Using Machine Learning to Explore our inventory in real time:

Process and Key Learnings.

Irati R. Saez de Urabain, PhD
July 16, 2018

What do we do at Dalia?

What do we do at Dalia?

Dalia enables **people all over the world** to **share their voice** through mobile surveys.

We deliver knowledge to decision makers in **business, politics & academia.**

Dalia's Audience Profiling Process

Millions of casual survey respondents are sourced through a network of over 40k apps and websites

Every user is dynamically profiled across key demographic and behavioral attributes

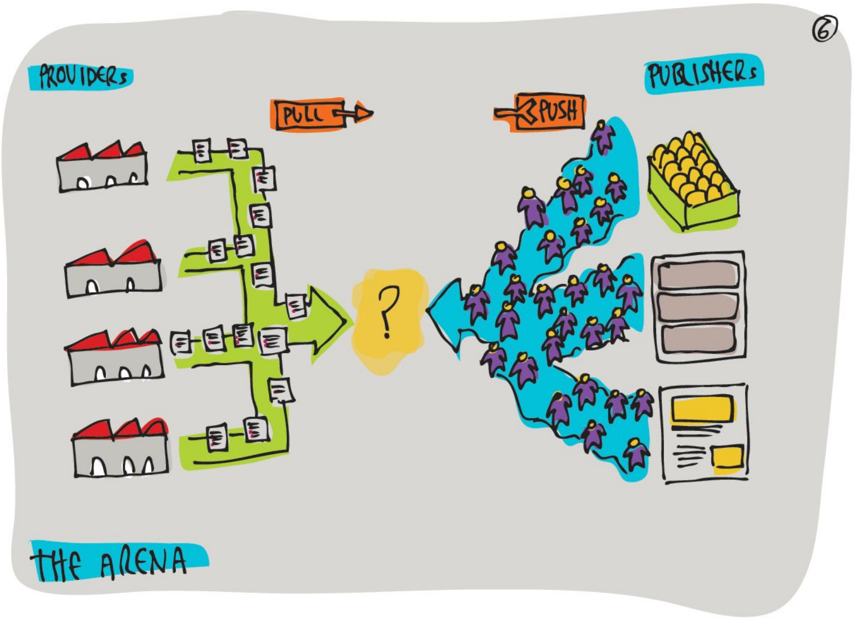
A self-learning quality assurance system, based on active & passive info, generates a unique trust score for each user

A targeted attribution matches high quality, verified users to the appropriate surveys

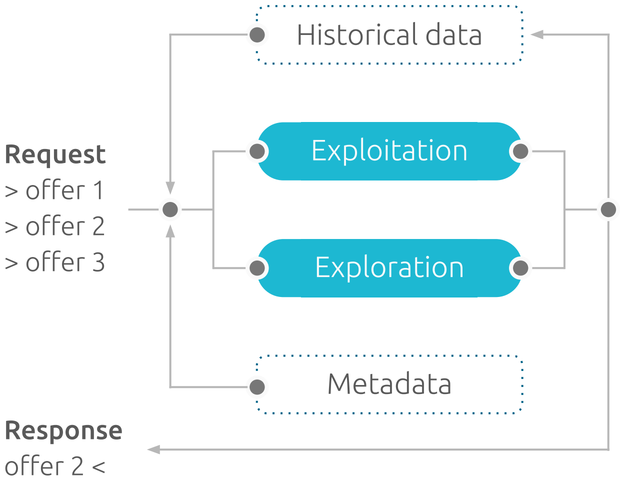
An instant reward (e.g. virtual currencies, prepaid credits, access to premium content) is awarded to users who complete their survey

Finding the right survey for the right user

What is the problem we are trying to solve?



The Explore - Exploit dilemma



Exploring the right surveys

We don’t want to waste traffic exploring bad surveys

We started to test very simple business ideas to explore new surveys:

- New surveys are better
- Short surveys are better
- Beta distribution

At some point, we decided to start experimenting with Machine Learning

Building the Exploration ML model

Real time service

- Receives user requests
- Loads the model and responses real time
- Caching with redis
- Fast (<100ms)

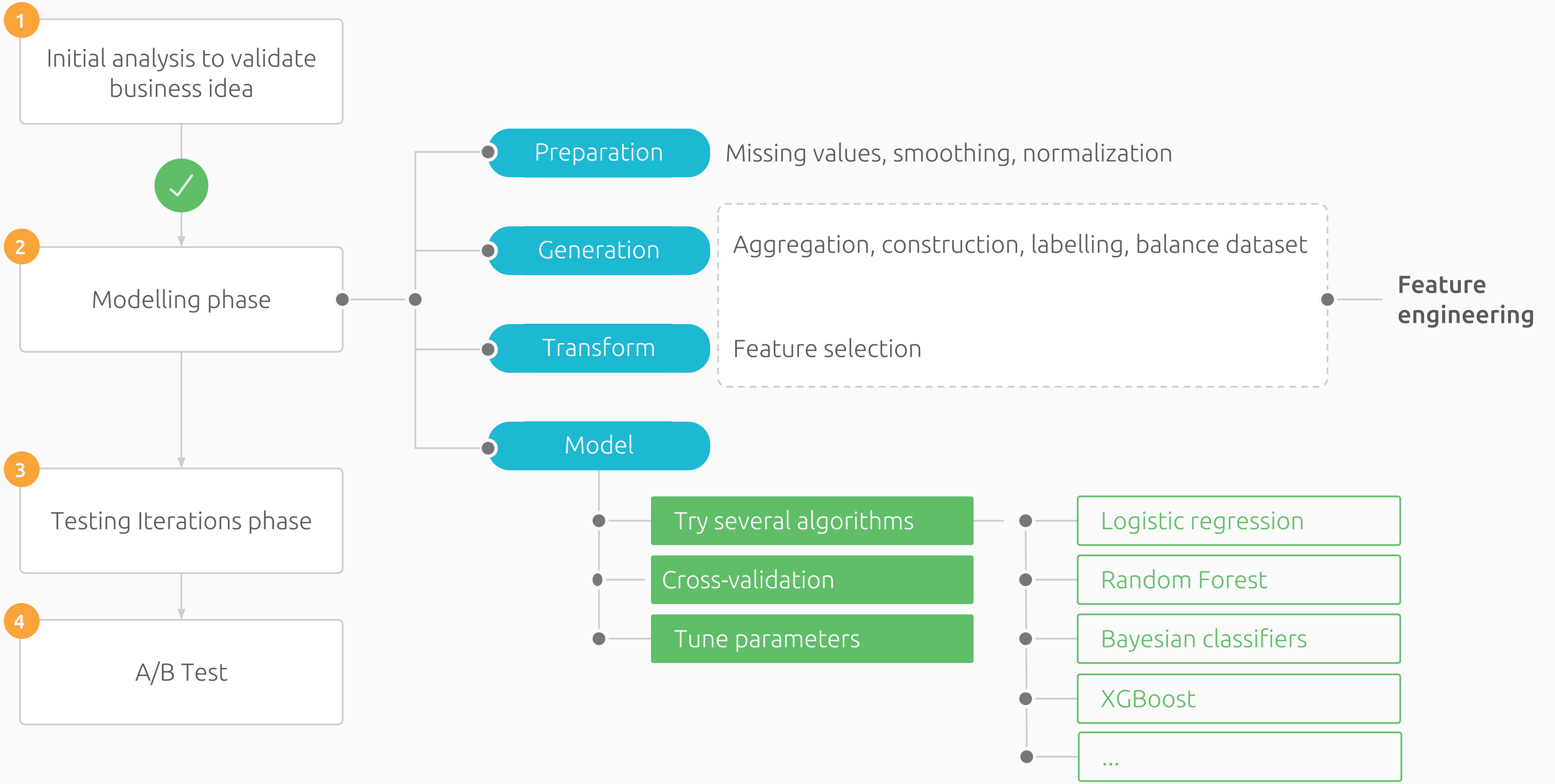
Backend service

- Classification algorithm
- Runs every week
- Updates the model, if the new model is good

MySQL mongoDB druid

redis

Process for building ML models in production



Modelling phase

Tools

python™

jupyter Notebook

scikits learn

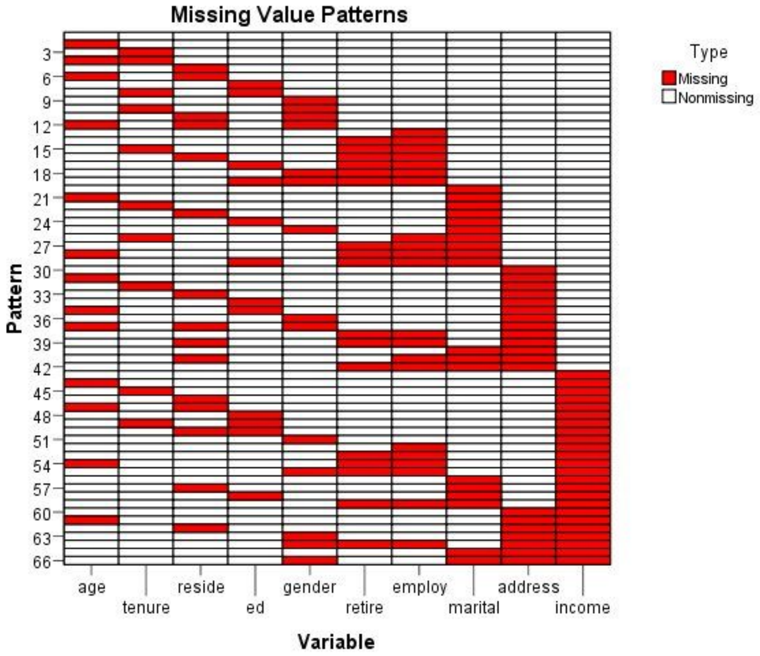
MySQL®

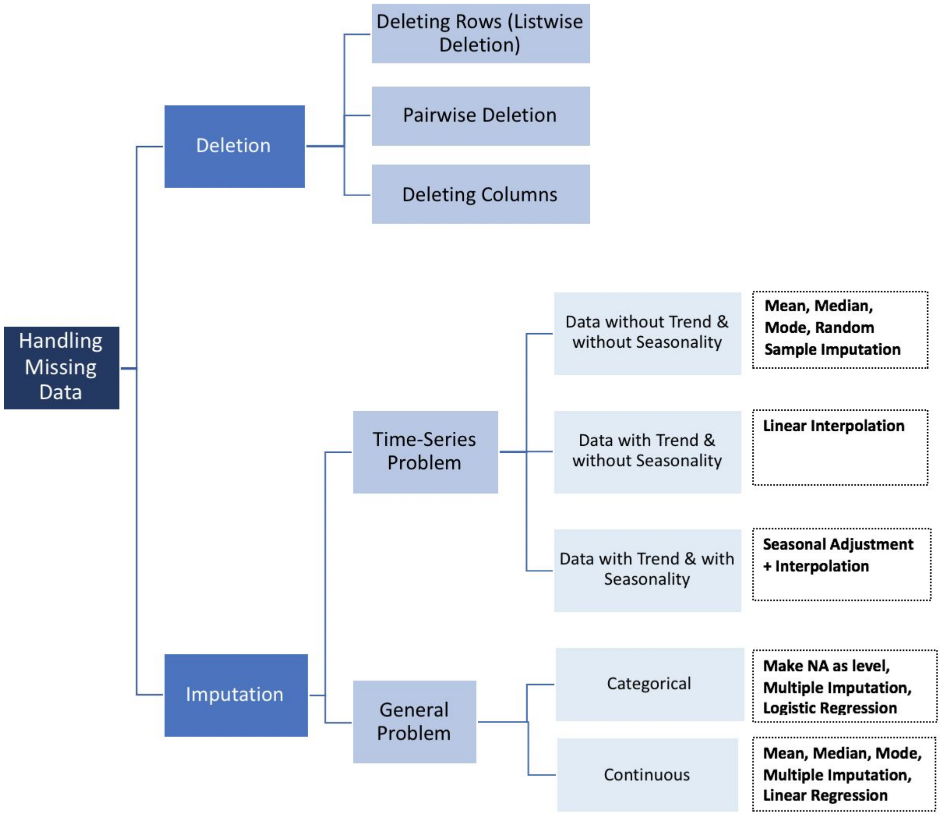
mongoDB

druid

Preparation

Missing values

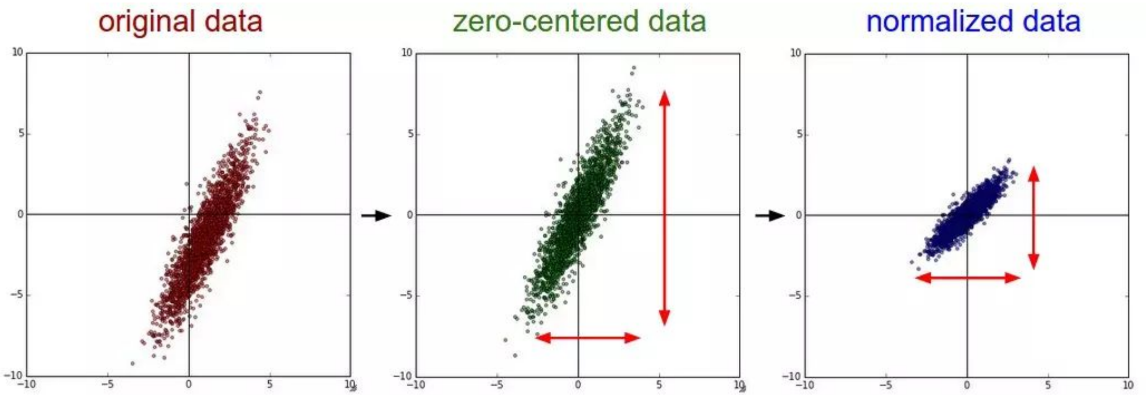




Learn more: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Preparation

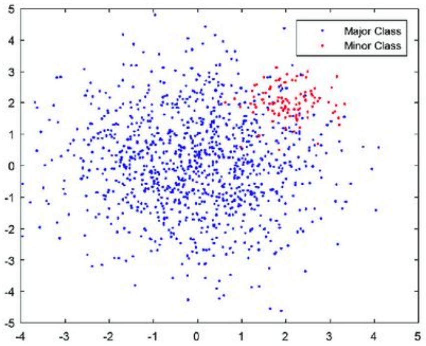
Normalization



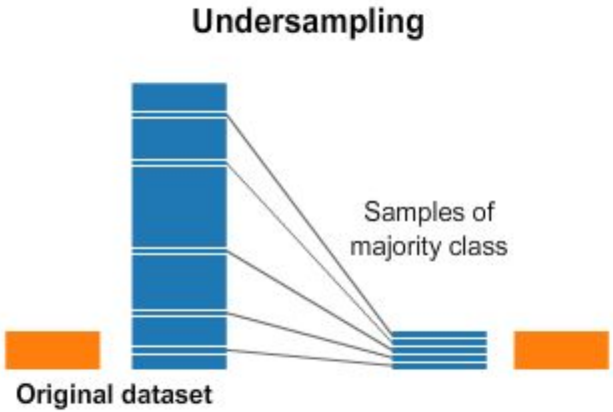
- For some algorithms normalization is very important (eg., Clustering)
- For others, we don’t need to normalize (eg., Decision Trees, Random Forest)

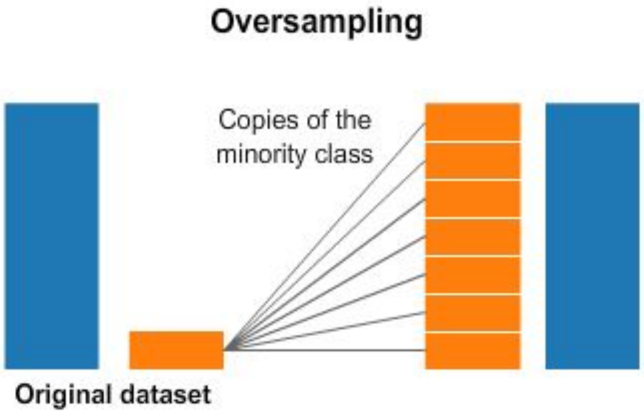
Generation - Feature engineering

How do we handle unbalanced datasets in a classification problem?



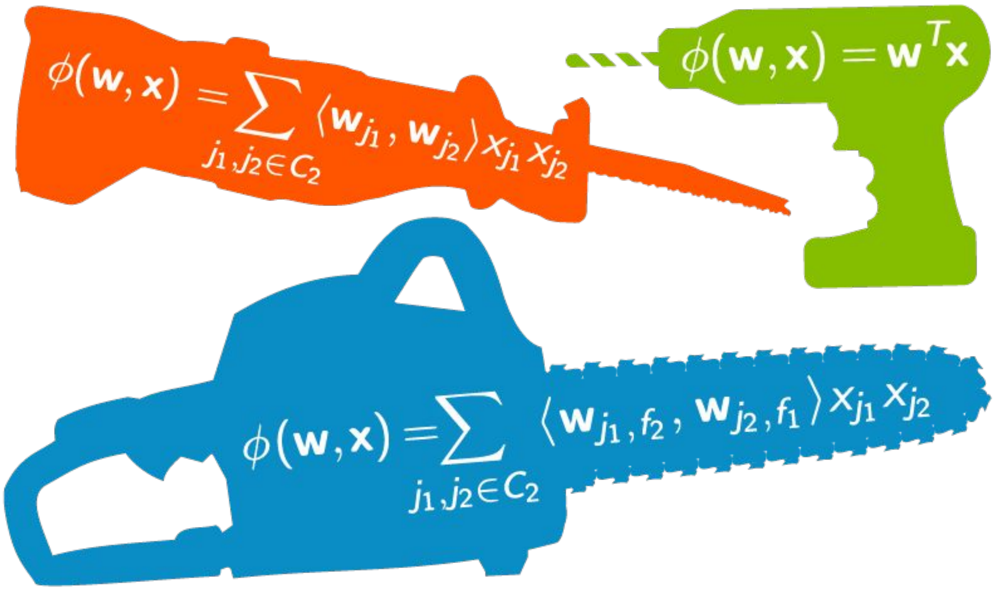
Info about different sampling methods:
http://contrib.scikit-learn.org/imbalanced-learn/stable/





Generation - Feature engineering

Aggregation, construction, labelling,.... Just get creative with your data!



Transform : Feature selection

Top reasons to use feature selection are:

- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It improves the accuracy of a model if the right subset is chosen.
- It reduces overfitting.

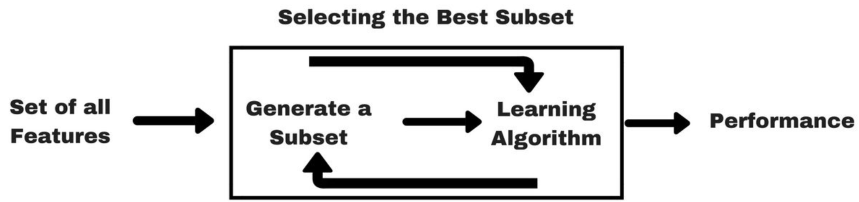
Filter methods

- Features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.

For more information:
https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

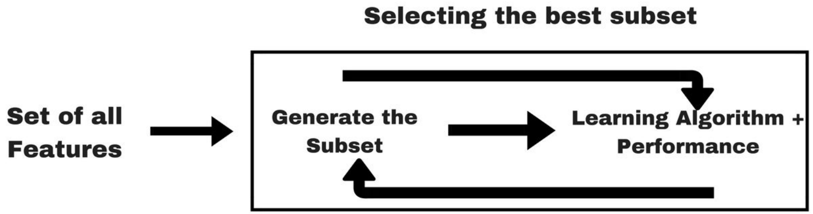
Transform : Feature selection

Wrapper methods



- We try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset.
- *Forward selection, backward selection, recursive feature elimination*
- Computationally expensive.

Embedded methods



- Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.
- *Lasso and Ridge regression*

Finding the right classification algorithm

- Logistic regression
- Random Forest
- Bayesian classifiers
- XGBoost
- ...

Grid Search

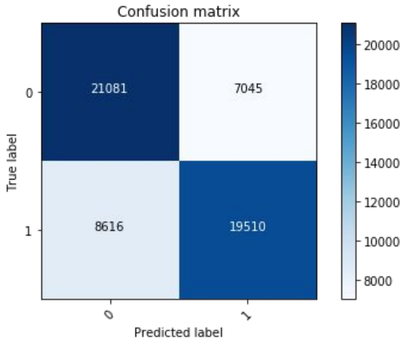
p2 \ p1			
p2	(1, 1e3)	(10, 1e3)	(100, 1e3)
	(1, 1e4)	(10, 1e4)	(100, 1e4)
	(1, 1e5)	(10, 1e5)	(100, 1e5)

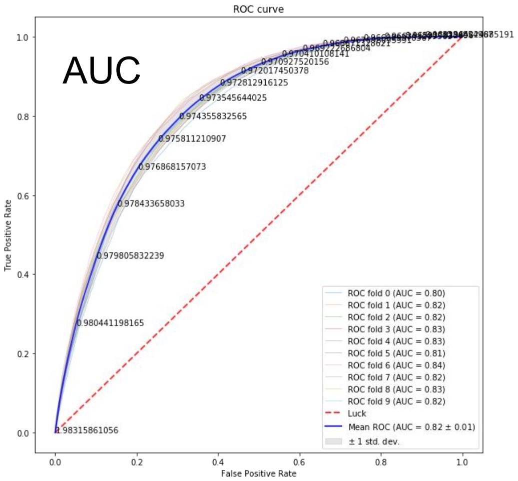
Cross Validation

	FOLD 1	FOLD 2	FOLD 3	FOLD 4	FOLD 5
ITERATION 1	TRAIN	TRAIN	TRAIN	TRAIN	TEST
ITERATION 2	TRAIN	TRAIN	TRAIN	TEST	TRAIN
ITERATION 3	TRAIN	TRAIN	TEST	TRAIN	TRAIN
ITERATION 4	TRAIN	TEST	TRAIN	TRAIN	TRAIN
ITERATION 5	TEST	TRAIN	TRAIN	TRAIN	TRAIN

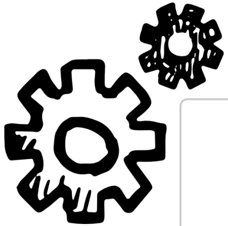
DATASET PARTITIONED INTO FOLDS

What metrics do we use to evaluate the algorithms' performance?

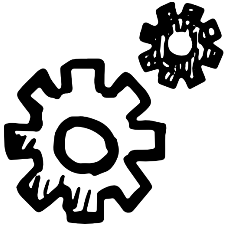




Sensitivity tests

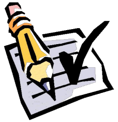


Machine Learning model



Machine Learning model **Training**

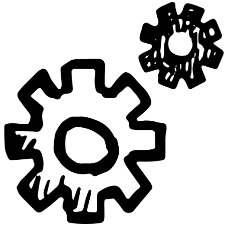
TESTS



Model goes to production!

Model is not updated in production

What did we end up doing?



RANDOM FOREST MODEL

Output

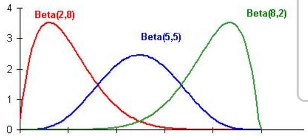
Converts (1) / Does not convert (0)

Input

SURVEY METADATA

- Length of interview
- Performance for past surveys for the same provider
- Provider CR
- Price for the survey
- Survey type
- ….

Probability to convert



BETA DISTRIBUTION

Input

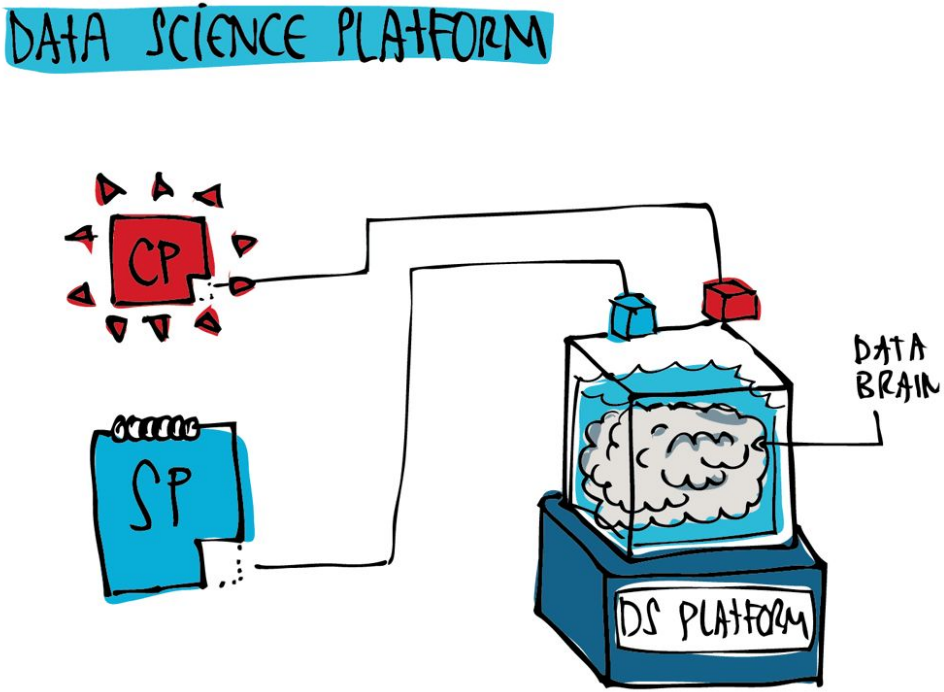
HISTORICAL DATA

- # conversions
- # non conversions

BETA (ML_prior + conversions, ML_prior + non_conversions)

Real Time service: The Data Science Platform

Real time service: The Data Science Platform

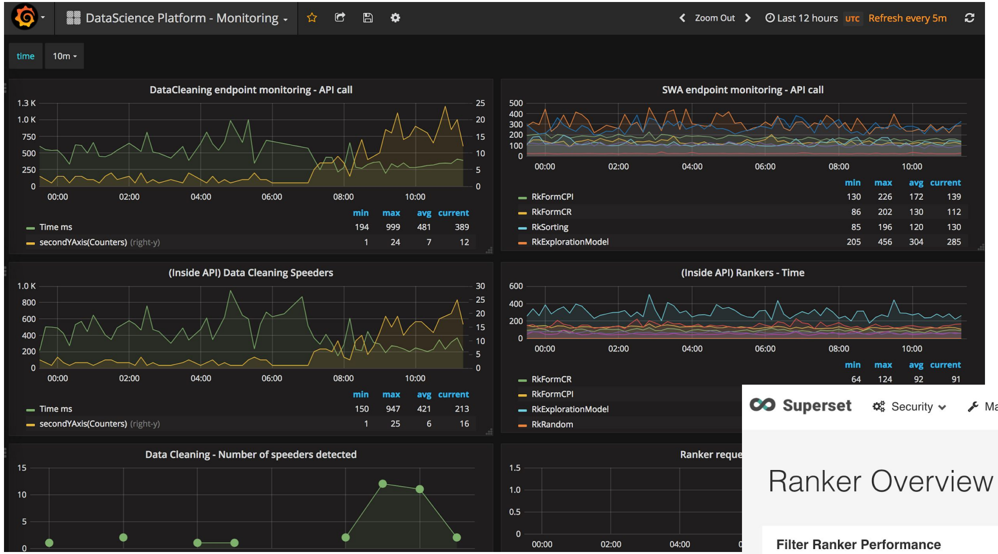


What is the Data Science Platform?

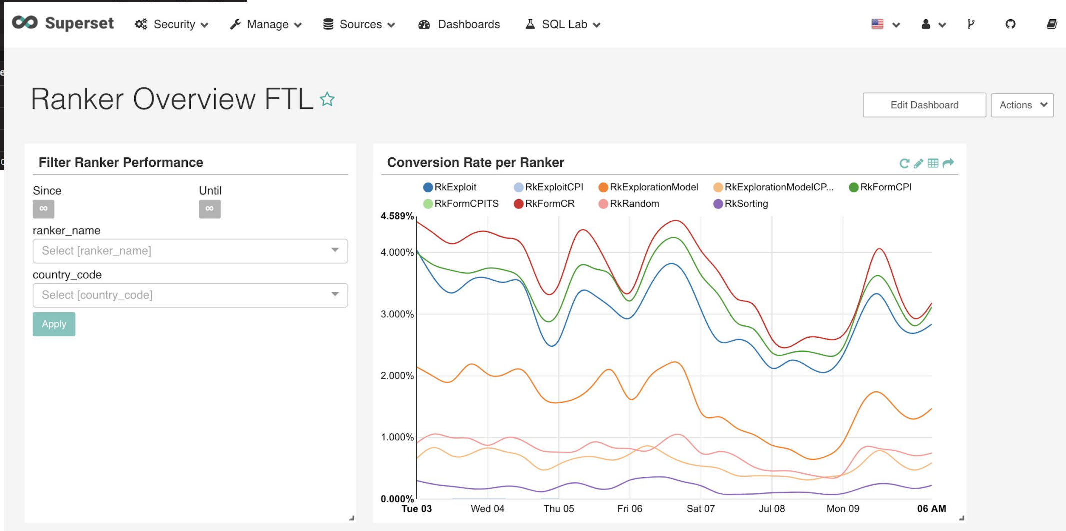
- Python REST API
- Data Science production code
- Different end points for each project
- Receives a request and returns a response using DS algorithms and Machine Learning
- Loads Models in memory to speed up response times

Performance monitoring

Performance Monitoring



Grafana

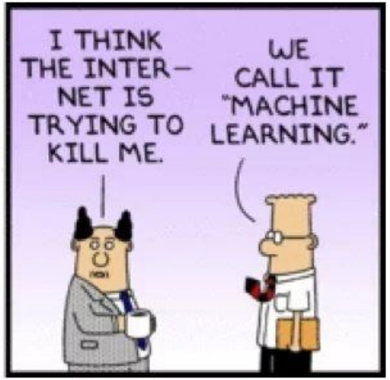


Superset

Key learnings

What did we learn?

- The whole process - from idea conception to model in production - …. TAKES SOME EFFORT!
- It is essential to understand the data and environment you are dealing with, before jumping into complex models
- Technical debt of ML can be a lot higher than for regular algorithm, plan ahead how you are planning to maintain and monitor your models!
- ML is not the magic tool that fits all. Evaluate whether this is really what you need.
- But… **if used correctly, it can be very effective!**



Any questions?

Thanks!

Kostas Christidis, PhD

Irati R. Saez de Urabain, PhD

Jakob Ludewig

Korbinian Oswald