Kaggle: platform for data scientists to meet and compete

Lukáš Drápal (lukas.drapal@cgi.com) &
Jana Papoušková (jana.papouskova@cgi.com)

CGI

Experience the commitment®

Kaggle.com: Allstate purchase prediction challenge

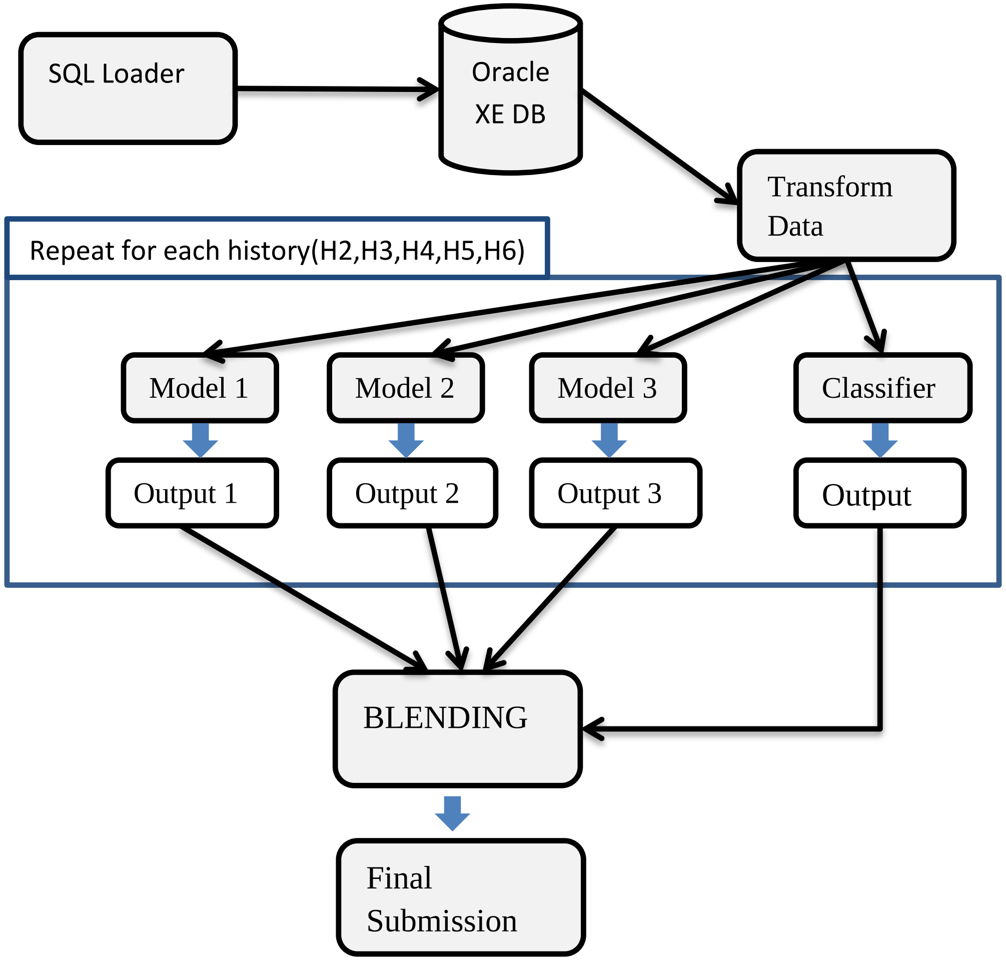
- **Kaggle.com**
  - Platform for data scientists
  - Cooperates with the biggest international companies: Facebook, GE, MasterCard, Merck, NASA, Deloitte and also with top universities
- **Allstate**
  - Insurance company
  - The largest property and casualty company in US

- **Competition**
  - Prediction of which insurance products will be purchased by the customers
  - Based on their demographical profile and history of recently view insurance products
- **CGI Prague Big Data team**
  - **1st from 1500+** worldwide competing teams

Allstate purchase prediction challenge

- R gradient boosting library (gbm)
- Task: predict 7 types of policies that a customer will buy
- Evaluation metric: all or none
- Benchmark – pick the last customer’s choice – worked really well
- Changes costly
- Final model – belt and braces approach:
  - Three models of different complexity
  - Classifier whether last customer’s choice should be used

Scheme of the solution



- On different levels of classifier use models when they coincide in a different way: three, two or never

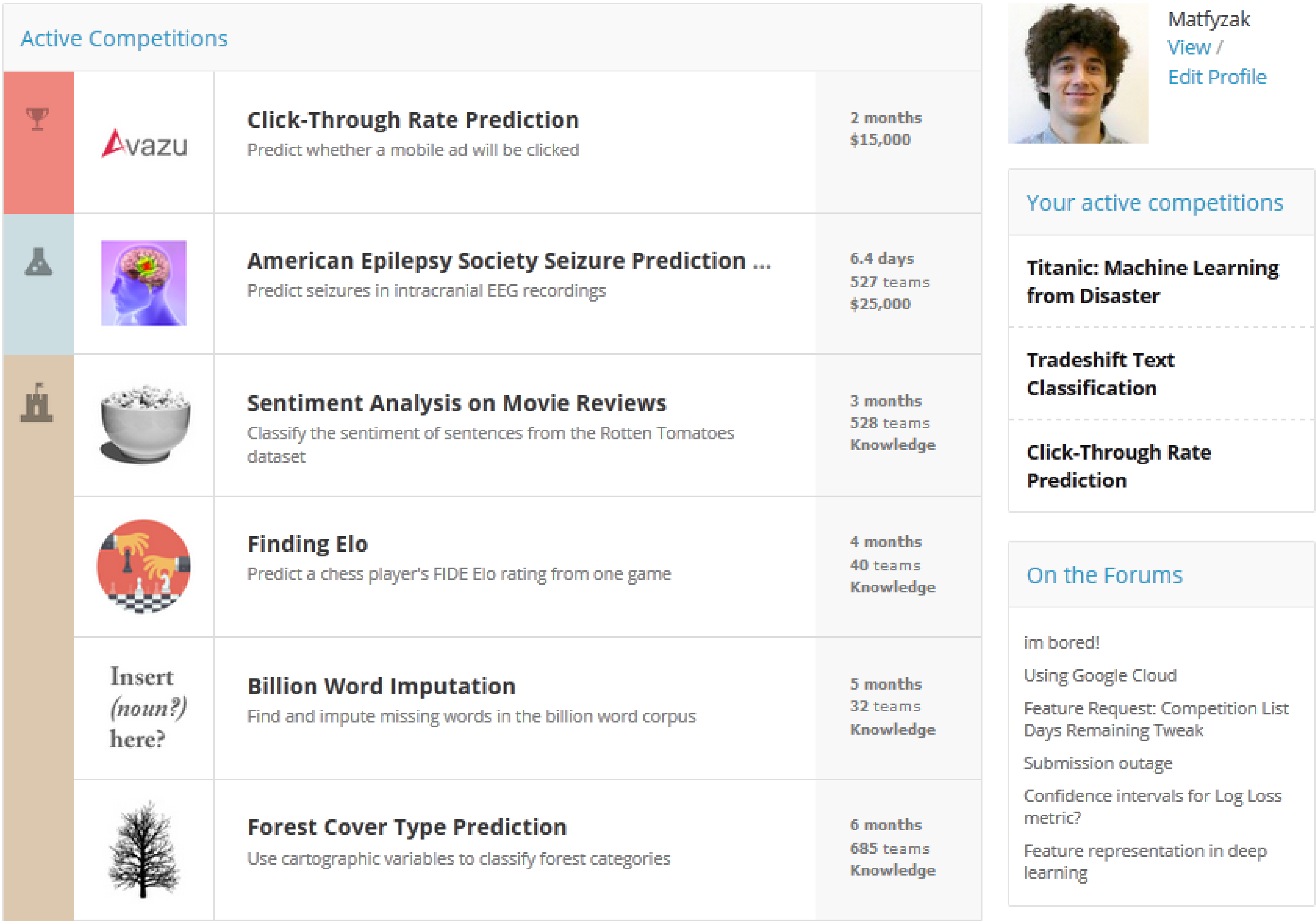
Lessons learnt

- Kaggle is great at knowledge sharing
- It is (almost) all about feature engineering
- Cross validate, cross validate, cross validate…

… but know the differences between test set & train set

- Caret is a great R package for parameters tuning
- Ensembling (blending) models together can bring the needed edge
- Optimize towards the right evaluation metrics

Warning: competing @ Kaggle is addictive



Active Competitions

Click-Through Rate Prediction Predict whether a mobile ad will be clicked	2 months $15,000
American Epilepsy Society Seizure Prediction ... Predict seizures in intracranial EEG recordings	6.4 days 527 teams $25,000
Sentiment Analysis on Movie Reviews Classify the sentiment of sentences from the Rotten Tomatoes dataset	3 months 528 teams Knowledge
Finding Elo Predict a chess player's FIDE Elo rating from one game	4 months 40 teams Knowledge
Billion Word Imputation Find and impute missing words in the billion word corpus	5 months 32 teams Knowledge
Forest Cover Type Prediction Use cartographic variables to classify forest categories	6 months 685 teams Knowledge

Matfyzak
View /
Edit Profile

Your active competitions

Titanic: Machine Learning from Disaster

Tradeshift Text Classification

Click-Through Rate Prediction

On the Forums

im bored!

Using Google Cloud

Feature Request: Competition List Days Remaining Tweak

Submission outage

Confidence intervals for Log Loss metric?

Feature representation in deep learning