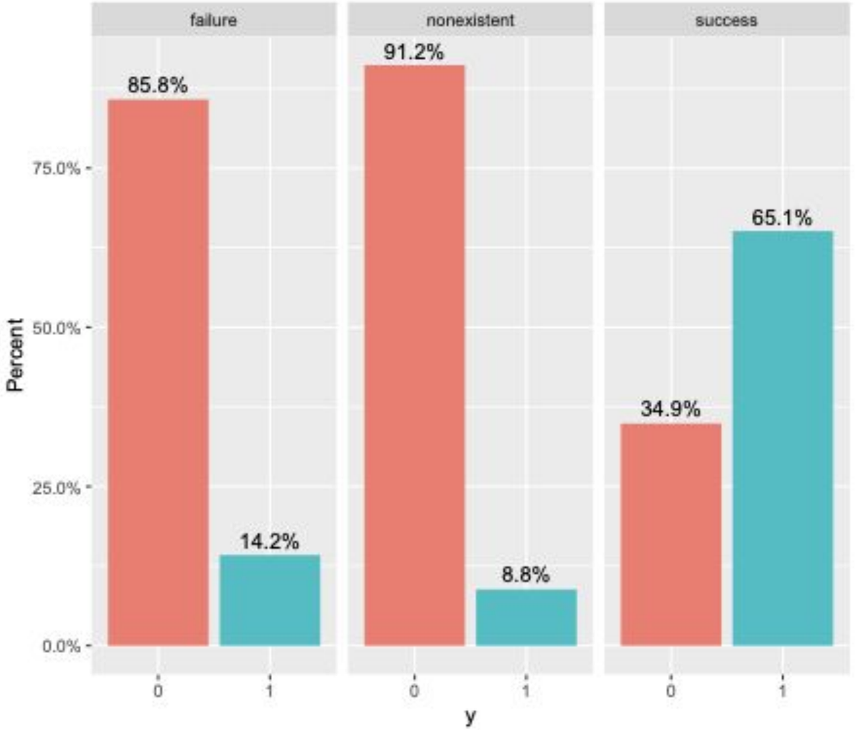
❑ Data Exploration

- **Several Histograms**

Take out your phone and call the cellulars of the previous buyers!



Contact

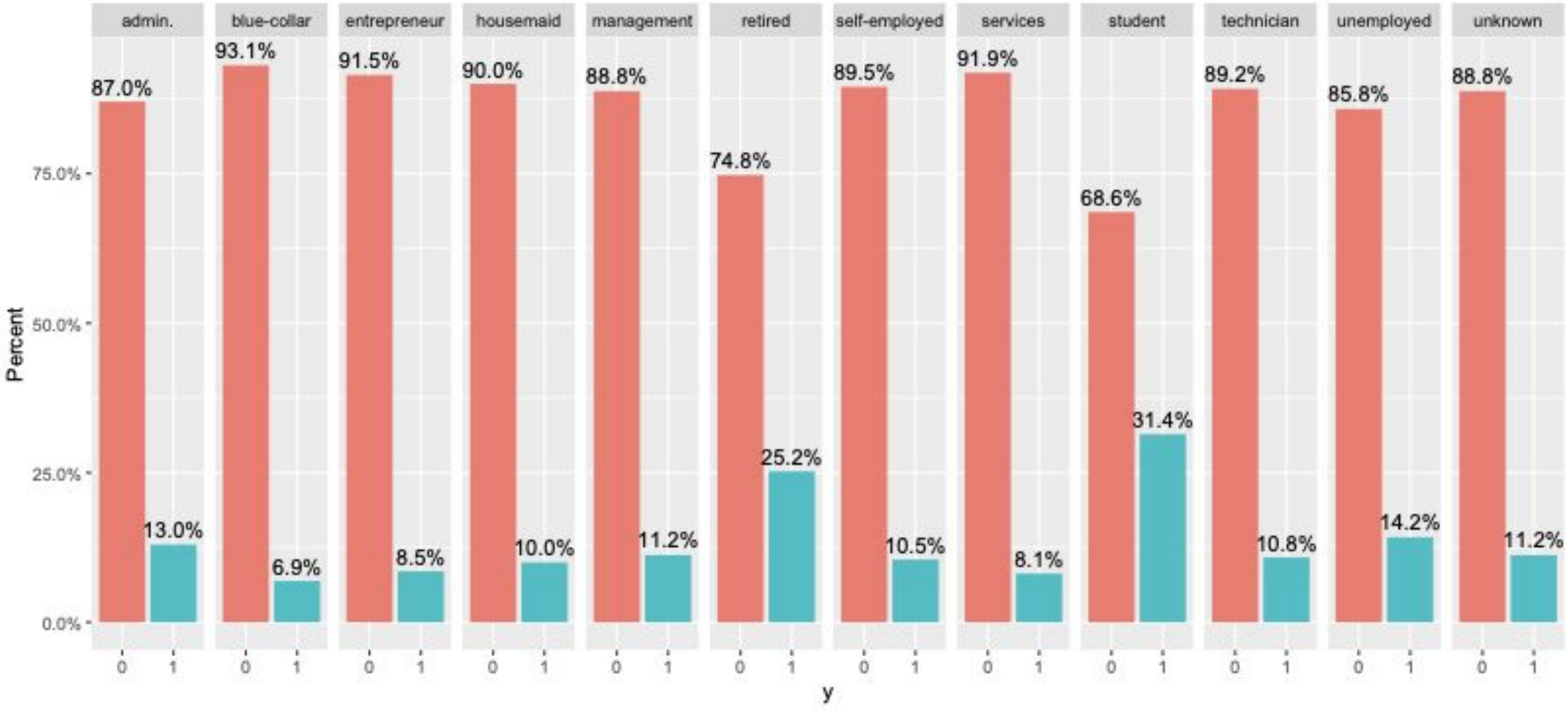


Poutcome

❑ Data Exploration

- **Several Histograms**

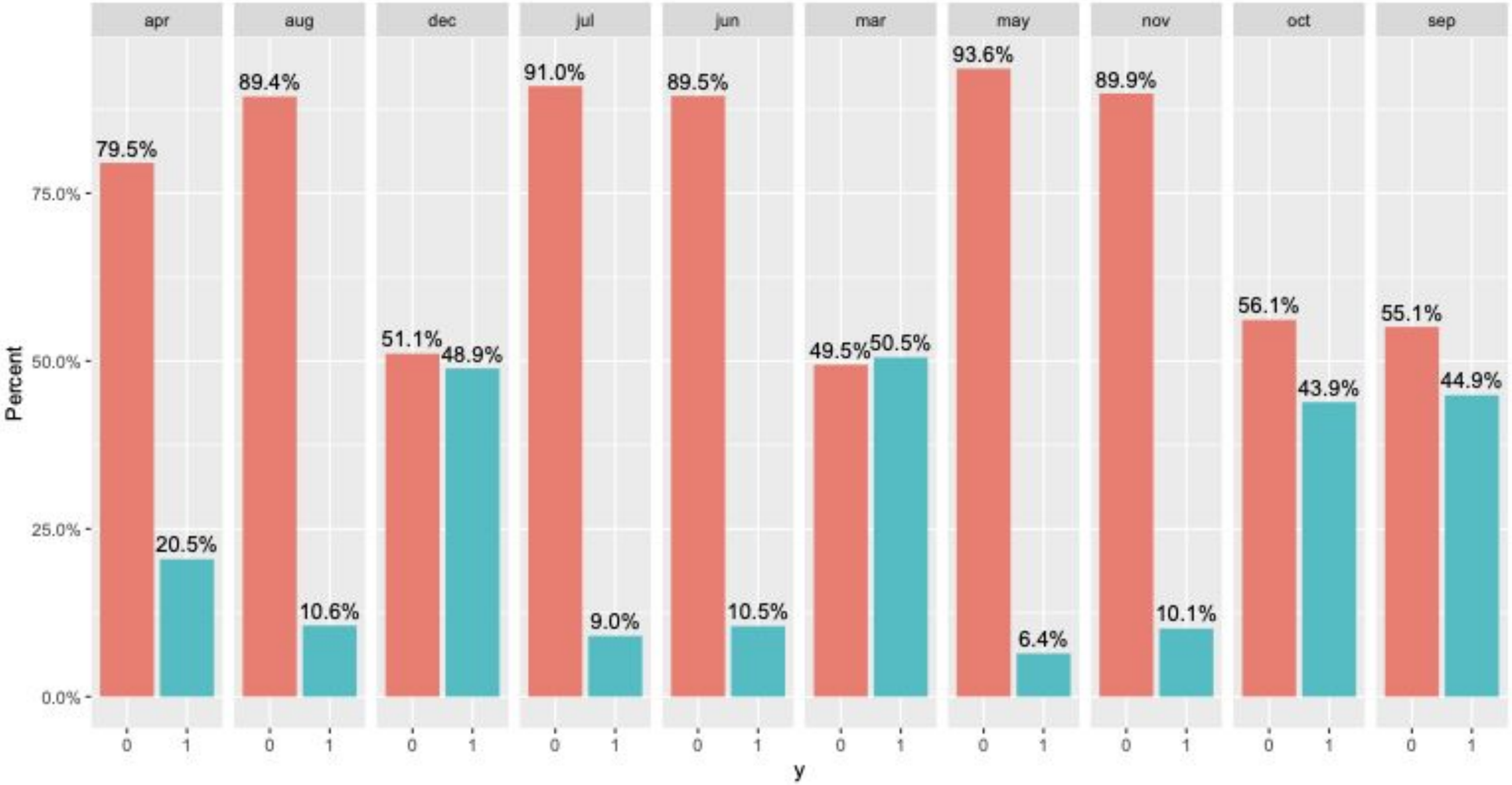
Retired and student are more likely to say ‘yes’.



❑ Data Exploration

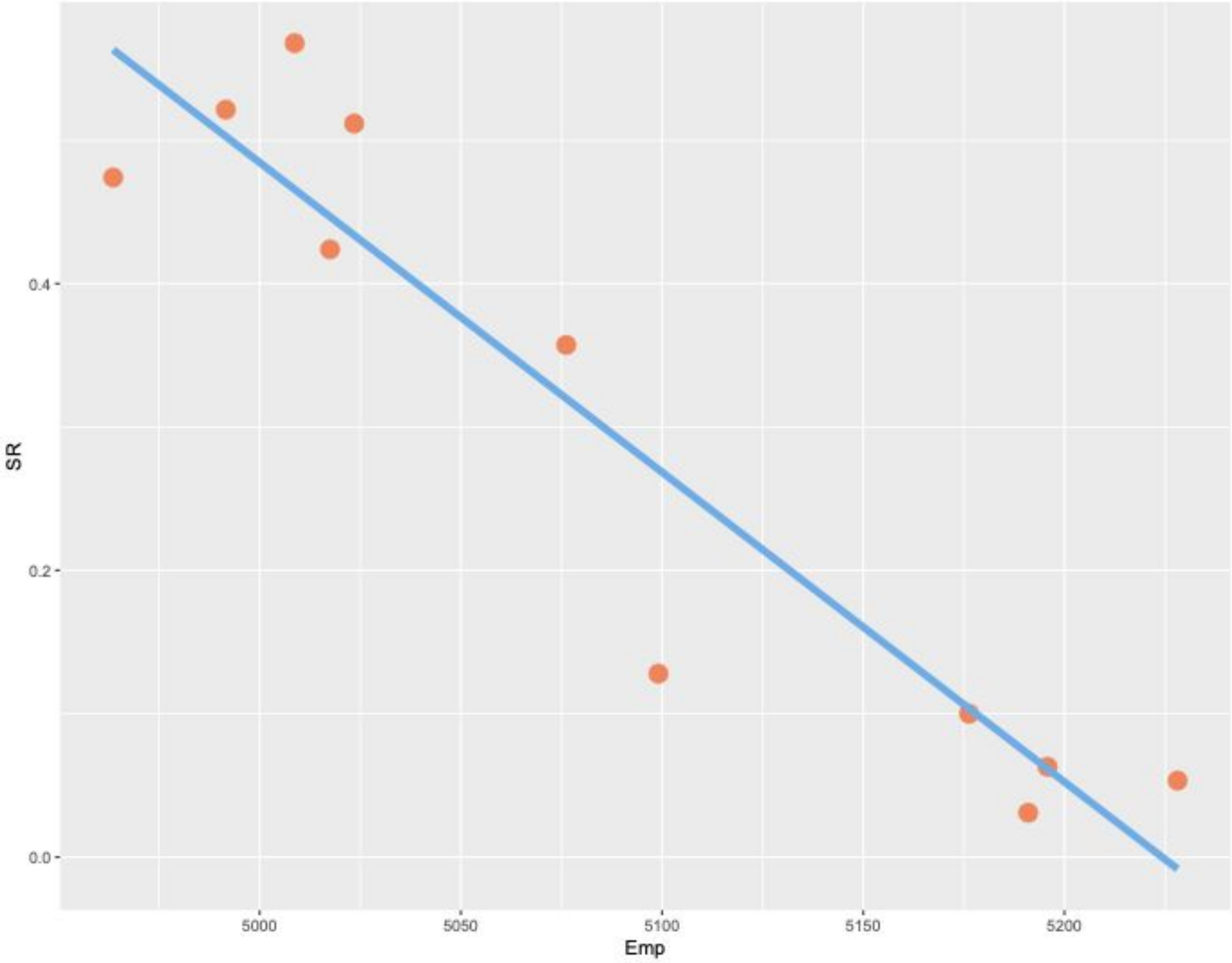
- **Several Histograms**

The most successful months!



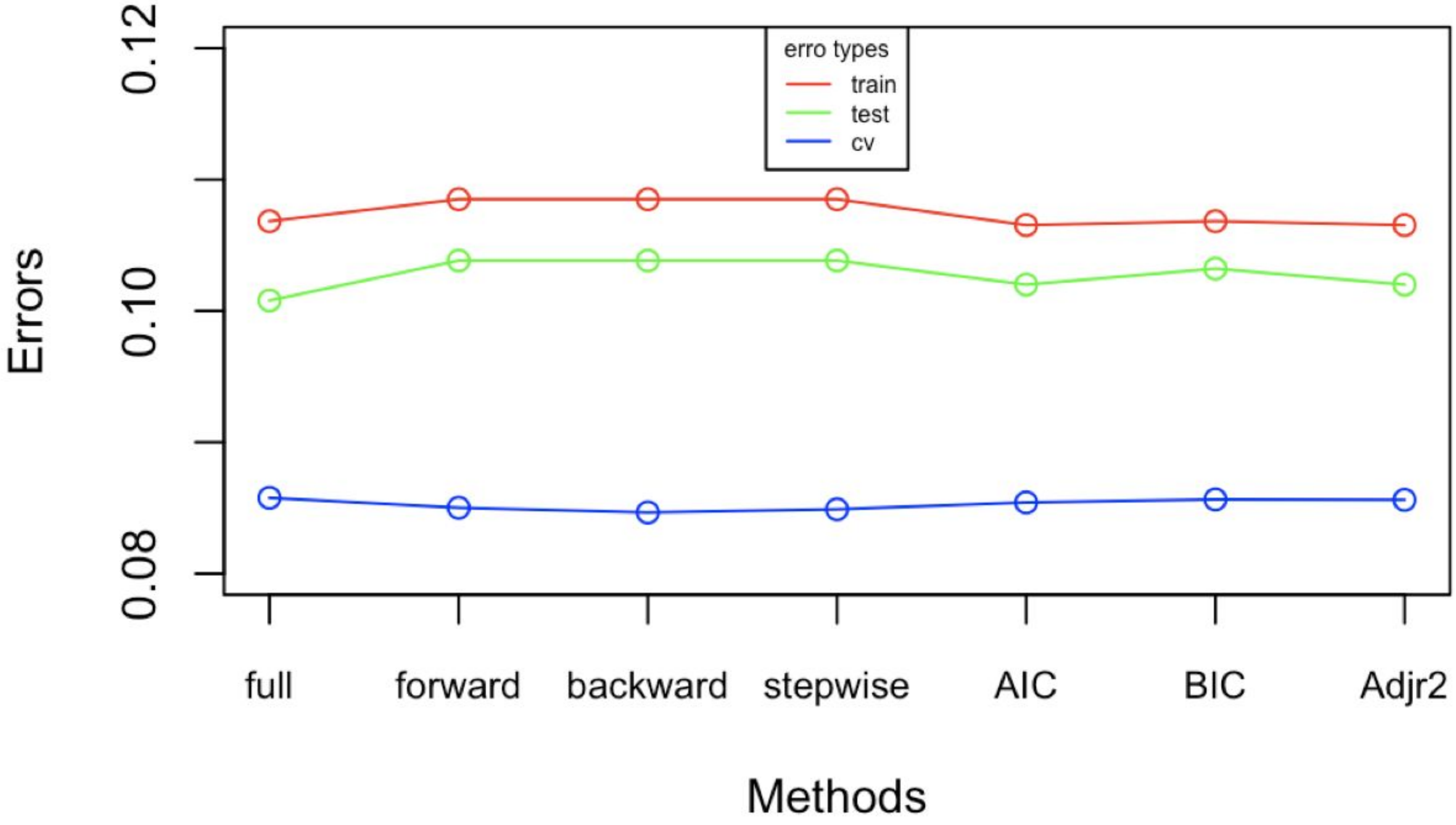
❑ Data Exploration

- **Importance of Employment**



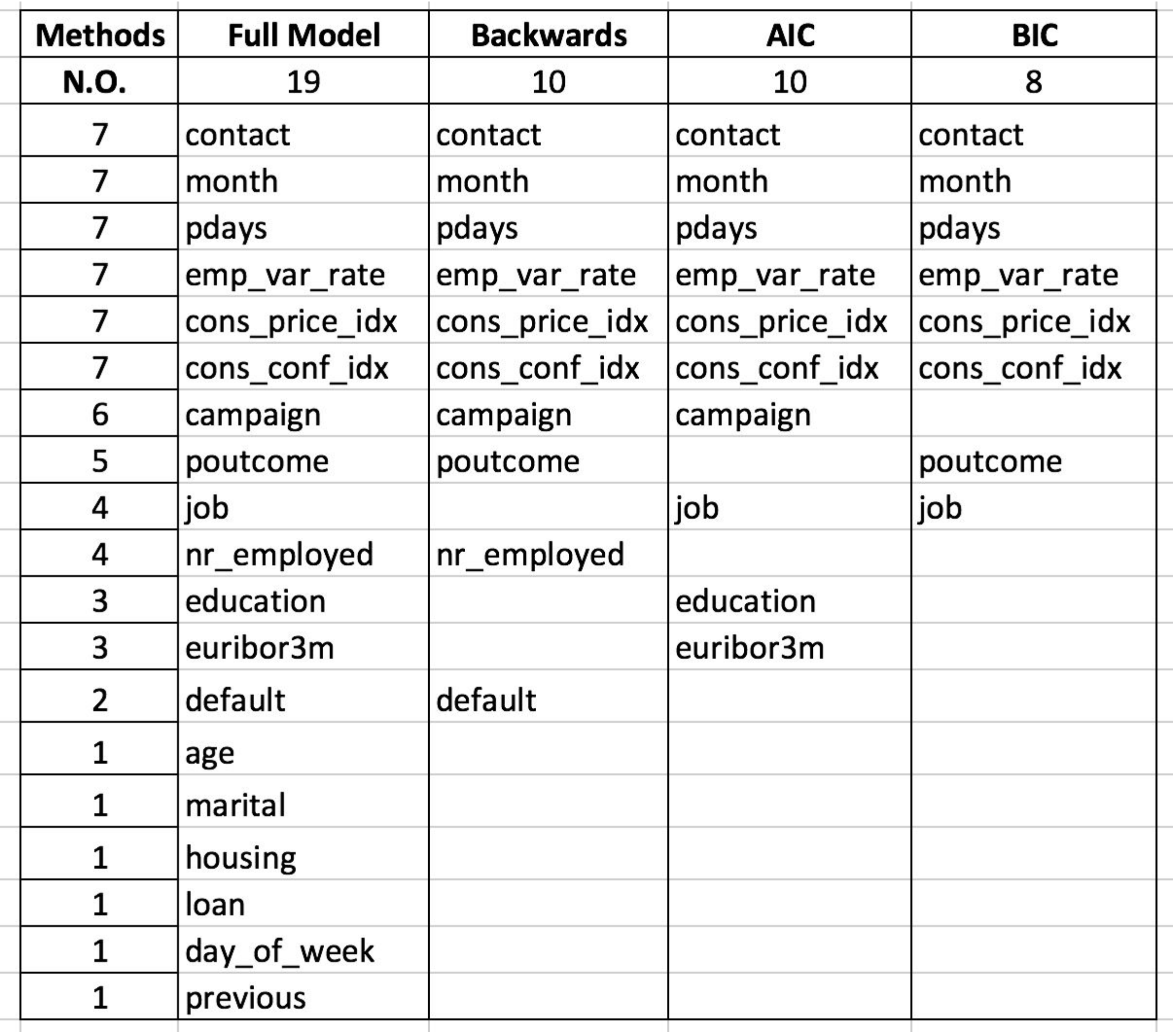
❑ Classification: Logit

- **Lowest Train Error: AIC**
- **Lowest Test Error: Full**
- **Lowest CV Error: Backwards**
- **Lowest N.O. of predictors: BIC**



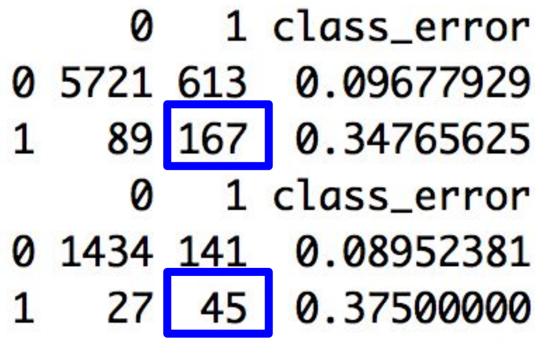
❑ Classification: Logit

- **Lowest class error for both train & test: AIC**



Methods	Full Model	Backwards	AIC	BIC
N.O.	19	10	10	8
7	contact	contact	contact	contact
7	month	month	month	month
7	pdays	pdays	pdays	pdays
7	emp_var_rate	emp_var_rate	emp_var_rate	emp_var_rate
7	cons_price_idx	cons_price_idx	cons_price_idx	cons_price_idx
7	cons_conf_idx	cons_conf_idx	cons_conf_idx	cons_conf_idx
6	campaign	campaign	campaign	
5	poutcome	poutcome		poutcome
4	job		job	job
4	nr_employed	nr_employed		
3	education		education	
3	euribor3m		euribor3m	
2	default	default		
1	age			
1	marital			
1	housing			
1	loan			
1	day_of_week			
1	previous			

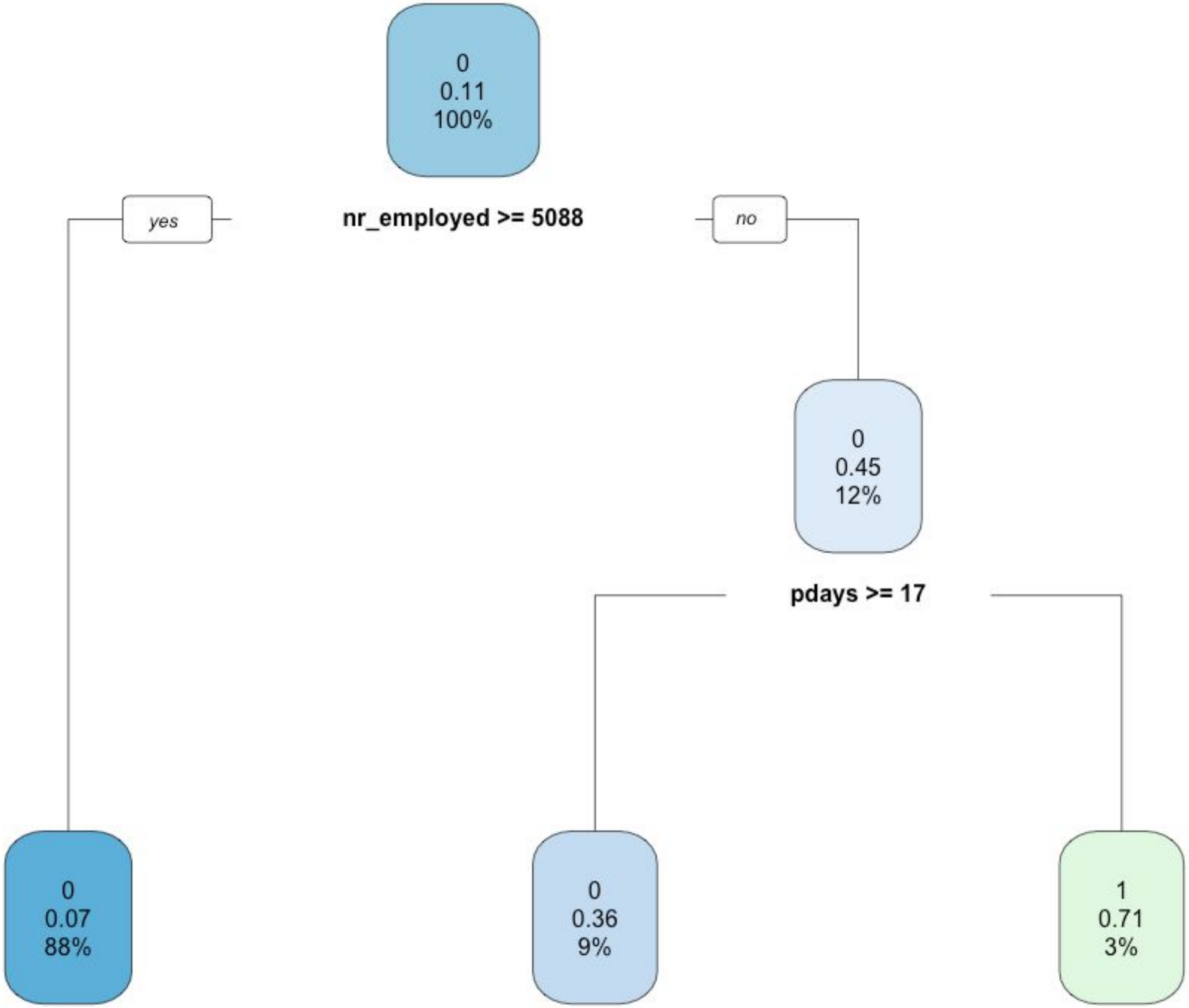
Confusion Table:



```
     0   1 class_error
0 5721 613  0.09677929
1   89 167  0.34765625
     0   1 class_error
0 1434 141  0.08952381
1   27  45  0.37500000
```

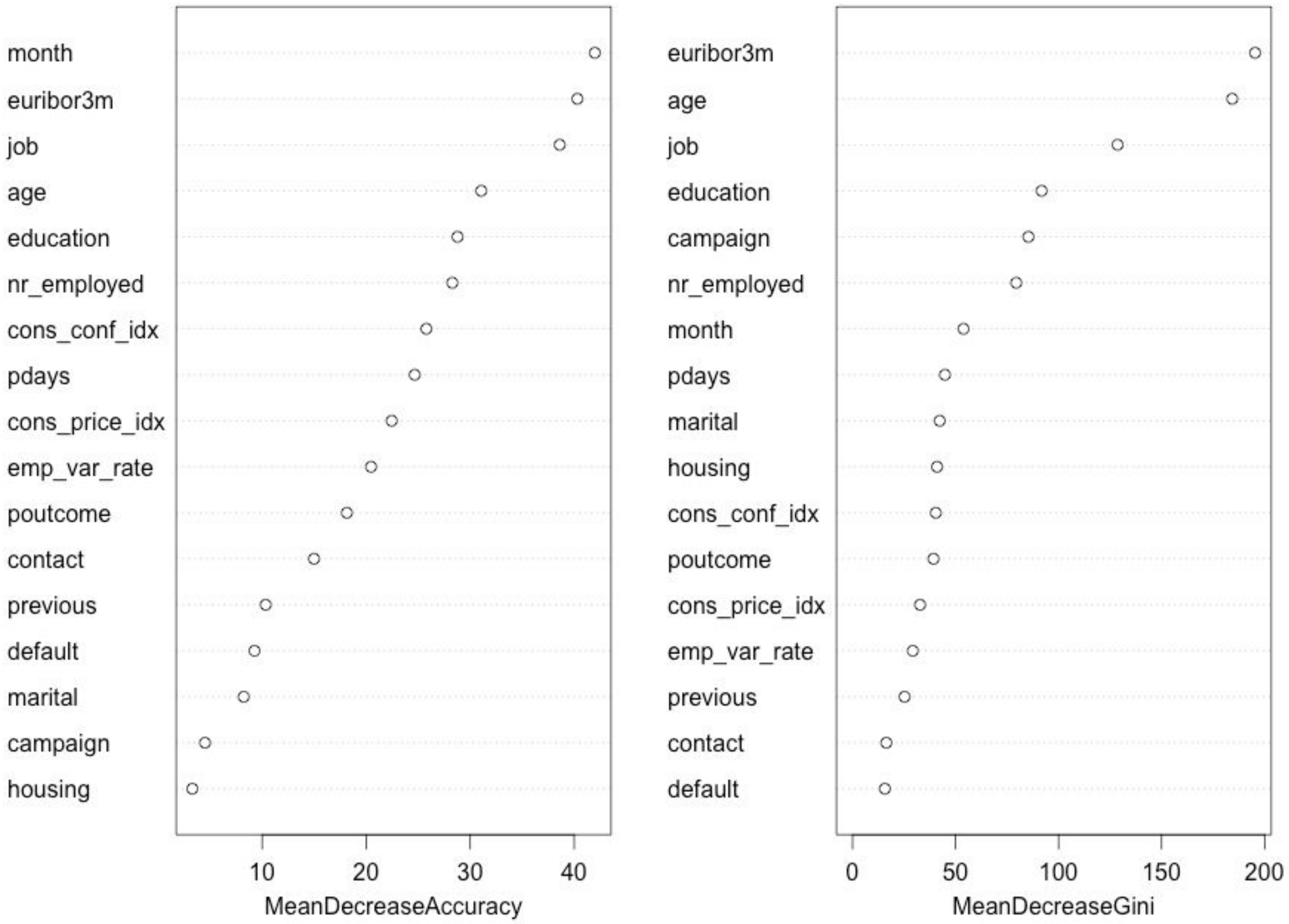
❑ Classification: Tree

- **Simple Tree Model (Pruned)**



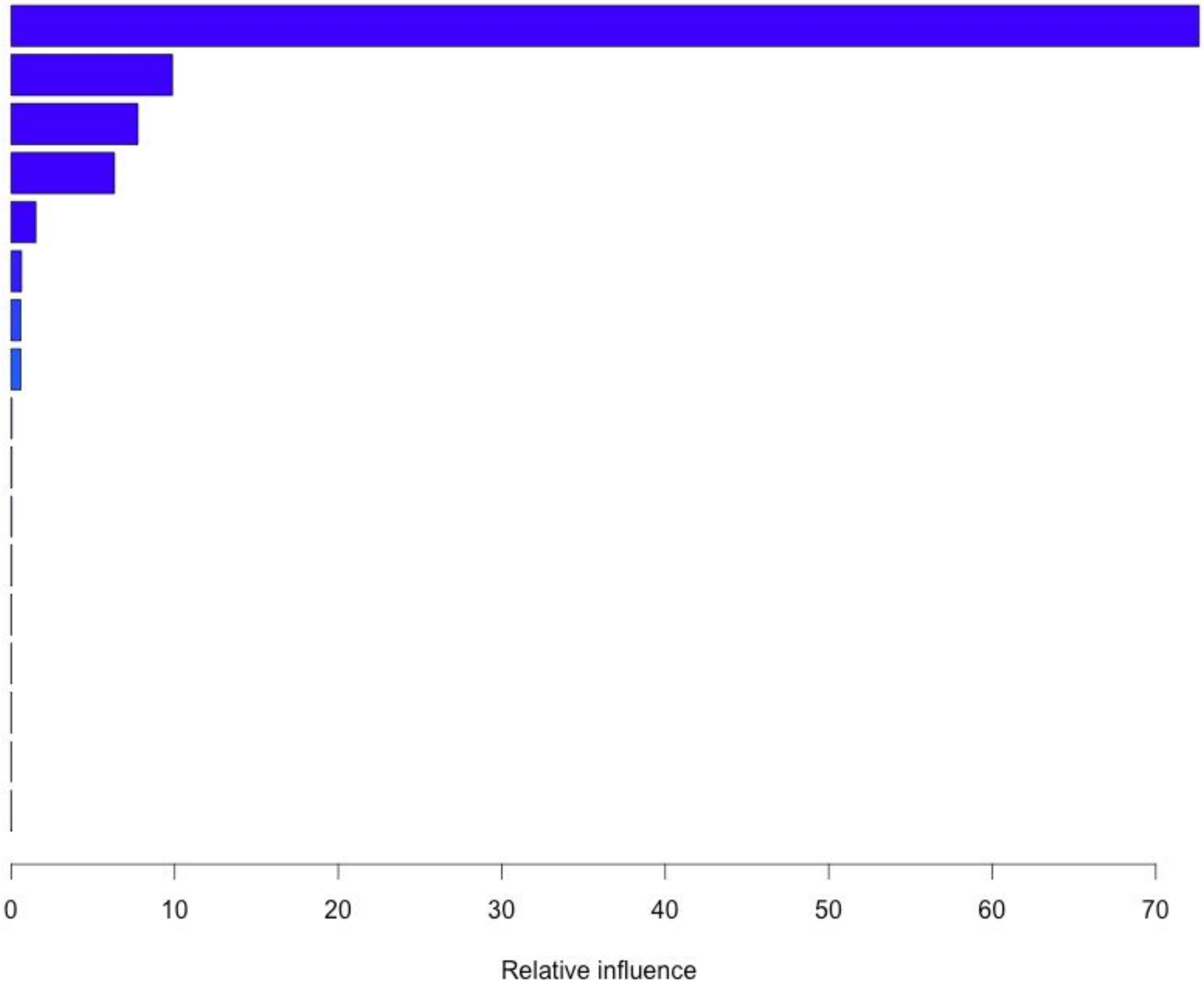
❑ Classification: Tree

- ## Random Forest



❑ Classification: Tree

- AdaBoost



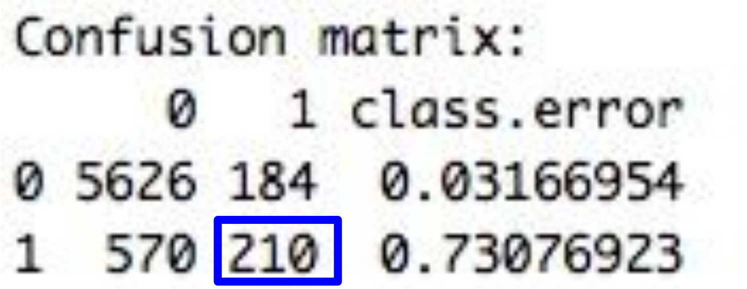
var (rel.inf)

nr_employed (72.643525009)
month (9.858543689)
pdays (7.745866460)
poutcome (6.313556929)
euribor3m (1.507383196)
job (0.628293087)
age (0.586199042)
cons_conf_idx (0.584907609)
cons_price_idx (0.039819632)
campaign (0.033449676)
education (0.030060442)
contact (0.025386936)
emp_var_rate (0.001840013)
marital (0.001168279)
default (0.000000000)
housing (0.000000000)
previous (0.000000000)

❑ Classification: Tree

- **Comparison and Some Issues**

	Simple Tree	Random Forest	AdaBoost
Training	10.75%	11.45%	11.80%
Testing	10.57%	11.06%	11.30%
C.V.	10.67%	-	-



```
Confusion matrix:
     0   1 class.error
0 5626 184  0.03166954
1  570 210  0.73076923
```

❑ Classification: SVM

Kernel	Parameters		Test	Class ‘1’ Errors
	Cost	Other		
Linear	1	-	10.40%	81.65%
Polynomial	1	d = 1	10.40%	81.65%
Radial	1	γ= 0.06	9.82%	76.38%

- We will change our classification probability threshold to 0.09.
- Improved SVM model gives 54% class error, which is not bad to predict a rare case.

❑ Conclusion

- Conclusion

  Based on train, test, CV errors:

  ❖ Logistic method: AIC is better;

  ❖ Tree method: simple tree is better;

  ❖ SVM method: radial kernel is better.

  Given confusion matrix: Logistic is better

- Limitation

  Imbalance class problem: where one response class outnumbered the other class.
  In our case, people finally agree to buy the product is only 10%.