TARAN
ADVISORY IN DATA & ANALYTICS

Data Science 2 - Intro

February 17, 2025

Data Science 2 - Intro

Faculty of Mathematics and Physics

About us

Václav Kozmík, Contact: vkozmik@taran.ai

- PhD in Econometrics at MFF UK
- One year at University of Texas

Chance Loterie Korunka

- Lead developer
- Head of analytics team

捷信 HOME CREDIT

HOME CREDIT BANK

- Underwriting analyst
- Head of Scoring, Big Data and Innovations

TARAN

- Co-founder
- Managing Partner Europe

About us

Karel Kozmík, Contact: kkozmik@taran.ai

- Studying PhD in Econometrics at MFF UK

TARAN

- Senior Data Scientist
- DevOps Engineer

About us

Ondřej Týbl, Contact: ondra@tybl.cz

- PhD in Probability Theory at MFF UK
- One year at King's College in London

- Postdoc
- Research in Computer Vision

Decision Making Evolution

- Simple rules
  - If outside temperature is lower than 15 => wear a jacket.
  - If applicant has sufficient income => grant a loan.
- Statistical model
  - Combines several predictors using trained weights.
  - Models such as linear regression or decision tree can be written in one simple equation.
  - From the model equation it is obvious how the included predictors affects the prediction.
  - Model can be manually modified (due to its simplicity).
- Machine learning
  - Provide algorithm, data and task and let the computer find the useful patterns for fulfilling the task.
  - Explaining the model might not be easy.

Machine learning basics

Basic principles of ML algorithms:

- Process huge amounts of data (in terms of both number of observations and attributes)
- High accuracy of the resulting models
- Full automation of the training process which should discover all relationships from the data

Useful languages and tools:

- Python / Jupyter
- Spark
- SQL
- Git

Machine learning models

There are variety of tasks which can be solved by application of machine learning methods:

MLTask	Type of relation	Business task example
Regression	$\mathbb{R}^n \to \mathbb{R}$	Forecasting salary
Binary classification	$\mathbb{R}^n \to \{0, 1\}$	Credit scoring, spam detection
Multi-class classification	$\mathbb{R}^n \to \{0, \ldots, m\}$	Classify support incidents by types
Multi-label classification	$\mathbb{R}^n \to \{c \vert c \in \{0, \ldots, m\}\}$	Email categorization
Ranking	…	Ranking search query results
Clustering	…	Find typical group of payments
Anomaly detection	…	Find exceptional customers, intrusions
Dimension reduction	…	Text annotating
Collaborative filtering	…	Movie recommendation

Examples from practice
Credit risk model

Task: predict if loan applicant will repay the debt.

- Data from application form, behavioral data, Credit Bureau data, transaction data, Telco data, etc.
- Binary target: 1 for non repaid loans, 0 for repaid loans
- Hundreds of thousands observations

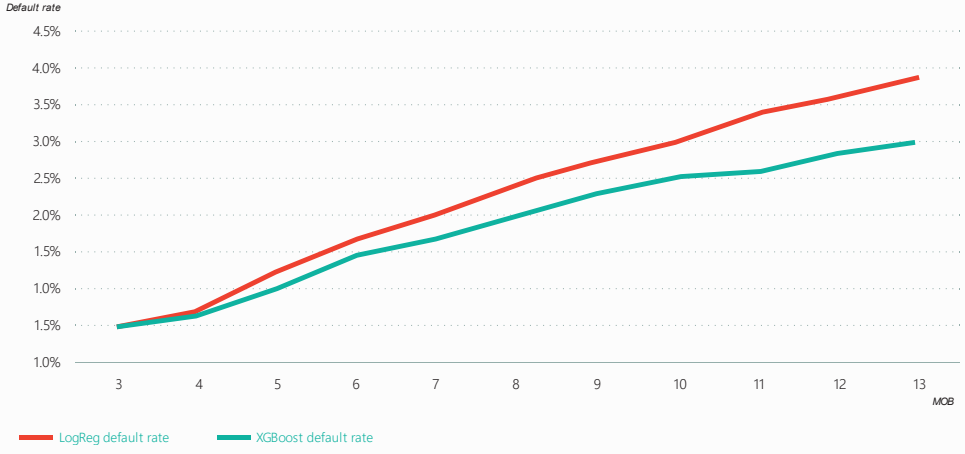
Common solution: logistic regression

ML solution: XGBoost with binary target, logistic loss and stopped by AUC

- Up to 15% better performance of ML algorithm on production A/B test

Examples from practice

Credit risk model



Examples from practice

Digital footprint model

Task: predict propensity or financial default based on digital footpring

- Data: list of URL (visited websites) together with timestamp for each client
- Tens of thousands of distinct domains
- Binary target

Challenges:

- Correlation
- Number of predictors

ML solution: Logistic regression with L2 regularization

Examples from practice

Price sensitivity model

Task: Determine the best interest rate to be offered for mortgage refixing

- Data about historical process: interest rates offered, client feedback, changes in the offer, accepted or not
- Market data: average interest rate

Challenges:

- No control group for manual discount process
- Each client can be only in treatment (discount) or control (no discount) group

ML solution: Uplift transformation of target + XGBoost

Examples from practice

Customer segmentation

Task: Segment customer of online grocery shop to allow for better targeting of campaigns and promotions

- Customer data is limited: basic info and addresses
- All historical transactions are available, as well as prices, SKU catalog, etc.
- Segmentation can be also used to power recommendation (what to buy next)

Challenges:

- SKU catalog structure cannot represent all dimensions (e.g. BIO/farmers vs. fruits/vegetables)
- Outliers in spendings - zero spending in the category as well as extra high spending
- Censor: people buy the necessary goods in other shop

Solution: K-means clustering

- Normalization of predictors is essential to form reasonable clusters

Examples from practice
Ranking model

Task: Rank items by popularity for e-commerce marketplace iPrice

- 7 countries, for example `https://iprice.sg/`
- Data about each product: price, category, brand, size, etc.
- Historical clicks on each product
- 500 million products

Challenges:

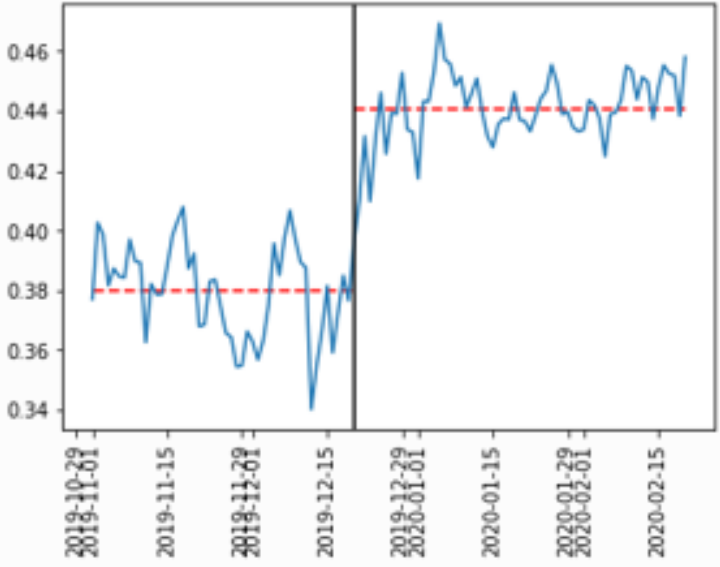
- Data size
- Presentation bias: people click only what is shown
- Position bias: people look more carefully at first results

ML solution: XGBoost with tweedie regression (we have lot of items with zero clicks)

Examples from practice

Ranking model

Increase in click-through-rate from 38% to 44%



Examples from practice
Fraud model

Task: Detect fraudsters in sport betting

- Some clients use insider info to beat the odds
- Data about each bet, client registration, cash flows, etc.
- Statistical approach can detect them after 50-100 bets, but that is too late

Challenges:

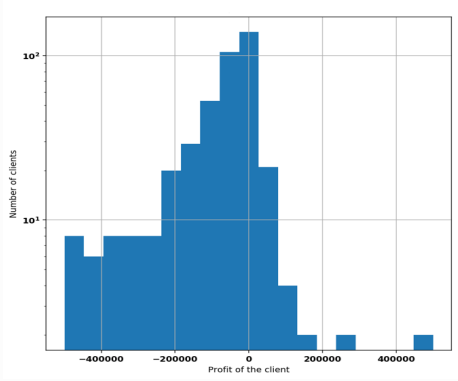
- What is the right target? 0/1 for fraudsters? Profit?
- Noise in profits given by randomness
- Target censor: existing manual process blocks detected fraudsters

ML solution: XGBoost regression with observation reweighting and reject inference

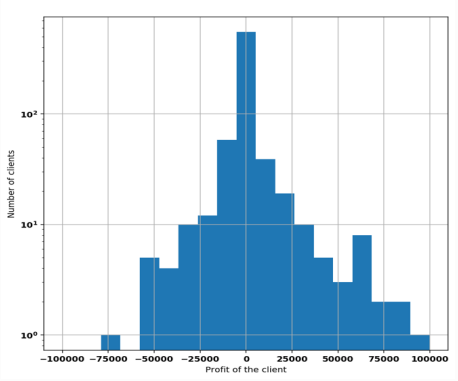
Examples from practice

Fraud model

Good clients



Bad clients



Examples from practice

Demand forecasting model

Task: Predict demand for grocery stores chain

- Demand forecast is used for automated replenishment of stocks as well as staffing
- Goal: increase sales (no out of stock) and reduce shrinkage (expiration)
- All historical transactions are available, as well as prices, SKU catalog, etc.

Challenges:

- Data size - millions of SKUs, billions of transactions
- Sparsity - some products are boughts few times a week or month, but we need daily prediction
- Promotions and cannibalization
- Covid closures affect demand

Common solution: time series algorithms like Holt-Winters

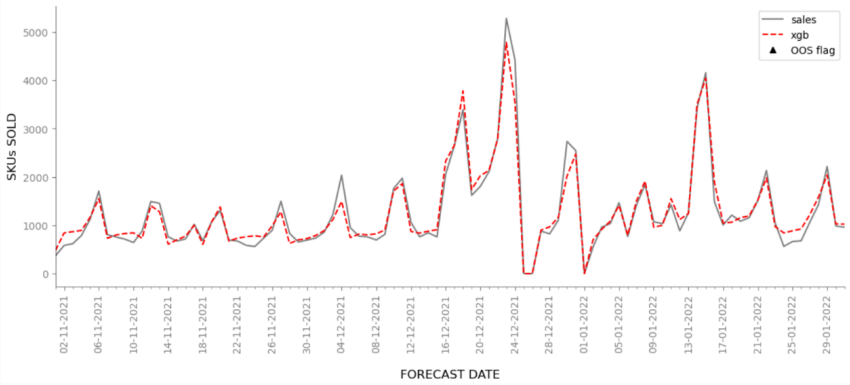
ML solution: XGBoost regression with artificial parameter (number of days ahead)

- Feature engineering is the key to model seasonality

Examples from practice

Demand forecasting model

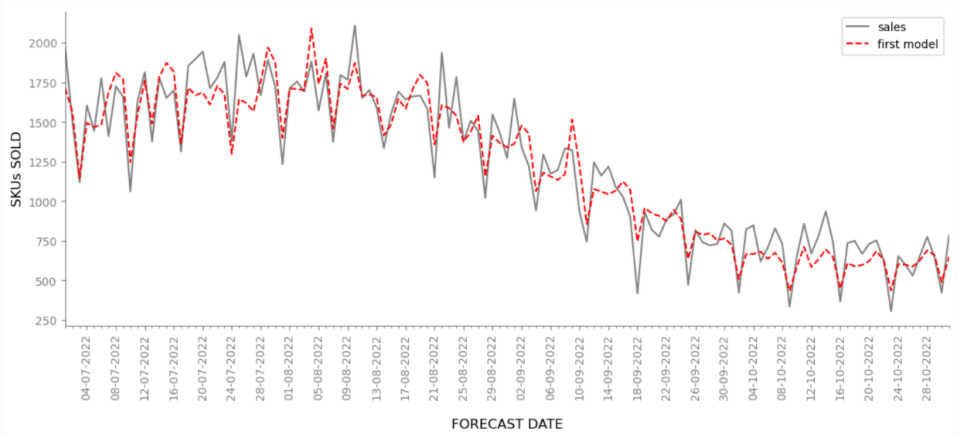
- Strong weekly pattern
- Strong effect of promotions



Examples from practice

Demand forecasting model

- Strong seasonal effect



Examples from practice

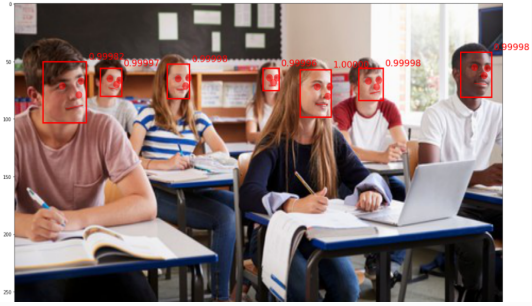
Face recognition

Task: Validate identity of an applicant

- Client loads photo of his/her ID to mobile application and then is requested to take selfie
- Face is detected both on ID and selfie and compared

Challenges:

- Face extraction from the picture
- Liveness detection
- Required high penalisation for false positive



ML solution: Convolutional neural network for face recognition, another network for face matching

Examples from practice
SMS content categorization

Task: Group together similar content SMSes

- Find financial related SMSes (payment reminders, overdue payments, etc.)
- Assess clients credit risk based on content of SMSes

Challenges:

- Processing text (unstructured) data
- Millions of observations

ML solution: Group detection based on word encoding (word2vec)

Examples from practice

Voice to text

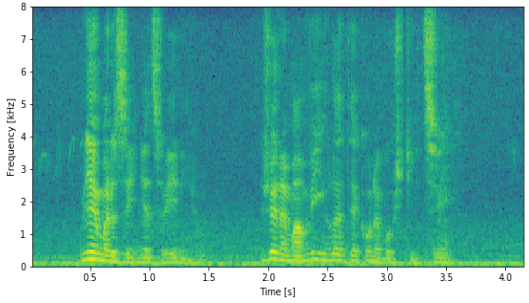
Task: Translate recorded voice to text

- Incoming calls are recorded and translated to text for further analysis

Challenges:

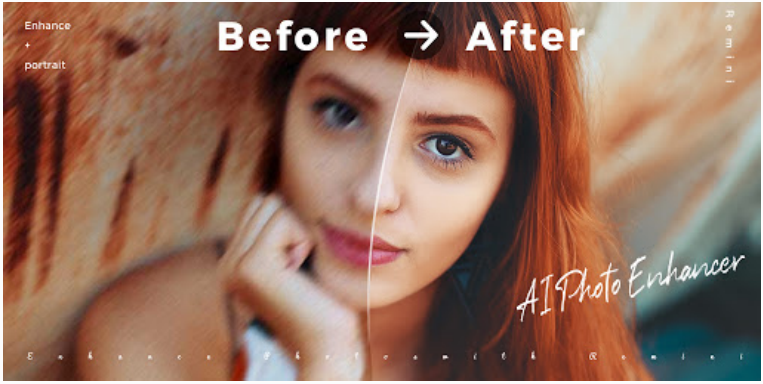
- For model training you are required to have labels
- How to pre-process input signal to be fed to neural network?
- What loss function should be used?

ML solution: LSTM powered neural network wit CTC loss function.



Other ML powered solutions

Photo enhancers



Other ML powered solutions

Detection of generated photo

Can you tell which face is real and which is AI generated?

Other ML powered solutions

Detection of generated photo

Can you tell which face is real and which is AI generated?

The face on the right is real. The photo on the left was generated by an AI application.

Contemporary ML powered solutions

Examples

- Autonomous driving
- Superhuman game playing
- Voice assistant
- Chatbots, work assistants, agents
- Image or video generation
- Translation
- ...

Lecture content

Models:

- decision trees, random forest, gradient boosting
- simple neural networks, convolutional neural networks, transformers

General topics:

- model quality metrics
- train-test split, oversampling, bootstraping
- over-fitting, regularization
- feature engineering, seasonality
- parallelization, programming languages

Thank you!

TARAN

ADVISORY IN DATA & ANALYTICS