knowledgePit
A Data Challenge Platform

Analysing competition results - Predicting Escalations in Customer Support Cases

Andrzej Janusz

University of Warsaw
and
QED Software

Warsaw, 23.10.2020

About me

- PhD in 2014 from University of Warsaw
- currently:
  - assistant professor at Institute of Computer Science, Univ. of Warsaw
  - Chief data scientist at QED Software
- co-founder of the KnowledgePit platform

The plan

- Introduction
  - a few words about Knowledge Pit
  - the BrightBox project
- IEEE BigData 2020 Cup
  - the competition task
  - summary of results
- Analysis of solutions from DM challenges
  - finding neighborhoods of test cases
  - diagnostic rules
- Conclusions

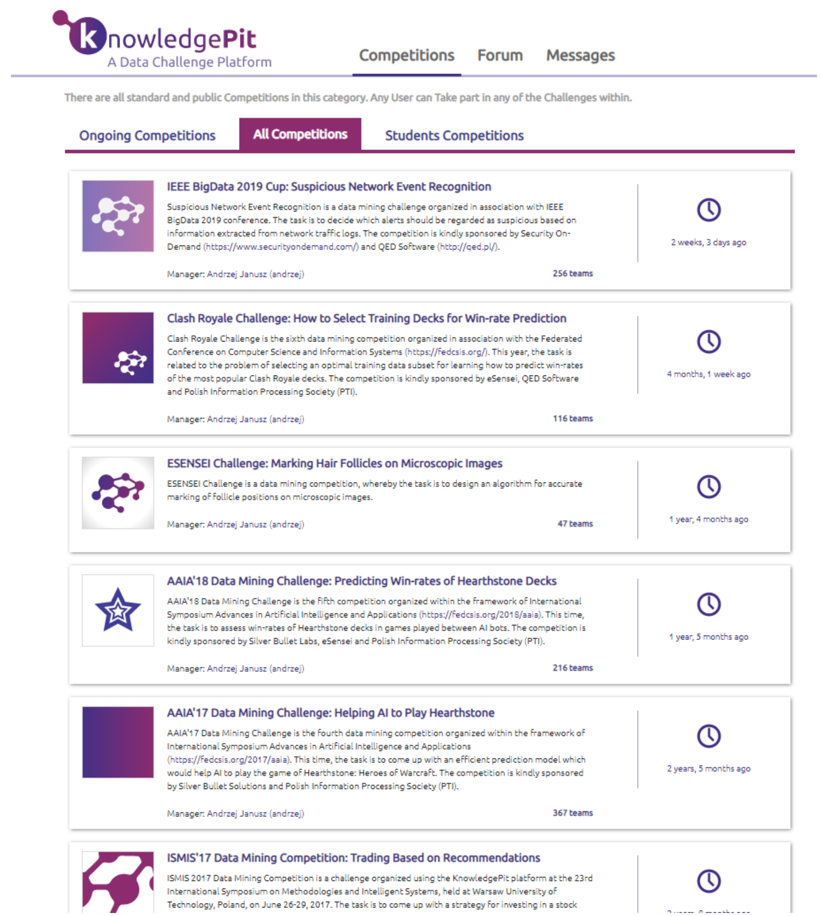
KnowledgePit.ml

Our objectives:

- stimulating data mining research
- attracting students/new researchers
- sharing “insights” and knowledge about data mining practices
- establishing connections between industry and academia
- promoting interesting events and conferences
- providing commercial services to companies that seek state-of-the-art in ML

knowledgePit
A Data Challenge Platform

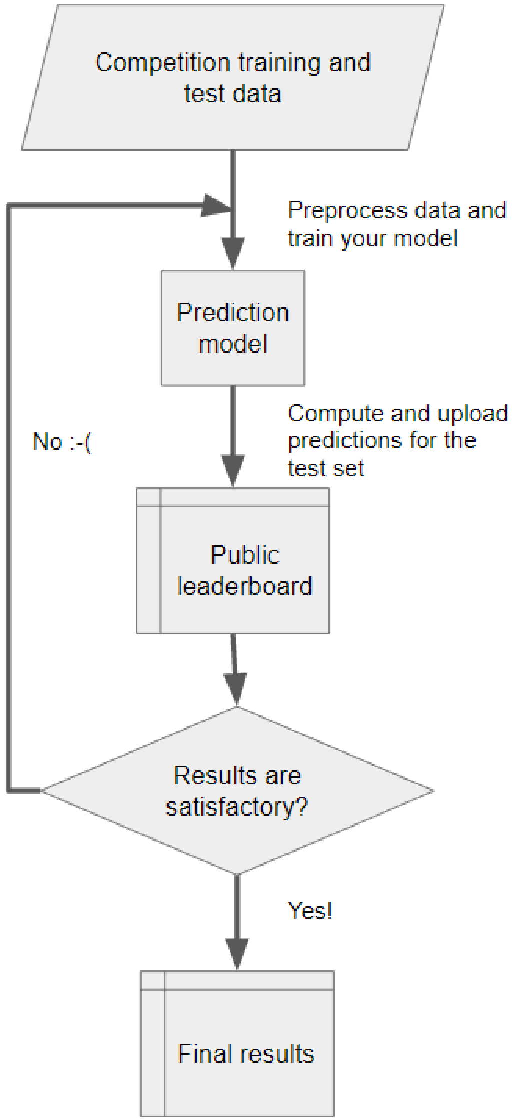
QED SOFTWARE



How does it work?

A typical competition schema:

1. The available data set is divided into the training and test parts.
2. Target values (e.g. labels) for the test set are hidden from participants – they have to be predicted.
3. Participants submit solutions which are assessed on a sample from the test set.
4. Participants select their most reliable models and write short reports.
5. The final solutions are evaluated on the remaining test data.



How to get more insights about solutions?

Rank	Team Name	Preliminary Score	Final Score	Submissions
1	Team	0.0710	**0.046300**	65
2	competition baseline	0.0415	**0.039400**	13
3	Debojit Mandal	0.0417	**0.035300**	100
4	shubham	0.0330	**0.029500**	54
5	sunsoul	0.0239	**0.028600**	100
6	Emi	0.0487	**0.028100**	68
7	Chopin	0.0428	**0.011300**	63
8	AMC_JTJ	0.0318	**0.004700**	47
9	victorkras2008	0.0000	**0.000000**	1
10	hieuvq	0.0811	**-0.019900**	83

?

?

Why?

Please, explain!

The BrightBox project

Label in the Loop

BRIGHTBOX

SMALL BIG DATA

CHALLENGES

QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models

SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows

Scope of the project

The three main goals of BrightBox:

- Explaining to humans why AI/ML models… are not certain
- Explaining to humans why AI/ML models… make mistakes
- Explaining to AI/ML models what humans want from them

AI/ML model diagnostics

Explaining why AI/ML models make mistakes

(examples of so-called diagnostic rules)

- IF there were no similar objects in the training data set, THEN the mistake is most likely because the model is not ready for such cases
  - *(but maybe it doesn't need to?)*
- IF mistakes happen quite often for similar objects in the training data set, THEN the model is not sufficiently tuned for such cases
  - *(but maybe it doesn't need to?)*
- IF there was a single (or a few) similar object and there was no mistake, THEN maybe there was something wrong with that object?
  - *(maybe it was incorrectly labeled?)*

Predicting Escalations in Customer Support

https://knowledgepit.ml/predicting-escalations-in-customer-support/

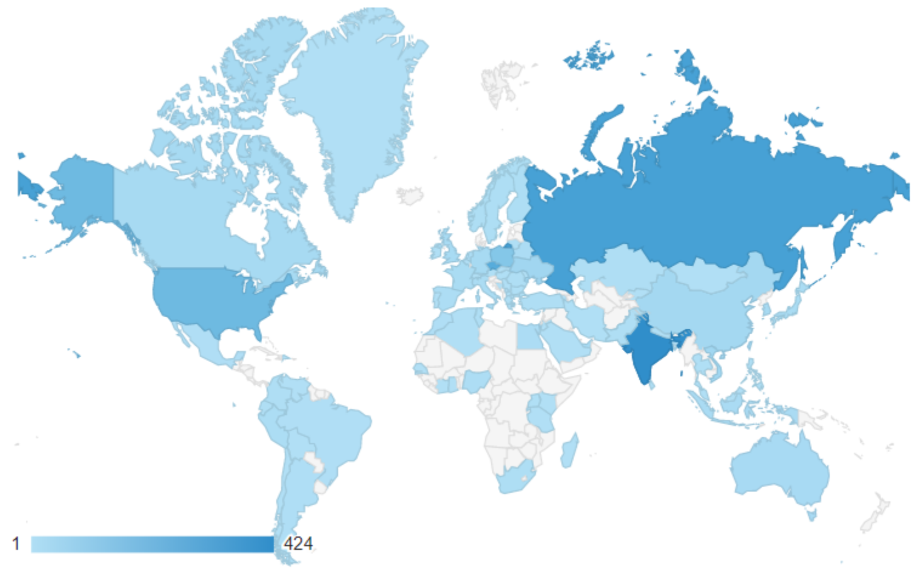
Data provided by ibi:

- History of all events and communication related to 52968 cases from the customer support department
  (40244 in the training set and 12724 cases for the evaluation)
- Case meta-data, severity logs, system milestones, and natural language communication between operatives and ibi's clients
- Anonymized data was augmented with NLP features

The task: *learn to predict the time to escalation of Information Builders's customer support cases.*

Competition overview

- 254 registered teams
- participants from 50 countries
- over 1000 submitted solutions
- the winners exceeded our baseline by more than 10%



Secondary dimension: Page

Advanced Filter ON | edit

	Country	Page	Users	Users
			2,014 % of Total: 54.01% (3,729)	2,014 % of Total: 54.01% (3,729)
1.	India	/predicting-escalations-in-customer-support/	422	20.48%
2.	Russia	/predicting-escalations-in-customer-support/	319	15.48%
3.	Czechia	/predicting-escalations-in-customer-support/	262	12.71%
4.	United States	/predicting-escalations-in-customer-support/	191	9.27%
5.	Poland	/predicting-escalations-in-customer-support/	127	6.16%
6.	Ukraine	/predicting-escalations-in-customer-support/	45	2.18%
7.	Vietnam	/predicting-escalations-in-customer-support/	44	2.13%

Final results

Validation of results using **R^2 measure:**

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_i (y_i - f_i)^2 \ ,$$

$$TSS = \sum_i (y_i - \bar{y})^2$$

Rank	Team Name	Preliminary Score	Final Score	Submissions
1	Team	0.0710	**0.046300**	65
2	competition baseline	0.0415	**0.039400**	13
3	Debojit Mandal	0.0417	**0.035300**	100
4	shubham	0.0330	**0.029500**	54
5	sunsoul	0.0239	**0.028600**	100
6	Emi	0.0487	**0.028100**	68
7	Chopin	0.0428	**0.011300**	63
8	AMC_JTJ	0.0318	**0.004700**	47
9	victorkras2008	0.0000	**0.000000**	1
10	hieuvq	0.0811	**-0.019900**	83

KnowledgePit + BrightBox = KnowledgeBox

A few challenges arise:

- Can we diagnose errors in a submission without access to the model?
- Can we meaningfully visualize submissions, their errors, and diagnoses?
- Can we explain our diagnoses to non-technical customers/competition organizers?

Similarity is the key!

- IF there were no **similar objects** in the training data set, THEN the mistake is most likely because the model is not ready for such cases
  - *(but maybe it doesn't need to?)*

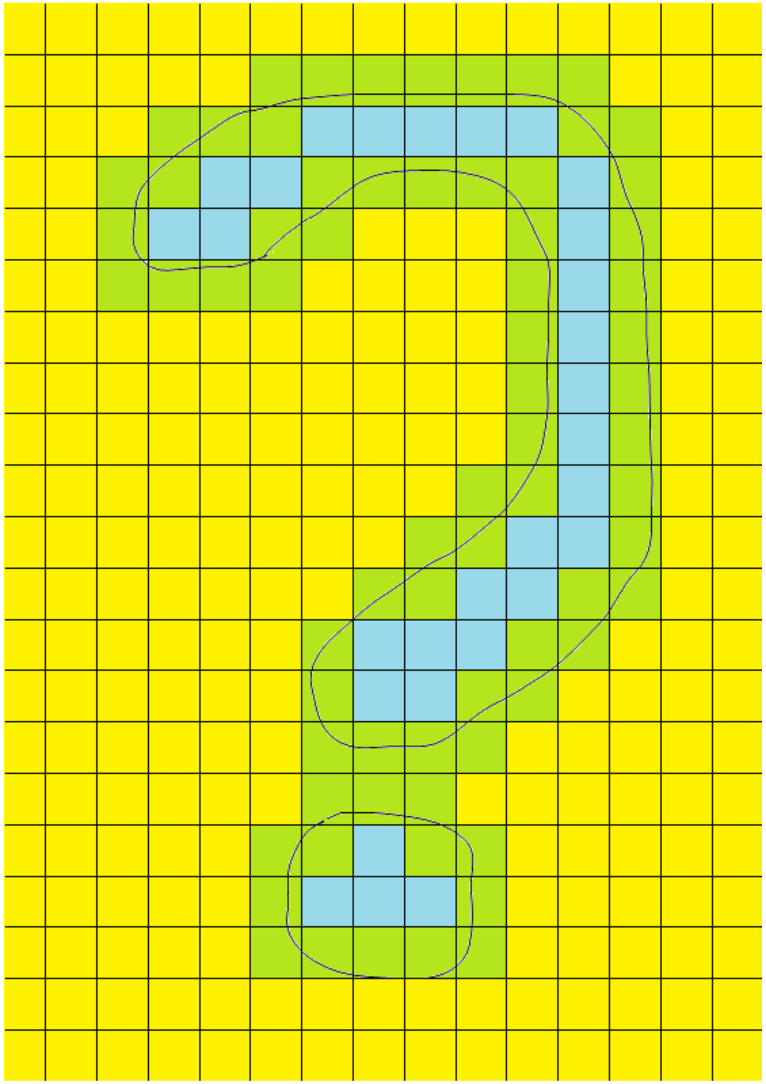
- IF mistakes happen quite often for **similar objects** in the training data set, THEN the model is not sufficiently tuned for such cases
  - *(but maybe it doesn't need to?)*

- IF there was a single (or a few) **similar object** and there was no mistake, THEN maybe there was something wrong with that object?
  - *(maybe it was incorrectly labeled?)*

But how can we select „similar objects"?

Model approximations – rough sets

- We don’t have the model – but we know how it works
- We can construct an approximation and use it instead of the model
- Closed world assumption – we only want to make diagnoses for the data that we currently have
- Rough sets seem to be the right tool for the job!



Decision reducts

The notion of a reduct is one of the most prominent contributions of the rough sets into knowledge discovery and data mining.

Information system:
$\mathbb{S} = (U, A)$

$B \subseteq A$ is an **information reduct** for $\mathbb{S} \Leftrightarrow B$ is irreducible and B discerns all objects that are discerned by A.

Decision system:
$\mathbb{S}_d = (U, A \cup \{d\})$

$B \subseteq A$ is a **decision reduct** for $\mathbb{S}_d \Leftrightarrow B$ is irreducible and B discerns all objects $i, j \in U$ such that $d(i) \neq d(j)$ and i, j are discerned by A.

- Each decision reduct corresponds to a set of rules.
- Extensions: approximate reducts, dynamic reducts, bireducts, DAAR.

A possible approach (work in progress)

- Discretize data and the analyzed model outputs (predictions)
- Create a large ensemble of diverse (approximate) reducts
  - we use the competition test data
  - the approximation accuracy can be arbitrary high (closed world…)
- Use the reducts to compute neighborhoods of test cases
  - we search for similar objects from the training data
- Apply a predefined set of diagnostic rules based on the case error, size of the neighborhood, distribution of targets, and consistency of the approximations
- **Visualize the results!**

Visualizations

We currently consider four visualization types:

- progression of solutions in time (for all, a team, etc.)
- target vs. prediction scatter plots
- target vs. diagnostic rule outcomes bar plots,
- UMAP visualization of diagnosis outcomes, errors, etc.

Some early experimental results

Conclusions and future works

- To make informed decisions, AI/ML model outcomes often require analysis/interpretation
- DM competitions are a great source of results that need interpretation ☺
- There is a lot of things that might still change in our approach – more research and experiments to do ☺
  - ➢ other approximation methods?
  - ➢ other similarity measures/neighborhood finding methods?
  - ➢ can we estimate models' uncertainty?

knowledgePit
A Data Challenge Platform

Thank You!

(any questions???)

https://knowledgepit.ml/