DSPT

DATA SCIENCE PORTUGAL


e-goi

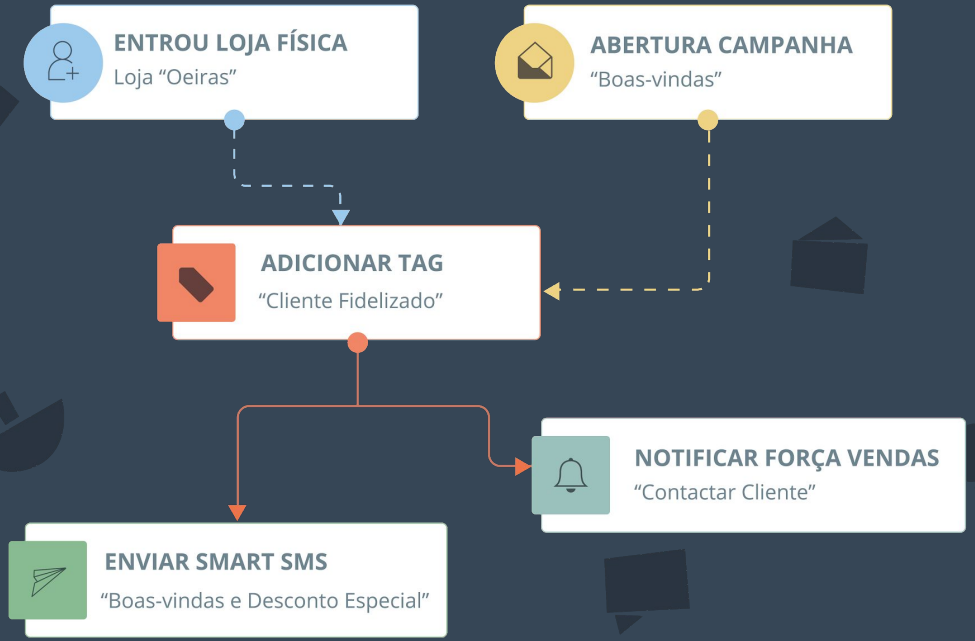
# Intelligent Analysis of Contact Databases' Importation for Spam Prevention

# What is E-goi?

DSPT
DATA SCIENCE PORTUGAL

e-goi

- E-goi is an omni channel marketing platform
- 350.000 users all over the world
- Millions of messages sent daily
- Used by various renowned brands




E-mail | Smart SMS | IVR | Push | WebPush | InStore Behaviour Tracking | Web Behaviour Tracking


**15 Years**
of experience

**600.000.000**
sent SMS in 2017

**350.000**
users

**90.000.000.000**
Sent e-mails in 2017

# Renowned Partners


DSPT
DATA SCIENCE PORTUGAL

e-goi

# Problem

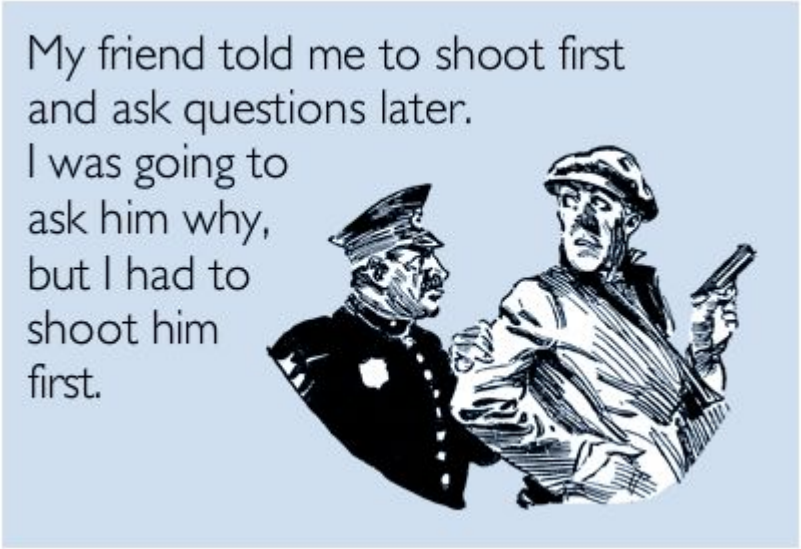
## Campaigns

- Users can create campaigns to send to their subscribers
- One way to do this is to import contact databases
- This functionality is sometimes used by spammers in order to make their work easier

# Problem

## ISP's Attitude

- ISP's are normally also responsible for providing e-mail services
- To put it lightly... They don't like spammers
- To put it in another way. You better run if there is even a reason to suspect you
- This leads to domains being blocked by ISP's or at least being treated cautiously




DSPT
DATA SCIENCE PORTUGAL

e-goi

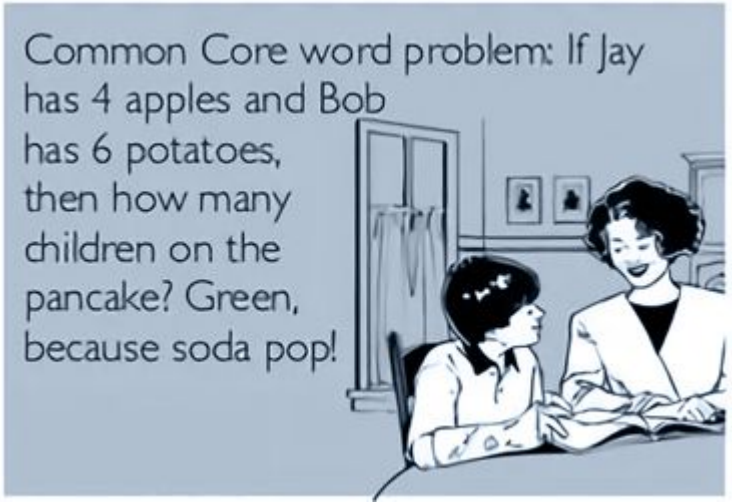
# How can we solve it?


DSPT
DATA SCIENCE PORTUGAL

e-goi

# Solution

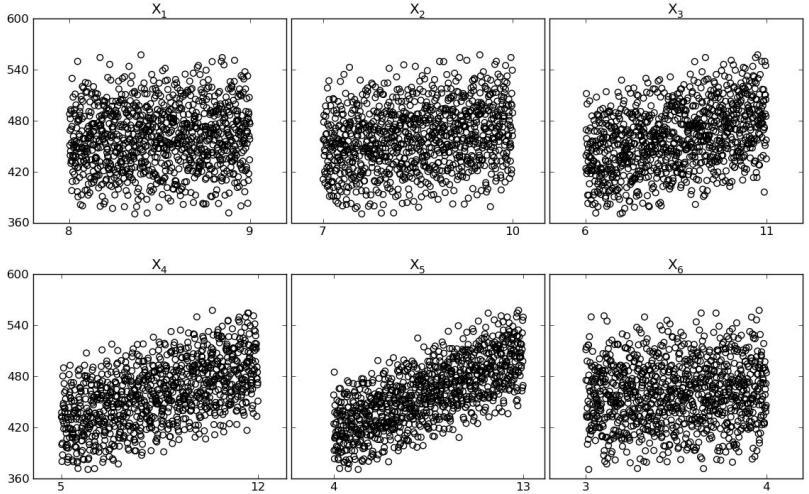
- The only reliable way to solve this problem is to block spammers before campaigns can be started
- However this means less information to be evaluated
- Time and experience are required to detect spam based on solely this amount of information
- In order to maintain such a solution automatization is necessary



# Solution

## Information

- Information is important
- However too many information can be problematic
- Although features analysed in this problem are not those normally associated with spam classification, their number is still significant
- Which ones are more relevant?



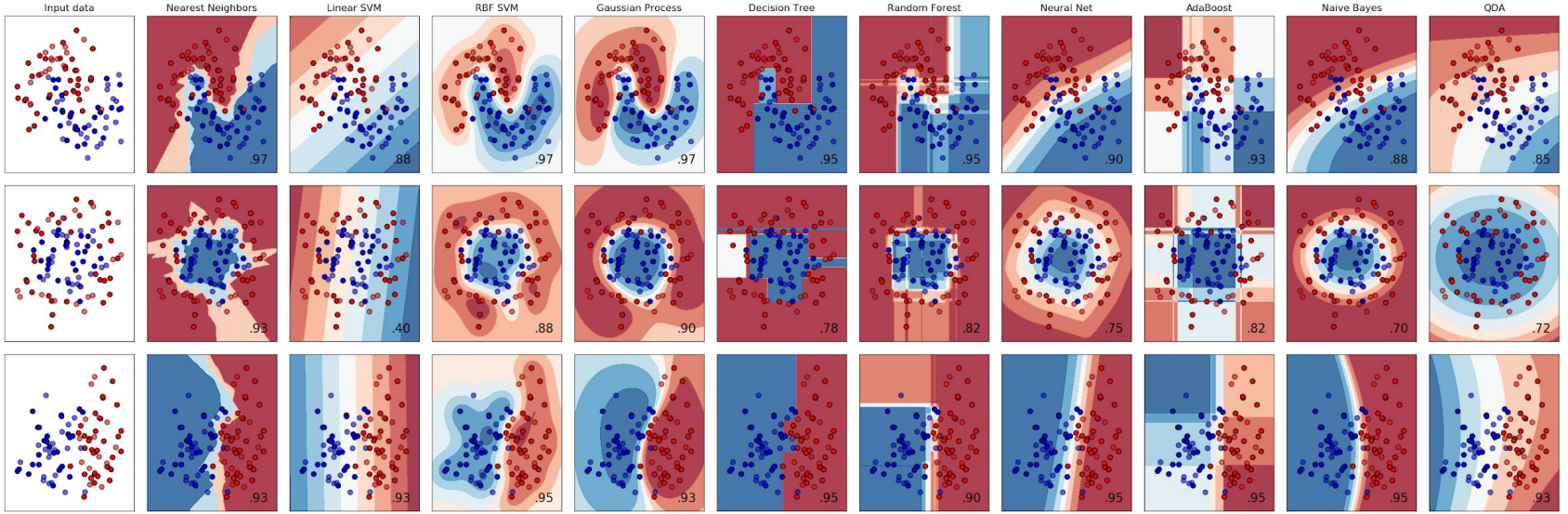

DSPT
DATA SCIENCE PORTUGAL

e-goi

# Solution

## Classifiers

- There are many types of classifier to be used
- We will focus on Random-forest, SVC and AdaBoost



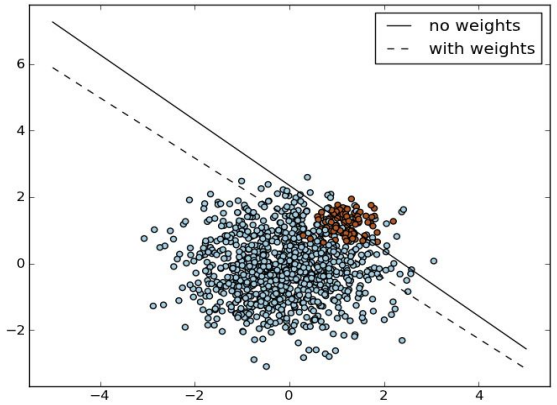

DSPT
DATA SCIENCE PORTUGAL

e-goi

# Solution

## Unbalanced Data

- Ideally a classifier should always receive a balanced set of data
- This is not realistic
- Example-set: 24000 successful importations to 600 blocked
- Solutions include:
  - Under-sampling the majority class
  - Over-sampling the minority-class
  - Altering the relative weights of each class



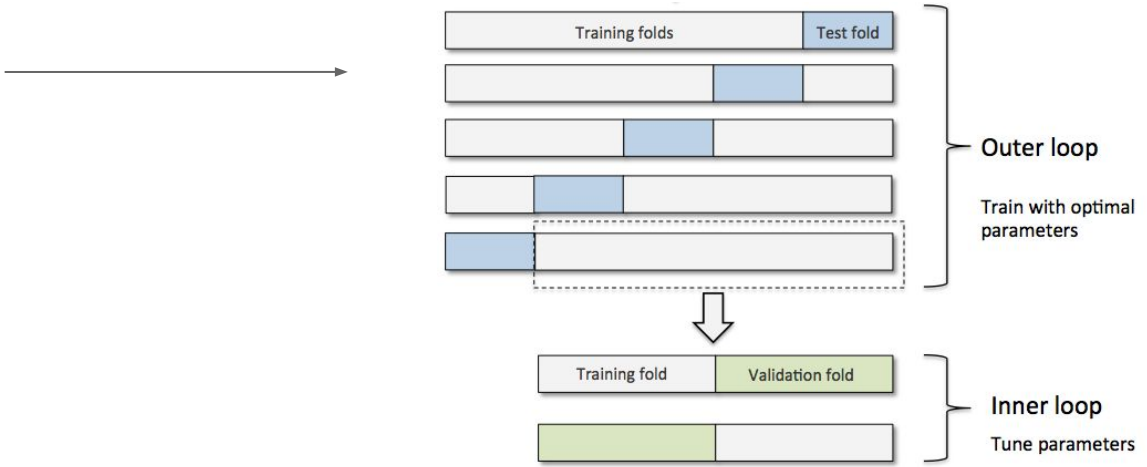

DSPT
DATA SCIENCE PORTUGAL

e-goi

# Solution

## Tuning

- Classifiers are affected by hyper-parameters
- For-each one there is a combination that offers the best classification results
- How to find it?



# Prototype Comparison

DSPT
DATA SCIENCE PORTUGAL

e-goi

# Evaluation

## Metrics

- Accuracy is not an adequate metric
- Precision and recall are more adequate
- Taking into account the context which metric should be more relevant?

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

# Evaluation

| | Precision | Recall | Support (Total Instances) |
|---|---|---|---|
| Positive Class | 0.733 | 0.726 | 609 |
| Negative Class | 0.993 | 0.993 | 24144 |
| Avg / Total | 0.987 | 0.987 | 24753 |

— SVC

| | Precision | Recall | Support (Total Instances) |
|---|---|---|---|
| Positive Class | 0.893 | 0.819 | 609 |
| Negative Class | 0.995 | 0.998 | 24144 |
| Avg / Total | 0.993 | 0.993 | 24753 |

— RF

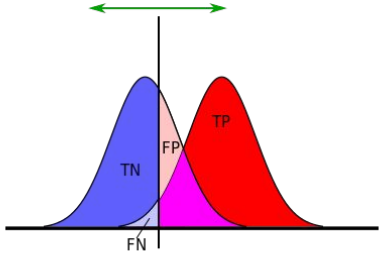
| | Precision | Recall | Support (Total Instances) |
|---|---|---|---|
| Positive Class | 0.867 | 0.844 | 609 |
| Negative Class | 0.996 | 0.997 | 24144 |
| Avg / Total | 0.993 | 0.993 | 24753 |

— AB

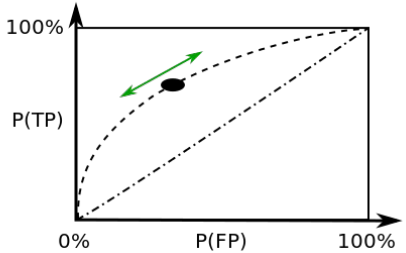
# Evaluation

## ROC Curves

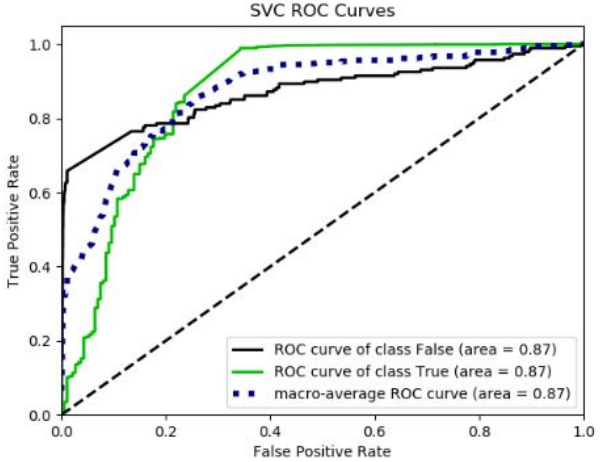
- Show a graphical representation of the connection/trade-off between sensitivity and specificity
- Sensitivity (true positive rate or recall)
- Specificity (also called the true negative rate)

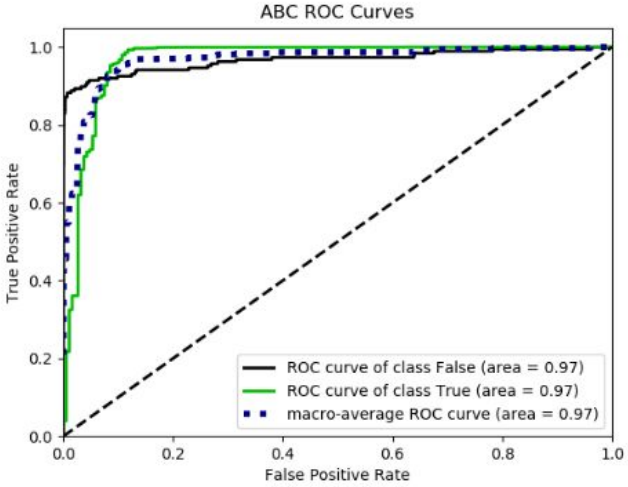


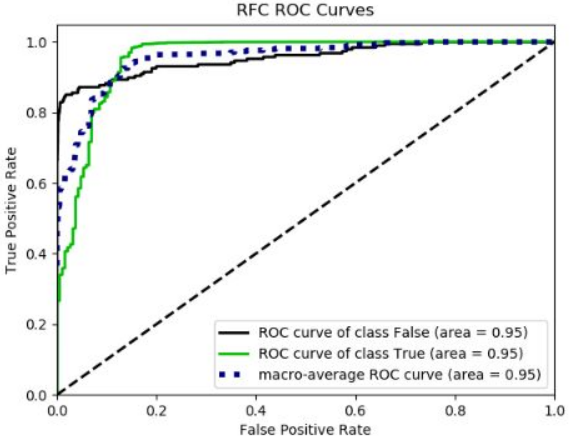

| | |
|---|---|
| TP | FP |
| FN | TN |



# Evaluation

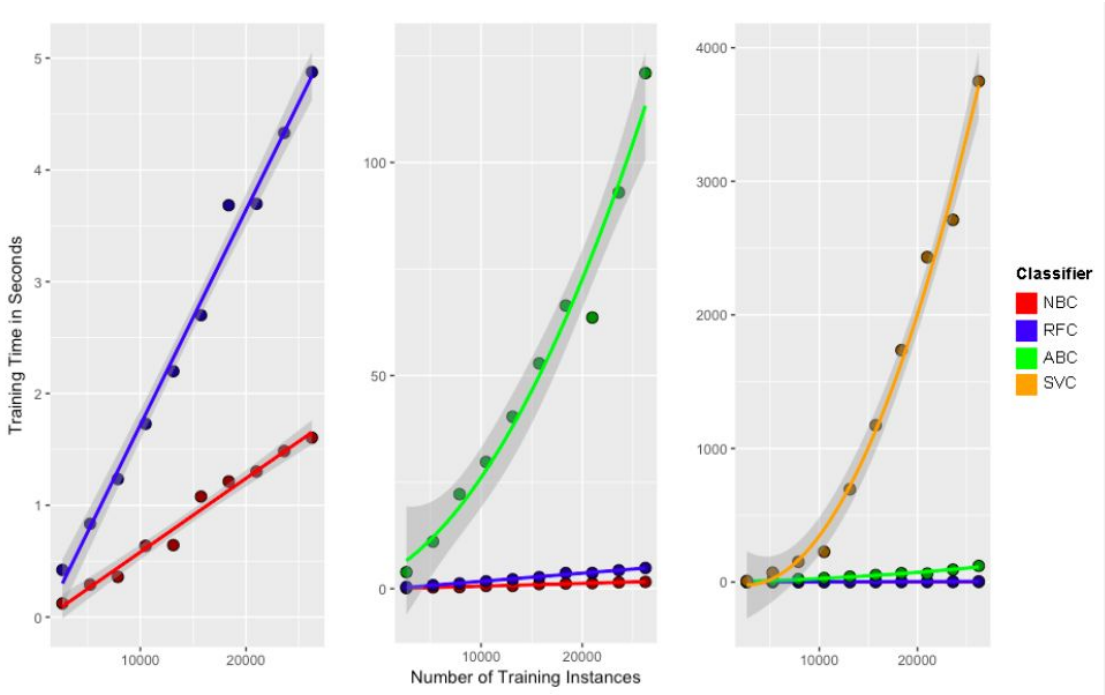

SVC ROC Curves
True Positive Rate
False Positive Rate
ROC curve of class False (area = 0.87)
ROC curve of class True (area = 0.87)
macro-average ROC curve (area = 0.87)


ABC ROC Curves
True Positive Rate
False Positive Rate
ROC curve of class False (area = 0.97)
ROC curve of class True (area = 0.97)
macro-average ROC curve (area = 0.97)


RFC ROC Curves
True Positive Rate
False Positive Rate
ROC curve of class False (area = 0.95)
ROC curve of class True (area = 0.95)
macro-average ROC curve (area = 0.95)

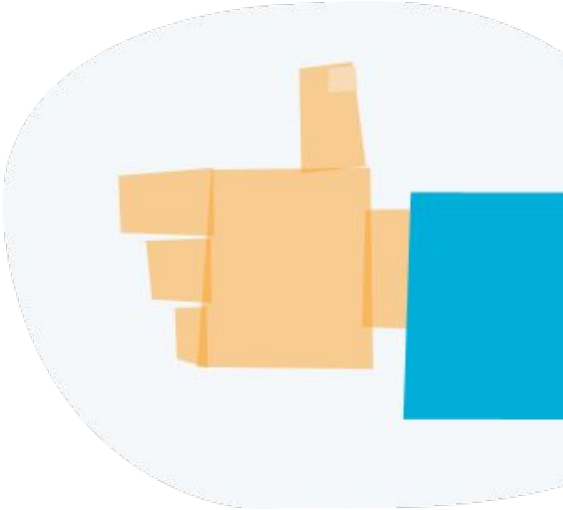
# Evaluation

## Training Time



# Conclusion

- Spam Prevention
- Easier Editability
- Less Work-load for the  Employees in charge of Deliverability
- Easily available through documented API



# QUESTIONS

DSPT
DATA SCIENCE PORTUGAL

e-goi