



TeamViewer

How To Do Cool Feature Engineering In Python

@FilipVitek, Director Data Science

Who the hell is Filip Vitek ?



Mr. Filip Vitek

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



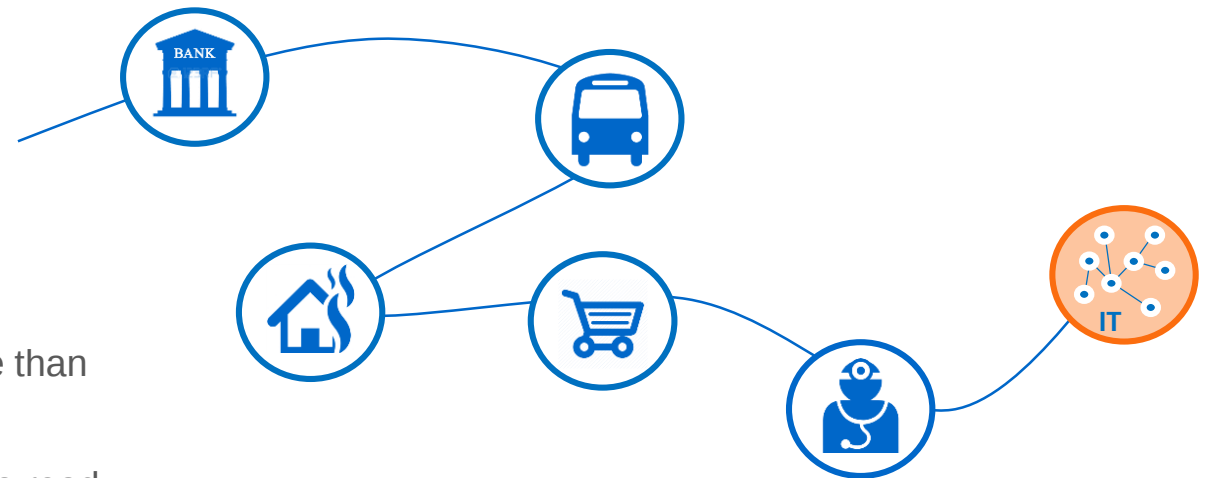
Data mining is my hobby and passion, I wrote more than

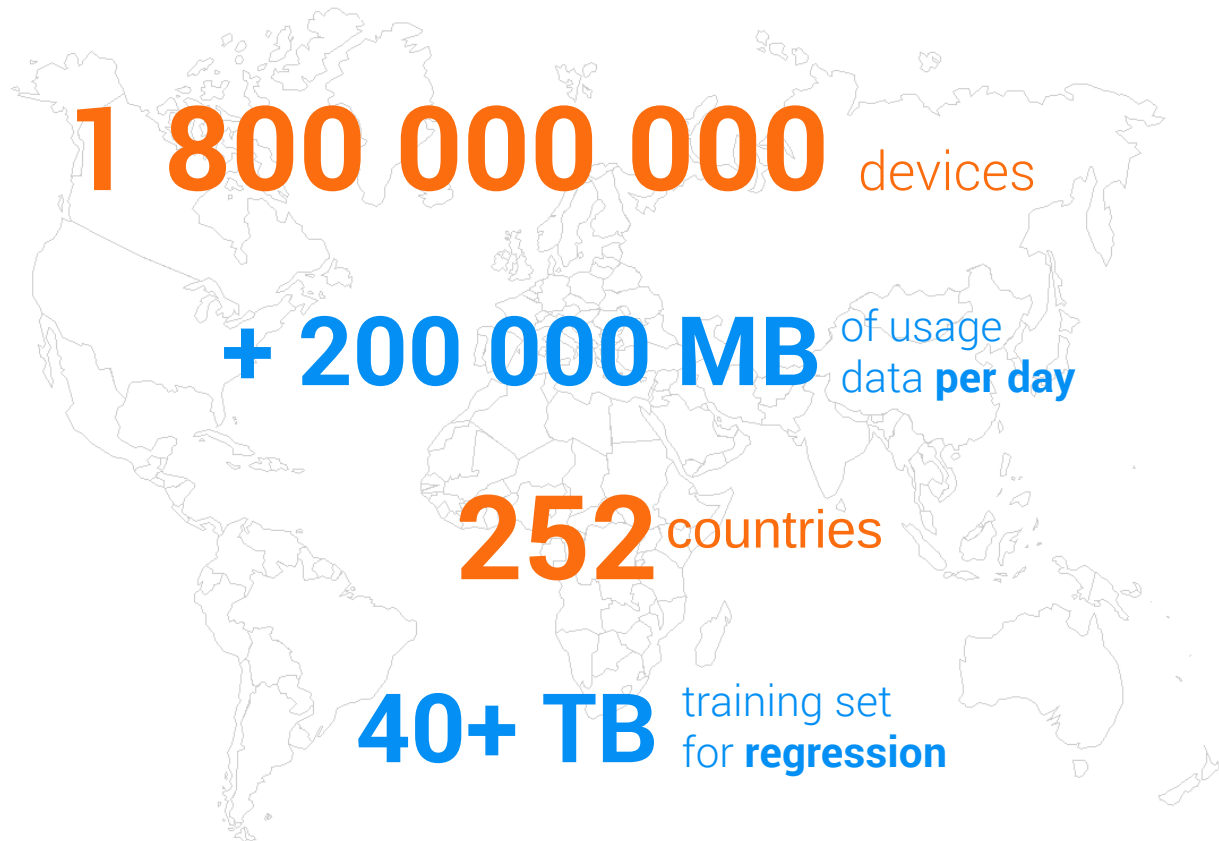
300+ expert blogs

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



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





TECH STACK



2 reasons why Feature engineering is king

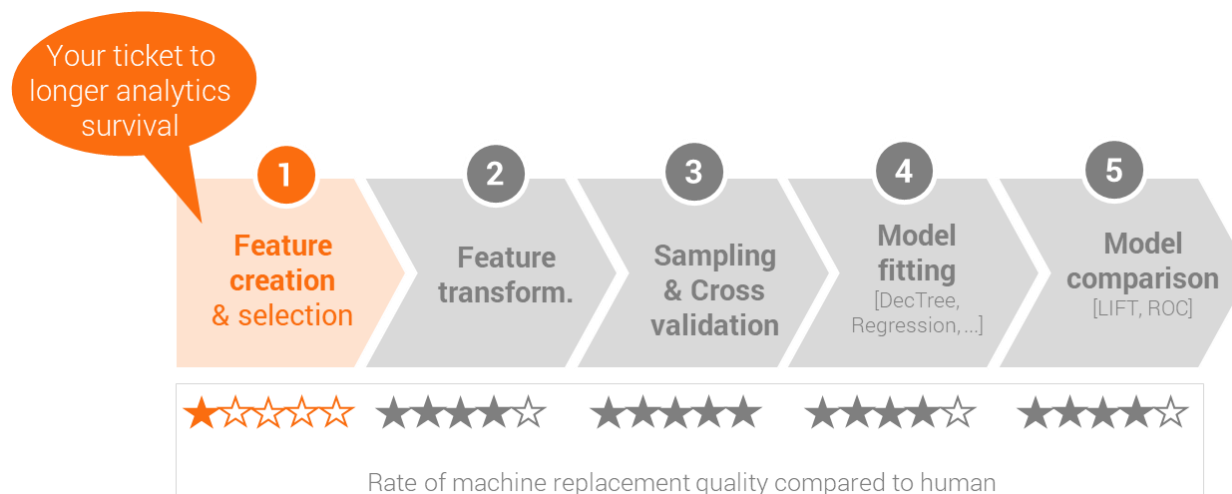
kaggle™

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, ...)

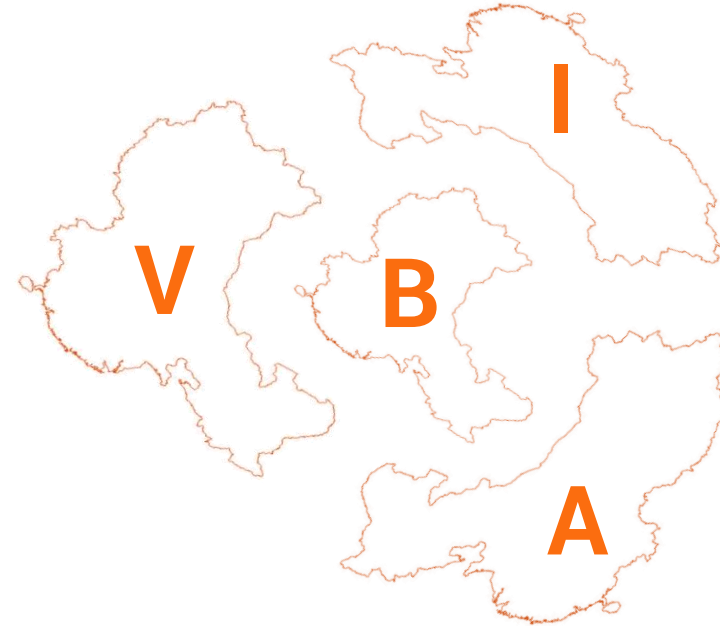


... but you need to be true to your self on WHO YOU ARE

Analytics as **Single continent**



Analytics **Archipelago**



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/>



What do I expect ... of COOL Feature Engineering

A Extending set of ingredients

(= insights available)



B Sorting out the hopeless cases



C Prune to have more efficient model training & operation



D



Signal, if I missed anything relevant



Python is ... the Microsoft Excel™ of our era



It Became standard. There are options, but why to bother to even try. (Think Lotus 1-2-3)

Everybody claims knowledge of it, but knowledge of most people is **very shallow**.

(Frustrating to test Data Science hires for basic Python ... and see them struggle with RegEx)

Tool is really powerful, but you need to possess certain skills beyond elementary use.

To rely hone the power of it, you need to know more than the default options/libraries.
(... which most people don't)



$=1*(0.5-0.4-0.1)$

VLOOKUP blinds
(back search, MIX, Case Sensitive)

Z-score glitch

"Don't CHALLENGE or REVIEW, just CONSUME."

(Have you ever checked what Excel calculates? / Have you challenged any Sci-kit learn routines?)



SciKit Learn ... Our U-bahn of the Machine Learning (?!)

✓ Supervised Learning

(GLM, LinDiscAnal, KernelRidge, SVM, StochGradDescent, NearestN, NaiveBayes, DecisionTrees, **Feature selection**, Ensemble methods, MulticlassAlgor, Isotonic Regress., Prob. Calibration, NeuralNetworks, ...)

✓ Unsupervised Learning

(GaussianMixture, ManifoldLearning, Clustering, Biclustering, MatrixFactorization, Covariance estimation, OutlierDetect, DensityEstimate, NeuralNet, ...)

✓ Model selection

(CrossValid, HyperParameters, Model evaluation, Model persistence, Validation curves.)

✓ Dataset Transformations

(Pipelines, Feature extraction, Preprocessing, Impute, DimensionReduction, Projections, KernelApprox, PairWiseMetrics, TargetTransform)

✓ Datasets, Loading & Scaling

(Toy/Real datasets, Generated datasets, Loading, Incremental learning, PredictionThrouput, Parallelism)



Feature selection tools

- Low Variance Removal
- Univariate feature selection
(Select K-Best)
- Recursive Feature Elimination
(only backwards)
- SelectFrom Model (Tree)
- Including into Pipeline

- Principal Component Analysis
- Independent Component Analysis



How does SciKit Learn ... meet our expectations?

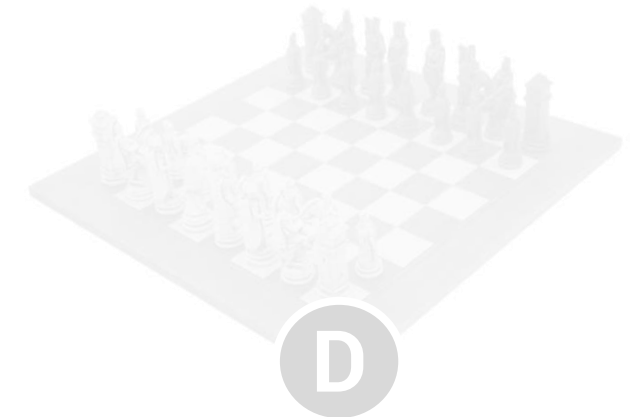
A Extending set of ingredients
(= insights available)



B Sorting out the hopeless cases



C Prune to have more efficient
model training & operation



D Signal, if I missed
anything relevant

How to compensate for that in Python space ...

 Pre-cooked Lambda on
always done transformations



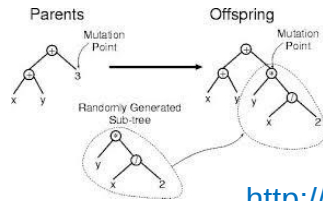
<https://www.featuretools.com/>

tsfresh

<https://github.com/blue-yonder/tsfresh>



<https://epistasislab.github.io/tpot/>



Genetic
Programming

<http://www.philipkalinda.com/ds8.html>

Go via
Deep
Learning

- Calculation costs
- Team skills
- Time-to-market
- Explain-ability

Build your **OWN FEATURE** engine

- 1 **Calculate** variable statistics
[see also transformations slide, ...]
- 2 **Generate obvious suspects**
[aggregations, time windows, ...]
- 3 **Indicate** missing info categories
[compare to dictionary, Expl. score ...]
- 4 **Hard criteria** knock-out
[Variance, NonNulls, distinct X, ...]
- 5 **Binning & Categorical** decomposition
[Forced binning if > N]
- 6 **Univariate** correlation & Log -P
[Simple tree is enough]
- 7 **Bivariate** relations
[Cut off for Categorical dummies by Support]
- 8 **Decision on ranking** of parameters
[Simple, Stage based, ...]



Thanks for Your attention and I am ready to answer
YOUR QUESTIONS



www.TheMightyData.com

**Join the
community**



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

**Feel free
to contact
me ...**

Mgr. Filip Vitek

Data Science Director
TeamViewer, Berlin



+ 49 1525 309 8505

filip.vitek@teamviewer.com



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

@FilipVitek



BACK UP SLIDES



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]. **Reduction** ≠ **selection**
- **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



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]



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)

- ☐ 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

Variable Transformations ... simply & shortly

Variance X_i, Y
Kurtosis, Skewness
UNI Pearson correlation
5th /95th percentile

