Machine Learning for Big Data: how to predict customers loyalty

Dmitry Efimov

American University of Sharjah

November 18, 2016

Outline

Problem formulation

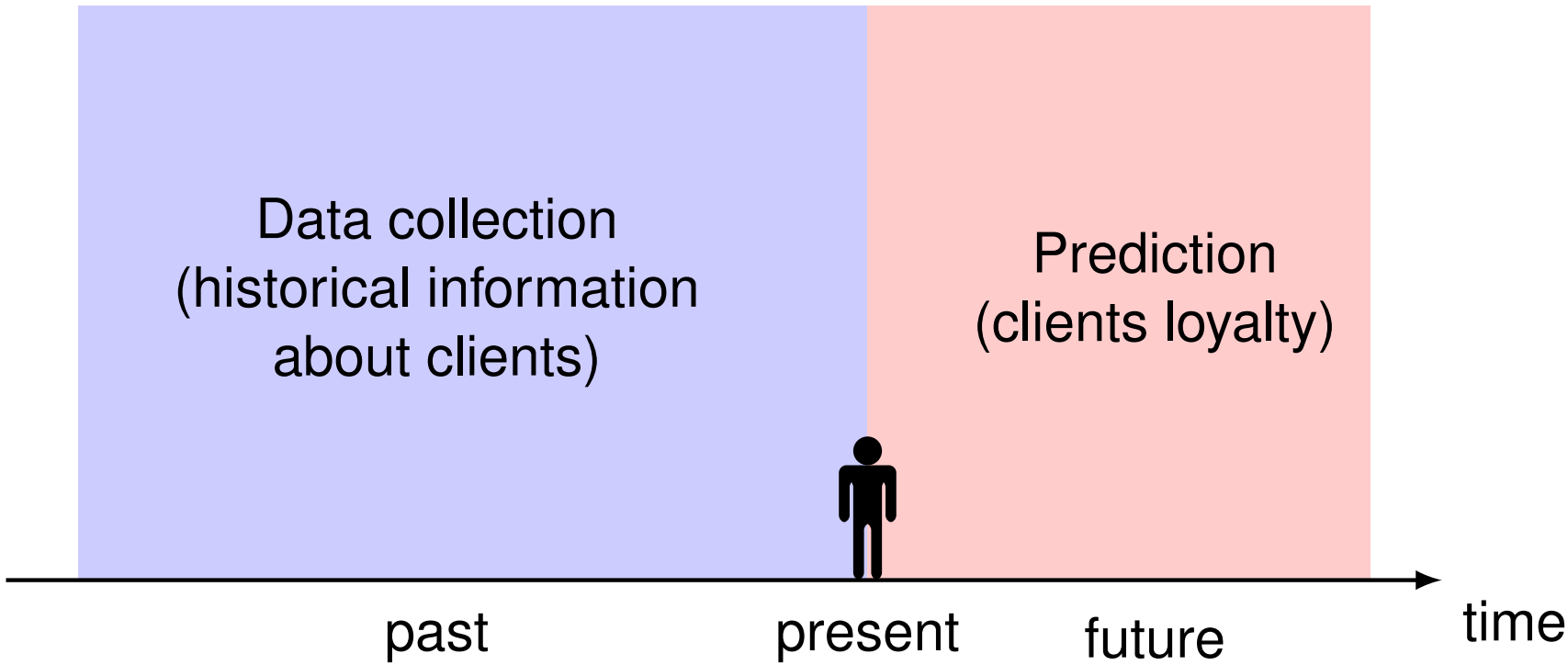
Cross validation and loss functions

Feature engineering

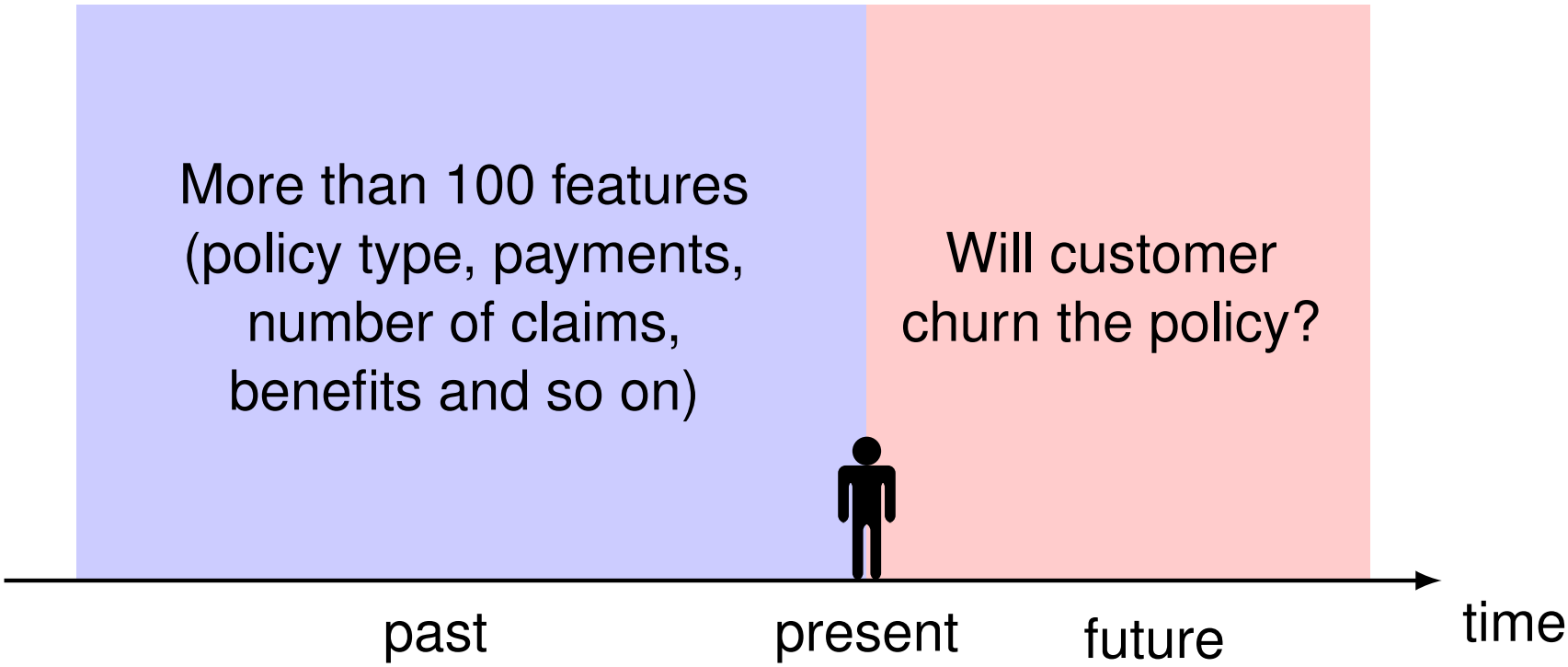
How to handle big data?

Scope of models

Problem formulation



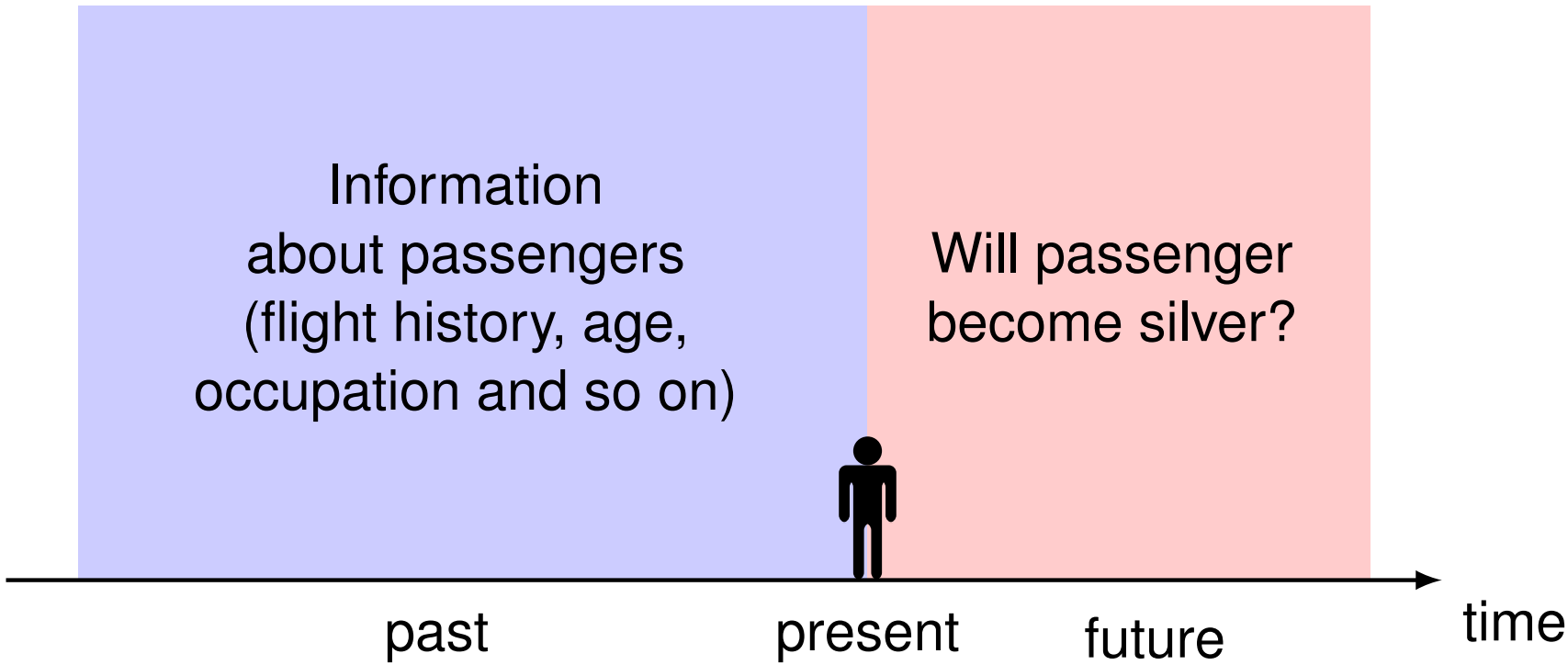
Example 1 (Deloitte competition)



- about 400 000 policies
- about 11 000 000 records

kaggle.com/c/deloitte-churn-prediction

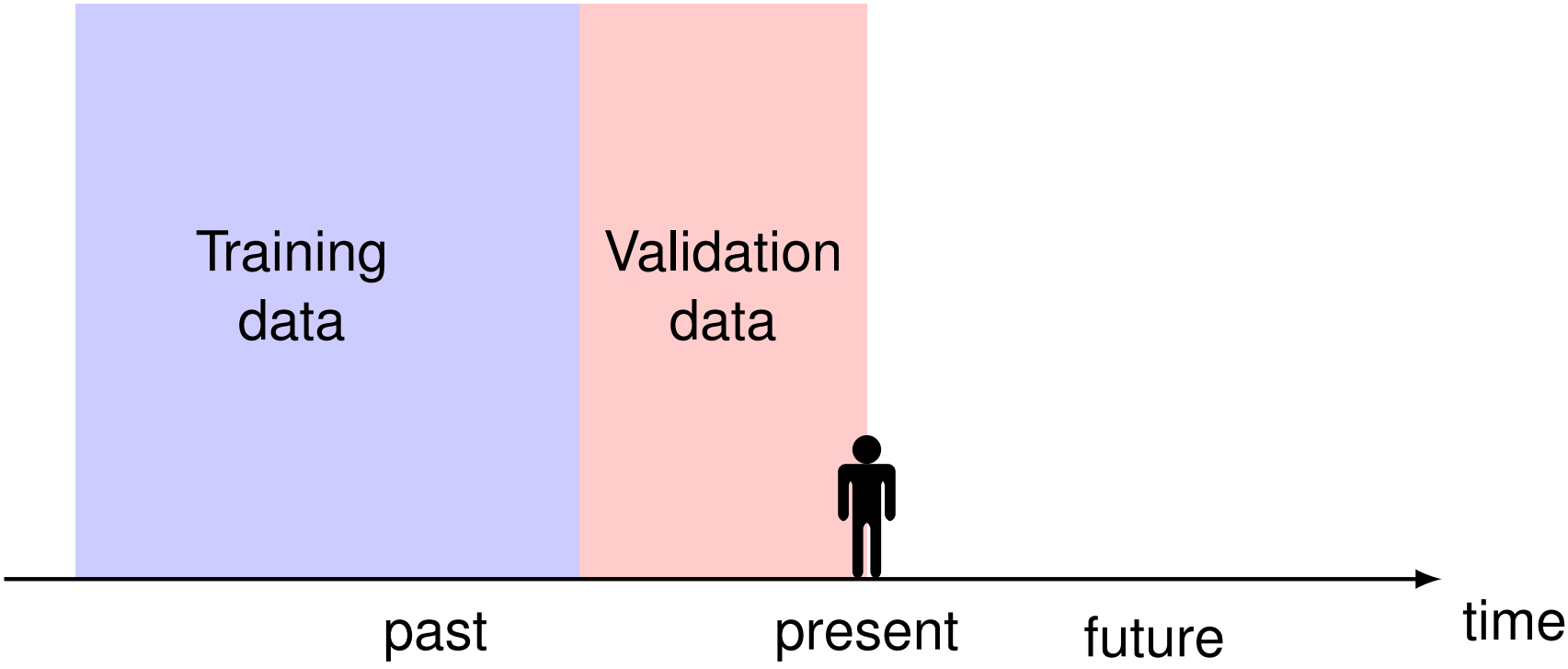
Example 2 (Etihad Airline)



Related HOWTO

- How to estimate the accuracy?
- How to work with different historical depths?
- How to handle huge amount of historical information?
- How to choose the predictive model?

How to estimate the accuracy?



- **cross validation** is a key procedure
- choose an appropriate loss function
- keep distributions:

$$p(x) \sim p_{train}(x) \sim p_{validation}(x) \sim p_{test}(x)$$

Example of loss functions

y - target variable

$x = (x_1, \ldots, x_n)$ - vector of features

$\hat{y} = f(x)$ - predictive model

- Mean squared error

$$E\left[(y - \hat{y})^2\right]$$

- LogLoss

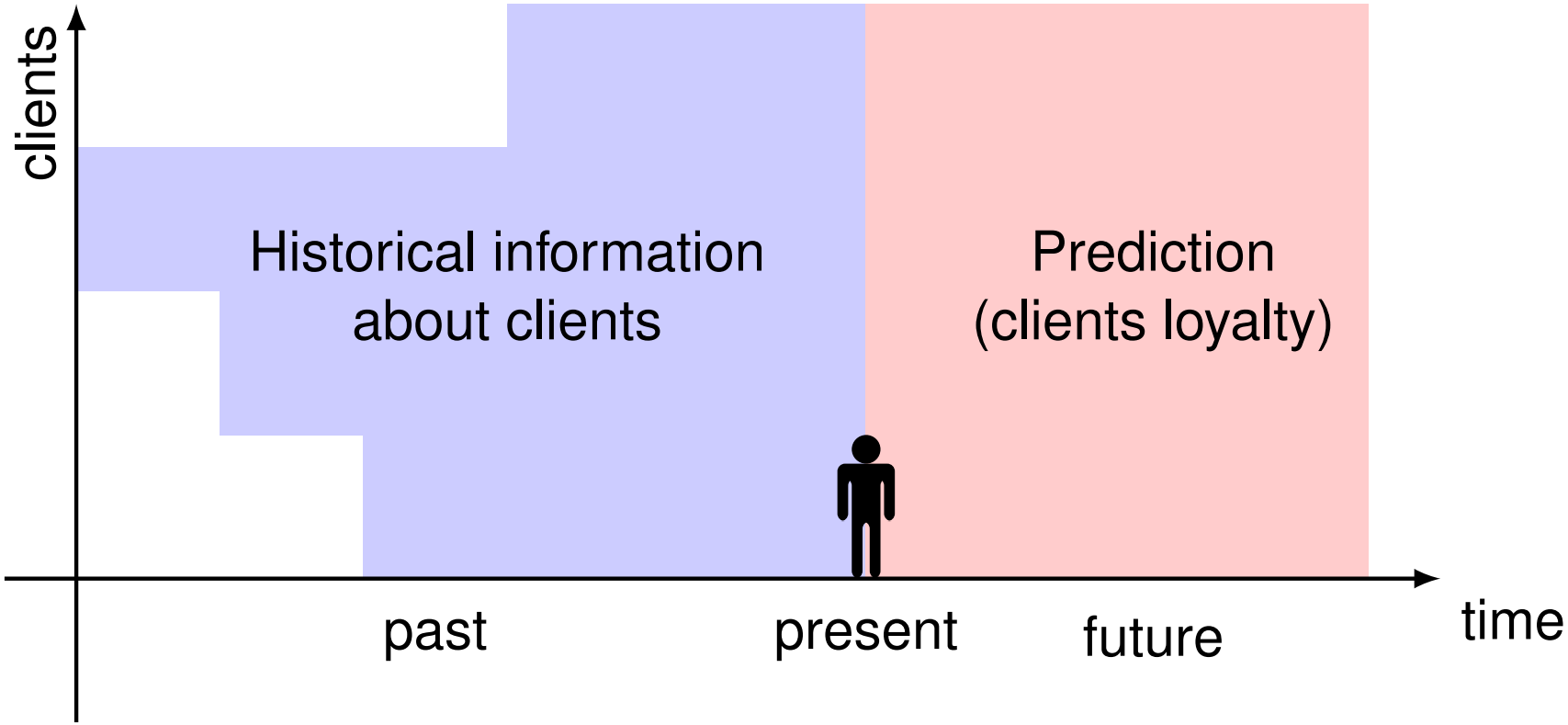
$$E\left[y \cdot \ln \hat{y} + (1 - y) \cdot \ln(1 - \hat{y})\right]$$

- Area under the curve (AUC)

P(positive example is higher than negative example)

T.Hastie, R.Tibshirani and J.Friedman "The elements of statistical learning." *Springer*, 2009

How to work with different historical depths?



- **feature engineering** is a key procedure
- unsupervised technique is very useful
- visualize your data

Example of features

- Statistics by the historical features with ”sliding window”
  - maximum during the last month
  - the average change during the last year
- Unsupervised features
  - t-distributed stochastic neighbor embedding (t-SNE)
  - principal component analysis (PCA)
  - autoencoders
- Other ideas
  - binary feature by discretized continuous features
  - ...

[1] lvdmaaten.github.io/tsne/

[2] www.deeplearningbook.org

Feature engineering in Deloitte

Completed • $70,000 • 37 teams

As the World Churns

Tue 22 Oct 2013 – Sat 21 Dec 2013 (2 years ago)

Dashboard

Public Leaderboard - As the World Churns

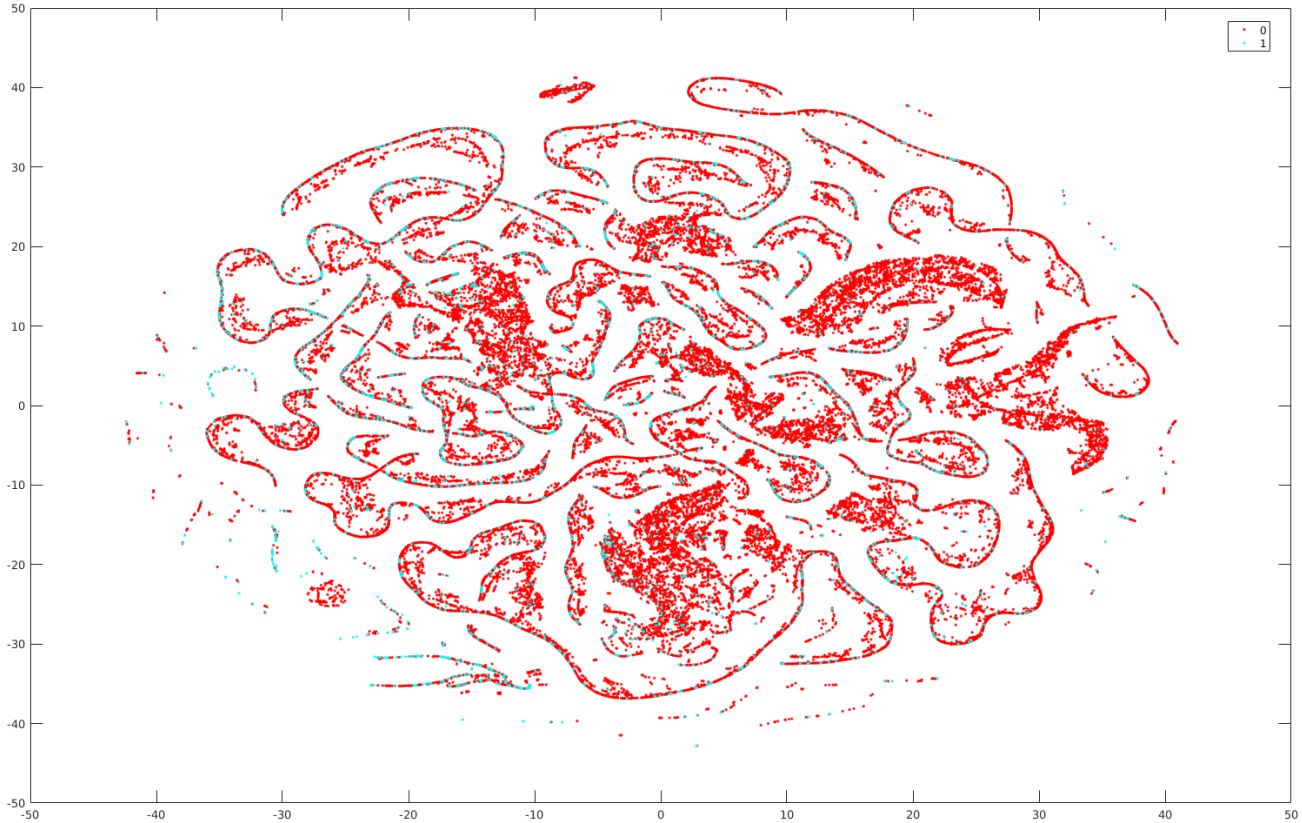
This leaderboard is calculated on approximately 25% of the test data.
The final results will be based on the other 75%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1w	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	**Dmitry Efimov** *	**0.81917**	**155**	**Sat, 21 Dec 2013 14:21:48 (-20h)**
2	—	Leustagos & Gxav *	0.81869	78	Sat, 21 Dec 2013 22:04:03 (-6.3h)
3	↑6	Michael Jahrer & Jeong-Yoon Lee *	0.81721	73	Sat, 21 Dec 2013 22:26:29 (-0.1h)
4	↑1	ivo and BreakfastPirate	0.81457	174	Sat, 21 Dec 2013 22:05:08 (-4.4h)
5	↓2	Datrik Intelligence	0.81442	7	Sat, 21 Dec 2013 23:46:07 (-0.3h)
6	—	FAndy & Sen	0.81326	72	Sat, 21 Dec 2013 19:36:50
7	↓3	An apple a day	0.81237	75	Sat, 21 Dec 2013 23:47:07 (-1.6h)
8	↑12	agdavis ‡	0.81176	10	Sat, 21 Dec 2013 16:53:22
9	new	alegro	0.80947	5	Sat, 21 Dec 2013 23:26:47 (-19.7h)
10	↓3	S&B500	0.80918	144	Sat, 21 Dec 2013 22:30:42 (-0.6h)

Example of visualization using t-SNE features

- Visualization helps to catch important facts about data



kaggle.com/c/santander-customer-satisfaction

How to handle huge amount of historical information?

- downsampling (remember to keep distributions)
- batch optimization
- online algorithms
- parallel computing
- non-standard ideas

How to choose the predictive model?

- Parametric
  - Regressions
  - Kernel methods (SVM)
  - Bayesian approach
  - Neural networks
- Non parametric
  - Decision trees
- Ensembling
  - Boosting

[1] github.com/diefimov/MTH594_MachineLearning
[2] Lectures by Andrew Ng on YouTube

Regressions (general framework)

- Predictive model depends on parameters θ

$$\hat{y} = f(x, \theta)$$

- To find θ we formulate an optimization problem

$$\hat{\theta} = \arg\min_{\theta} L(y, f(x, \theta)),$$

  where L is a loss function
- Use optimization algorithm (e.g., SGD) to find the best values for θ

Bayesian approach

- Predictive model is a parametric family of distributions

$$p(x, y; \theta) = p(x, y \mid \theta) \cdot p(\theta) = p(\theta \mid x, y) \cdot p(x, y)$$

- To find θ we formulate an optimization problem

$$\hat{\theta} = \arg\max_{\theta} p(\theta \mid x, y),$$

- Use Bayes rule to solve it

$$p(\theta \mid x, y) = \frac{p(x, y \mid \theta) p(\theta)}{p(x, y)} \propto p(x, y \mid \theta) \cdot p(\theta)$$

$$(\text{posterior} \propto \text{likelihood} \cdot \text{prior})$$

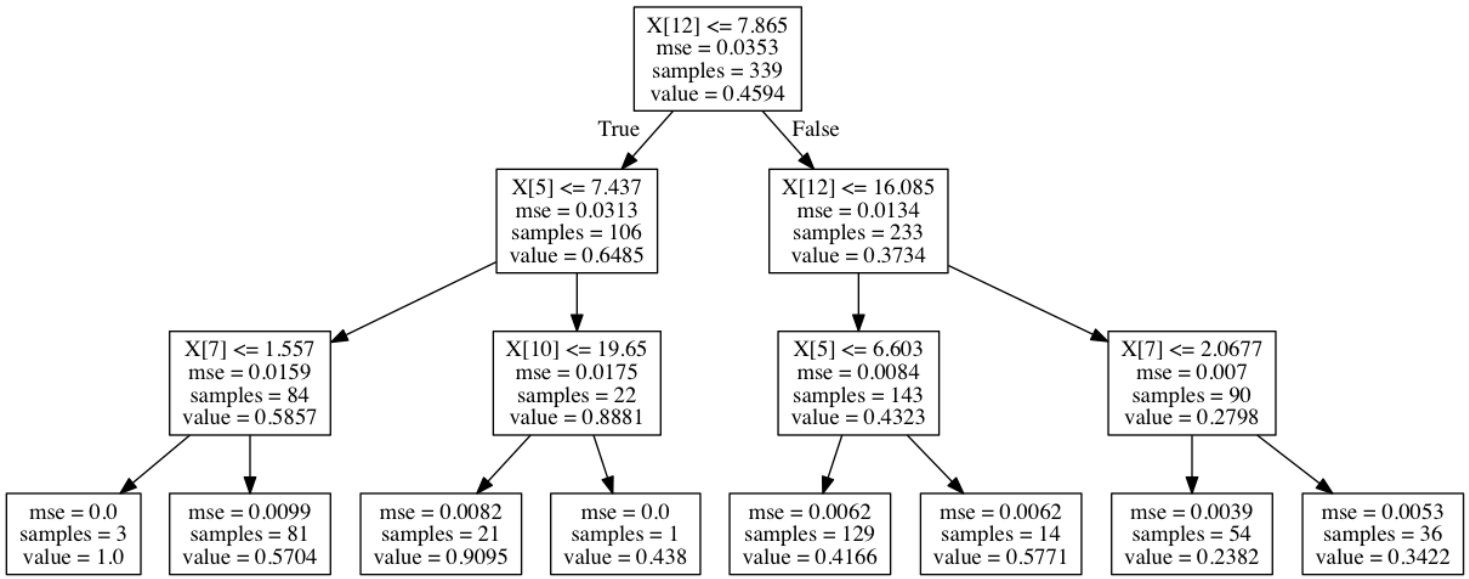
E.T.Janes "Probability theory: logic of science." *Cambridge University Press*, 2003

Decision trees

- Predictive model is non-parametric

$$\hat{y} = f(x)$$

- The resulted model can be visualized as



T.Hastie, R.Tibshirani and J.Friedman "The elements of statistical learning." *Springer*, 2009

Boosting

$$\hat{y} = f(x_1, \ldots, x_n) = \sum_{k=0}^{N} f_k(x_1, \ldots, x_n)$$

Algorithm 1 General boosting algorithm

1: $f_0(x) = E[y]$
2: **for** $k = 1$ **to** N **do**
3: evaluate current errors $z = y - \sum_{s=0}^{k-1} f_s(x_1, \ldots, x_n)$
4: train model f_k to predict z
5: **return** $\sum_{k=0}^{N} f_k(x_1, \ldots, x_n)$

General strategy

- Investigate the data manually
- Choose loss function
- Define the cross validation scheme
- Generate features
- Choose the algorithm

Thank you! Questions?

Dmitry Efimov

diefimov@gmail.com

kaggle.com/efimov

github.com/diefimov