Feature engineering

is Your ticket to survival in Analytics

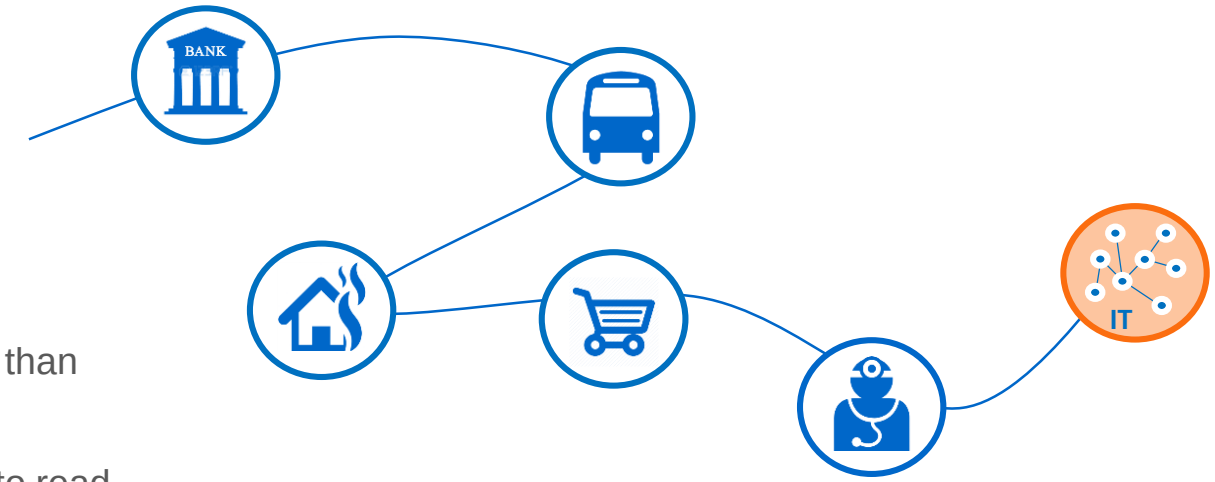
@FilipVitek, Director Data Science

Who the hell is Filip Vitek ?

Mr. Filip Vítek

15 years building business strategies, Data Science, CRM systems development and BigData projects

Built **analytical units** in **6 different industries,** now working for **Teamviewer (IT) :**



MORE INFO

Data mining is my hobby and passion, wrote more than

200+ expert blogs

If no time to go into details, I will leave you an link to read further on given topic.

Mighty data
www.TheMightyData.com

TeamViewer

TeamViewer | ITbrain | TeamViewer IoT | Blizz

1 800 000 000 devices

+ 200 000 MB of usage data **per day**

252 countries

40+ TB training set for **regression**

TECH STACK

Flume | kafka | amazon REDSHIFT | Microsoft SQL Server | Amazon S3 | AVRO | Parquet | Scala | python | tableau

2 reasons why Feature engineering is king

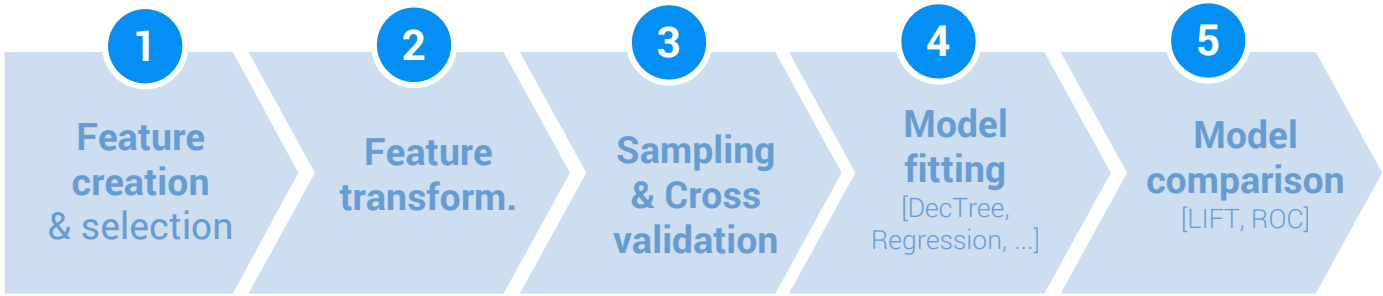
kaggle

XGBoost

winning many of the competitions on kaggle

Algorithms are becoming commodities, you barely can beat others by "better algorithm"

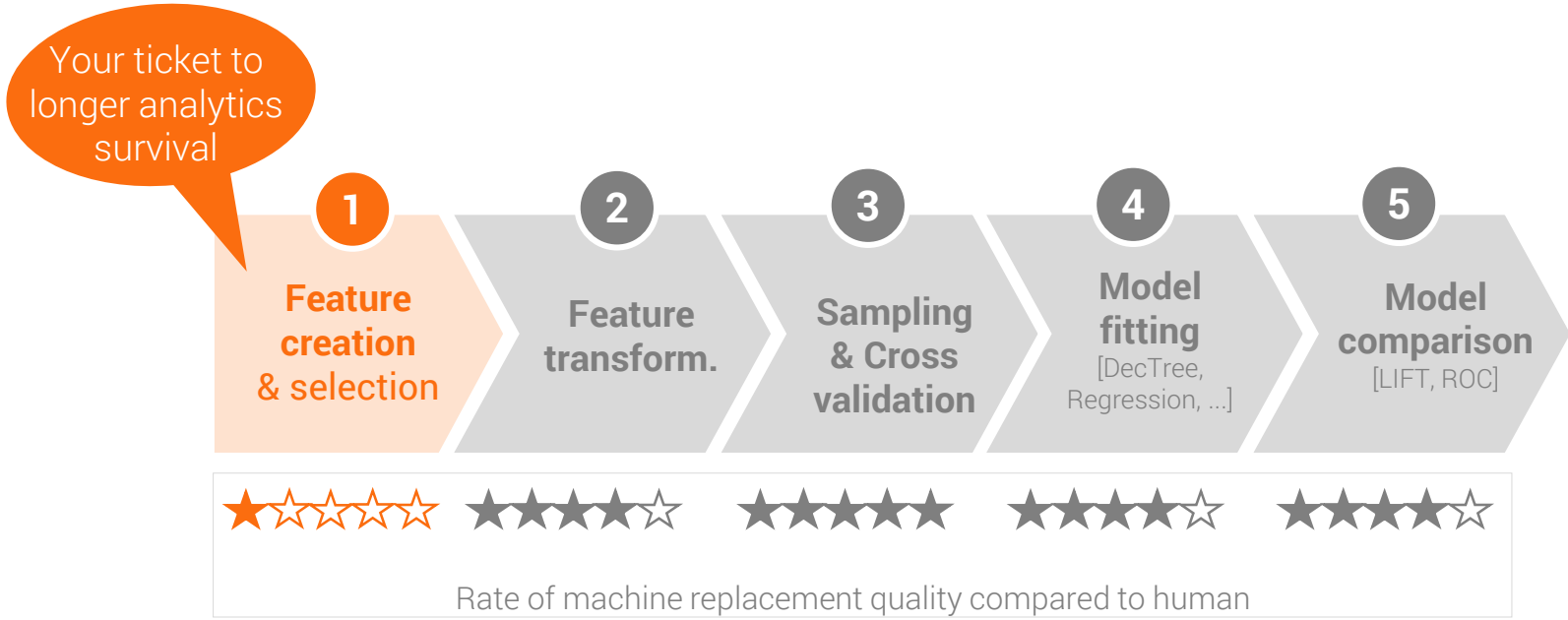
Think of last time you were designing the Machine learning model. In which of the following steps **did you rely on some of the ready-made packages?** (e.g. SciKit Learn, ...)



MORE INFO

"What alternative to AI do we, humans, have?"
http://mocnedata.sk/en/what-alternative-do-we-have-to-ai/

Feature engineering is your ticket to longer survival in analytics



How many features **is enough?**

Avoidable mistakes in traditional feature engineering

Unconventional approach to generating features

Taylor Polynomial …

How many features is enough?

$$T_n f(x;a) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$$
$$= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Same principle	Example: **Autonomous car**	Example: **Client churn probability**
0] **the variable itself**	What is our actual speed?	What was his service usage lately?
1] **relative change** (in time)	Are we already breaking or do we accelerate?	Was that regular miss or rather surprising one given the history?
2] **changes tendencies**	With full push to brakes how strong negative acceleration can we still achieve?	Does this behavior overachieves tempo of churn or it is serving as slow-down of it?

*CAUTION! It is second partial derivative so it needs not to be dFx * dFx, but can be dFx*dFy as well.*

MORE INFO *Which brand is most advanced in autonomous car research*
http://mocnedata.sk/en/autonomous-cars/

Common FE approaches in business (and their pitfalls)

Avoidable mistakes in traditional feature engineering

Demographics

- Client gender
- Client age
- Family status/size
- Country of seat
- Geographical region
- Income group
- …

Time based

- Time since 1st transaction
- Frequency of purchase
- Time since most recent transaction made
- Usual day/time of purchase
- …

Client behavior

- Most common way of payment
- Usual delivery method
- Return rate of goods
- Satisfaction / NPS
- Number of repeating's of the same behavior…

Outliers / Specials

- Entry/First product bought
- Max amount paid
- Longest pause between two purchases
- Most often bought unit/category
- …

STOP

When modelling client behavior, Demographics holds NO behavior, just proxies to it

Often selected as , default features, no matter what

Yes, data probably has some time seasonality in it, but you have detect it, not just assume it (last 12M)

Remember at least two degrees of Taylor polynomial

Do not look just for Boolean features about behavior, rather use how many times pattern repeated

Extra caution with highly correlated behaviors (we only do things too similarly, if they are part of the same procedure)

Every feature has its outliers, do not ignore them [max ATM withdrawl]

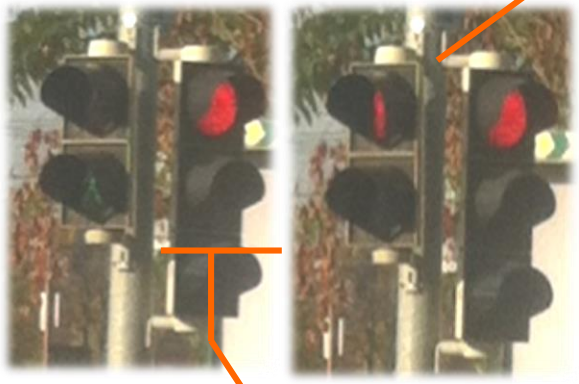
MORE INFO

Unusual features in sport analytics

http://mocnedata.sk/en/what-peter-sagan-learned-about-his-sprint-from-helicopter/

Data underdogs … and their impact

Who will win the car race to nearest lights?



Has originally **other informational role**

Indicates **client behavior**
[or its change]

- ☐ Data fields that are "just identifiers"
- ☐ Contact & Transactional data
- ☐ No obvious relations as champion challengers (Joker cards)
- ☐ Unusual aspect of usage
- ☐ "Ryanair-like" data test

Social impact on other clients in portoflio

Jane.Angry@teamviewer.com
Martin.Neutral@teamviewer.com

1

2

Tone of voice

John.Warton @hotmail.com

Johny_geek @hotmail.com

Bank preference

(Online bank vs. Postal bank)

3

Relationship proxy

4

(133333333 /xxxx
133353333 /xxxx)

Unconventional approach to generating features

MORE INFO

Apologize to your database
https://blog.etrend.sk/filip-vitek/data-underdogs.html

Real examples of Unconventional Feature Generation

Unconventional approach to generating features

How old are you, **Bernard?**

- Nothing like "National ID" for German insurance companies = they have no clue about age of customer
- Important for setting proper communication (web vs. call vs. paper letter)

- First name + Region predicting 92% accurately the decade when the customer was born

[cut/off point for approx. 25% Individuals]

Fee increase tolerance

- Fee increase sensitivity for retail bank
- In search for metric that would tell: How "lazy" user is?

- Limited space, banking feels very un-emotional
- Lowest amount ever withdrawn from the ATM

[worked surprisingly well, due to large coverage]

Detecting **commercial** customer

- Quite a few small companies without license
- Too small to detect via IP address range
- Using standard desktop OS versions

- Pattern of use strong within working hours, weak outside

[nightmare of time zones from UTC]

Zodiac, are you kidding me?

- Probability to have car accident
- As Joker card for model
- Strong objection from Data Scientists: "*This is not serious work, we protest.*"
- Ended up as the Second strongest parameter in model.
- Later confirmed in 4 other countries in same issue

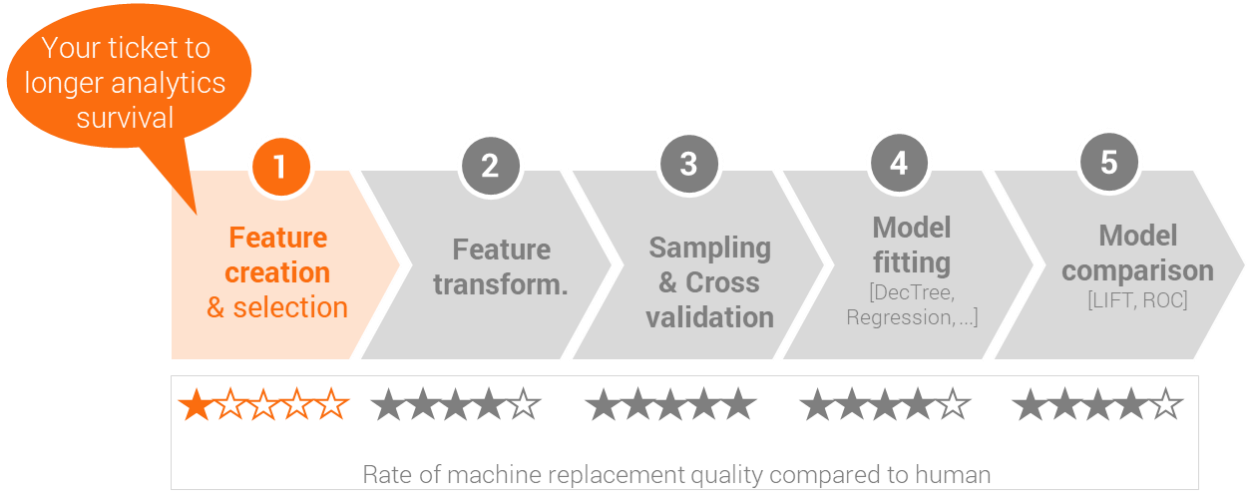
[I have a hypothesis why it works]

MORE INFO

Find more of interesting feature examples here:
http://mocnedata.sk/en/what-to-read-in-english-here/

Feature engineering is your ticket to longer survival in analytics

X



… but also depends on what LETTER data analyst you are …

V - I - B - A

… the MBTI of the analytics. Find out which type of Data Scientist YOU are:

http://mocnedata.sk/en/**VIBA-type-of-analyst/**

Thanks for Your attention and I am ready to answer

YOUR QUESTIONS

Mighty data

www.TheMightyData.com

Join the community

http://mocnedata.sk/en/lets-keep-in-touch/

Feel free to contact me ...

Mgr. Filip Vítek

Data Science Director

TeamViewer, Berlin

+ 49 1525 309 8505

filip.vitek@teamviewer.com

https://sk.linkedin.com/in/vitekfilip

@FilipVitek

Data are fuel of the New Economy. Or is it?

1995	2004	2012	2016
			

Is Principal Component Analysis your Friend or Foe?

PCA

- Pure **Machine to Machine** interface
- Data-space **visualization** required
- Overcomes **mutual correlations** of features without even explicitly checking for them

- **Feature selection** procedure [even in SciKit Learn]
- **Humans using** the result of predictions
- **Had to do** oversampling in process of the model preparation
- **Neural network** one of the rival models
- **Non linear** effects of the variables