kaggle

WiDS Datathon 2023
Team zn_k

Zeyneb N. Kaya

Agenda

Agenda

1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Background

Zeyneb N. Kaya

- Junior, Saratoga High School (CA, USA)
- WiDS Student Ambassador
- National Winner, NCWIT Aspirations in Computing Award
- 2022 Datathon competitor
- Data Science Certifications
  - Data Science and Machine Learning Certificate
  - NLP Specialization Certificate
- Machine Learning Research
  - NLP, Data analysis (published @ EACL, UCB TextXD...)

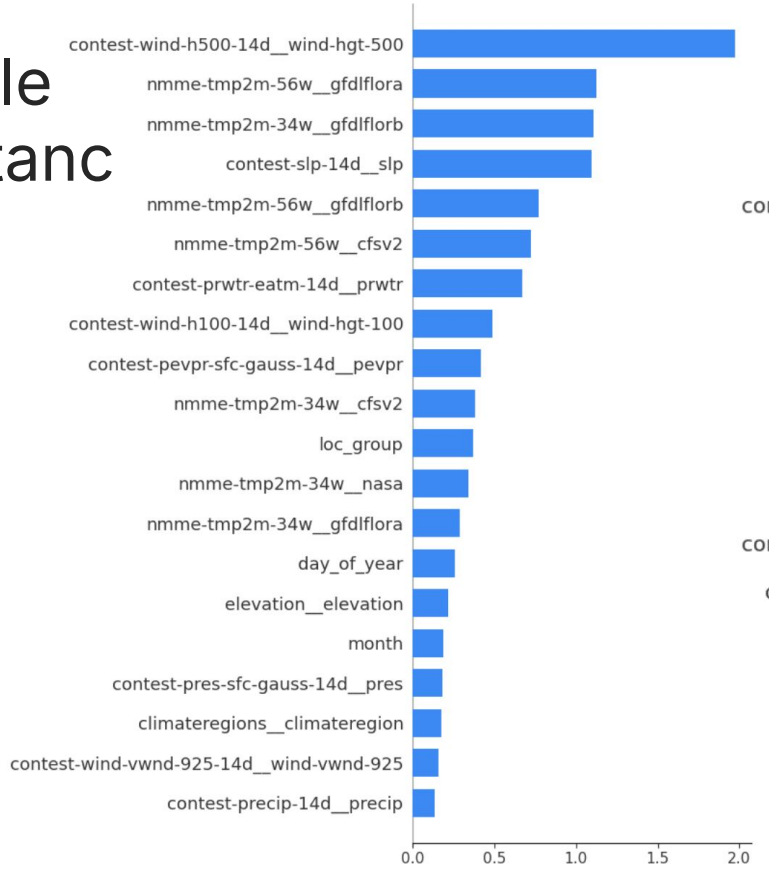
Overview

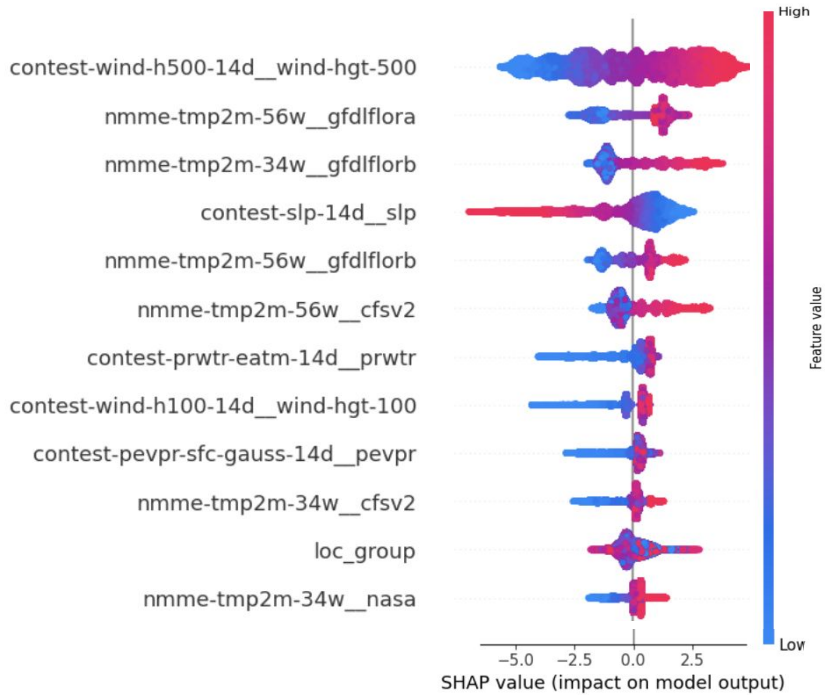
- Models: Gradient Boosting
  - CatBoost, LightGBM
- Key Features: Forecasts
  - nmme0-tmp2m-34w___
- Runtime: ~1 hour
- Key Method: “Iterative Pseudolabeling”

Features Selection/ Engineering

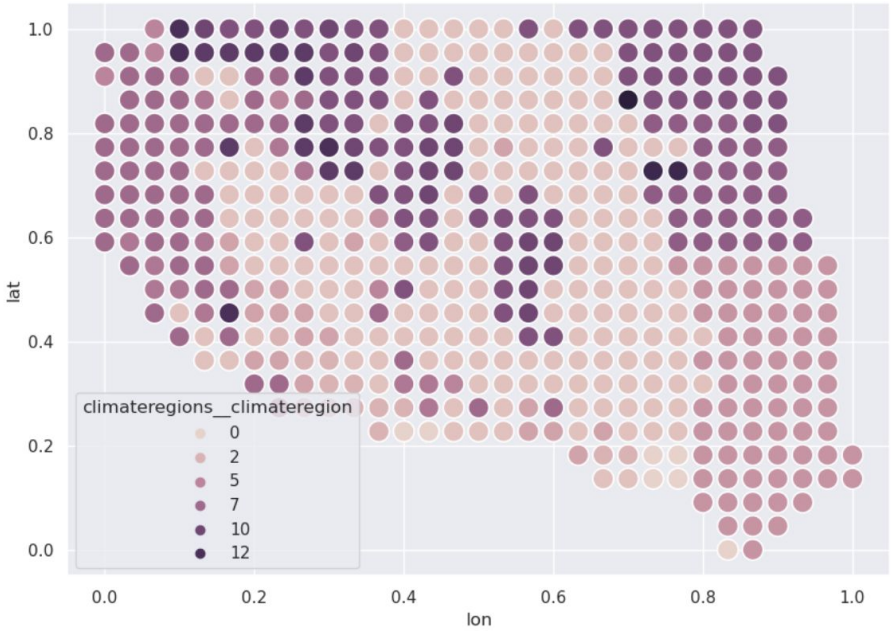
- Most important feature types
  - Forecast features
  - Wind
  - Location
- Feature Engineering
  - loc_group: number each lat-lon location
  - label encoded climate region
  - year/month/day from startdate
- Feature Selection
  - Drop highly correlated features
  - Note: categorical features *not* indicated in CatBoost

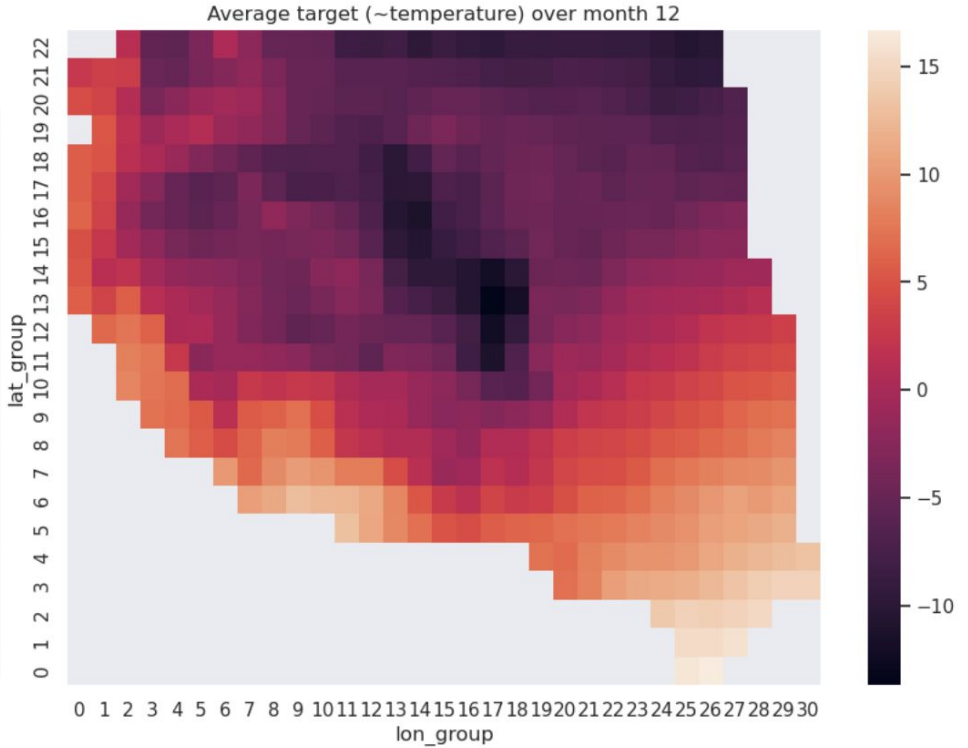
Variable Importance



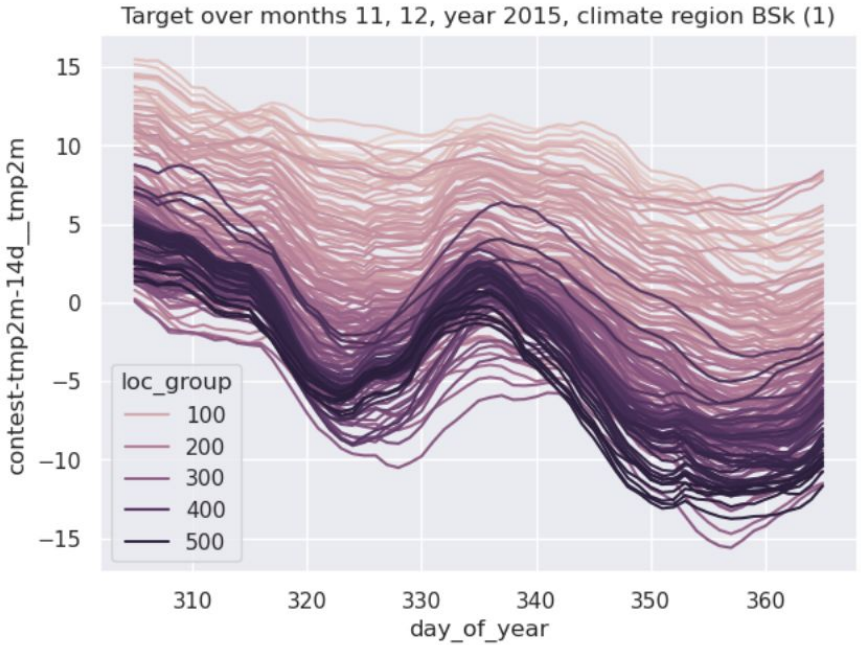


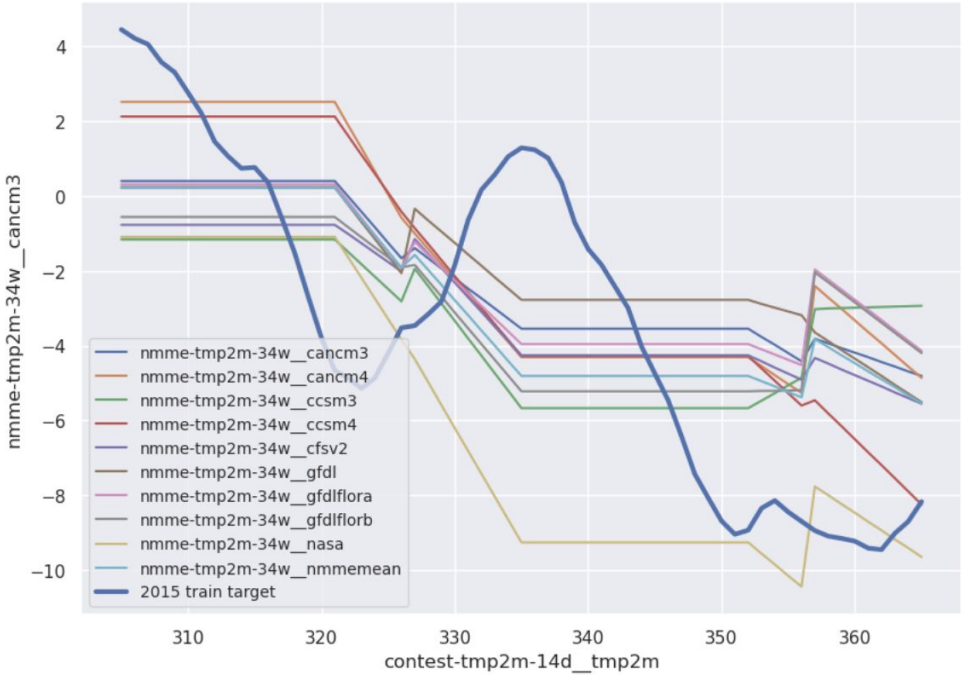
Features Selection/Engineering





Features Selection/Engineering

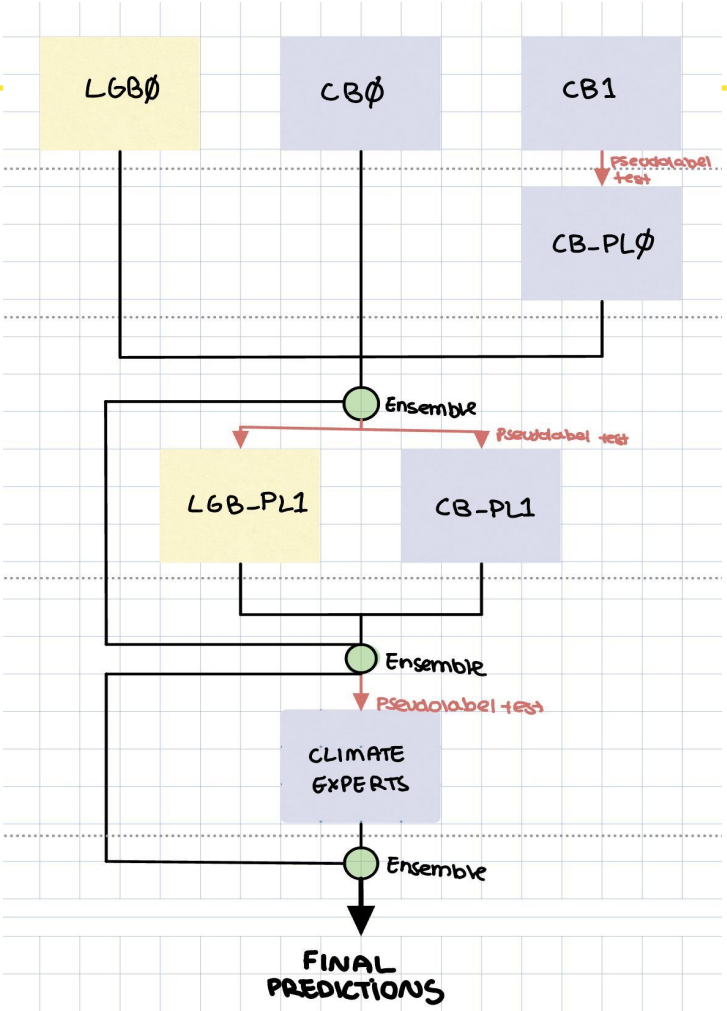




Training Methods

- Climate Region Experts
  - One set of models ensembled are “experts” in only their particular climate region. They train with *all* the data, but the points from their own “expert” region have a higher sample weight. When predicting, use predictions from the corresponding expert model.
- “Iterative Pseudolabeling”
  - CatBoost + LGBM ensembled, then predictions are generated on test set and used as pseudolabels
  - A threshold is set in ensemble (only ensemble if the absolute difference to the predicted values in previous round is close). If they differ a lot then use the new model. The idea is that it favors the most recent model.

Training Methods



Important and Interesting Findings

What sets me apart?

- Interest to both learn and apply!
  - Build upon ideas while also being creative to further it ☺️

Interesting thing found while exploring the data?

- (drumroll, please)...prediction is that the anonymized region is mid + west USA
  - From climate region, observe a very peculiar pattern in the climate regions—very similar to that in mid+west US
  - Scale lat + lon between this area and plot true climate regions
  - Seemed to match?...hmmm
  - *P.S. not used at all, just a bonus observation!*

Other Experiments

- TabNet, RNN
- Data augmentation (GAN, noise)
- Predicting forecast error

Simple Model

Simple Model

- CatBoost + Feature Engineering + Tuned Hyperparameters
- RMSE ~0.8

Question and Answer


kaggle