rapidminer

Selling Data Science: Unboxing the Black Box

Vladimir Mikhnovich
Friendly Finance Group

© RapidMiner, Inc. All rights reserved.

Housekeeping

- Today's session is being recorded
- Trouble with audio or video? Try logging out and back in
- Q&A via the questions panel

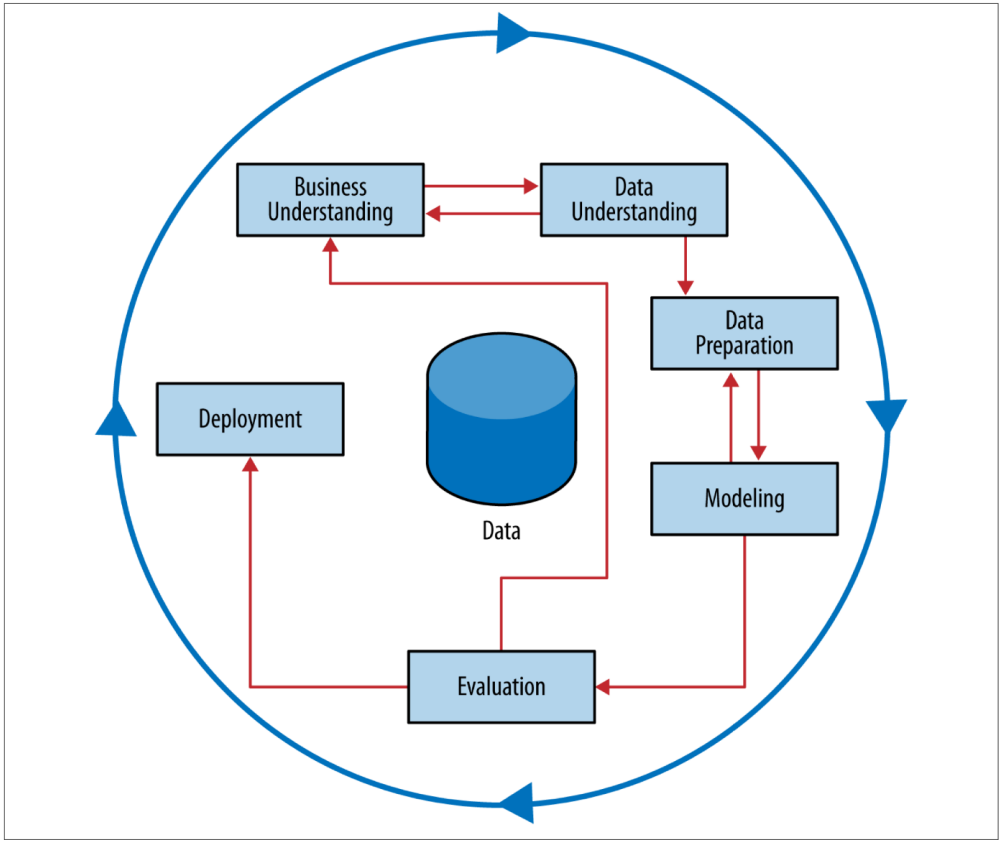
©2017 RapidMiner, Inc. All rights reserved.

Today's Discussion Points

- Why simple models outstand complex ones
- Why communicating is different when talking to data scientists, analysts, and stakeholders
- Why business metrics matter

©2017 RapidMiner, Inc. All rights reserved.

CRISP-DM process



Important consideration: business understanding is the very first step of the process.

But also it's at the logical end of the cycle in case we repeat it.

©2017 RapidMiner, Inc. All rights reserved.

Doing science vs. doing business

- You build a model
- You measure the performance
- You check its fitness on a real problem

- Why? Because you most likely ought to solve the real problem, not do academic research

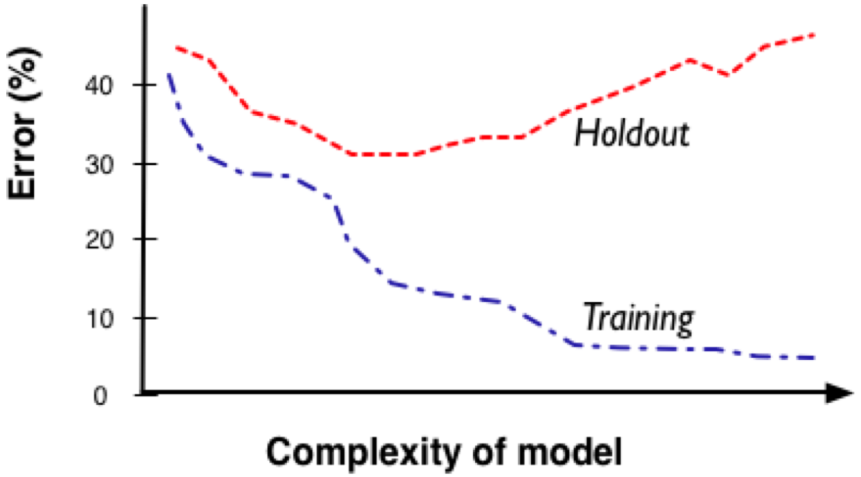
©2017 RapidMiner, Inc. All rights reserved.

rapidminer

Why simplicity is important

© RapidMiner, Inc. All rights reserved.

Why should we make it simple?



- Well-known metric: model complexity vs. training and testing errors
- But what about interpretability and ease of understanding?

©2017 RapidMiner, Inc. All rights reserved.

Why should we make it simple?

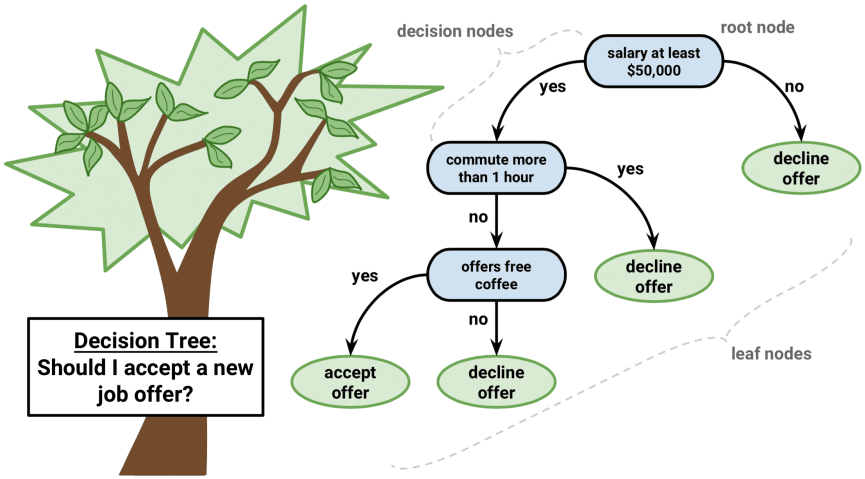
- Simpler models show more stable performance and more robust results
- Simpler models help better understand how data is used, thus it is easier to built trust in the proposed ML method
- Once it has been made clear, you surely can move towards more complex models

- Still not convinced?... Well, try to explain it to parents or children :)

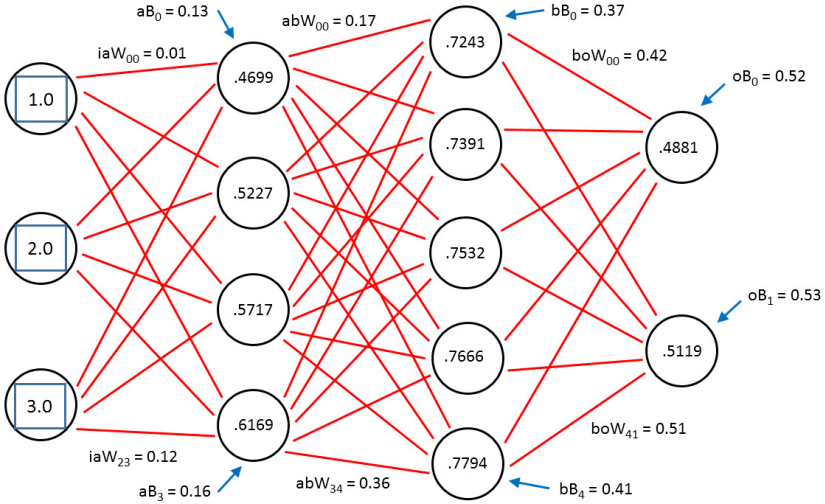
©2017 RapidMiner, Inc. All rights reserved.

Simple example: algorithms in pictures

Decision tree



Neural network



©2017 RapidMiner, Inc. All rights reserved.

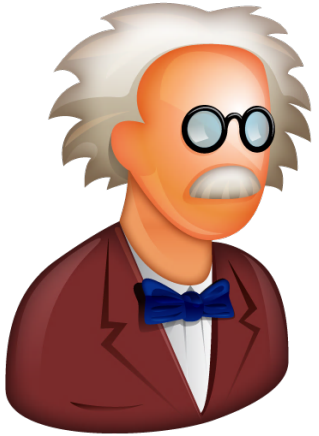
rapidminer

How to communicate results and sell solutions

© RapidMiner, Inc. All rights reserved.

Communicating with data scientists

- With fellow data scientists, you can freely discuss C parameter tuning in SVM, or how to choose optimal number of clusters, and what type of feature normalization works better with GBT



©2017 RapidMiner, Inc. All rights reserved.

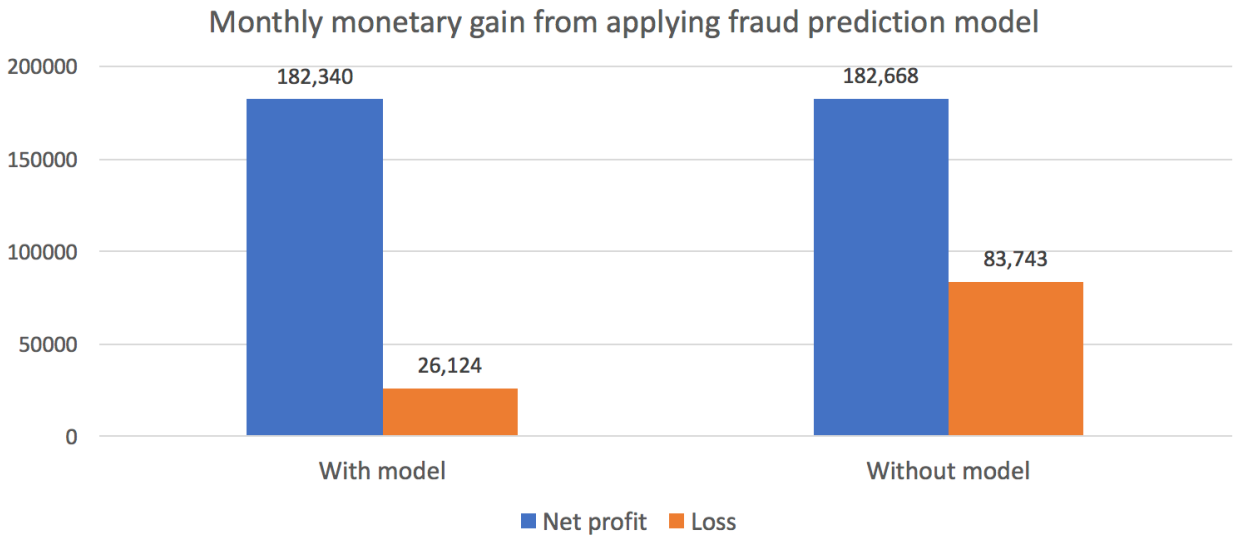
Communicating with analysts

- With analysts, it is important to explain how the results are to be interpreted to make inferences and conclusions. Possible deep knowledge of algorithms might be beneficial, but at the end analysts are the first to *interpret* and *use* the results.

©2017 RapidMiner, Inc. All rights reserved.

Communicating with the business

- With stakeholders, **business impact and $ figures matters the most** (regardless of an actual algorithm).



©2017 RapidMiner, Inc. All rights reserved.

Rule # 1

Sell results, not algorithms

This means it is always an advantage not only to explain and demonstrate theoretical possibilities of machine learning, but also to have a prototype built which allows you to assess potential advantages and gains

©2017 RapidMiner, Inc. All rights reserved.

Rule # 2

Showcase in a form understandable by the business

Use metrics that produce meaningful business impact (revenue, ROI, loss, savings…) it all depends on type of problem and also business model and map performance metrics to business values.

©2017 RapidMiner, Inc. All rights reserved.

Rule # 3

Make comparisons

A/B testing is one of the most widely used analytic evaluation methods in marketing and web design, but is also applicable in many other domains.

For business, it’s useful to consider border case: what if we had nothing instead of the proposed solution?

©2017 RapidMiner, Inc. All rights reserved.

Other rules to keep in mind

- Don’t hesitate to **make allies**: financial analysts, risk analysts, business analysts etc. They can provide great help in building correct numbers.
- Do not fear to **show best practices and success stories of other companies** as an example. Even the competitors’.

©2017 RapidMiner, Inc. All rights reserved.

rapidminer

Use case:
Business value of fraud detection model

© RapidMiner, Inc. All rights reserved.

Objective

- Build and evaluate a model for detecting fraud in online payment transactions of a merchant.
- Assess model performance from a business point of view.

©2017 RapidMiner, Inc. All rights reserved.

Transactions dataset

We have data on a merchant with 12k transactions and ~1% fraud level

ExampleSet (12456 examples, 2 special attributes, 9 regular attributes)

Filter (12,456 / 12,456 examples): all

id	Fraud	card_bin	issuer_bank	card_type	user_city	user_country	user_ISP	sum	country_mismat...	country_code
215008693	false	520373	PROMSVYAZBANK	MasterCard World Card	missing	IL	013 NetVision Ltd.	800	true	643
215072636	false	525477	CREDIT UNION ...	MasterCard World Card	missing	LT	TEO LT, AB	400	true	643
214754715	false	512762	Citibank	MasterCard World Card	Amsterdam	NL	Digital Ocean, Inc.	1500	true	643
214618847	false	548673	ALFABANK	MasterCard Standard ...	Amsterdam	NL	Websense Hoste...	900	true	643
214550431	false	546938	SBERBANK OF R...	MasterCard Standard ...	Amsterdam	NL	LeaseWeb B.V.	950	true	643
214494212	false	427229	VTB24	Visa Classic	Craiova	RO	Voxility S.R.L.	400	true	643
214557016	false	415429	ALFABANK	Visa Gold	Craiova	RO	Voxility S.R.L.	900	true	643
215065267	false	548673	ALFABANK	MasterCard Standard ...	Craiova	RO	Voxility S.R.L.	1400	true	643
214481959	false	440503	ROSBANK	Visa Classic	missing	RO	Luxoft Profession...	280	true	643
221447892	true	521324	TINKOFF CREDI...	MasterCard World Card	Saint Petersburg	RU	OJSC MegaFon	6185	false	643
214281248	true	479004	ALFABANK	Visa Platinum	Moscow	RU	OJSC MegaFon	1400	false	643
208488146	true	427681	SBERBANK OF R...	Electron	Irkutsk	RU	OJSC MegaFon	2900	false	643
202509881	true	546955	SBERBANK OF R...	MasterCard Standard ...	Saint Petersburg	RU	OJSC MegaFon	5000	false	643
200340445	true	527594	Citibank	Standard MasterCard ...	Saint Petersburg	RU	OJSC MegaFon	929	false	643
198825345	true	453450	missing	missing	missing	RU	OJSC MegaFon	4000	true	392
198979814	true	527594	Citibank	Standard MasterCard ...	missing	RU	Continental Ltd.	500	false	643
199025502	true	527594	Citibank	Standard MasterCard ...	missing	RU	Continental Ltd.	300	false	643

©2017 RapidMiner, Inc. All rights reserved.

Traditional validation method

Splitting dataset from the past into training and test subsets: estimate average performance on 20% holdout subset

	January	February	March	April	May	June	…	November
Training split	80% → model training and validation							<we are here>
Testing split (holdout)	20% → evaluating performance							

©2017 RapidMiner, Inc. All rights reserved.

More practical way to test

”What if we already deployed the model few months back” approach: estimate performance on whole month next to training period

	January	February	March	April	May	June	…	November
Training data	**100%** → model training and validation							<we are here>
Testing data (holdout)						**100%** → testing		

©2017 RapidMiner, Inc. All rights reserved.

Model performance in numbers

Looks pretty good… but how well it is aligned with business objectives?

accuracy: 99.58%

	true true	true false	class precision
pred. true	27	4	87.10%
pred. false	9	3073	99.71%
class recall	75.00%	99.87%	

©2017 RapidMiner, Inc. All rights reserved.

A little trick

Sum of transaction is exactly the financial metric to consider, so for testing data we **convert sums to example weights**

ExampleSet (3113 examples, 6 special attributes, 8 regular attributes)

Filter (3,113 / 3,113 examples): all

sum	card_bin	issuer_bank	card_type	user_city	user_country	user_ISP	country_co...	country_mi...
2900	427681	SBERBANK OF RUSSIA	Electron	Irkutsk	RU	OJSC MegaFon	643	false
350	533736	MTS BANK	Standard MasterCard ...	Boardman	US	Amazon.com, Inc.	643	true
5000	427640	SBERBANK OF RUSSIA	Visa Classic	Moscow	RU	OJSC MegaFon	643	false
3300	521172	Rosgosstrakh Bank	Standard MasterCard ...	Moscow	RU	OJSC MegaFon	643	false
800	476367	missing	missing	Kazan	RU	OJSC VimpelCom	826	true
485	440505	ROSBANK	Electron	Moscow	RU	OJSC MegaFon	643	false
430	427685	SBERBANK OF RUSSIA	Electron	Saint Petersburg	RU	OJSC MegaFon	643	false
500	557621	missing	missing	missing	RU	CJSC Comstar-Regions	840	true
1000	512107	missing	missing	Penza	RU	OJSC MegaFon	840	true
5100	490855	UNICREDIT BANK	Visa Classic	Saint Petersburg	RU	OJSC MegaFon	643	false
500	427638	SBERBANK OF RUSSIA	Visa Classic	missing	missing	missing	643	true

©2017 RapidMiner, Inc. All rights reserved.

Model performance in money - I

Now we have the same performance but mapped to real sums involved:

accuracy: 97.86%

	true true	true false	class precision
pred. true	40392	21010	65.78%
pred. false	19425	1805446	98.94%
class recall	67.53%	98.85%	

But in reality, is it actually like that –
$ 61,402 stopped vs.
$ 1,805,446 let thru as good transactions vs.
$ 21,010 lost due to undetected fraud?

©2017 RapidMiner, Inc. All rights reserved.

More details on business model

- **How much do you actually profit?**
  If you sell for $100, you put into your pocket only $10
  Hence, **profit margin** = 10% of revenue (turnover)
  (the rest is cost of goods sold, ad budget, payroll, etc)

- **How much do you actually lose due to fraud?**
  As a merchant, you are 100% liable for fraudulent transactions
  If fraud = $100, you lose **minimum** $100 + cost of goods = $140

- **Let's finaly assume**:
  Average cost of goods = 40% → you lose **140%** from fraud sum
  Profit margin = 10% → you net earn **10%** of total revenue

©2017 RapidMiner, Inc. All rights reserved.

Adjustments to be made

accuracy: 97.86%

x 140%

x 10%

	true true	true false	class precision
pred. true	40392	21010	65.78%
pred. false	19425	1805446	98.94%
class recall	67.53%	98.85%	

©2017 RapidMiner, Inc. All rights reserved.

Model performance in money - II

Total revenue for test period: $ 266,411

accuracy: 90.19%

	true true	true false	class precision
pred. true	57947	328	99.44%
pred. false	25796	182340	87.61%
class recall	69.20%	99.82%	

Which means that our model…

- Luckily saved us **$ 57,947** of detected fraud
- Unfortunately didn't detect **$ 25,796** of real fraud value (which we lost)
- Also lost **$ 328** for false positives
- Resulted in net profit margin **$ 182,340**

©2017 RapidMiner, Inc. All rights reserved.

What if there were no model at all?

No model = default model which always predicts ‘no fraud’

accuracy: 68.57%

	true true	true false	class precision
pred. true	0	0	0.00%
pred. false	83743	182668	68.57%
class recall	0.00%	100.00%	

- Net profit margin = **$ 182,668**
- Fraud loss = **$ 83,743**

©2017 RapidMiner, Inc. All rights reserved.

Final comparison

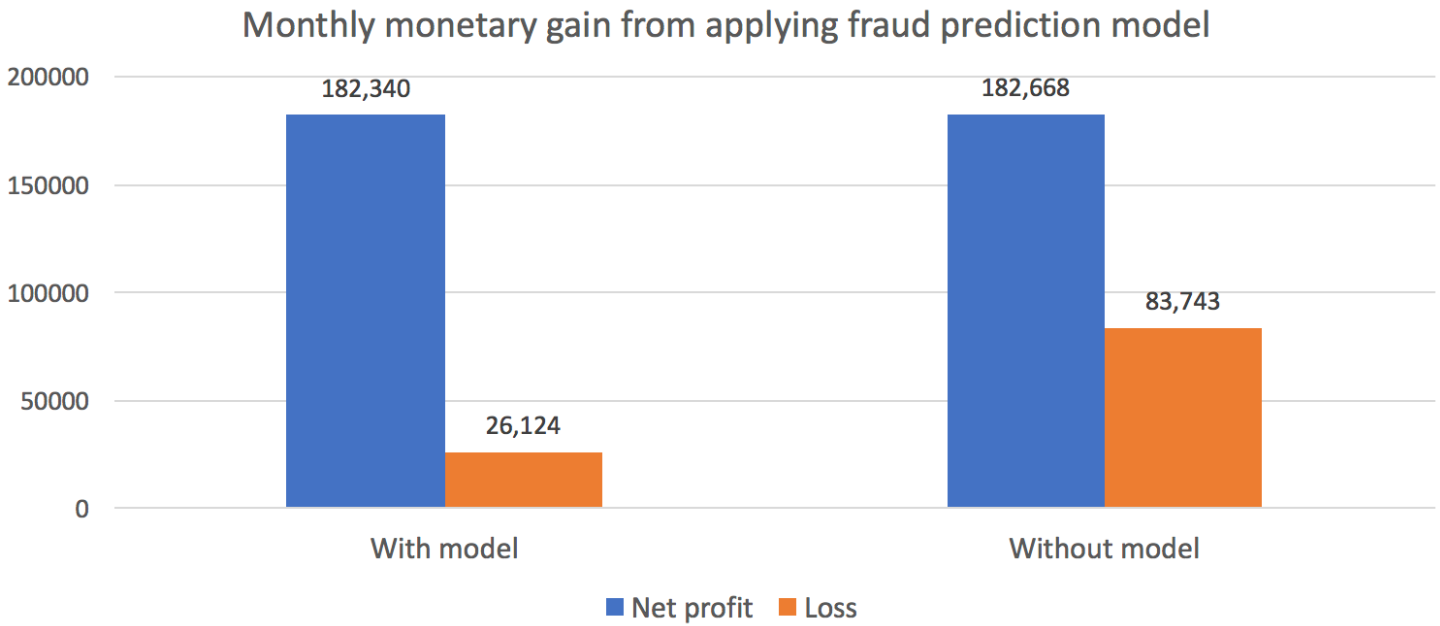
Total revenue for test period: **$ 266,411**

	With model	Without model
Net profit	182 340	182 668
Loss	26 124	83 743

Applying data science methods saves us **$ 57,619 per month**, which makes on average **$ 690,000 per year**

©2017 RapidMiner, Inc. All rights reserved.

Money plot



Applying data science methods saves us **$ 57,619 per month**, which makes on average **$ 690,000 per year**

©2017 RapidMiner, Inc. All rights reserved.

Q & A

Vladimir Mikhnovich
mikhnovich@gmail.com

Ryan Johansen
rjohansen@rapidminer.com