Churn Prediction with Machine Learning

Hello!

I am Zsolt Fekete

I’m here today to introduce Antavo’s solution for churn prediction.

You can find me at:

Three things to talk about:

The problem

What do we call as churn? Why is it so important? How can we predict it accurately?

The solution

Antavo's solution should be scalable, independent and real-time. Can we achieve this somehow?

The future

How can we improve our solution? What are the next steps?

1.
The problem itself

What do we call as churn? Why is it so important? How can we predict it accurately?

What is customer churn?

*“Customer churn is the percentage of **customers that stopped using your company's product or service** during a certain time frame.*

You can calculate churn rate by dividing the number of customers you lost during that time period - say a quarter - by the number of customers you had at the beginning of that time period.”

Initial problems

IMBALANCED DATA

We already have a lot of customer and transactional data, but the customer-base is really imbalanced: more than **80% of users buy only once**.

If we handle this problem incorrectly, we will immediately face the problem of **overfitting**, which will result in all users being told that they have been churned.

COLD START PROBLEM

We have data for labeling, but how can we prove the alignment of our algorithm?

In the first round, it will be necessary to **shift the time-out**: for example, spin the time back to 3 months ago, so the test data can be the last 3 months.

Initial problems

THE COST

The recalculation of the predicted values is continuous, so we also have to **calculate the generated infrastructure costs**.

If we want to achieve real-time prediction, we must leave the original technology stack and perform the calculations on a **physical machine**. We then need to return the calculated values to the original stack.

GDPR AND DATA SECURITY

We live in the age of GDPR, so the issue of data security is not negligible either.

By migrating the data, we need to anonymize it, deliver it on a secure channel, and ensure our customers that their data is still secure.

85K
active users per month

45K
transactions made per month

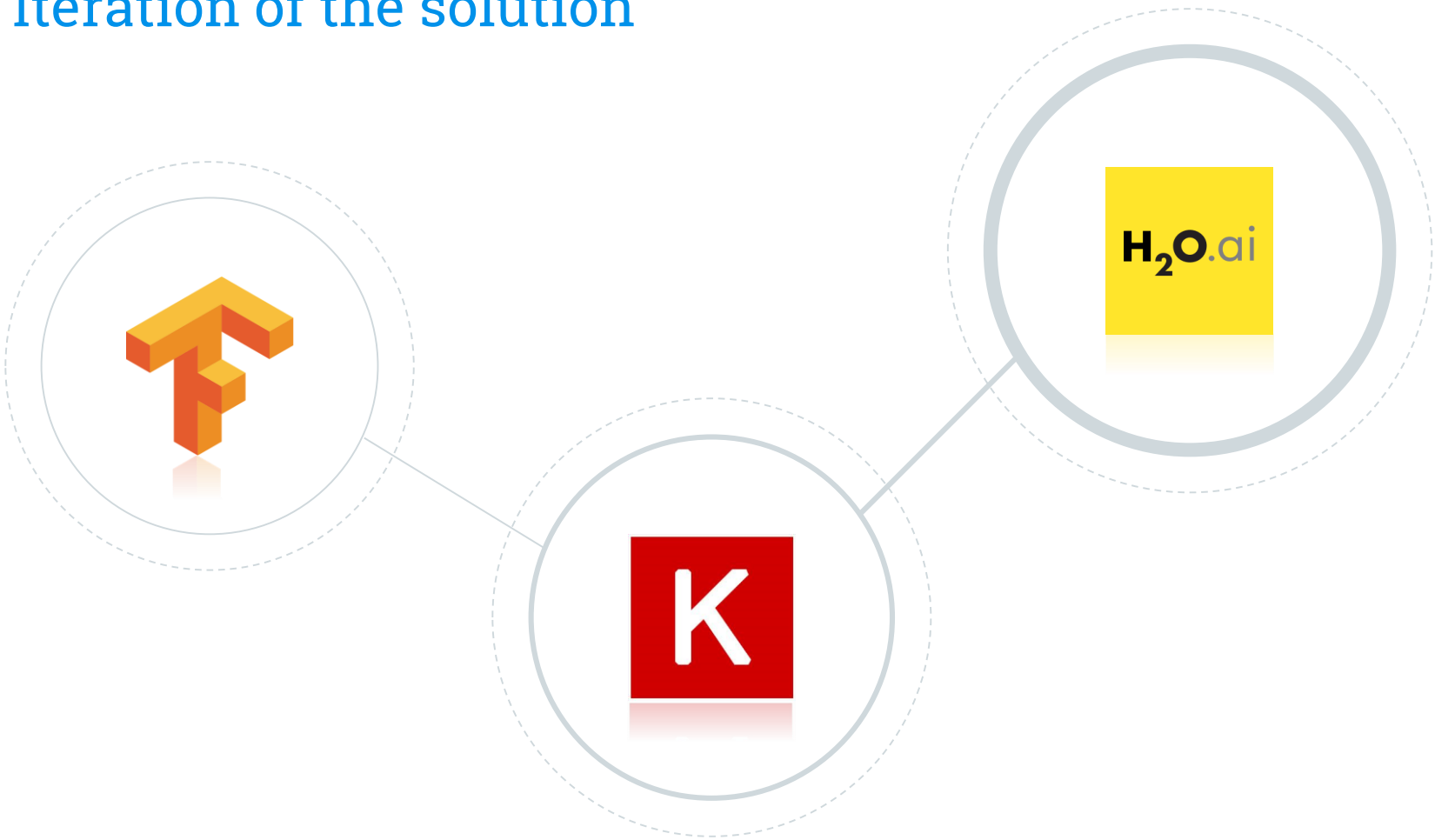
6M
actions performed per month

35M
predictions will be calculated in every month

2.
The solution

Antavo’s solution should be scalable, independent and real-time. Can we achieve this somehow?

Iteration of the solution



Tensorflow as the first solution

Advantages

- low-level API
- lot of popularity
- high performance on large datasets

Disadvantages

- poor readability
- no released stable version
- hard to use
- quickly changing API

Keras as the second solution

Advantages

- rapid prototyping
- easy-to-learn
- readable and concise architecture

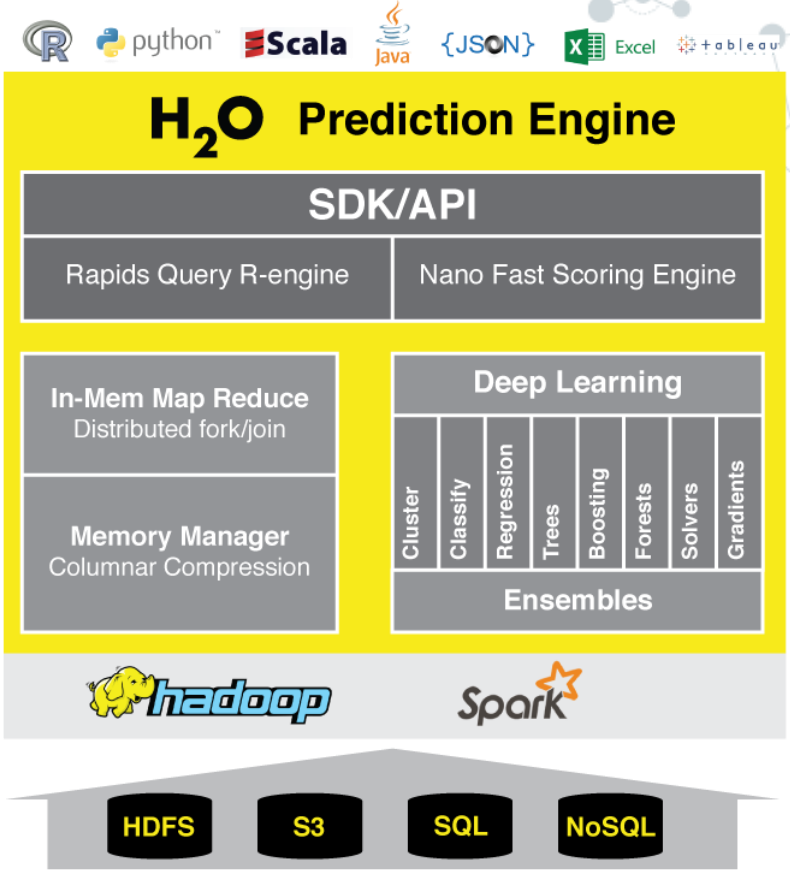
Disadvantages

- only deep learning is allowed
- performance is comparatively slower
- lack of important features

H2O as the final solution

Advantages

- vertically scalable
- AutoML & assembling methods
- built-in HDFS/Scala compatibility
- R & Python SDK
- complex and mature ecosystem

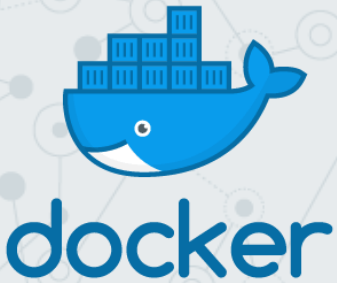


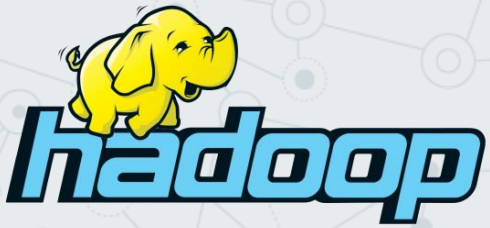

mongoDB

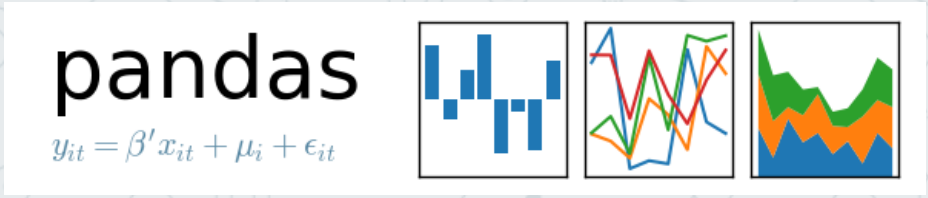

APACHE
kafka

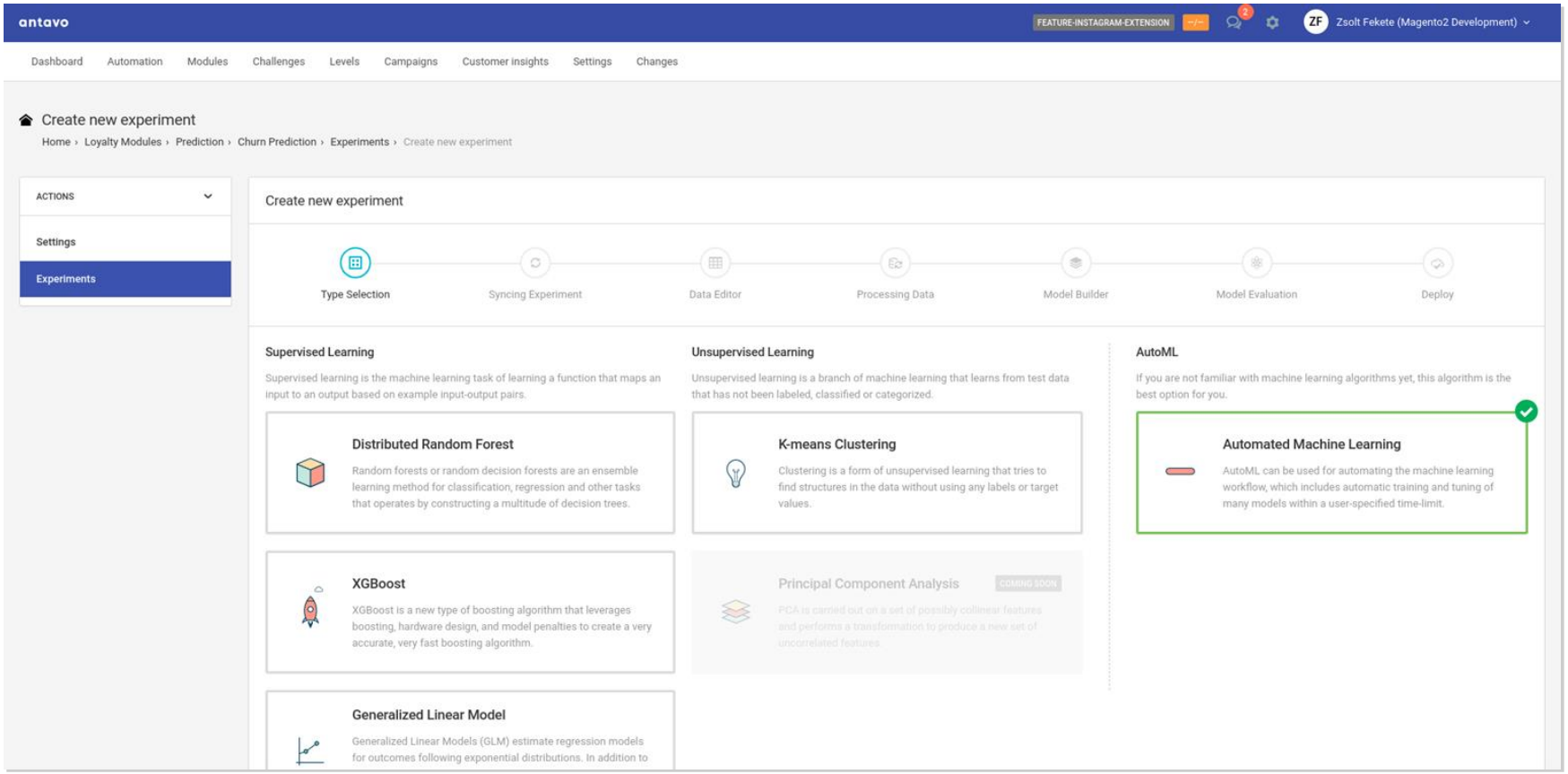

NumPy

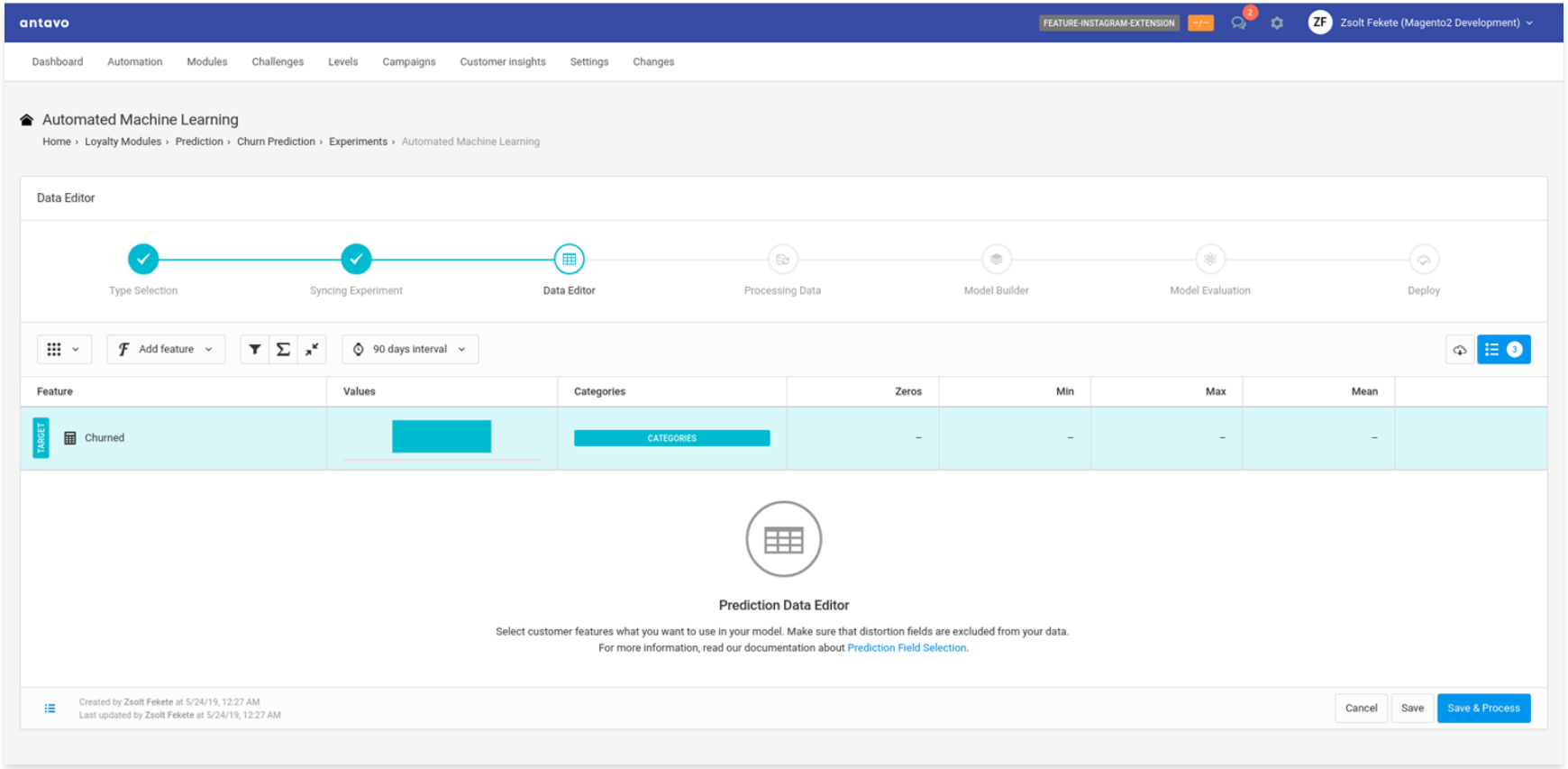

H2O.ai


docker


hadoop


pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$


antavo
FEATURE-INSTAGRAM-EXTENSION
ZF
Zsolt Fekete (Magento2 Development)
Dashboard
Automation
Modules
Challenges
Levels
Campaigns
Customer insights
Settings
Changes
Create new experiment
Home › Loyalty Modules › Prediction › Churn Prediction › Experiments › Create new experiment
ACTIONS
Settings
Experiments
Create new experiment
Type Selection
Syncing Experiment
Data Editor
Processing Data
Model Builder
Model Evaluation
Deploy
Supervised Learning
Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
Distributed Random Forest
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees.
XGBoost
XGBoost is a new type of boosting algorithm that leverages boosting, hardware design, and model penalties to create a very accurate, very fast boosting algorithm.
Generalized Linear Model
Generalized Linear Models (GLM) estimate regression models for outcomes following exponential distributions. In addition to
Unsupervised Learning
Unsupervised learning is a branch of machine learning that learns from test data that has not been labeled, classified or categorized.
K-means Clustering
Clustering is a form of unsupervised learning that tries to find structures in the data without using any labels or target values.
Principal Component Analysis
COMING SOON
PCA is carried out on a set of possibly collinear features and performs a transformation to produce a new set of uncorrelated features.
AutoML
If you are not familiar with machine learning algorithms yet, this algorithm is the best option for you.
Automated Machine Learning
AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models within a user-specified time-limit.


antavo
FEATURE-INSTAGRAM-EXTENSION
Zsolt Fekete (Magento2 Development)
Dashboard
Automation
Modules
Challenges
Levels
Campaigns
Customer insights
Settings
Changes
Automated Machine Learning
Home › Loyalty Modules › Prediction › Churn Prediction › Experiments › Automated Machine Learning
Data Editor
Type Selection
Syncing Experiment
Data Editor
Processing Data
Model Builder
Model Evaluation
Deploy
Add feature
90 days interval
Feature
Values
Categories
Zeros
Min
Max
Mean
TARGET
Churned
CATEGORIES
Prediction Data Editor
Select customer features what you want to use in your model. Make sure that distortion fields are excluded from your data.
For more information, read our documentation about Prediction Field Selection.
Created by Zsolt Fekete at 5/24/19, 12:27 AM
Last updated by Zsolt Fekete at 5/24/19, 12:27 AM
Cancel
Save
Save & Process

3.
The future

How can we improve our solution? What are the next steps?

Improvements & Usage

- Building a Next-Best-Action algorithm top on the churn prediction values
- More accurate customer targeting, preventive loyalty
- Automatic customer segmentation, clustering

Technology Upgrades

- Tensorflow 2.0 upgrades
- Keras ecosystem, Tensorflow Extended
- Predicting values in batches
- Introducing a distributed computing software natively, like Hadoop

Thanks!

Any questions?

You can find me at:

@pjtuxe

zsolt.fekete@antavo.com