



Feature engineering is your ticket to survival

@FilipVitek, Director Data Science

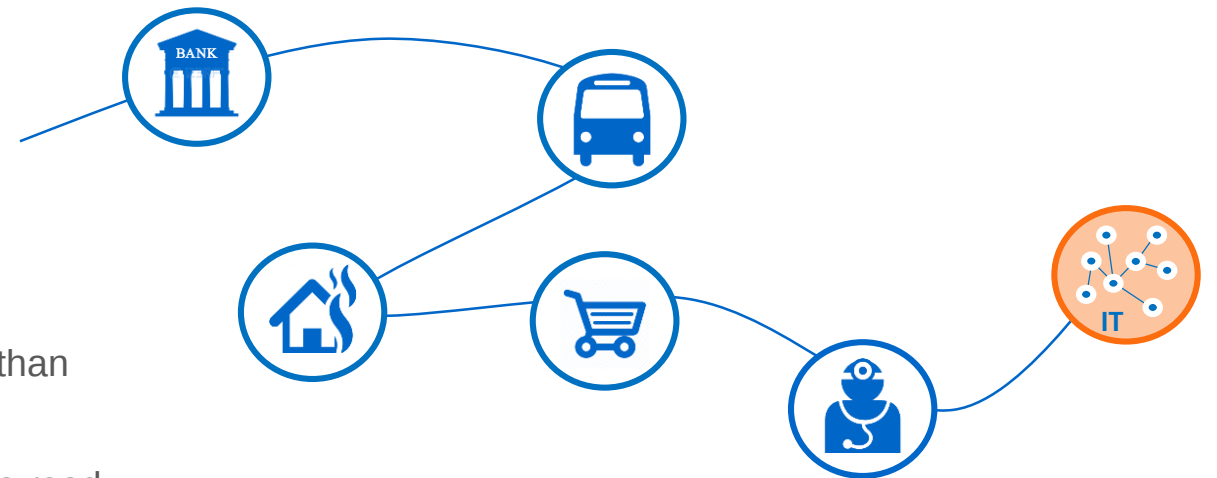
Who the hell is Filip Vitek ?



Mr. Filip Vitek

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

Built **analytical units** in **5 different industries**, now working for **Teamviewer (IT)** :

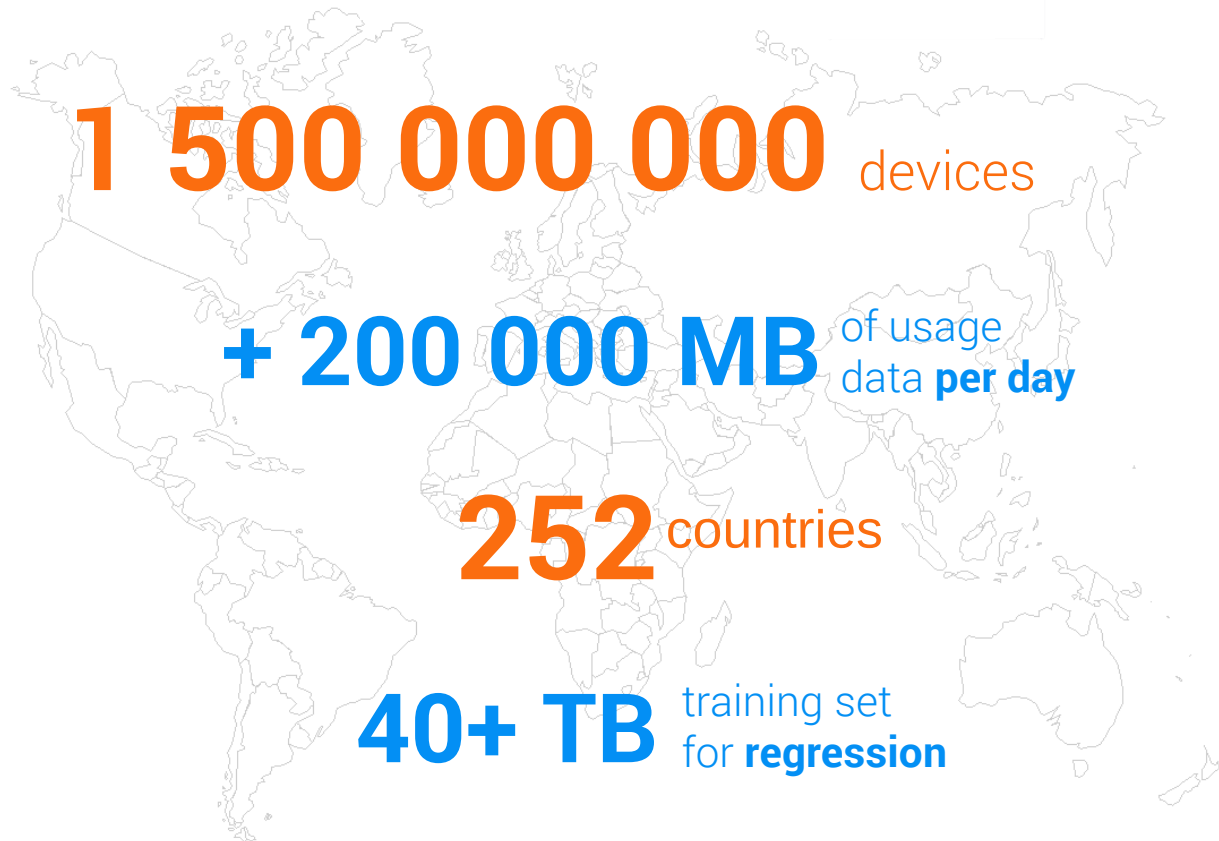


Data mining is my hobby and passion, wrote more than

200+ expert blogs

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





TECH STACK



2 reasons why Feature engineering is king

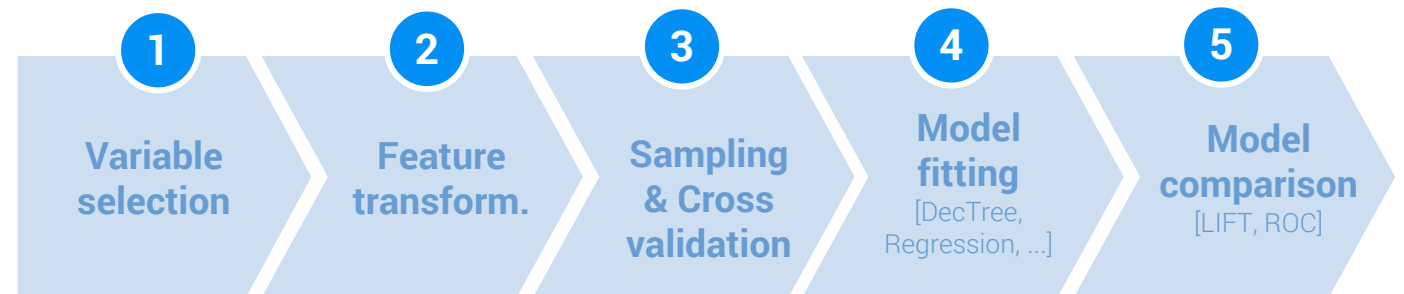
kaggle™

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, ...)



Most common FE approaches in business (coming soon)

Demographics

Time based

Client behavior

Outliers / Specials

STOP

■ ???

■ ???

■ ???

■ ???



Taylor Polynomial ... as inspiration for feature creating

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

Same principle

Example: Autonomous car

0] the variable itself

What is our actual speed?

1] relative change (in time)

Are we already breaking or do we accelerate?

2] changes tendencies

With full push to brakes how strong negative acceleration can we still achieve?

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

Common FE approaches in business (and their pitfalls)

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** [min ATM withdrawal]



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]

Social impact on other clients in portfolio

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



Tone of voice

John.Warton@hotmail.com

Johny_geek@hotmail.com

Bank preference
(Online bank vs. Postal bank)

Relationship proxy

(133333333 /xxxx
133353333 /xxxx)



Real examples of Unconventional Feature Generation



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 of approx. 20% Individuals]

Estimating age of car

- Modelling risk profile of the cars
- For many cars the age of the car was unknown in data
- Market relatively stable in past
- **Duration of the contract** with our company & age of owner

[even if a miss, still positively correlated]



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 Scientist "*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]



Thanks for your attention and I am ready NOW to answer
YOUR QUESTIONS



If you
have Q's
later ...

Mgr. Filip Vitek
Data Science Director
TeamViewer, Berlin

filip.vitek@teamviewer.com



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

@FilipVitek

