WiDS Datathon 2023

https://www.kaggle.com/competitions/widsdatathon2023

#wids

1

General tips

- Browse through the Datathon specific tutorials before it starts
- Check the Discussion and Code tabs early on
- Check to make sure your training data looks the same as test data
  - This year the test data had some rounding issue preventing some latitude and longitude from matching matching up

2

Special Challenges

Your task is to predict the arithmetic mean of the maximum and minimum temperature over the next 14 days, for each location (514) and start date.

- So many features (274) with many very highly correlated with each other.
- No actuals for historical data - we are forecasting based on forecasts, and "averages of averages" of forecasts.
- Relatively speaking small duration of training data.
- There is a long gap in when training period ends and test hold out starts
  - Training: Sept 2014 - Aug 2016
  - Scoring: Nov 2022 – Dec 2022 (5+ years later)
- Not sure what to do with location…leave lat and lon? Convert to categorical? Train model per location?

3

The Journey Begins

- Started with feature analysis and used Random Forest just to get a baseline and idea on feature importance. Various other Boosted Trees gave same performance or worse. (RMSE 1.701)
- CatBoostRegressor (1.332)
  - Really fast training, good defaults
  - Seemed to like all the variables!?!
  - Preferred all locations together, with location as a categorical variable
- More feature engineering, and parameter tuning (1.135)
  - 0.376 local RMSE on test split
  - Hyperparameters = {'iterations': 3000, 'learning_rate': 0.098, 'subsample': 0.744, 'l2_leaf_reg': 2.372, 'max_depth': 6, 'loss_function': 'RMSE', 'model_size_reg': 0.483}

4

Various new features based on the following didn't have much effect:

- Rolling averages, min/max temps
- StdDev
- Spreads

Shift and Diffs helped us…

```
# trying to predict next 14 day average might be good to take the 3-4 week prediction from 2-3 weeks ago
df['nmme-tmp2m-34w__nmmemean_2wprior'] = df.groupby(['loc_group'])['nmme-tmp2m-34w__nmmemean'].shift(-14)
df['nmme-tmp2m-34w__nmmemean_3wprior'] = df.groupby(['loc_group'])['nmme-tmp2m-34w__nmmemean'].shift(-21)

# trying to predict next 14 day average might be good to take the 5-6 week prediction from 3-4 weeks ago
df['nmme-tmp2m-56w__nmmemean_3wprior'] = df.groupby(['loc_group'])['nmme-tmp2m-56w__nmmemean'].shift(-21)
df['nmme-tmp2m-56w__nmmemean_4wprior'] = df.groupby(['loc_group'])['nmme-tmp2m-56w__nmmemean'].shift(-28)
```

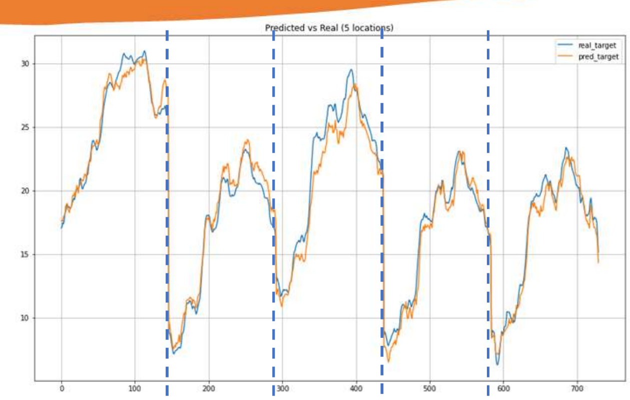
```
# add a feature to indicate the diff between days (is temp going up or down)
df['nmme-tmp2m-34w__nmmemean_5dayDiff'] = df.groupby(['loc_group'])['nmme-tmp2m-34w__nmmemean'] \
          .transform(lambda g: g.shift(5) - g)
```

- Forward filled the nans created from the shifts (sorted by location and time)
- Backfilled the nans created from the diff

5

The Journey Continues

- XGBoost Regressor (1.412)
  - More tentative approach to modeling: e.g. started w/ 1 location then expanded, similar with columns
  - Added 1-day lag features
  - Training time was reasonable ~10-20 minutes
- Performance appeared to degrade with addition of columns, but did not pursue
- Basic hyper-parameters
  - Estimators: 1800
  - Learning Rate: 0.1
  - Early-stopping rounds: 50
- Metrics
  - R2: 97.8%
  - RMSE: 0.938 local test split



Visualizing sample predictions: 5 locations

6

The Journey Continues

- SarimaX…. *seemed* like a good idea
  - Vast improvement over Univariate model (Sarima) but…
  - 5 year gap from training to scoring window
  - All exogenous variables need to be forecast during the gap
  - Nearly endless combinations of variables… which would all need to be forecast… with model parameters
  - Occasionally the models predicted wildly inaccurate temperatures

7

Inspect the AIC/BIC…

Less is more…

- AIC – prediction ability
- BIC – quality of model

There is interplay between the features based on presence of other features and model parameters

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:           contest-tmp2m-14d__tmp2m   No. Observations:                  105
Model:             SARIMAX(0, 0, 2)x(1, 1, [], 52)   Log Likelihood                   6.681
Date:                            Fri, 03 Mar 2023   AIC                              0.639
Time:                                    18:10:23   BIC                            -13.361
Sample:                                         0   HQIC                              -inf
                                            - 105
Covariance Type:                              opg
==========================================================================================
                                    coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------------------
contest-slp-14d__slp             -0.0193   4.71e-11  -4.09e+08      0.000      -0.019
contest-wind-h850-14d__wind-hgt-850   0.2249      9e-12    2.5e+10      0.000       0.225
nmme-tmp2m-34w__nmmemean_2wprior  0.0451   7.32e-13   6.16e+10      0.000       0.045
```

8

and Exogeneous Features Significance

Significant feature defined as:

- P>|z| less than 0.05

Model selected impacts features used

- 6 significant
- 4 not significant

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:           contest-tmp2m-14d__tmp2m   No. Observations:                  105
Model:             SARIMAX(0, 0, 2)x(2, 0, [1], 52)   Log Likelihood                -129.323
Date:                            Thu, 09 Mar 2023   AIC                            290.647
Time:                                    08:14:03   BIC                            333.110
Sample:                                         0   HQIC                           307.854
                                            - 105
Covariance Type:                              opg
==========================================================================================
                                      coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------------------
contest-wind-h500-14d__wind-hgt-500     0.0734      0.014      5.271      0.000       0.046
contest-slp-14d__slp                   -0.0033      0.001     -4.698      0.000      -0.005
contest-wind-h850-14d__wind-hgt-850    -0.0482      0.019     -2.602      0.009      -0.084
wind-vwnd-925-2010-15                   0.0094      0.008      1.187      0.235      -0.006
wind-uwnd-250-2010-18                  -0.0067      0.002     -2.782      0.005      -0.011
icec-2010-9                             2.2265      0.723      3.079      0.002       0.809
nmme0-prate-56w__cancm30               -0.0296      0.025     -1.195      0.232      -0.078
nmme-tmp2m-34w__nmmemean_2wprior       -0.1115      0.104     -1.076      0.282      -0.315
nmme-tmp2m-34w__nmmemean_3wprior        0.4467      0.095      4.682      0.000       0.260
month                                  -0.0270      0.082     -0.328      0.743      -0.189
```

9

and Exogeneous Features Significance

Model changed relevant features:

- **4 significant**
- **6 not significant**

Dropping / adding features will also change significance of other features

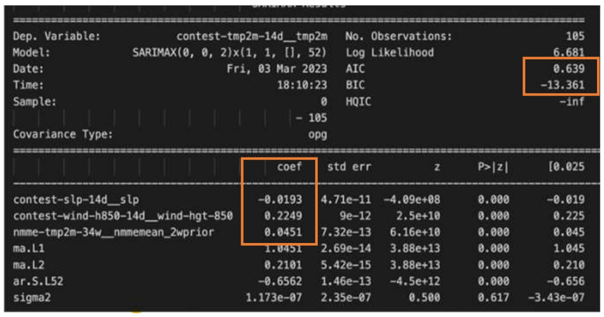
```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:           contest-tmp2m-14d__tmp2m   No. Observations:                  105
Model:             SARIMAX(0, 1, 0)x(1, 0, 0, 52)   Log Likelihood                 -76.679
Date:                            Thu, 09 Mar 2023   AIC                            177.358
Time:                                    08:23:06   BIC                            209.090
Sample:                                         0   HQIC                           190.214
                                            - 105
Covariance Type:                              opg
==========================================================================================
                                      coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------------------
contest-wind-h500-14d__wind-hgt-500    -0.0182      0.006     -2.968      0.003      -0.030
contest-slp-14d__slp                   -0.0220      0.001    -22.801      0.000      -0.024
contest-wind-h850-14d__wind-hgt-850     0.2852      0.016     17.488      0.000       0.253
wind-vwnd-925-2010-15                   0.0013      0.003      0.432      0.666      -0.004
wind-uwnd-250-2010-18                  -0.0014      0.001     -1.385      0.166      -0.003
icec-2010-9                             1.3513      0.425      3.181      0.001       0.519
nmme0-prate-56w__cancm30               -0.0186      0.012     -1.609      0.108      -0.041
nmme-tmp2m-34w__nmmemean_2wprior       -0.0014      0.063     -0.022      0.982      -0.124
nmme-tmp2m-34w__nmmemean_3wprior        0.0591      0.072      0.825      0.409      -0.081
month                                  -0.0057      0.037     -0.156      0.876      -0.078
```

10

Parameters for SarimaX influence weight of exogenous variables

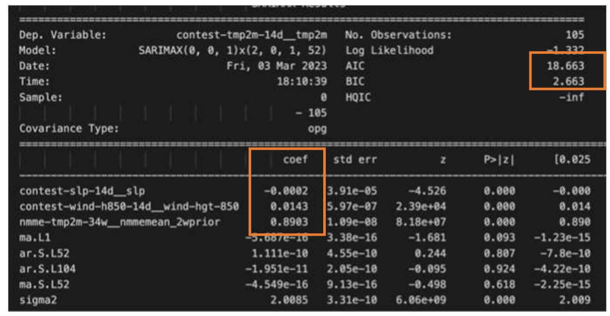
- Orders: (0,0,2) – (1,1,0,52)
- Highest impact: wind-hgt-850

- Orders: (0,0,1) – (2,0,1,52)
- Highest impact: 2wprior



Dep. Variable:	contest-tmp2m-14d__tmp2m	No. Observations:	105
Model:	SARIMAX(0, 0, 2)x(1, 1, [], 52)	Log Likelihood	6.681
Date:	Fri, 03 Mar 2023	AIC	0.639
Time:	18:10:23	BIC	-13.361
Sample:	0	HQIC	-inf
	- 105		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025
contest-slp-14d__slp	-0.0193	4.71e-11	-4.09e+08	0.000	-0.019
contest-wind-h850-14d__wind-hgt-850	0.2249	9e-12	2.5e+10	0.000	0.225
nmme-tmp2m-34w__nmmemean_2wprior	0.0451	7.32e-13	6.16e+10	0.000	0.045
ma.L1	1.0451	2.69e-14	3.88e+13	0.000	1.045
ma.L2	0.2101	5.42e-15	3.88e+13	0.000	0.210
ar.S.L52	-0.6562	1.46e-13	-4.5e+12	0.000	-0.656
sigma2	1.173e-07	2.35e-07	0.500	0.617	-3.43e-07

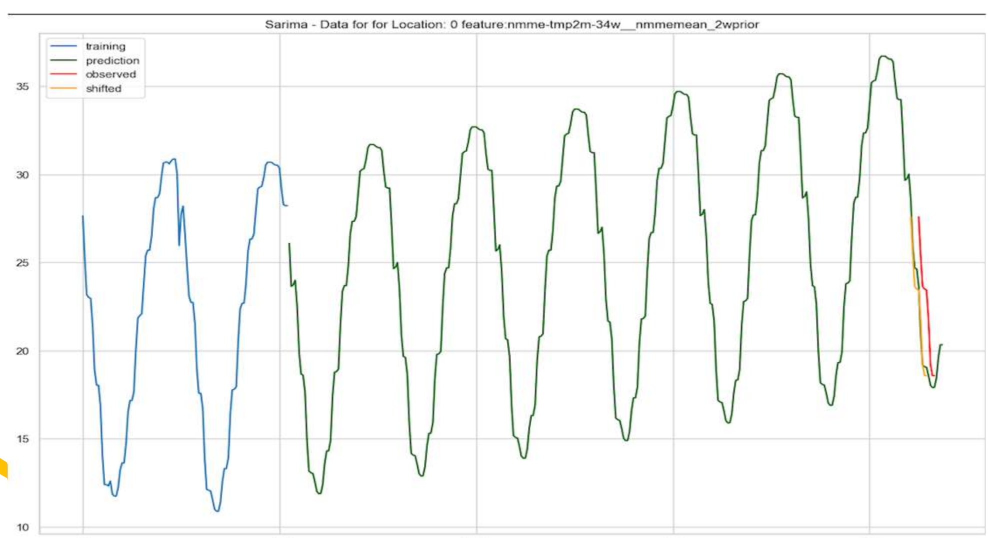


Dep. Variable:	contest-tmp2m-14d__tmp2m	No. Observations:	105
Model:	SARIMAX(0, 0, 1)x(2, 0, 1, 52)	Log Likelihood	-1.332
Date:	Fri, 03 Mar 2023	AIC	18.663
Time:	18:10:39	BIC	2.663
Sample:	0	HQIC	-inf
	- 105		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025
contest-slp-14d__slp	-0.0002	3.91e-05	-4.526	0.000	-0.000
contest-wind-h850-14d__wind-hgt-850	0.0143	5.97e-07	2.39e+04	0.000	0.014
nmme-tmp2m-34w__nmmemean_2wprior	0.8903	1.09e-08	8.18e+07	0.000	0.890
ma.L1	-5.687e-16	3.38e-16	-1.681	0.093	-1.23e-15
ar.S.L52	1.111e-10	4.55e-10	0.244	0.807	-7.8e-10
ar.S.L104	-1.951e-11	2.05e-10	-0.095	0.924	-4.22e-10
ma.S.L52	-4.549e-16	9.13e-16	-0.498	0.618	-2.25e-15
sigma2	2.0085	3.31e-10	6.06e+09	0.000	2.009

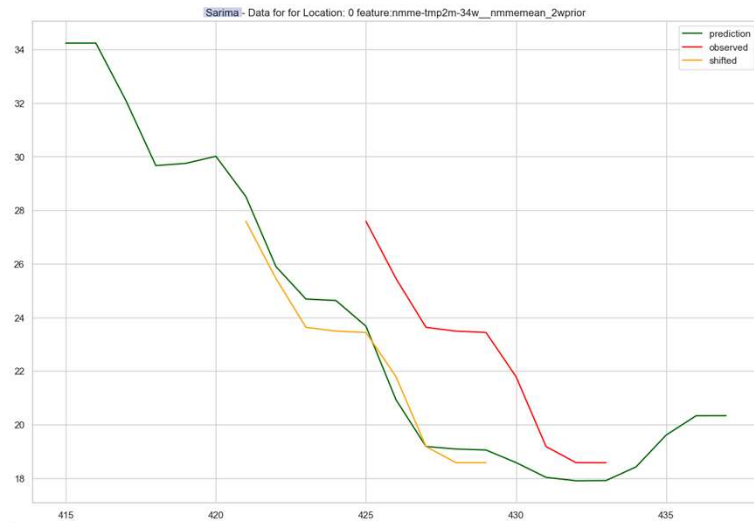
11

2wk Prior: Training + Forecasting + Observed



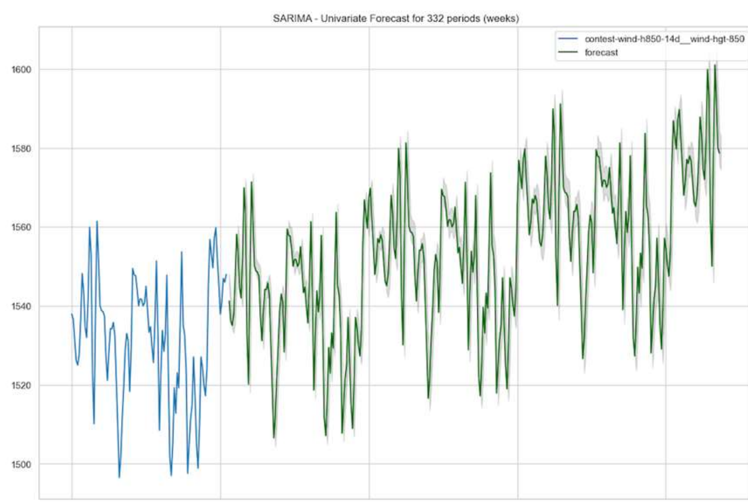
12

2wk Prior: maybe it is out of phase?



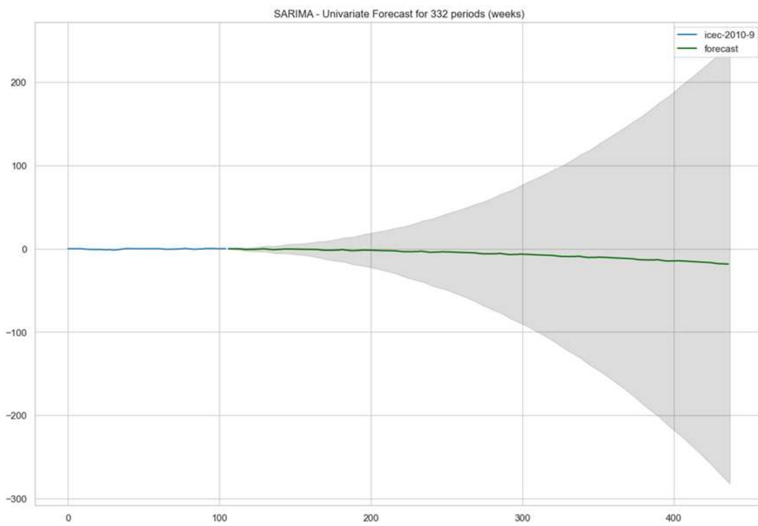
13

Confident in its wind prediction?

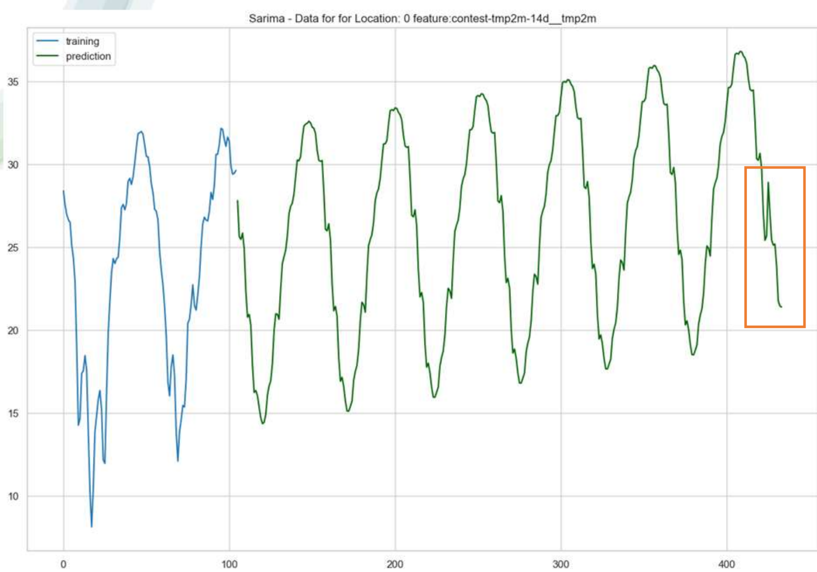


14

Not so confident on sea ice concentration



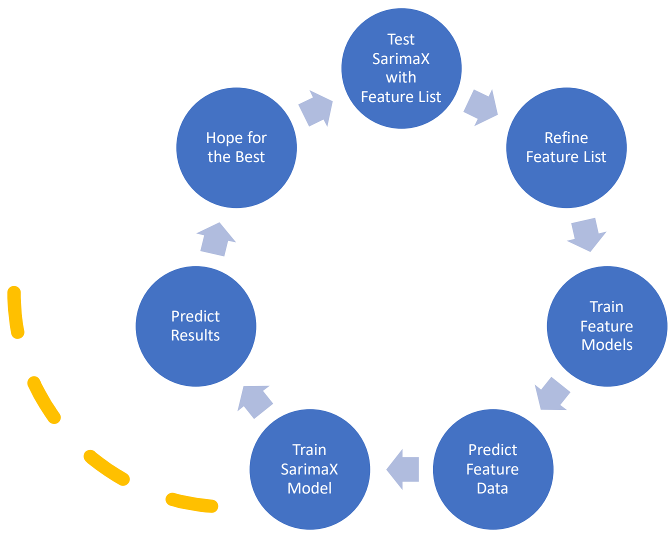
15



Prediction for Location 0

16

Process Flow



- Landed on weekly model (m=52)
- Training / Predicting occurred on each of the 514 Locations
- Model used to select features used for SarimaX training and prediction per location
- Combination of features and model orders almost infinite
- Should have instantiated a better pipeline for iterations

17

SarimaX – Take Aways

- Use auto_arima and your own *grid search* with caution
  - Model is invalid with BIC of:
    - -inf or Nan
  - Nearly identical models should be tried (vs best score only)
  - Add a check for reasonable prediction
  - Issues with exogenous variables on auto_arima
- Exogenous variables are valuable
  - Sarima = 13.7
  - Typical SarimaX = 2.9 to 3.8
  - Miracles = 1.33 / 1.28

18

Best practices we didn't practice ;)

1. Using a too many features which were highly correlated with each other (Cat Boost)
2. Leaving data "in time order" when doing train-test splits (for tree-based algorithms)

```
# Creating the Training and Test set from kaggle train data
X_train_loc, X_test_loc, y_train_loc, y_test_loc =
        train_test_split(X_loc, y_loc, test_size = 0.25, shuffle=False, random_state = 21)
```

```
# grid-search over a parameter grid
custom_cv = TimeSeriesSplit(n_splits=5)
search = GridSearchCV(estimator=estimator, param_grid=parameters, scoring='neg_root_mean_squared_error', \
                      cv=custom_cv, n_jobs=1)
search.fit(X_train, y_train)
```

3.

19

Destination – Rank 22/697 Top 3%

T-5 Days…. Best RMSE by approach (individually ranking ~350 – 50 percentile)

1. SarimaX – 1.284 (predicted highest temperatures)
2. CatBoost – 1.135 (predicted lowest temperatures)
3. XGB – 1.412

Ensemble Model – weighted average for values from above

- 45 / 35 / 20

Final Score / Ranking

- 0.727 (Public - data provided during the contest duration = 50% of data)
- **0.718 (Private – final scoring)**

20