Kaggle competition
Grupo Bimbo Inventory Demand
Winning solution by
“The Slippery Appraisals” team

Dmitry Larko, Sr. Data Scientist @ H2O.ai
dmitry@h2o.ai

January 19th, 2017

About me

Dmitry Larko

Sr. Data Scientist at H2O.ai

San Francisco Bay Area, CA, United States

Joined 4 years ago · last seen in the past day

http://h2o.ai

Competitions Grandmaster

Home | Competitions (33) | Kernels (0) | Discussion (31) | Datasets (0) | More | Edit Profile

Competitions Grandmaster		
Current Rank 40 of 53,474	Highest Rank 25	
9	8	6

Kernels Contributor		
Unranked		
0	0	0

Discussion Contributor		
Unranked		
0	4	12

Team

- **Alexander Larko** - MSc in Computer Science. 10 years in Data Mining.
- **Dmitry Larko** – Sr Data Scientist, H2O.ai
- **Bohdan Pavlyshenko** - Ph.D., Data Scientist at SoftServe, assoc.prof. at Lviv National University (Ukraine)
- **Philip Margolis** – Freelancer Data Scientist and Consultant
- **Stanislav Semenov** – Data Scientist and Quantitative Researcher

Team

Stanislav Semenov

Moscow, Russian Federation

Joined 3 years ago · last seen in the past day

in

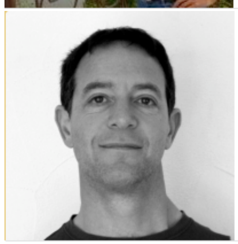
Competitions Grandmaster

Alexander Larko

Minusinsk, Krasnoyarsk region, Russia

Joined 7 years ago · last seen in the past day

Competitions Grandmaster



Silogram

Zurich, Switzerland

Joined 4 years ago · last seen in the past day

in

Competitions Grandmaster

Dmitry Larko

Sr. Data Scientist at H2O.ai

San Francisco Bay Area, CA, United States

Joined 4 years ago · last seen in the past day

in http://h2o.ai

Competitions Grandmaster

Bohdan Pavlyshenko

Lviv, Ukraine

Joined 3 years ago · last seen in the past day

in http://bpavlyshenko.blogspot.com/

Competitions Master

Solution overview

- XGBoost - main workhorse
- Interesting feature: Product cluster ID
- Tools: Python 2/3 and R
- Full training: ~ 2 week on 8 cores to train 1st level models and another 3-4 days to build ExtraTrees and linear models on top of that

Problem

- Goal:
  - Develop a model to accurately forecast inventory demand based on historical sales data
- Evaluation:
  - Root Mean Squared Logarithmic Error:
    - $\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$

Dataset

- Train.csv (74 million observations)
- Test.csv (7 million observations)
- Cliente_tabla.csv (Client Names)
- Producto_tabla.csv (Product Names)
- Town_state.csv (Town and State information)

Target variable

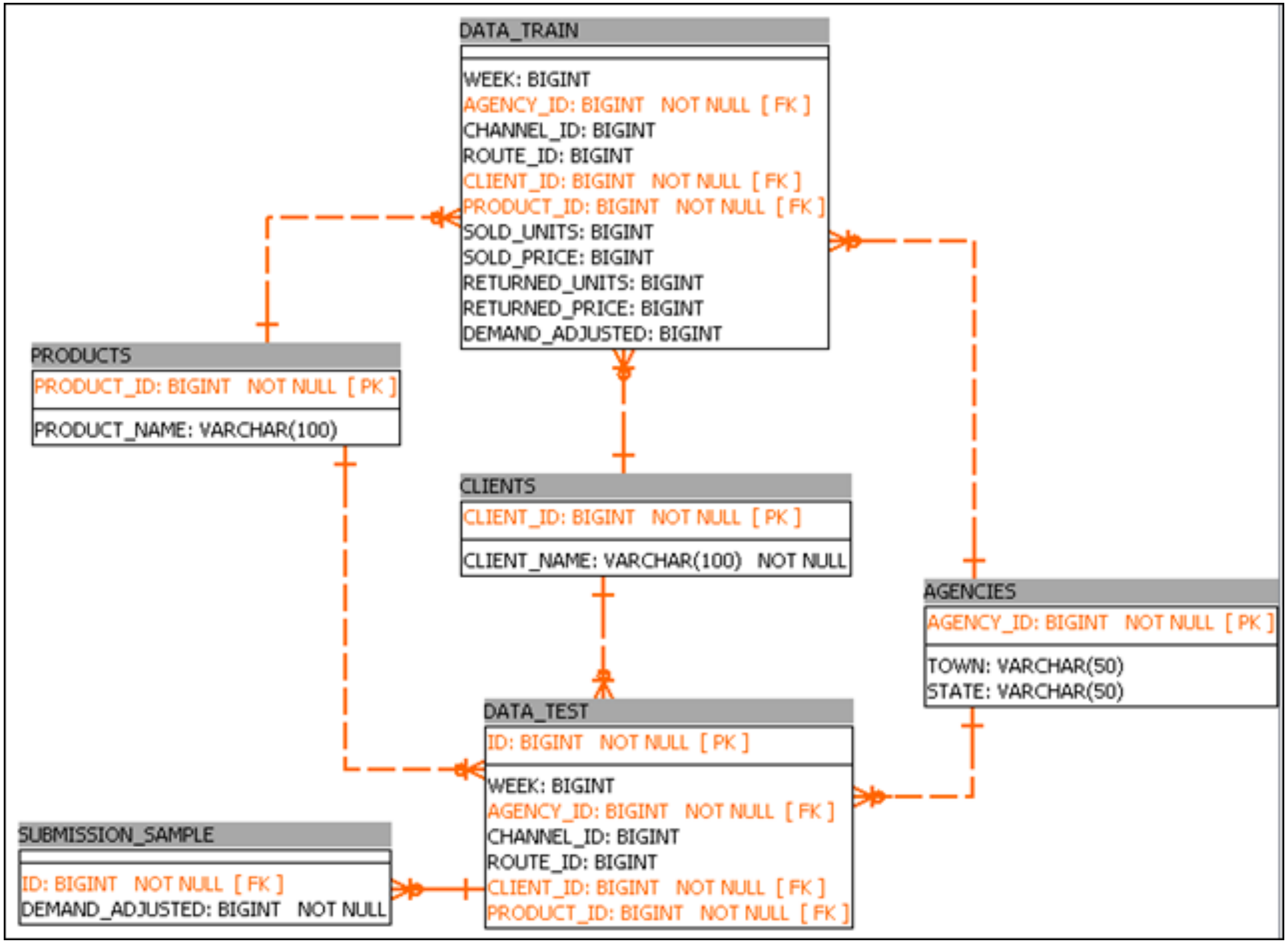
- Mean: 7.22
- Median: 3
- Min: 0
- Max: 5000
- 75% of data is between 0 and 6
- Right-skewed
- Most of ML models can optimize RMSE, to optimize RMSLE, log-transform target variable:
  - log(target+1)

Dataset

File	Column	Table	Column	Type
cliente_tabla.csv	cliente_id	CLIENTS	CLIENT_ID	NUMBER(10)
	nombrecliente		CLIENT_NAME	VARCHAR2(100)
producto_tabla.csv	producto_id	PRODUCTS	PRODUCT_ID	NUMBER(5)
	nombreproducto		PRODUCT_NAME	VARCHAR2(100)
sample_submission.csv	id	SUBMISSION_SAMPLE	ID	NUMBER(7)
	demanda_uni_equil		DEMAND_ADJUSTED	NUMBER(1)
test.csv	id	DATA_TEST	ID	NUMBER(7)
	semana		WEEK	NUMBER(2)
	agencia_id		AGENCY_ID	NUMBER(5)
	canal_id		CHANNEL_ID	NUMBER(2)
	ruta_sak		ROUTE_ID	NUMBER(4)
	cliente_id		CLIENT_ID	NUMBER(10)
	producto_id		PRODUCT_ID	NUMBER(5)
town_state.csv	agencia_id	AGENCIES	AGENCY_ID	NUMBER(5)
	town		STATE	VARCHAR2(50)
	state		TOWN	VARCHAR2(50)
train.csv	semana	DATA_TRAIN	WEEK	NUMBER(2)
	agencia_id		AGENCY_ID	NUMBER(5)
	canal_id		CHANNEL_ID	NUMBER(2)
	ruta_sak		ROUTE_ID	NUMBER(4)
	cliente_id		CLIENT_ID	NUMBER(10)
	producto_id		PRODUCT_ID	NUMBER(5)
	venta_uni_hoy		SOLD_UNITS	NUMBER(4)
	venta_hoy		SOLD_PRICE	NUMBER(9)
	dev_uni_proxima		RETURNED_UNITS	NUMBER(6)
	dev_proxima		RETURNED_PRICE	NUMBER(6)
	demanda_uni_equil		DEMAND_ADJUSTED	NUMBER(4)

Slide based on Mario Talavera Writes blogpost: https://mtalavera.wordpress.com/2016/06/30/kaggle-grupo-bimbo-inventory-demand/

Schema



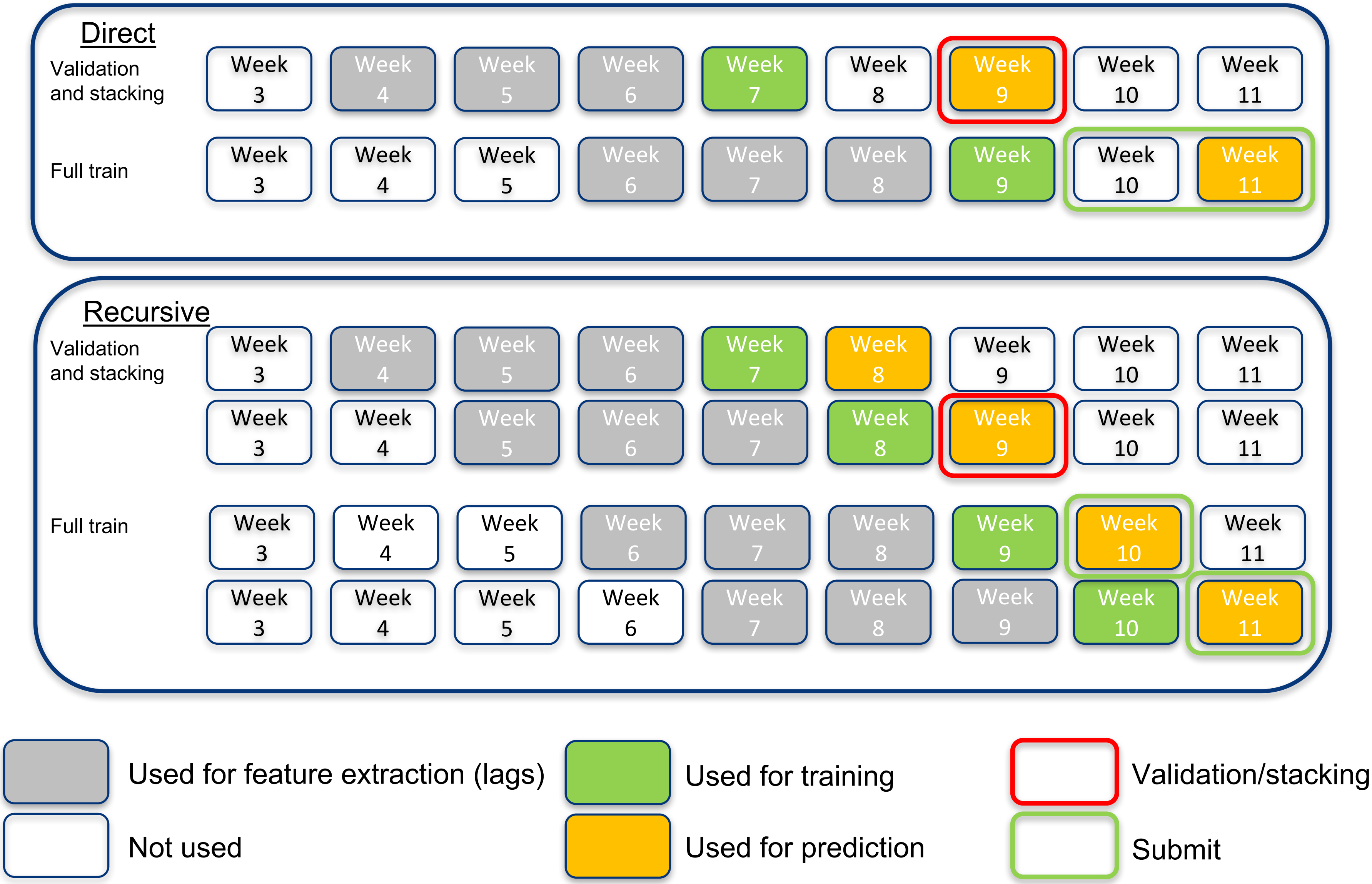
Slide based on Mario Talavera Writes blogpost: https://mtalavera.wordpress.com/2016/06/30/kaggle-grupo-bimbo-inventory-demand/

Stats

- 930,500 Clients. Of these clients, 9,663 show up in the test data set (the one to predict demand for) that do not exist in the train set.
- 2,592 Distinct Products. 34 new products in test data.
- 790 Agencies across 260 towns in 33 states in Mexico.
- Each of these agencies, also known as sales depots, contain several delivery routes.
- Each route serves multiple clients delivering and collecting returned products.
- 9 Sales Channels.
- 9 weeks of sales data broken into 7 weeks of sales data (from week 3 to week 9) and 2 weeks (week 10 and 11) of test data.
- 3,603 routes on train data, 2,608 routes on test data.
- For the 7 weeks of train data, 1,799 different products were delivered across 552 agencies on 3,603 routes to 880,604 clients.

Slide based on Mario Talavera Writes blogpost: https://mtalavera.wordpress.com/2016/06/30/kaggle-grupo-bimbo-inventory-demand/

Validation schemas



Features Selection / Engineering

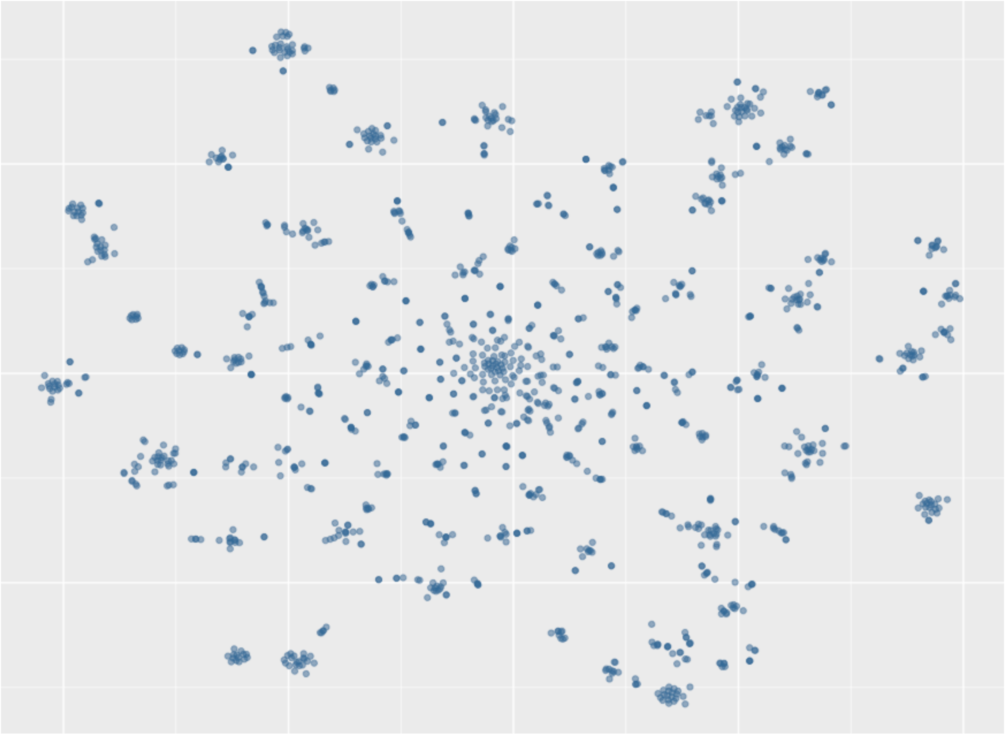
- Feature transformations / engineering
  - Value's frequency for categorical variables (e.g. Producto_ID, Cliente_ID, Agencia_ID, etc.) and different combinations of them
  - Target variable Demanda_uni_equil, grouped by factors variables (mean,median, max, min, sum)
  - Numeric features (Venta_hoy, Venta_uni_hoy, Dev_uni_proxima, Dev_uni_proxima), grouped by factors variables (mean,median, max, min, sum)

Features Selection / Engineering

- Feature transformations / engineering
  - Products clustering

Using product names to cluster products into 864 clusters

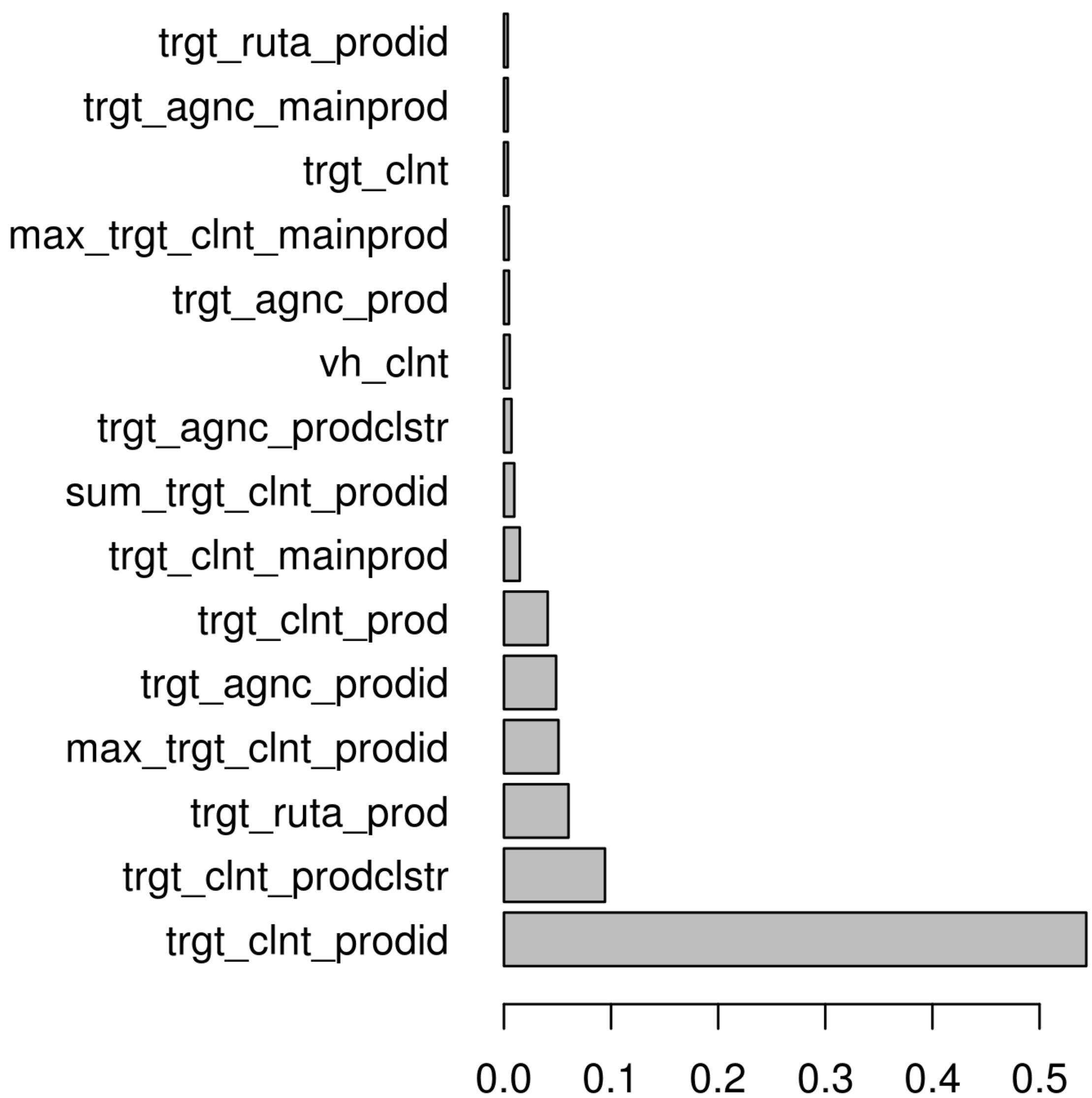
(3 products per cluster)



Features Selection / Engineering

- Best 5 features:
  - Mean target value per client and product
  - Mean target value per client and product cluster
  - Mean target value per route and product
  - Max target value per client and product
  - Mean target value per agency and product

Variable Importance Plot



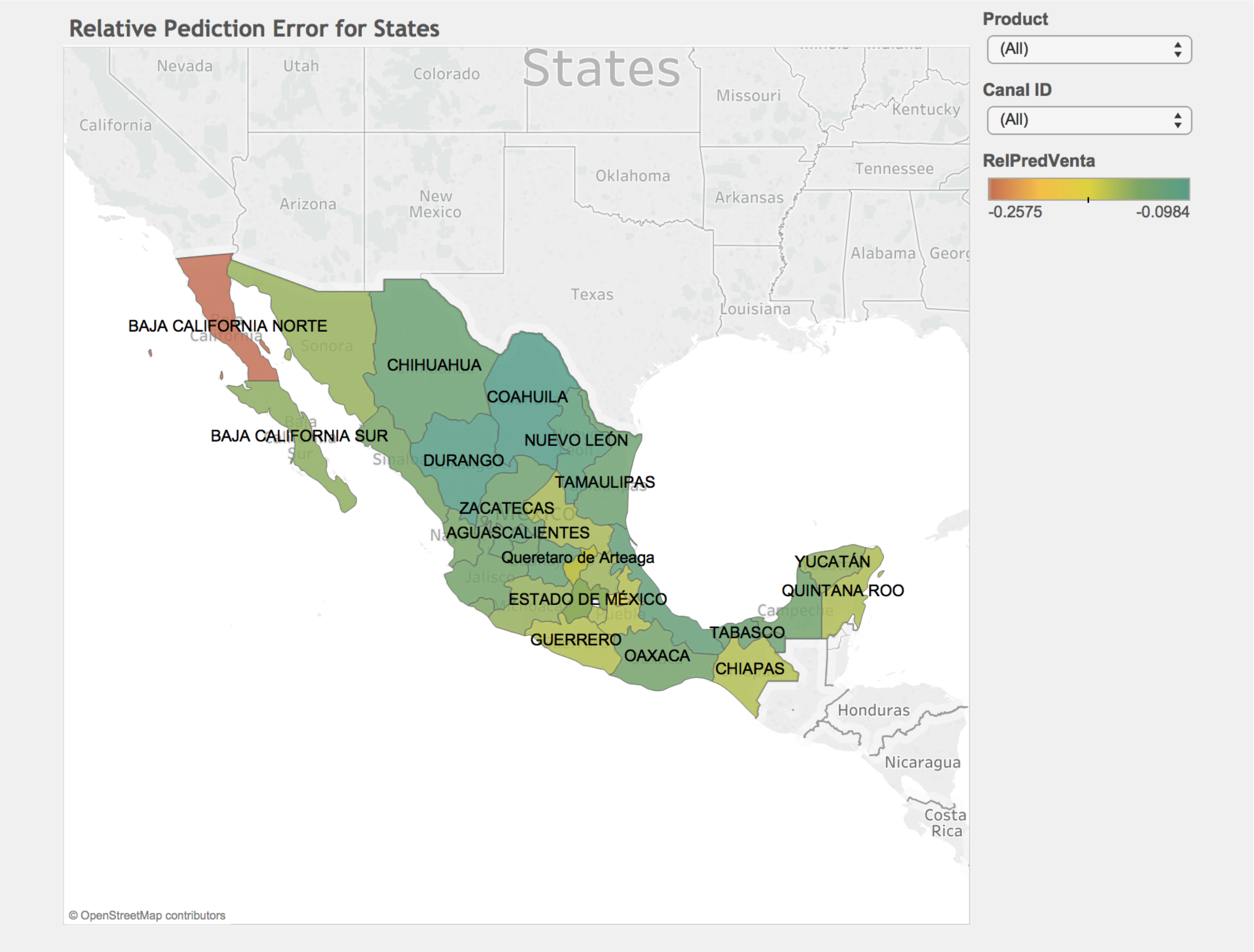
Training Methods

- 1st level: XGBoost build on full dataset and using features subsets and different target variables (Venta_hoy, Venta_uni_hoy, Dev_uni_proxima, Dev_proxima)
- 2nd level: Linear and ExtraTrees regressor
- 3rd level: Weighted average (weights based on LB feedback)

Some tricks for XGBoost

- After tuning your parameters you should adjust number of rounds (***nrounds***) for training on the whole dataset:
  - Validation ***nrounds*** = 1089 -> Full dataset train ***nrounds*** = 1903
- Reducing ***eta*** and increasing ***nrounds*** usually improve results:
  - ***eta*** = 0.025 -> ***eta*** = 0.0125
  - ***nrounds*** = 1903 -> ***nrounds*** = 3806

Important and Interesting Findings



Calculated for 9th week based on 7th week data

Simple Model

- XGBoost model can be build using only top 50 features without significant loss of quality
- Best single XGBoost:
  - 0.43794 / 0.45171 (17th place on private LB)
- XGBoost on 175 features:
  - 0.43487 / 0.45316 (19th place on private LB)

What else to try?

- Categorical embedding:
  - https://github.com/entron/entity-embedding-rossmann
  - https://arxiv.org/pdf/1604.06737v1.pdf
- FTRL and Factorization Machines

Thank you!

Q & A