May 2022

MI Data – Analytics Case Study for Miya Wang

Agenda

- Project goal
- Results & conclusions
- Solution (V1.0)
  - Workflow design
  - Assumptions
    - Features
    - Model
    - Metrics
- Error analysis
- Weather impact
- Production
- Next steps

Project Goal

Provide a solution to predict the daily 311 inbound calls percent change for the next 7 days

Validate predictive power of weather data

Results & Conclusions

Model performance

Test set: 2018 (365 days)

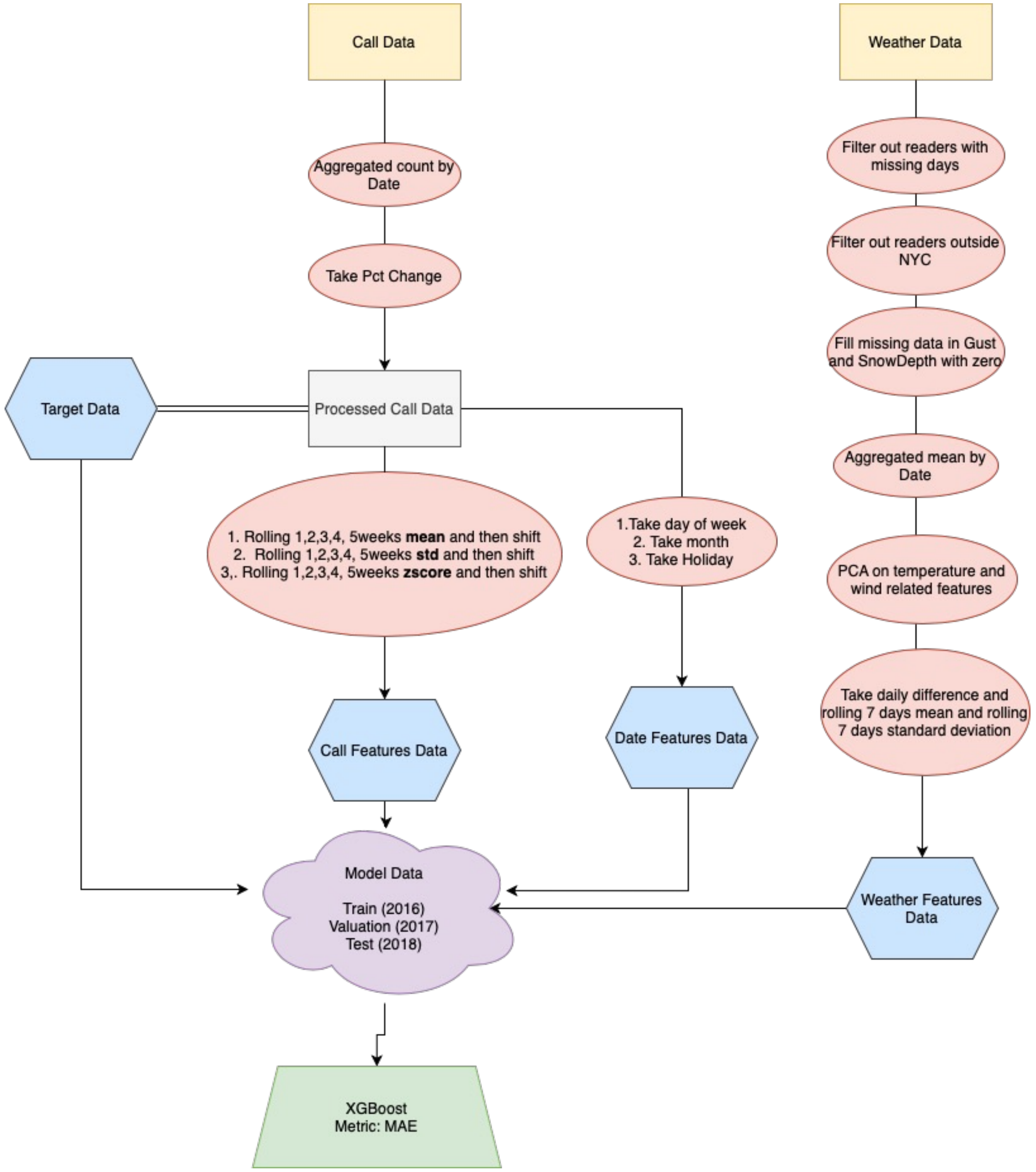
Performance:

- Mean absolute error (MAE) **0.03727** VS baseline 0.1391
- Mean squared error (MSE) **0.003156** VS baseline 0.03704
- Correctly predict direction of change **93.42%** of the time VS baseline 39.51%

Weather data

By reducing MSE by **at least 14.75%,** weather data has shown predictive power especially for "outlier" days

Workflow Architecture



Assumptions

- Data feed
- Features
- Model
- Metrics

DATA FEED

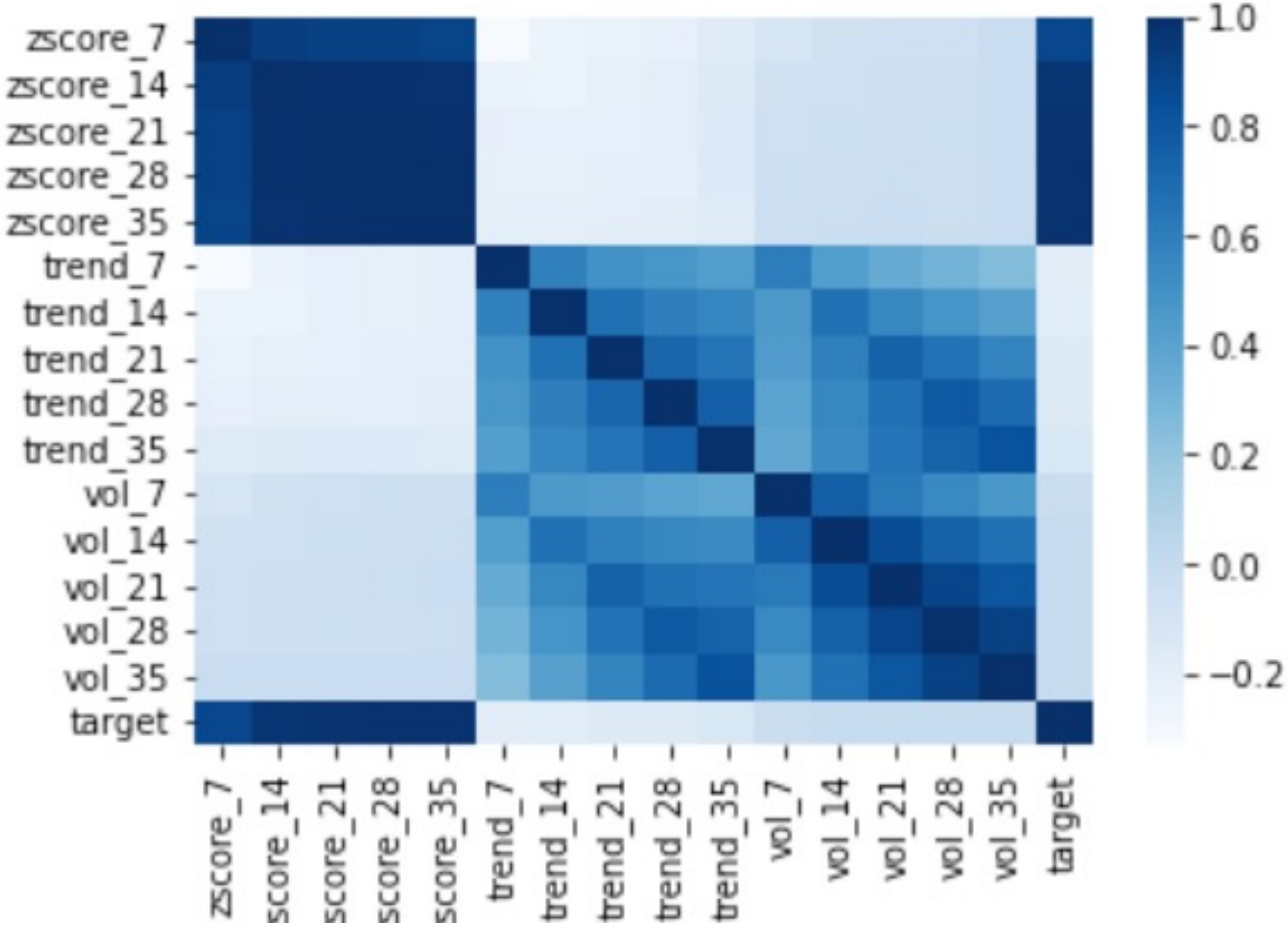
Data latency: 1 day

Input data for day d^t will be ready by EOD day d^{t-1}

Prediction for day d^t will be ready by EOD day d^{t-1}

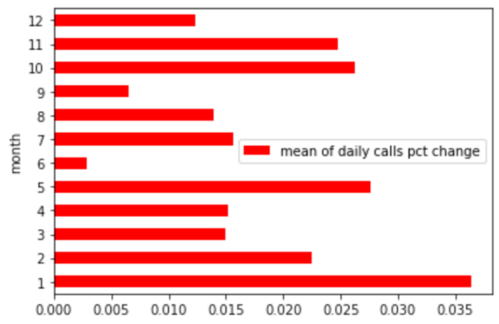
FEATURES: call

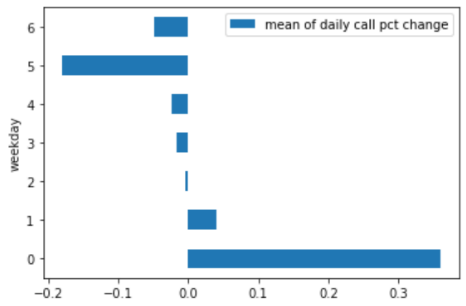
- We assume predictive power

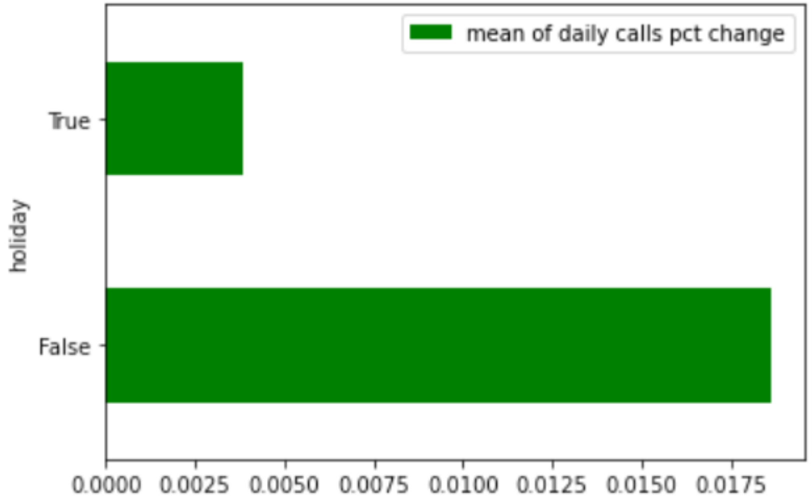


FEATURES: date

- We assume predictive power







FEATURES : weather

We assume predictive power

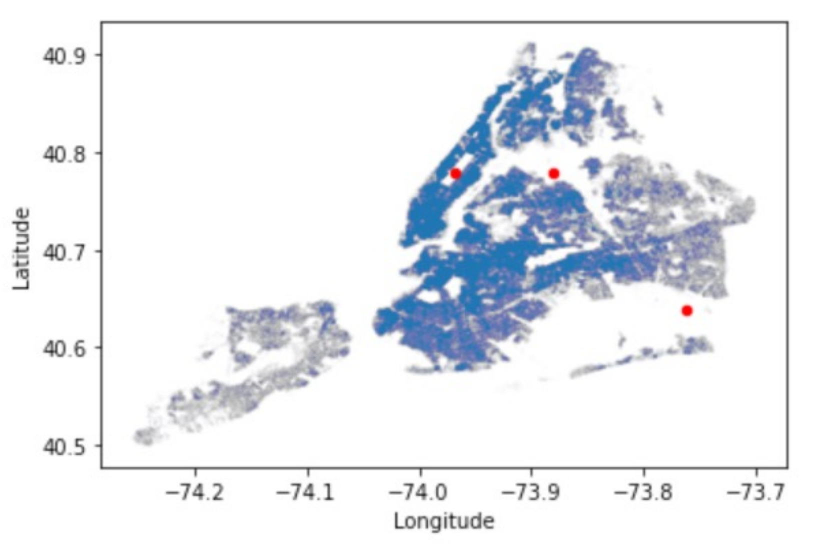
- Direct impact:
  - Heat/cold water (8.7% of calls)
  - Air quality
  - etc. *(we can probably figure those all out using NLP)*
- Indirect impact
  - Traffic
    - DOT (Department of Transportation) handles over 10% of calls

We assume good data quality

- 3 readers in NYC covering all the sample days (3238 days) will continue provide reliable data reads.
- Missing values won't impact significantly data predictive power
  - Among 3 readers, two miss Percipitation for 19 days and one miss WindSpeed and MaxSustainedWind for 53 days.
  - No day when all three readers miss.

We assume representative weather

- Mean of reads from 3 readers represent daily weather
- Those 3 readers provide presentative weather data for NYC area



MODEL - XGboost

Learn nonlinearity

Learn from noise data

- Complaints submitted to 30 agencies in 200+ categories

Learn from weak learners

Learn from non-sparse data

Robust to multicollinearity

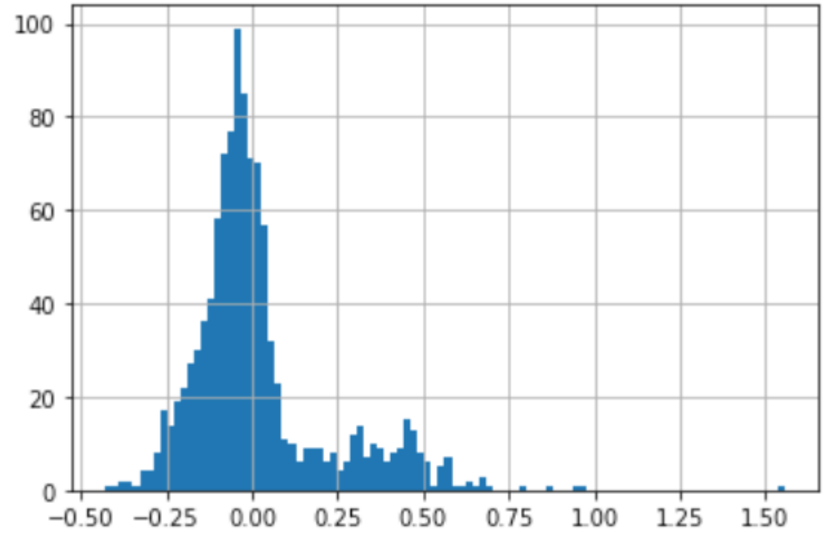
Generalize

- Distribution of daily change of calls is rightly skewed - Spikes are common
  - Regularization
  - Cross-validation
  - Shrinkage
  - Column subsampling
  - Learning curve
  - Etc.

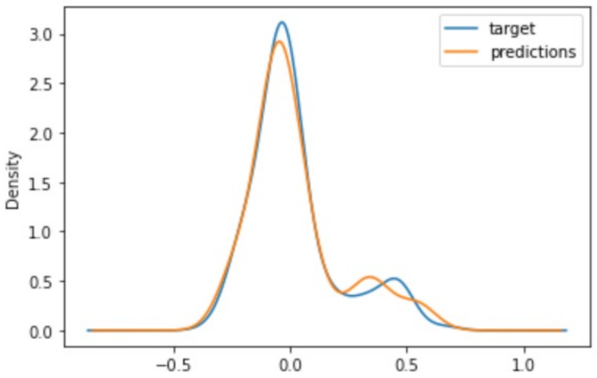
Fast execution

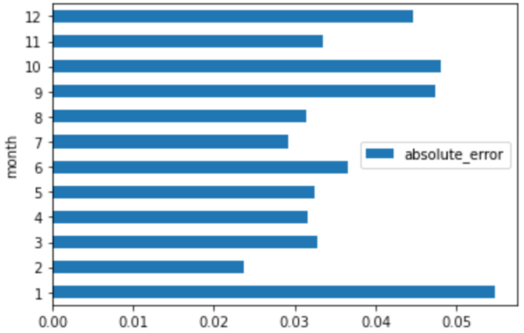
Metrics

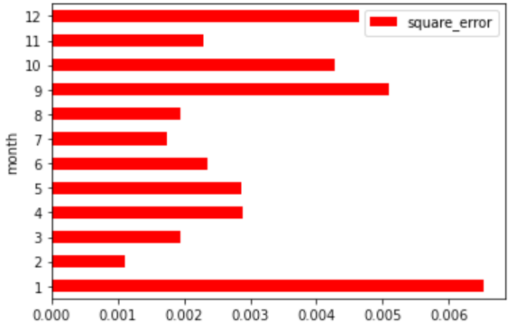
- Sensitive to outliers
  - Squared error as loss function & evaluation metric for validation sets

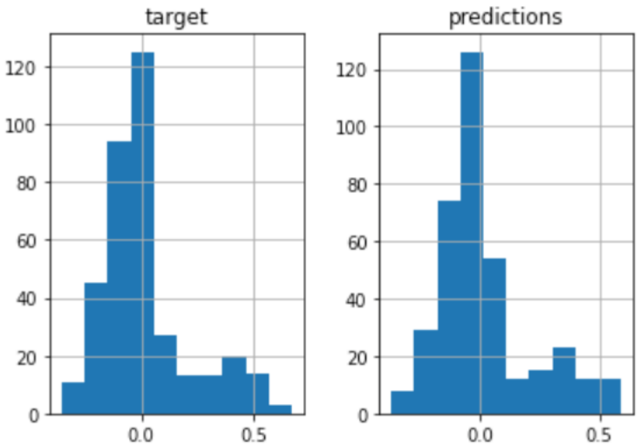


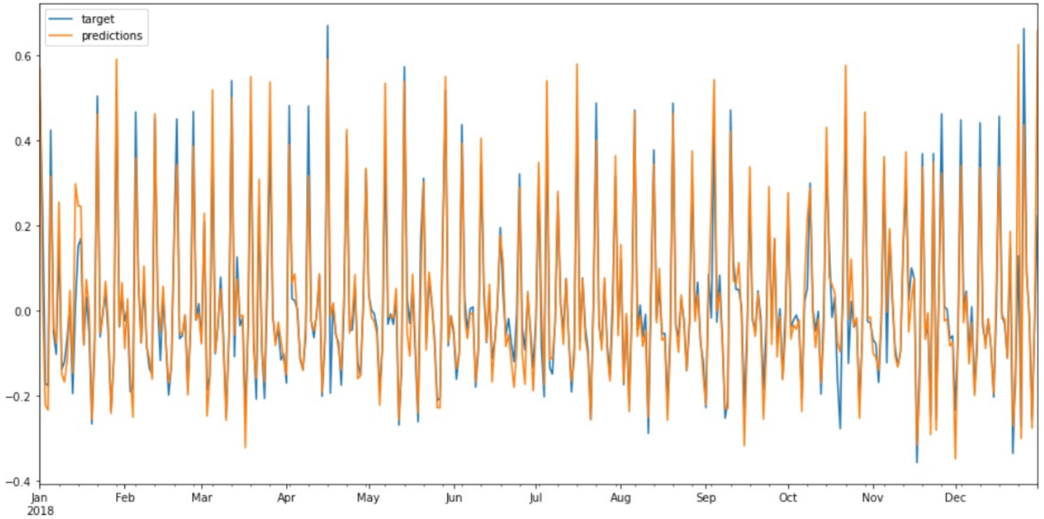
ERROR ANALYSIS

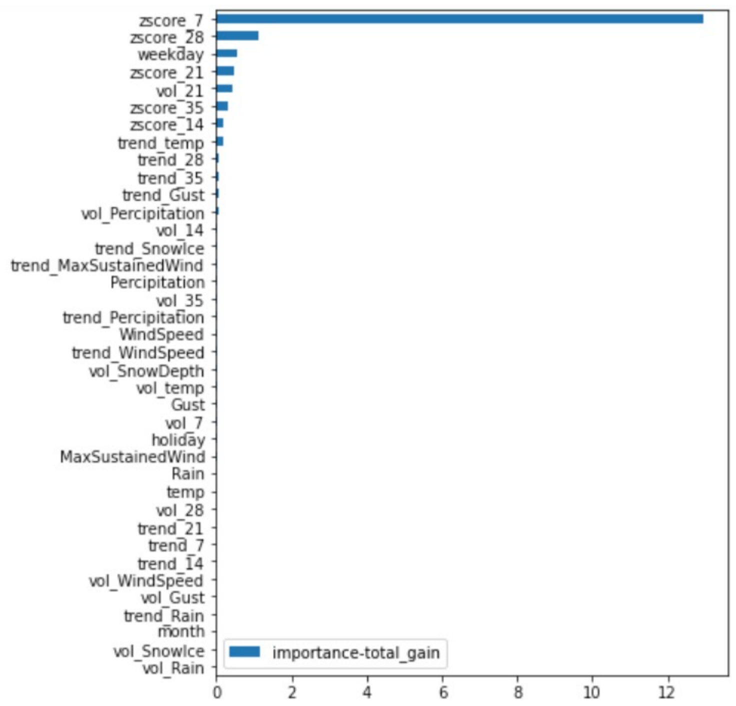


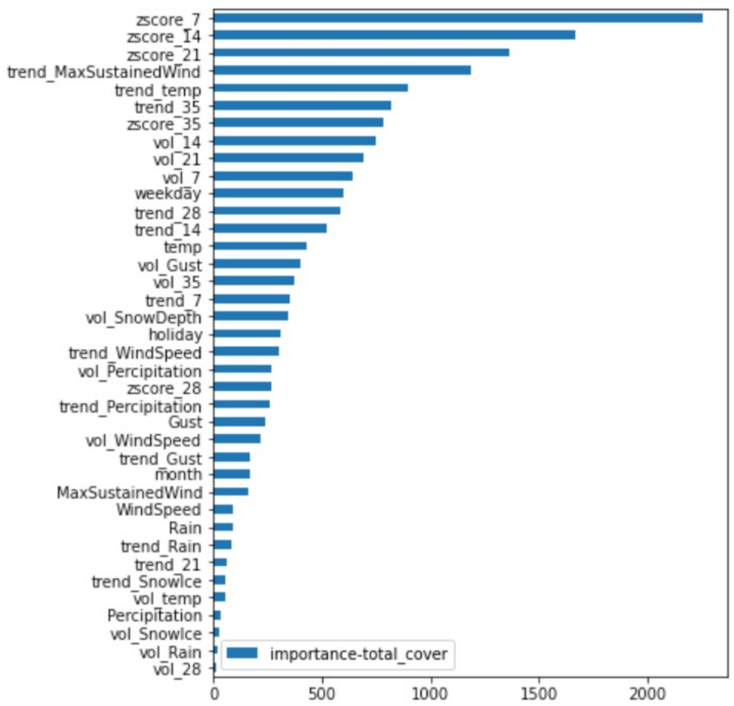


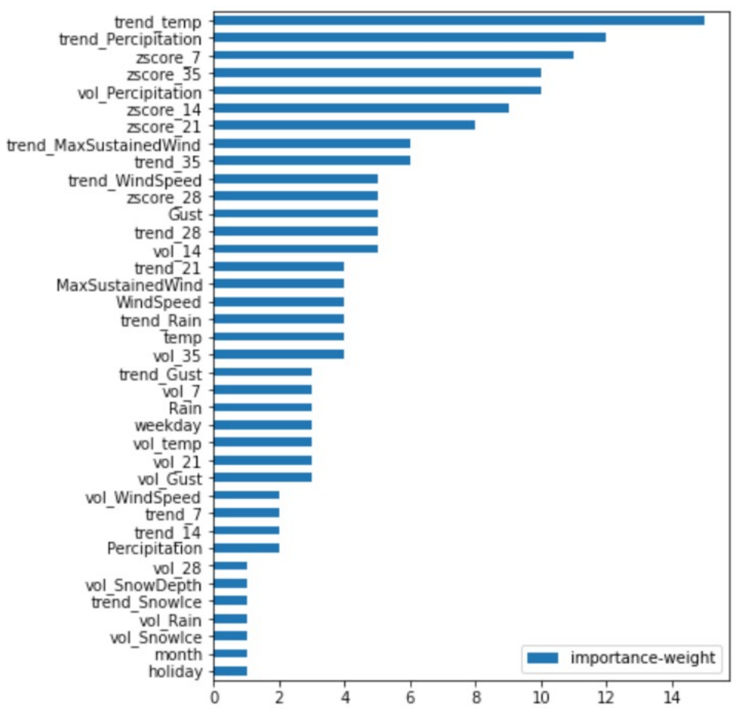












WEATHER IMPACT: model score

- Call data + date data:
    - MAE: 0.03039
    - MSE: 0.003702
- Call data + date data + **weather data** (REPORTED)
    - MAE: 0.03727 (-22.63%)
    - MSE: 0.003156 (+14.75%)
- Call data + date data + **selected weather data** (daily temperature change moving average & daily precipitation change volatility):
    - MAE: 0.03069 (-0.98%)
    - MSE: 0.002373 (+35.89%)

WEATHER IMPACT: performance

PRODUCTION

Monitor daily data feed

Refresh model/dynamic modeling

Multiple Run

Monitor output

NEXT STEPS

- More feature engineering for weather data
- Model optimization
- Different training/test time windows
- Region/Agency breakdown prediction