### **Table of Contents**

1 Notebook 2 - Introduction

- 1.0.1 Data Reduction Function1.0.2 Set up X and y1.0.3 Time Series Split1.0.4 Scaling Training Set and Test Set
  - 1.0.5 Initial MLPRegressor Evaluation
    - 1.0.5.1 Loss vs. Iteration Analysis
  - 1.0.6 MLPRegressor Grid Search Best Parameters and Run Time
  - 1.0.7 MLP Grid Search Evaluation
  - 1.0.8 MLPRegressor 200 Iterations with "Optimized Parameters"
  - 1.1 RFR Submission
    - 1.1.1 Transform Test\_Data
    - 1.1.2 Fit RFR on Full training\_data from WiDS
- 2 Notebook 2 Conclusion

# Notebook 2 - Introduction

Workflow Initial MLPRegressor --> MLPRegressor Optimization --> MLPR vs. RFR Comparison--> RFR Baseline WiDS Submission

We have refined our ability to prepare and evaluate models in this notebook. We adopted a data reduction strategy to reduce each parameter's data type to minimize memory load and run times. We will run our MLPRegressor on Google Collab and on my local machine to compare run times to establish more efficient methodology moving forward.

We will optimize a Multi-layer Perceptron Regressor(MLPR or MLPRegressor) as they have shown success in predictive modeling with times-series data. We will compare our scoring metrics from the MLPRegressor to our optimized Random Forest Regressor(RFR) from Notebook 1.

We will make a submission to WiDS Kaggle Submission page to give ourselves a true baseline to improve upon moving forward on this project.

```
In [1]:
#Import python libraries
import numpy as np
import pandas as pd
```

Out[4]:

```
import matplotlib.pyplot as plt
import seaborn as sns
scom datetime import datetime
scom math import sqrt
```

```
In [3]: # Read in clean training set with start date as datetime index
time_training_data =
pd.read_csv('data/time_training_data_clean.csv',
index_col='startdate')
```

```
In [4]: [time_training_data.head()
```

contest-

, d c [ 1 ] 1		index	lat	lon	pevpr-sfc- gauss- 14dpevpr	nmme0- tmp2m- 34wcancm30	nmme0- tmp2m- 34wcancm40	nm tm 34wcc
	startdate							
	2014- 09-01	0	0.000000	0.833333	237.00	29.02	31.64	
	2014- 09-01	290938	0.818182	0.633333	323.63	24.18	26.75	
	2014- 09-01	35819	0.227273	0.900000	385.92	31.16	32.19	
	2014- 09-01	290207	0.818182	0.600000	303.36	23.34	25.66	
	2014- 09-01	289476	0.818182	0.566667	319.97	22.50	24.57	

5 rows × 261 columns

```
In [5]: time_training_data.dtypes
```

```
In [6]: #Change index of startdate to date time
time_training_data.index = pd.to_datetime(time_training_data.index)
```

#### **Data Reduction Function**

- We will reduce run times of models using this data reduction function converting all data types to the data type of lowest memory usage.
- Function acquired from my esteemed teammate in the contest Daniel Logan.

```
In [7]:
        #Function acquired from Daniel Logan - WiDS teammate
         ef reduce mem usage(df, verbose=True):
            numerics = ['int16', 'int32', 'int64', 'float16', 'float32',
            start mem = df.memory usage().sum() / 1024**2
            for col in df.columns:
                col type = df[col].dtypes
                if col type in numerics:
                    c min = df[col].min()
                    c max = df[col].max()
                    if str(col type)[:3] == 'int':
                         if c min > np.iinfo(np.int8).min and c max <
        np.iinfo(np.int8).max:
                            df[col] = df[col].astype(np.int8)
                        elif c min > np.iinfo(np.int16).min and c max <</pre>
        np.iinfo(np.int16).max:
                             df[col] = df[col].astype(np.int16)
                         elif c min > np.iinfo(np.int32).min and c max <</pre>
        np.iinfo(np.int32).max:
                             df[col] = df[col].astype(np.int32)
                         elif c_min > np.iinfo(np.int64).min and c_max <</pre>
```

```
In [8]: time_training_data = reduce_mem_usage(time_training_data)

Mem. usage decreased to 355.46 Mb (52.7% reduction)
```

We were able to reduce our data set by 52.7% in memory usage.

### Set up X and y

- X All features but our target
- y Our target contest-tmp2m-14d\_\_tmp2m

```
In [9]: # Set up independent variables of all columns except target
   X = time_training_data.drop(['contest-tmp2m-14d_tmp2m'], axis = 1)

In [10]: # Set up array for target data - dependent variable
   y = time_training_data['contest-tmp2m-14d_tmp2m']

In [11]: # Check
   X.shape

Out[11]: (375734, 260)

In [12]: y.shape

Out[12]: (375734,)
```

## Time Series Split

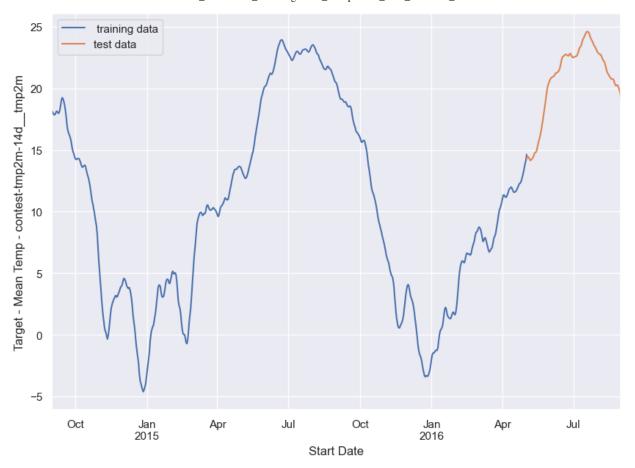
- We will run the TimeSeriesSplit to set up our training and test data or really our train
  and validation data as our real test set will be used for our submission to the WiDS
  Datathon.
- This TimeSeriesSplit method was used for our train test split for the Random Forest Regressor in the last notebook.

Out[64]:

		index	lat	lon	contest- pevpr-sfc- gauss- 14d_pevpr	nmme0- tmp2m- 34wcancm30	nmme0- tmp2m- 34wcancm40	nm tm 34wcc
S	tartdate							
	2014- 09-01	0	0.000000	0.833333	237.000000	29.020000	31.639999	29.57
	2014- 09-01	290938	0.818182	0.633333	323.630005	24.180000	26.750000	21.09
	2014- 09-01	35819	0.227273	0.900000	385.920013	31.160000	32.189999	33.2
	2014- 09-01	290207	0.818182	0.600000	303.359985	23.340000	25.660000	20.45
	2014- 09-01	289476	0.818182	0.566667	319.970001	22.500000	24.570000	19.67
	•••							
	2016- 05-02	285699	0.818182	0.366667	235.639999	3.780000	4.280000	30.0
	2016- 05-02	49586	0.272727	0.733333	573.750000	16.160000	18.850000	15.1 <sup>-</sup>
	2016- 05-02	306167	0.863636	0.400000	202.020004	3.760000	3.940000	0.68
	2016- 05-02	34235	0.227273	0.800000	555.150024	18.139999	20.209999	17.32
	2016- 05-02	107335	0.409091	0.800000	522.510010	15.030000	17.459999	14.5(

313112 rows × 260 columns

```
In [65]: #Visualize test set - Make sure testing data is after training data.
   plt.figure()
   sns.set(rc={'figure.figsize':(10,7)})
   y_train.groupby('startdate').mean().plot(label=' training data')
   y_test.groupby('startdate').mean().plot(label = 'test data')
   plt.xlabel('Start Date')
   plt.ylabel('Target - Mean Temp - contest-tmp2m-14d__tmp2m')
   plt.legend()
   plt.show()
```



## Scaling Training Set and Test Set

- Unlike the Random Forest Regressor model in Notebook 1, we will need to scale our data for the MLPRegressor.
- We will move ahead with a standard scale.
- Scaling was carried out after the train test split or TimeSeriesSplit to prevent data leakage caused by scaling data set together.

```
teration 34, loss = 0.14837674
teration 47, loss = 0.13756303
```

```
In [59]:
          pred = MLPmodel.predict(X testscaled)
In [60]:
          results MLP = pd.DataFrame(data = {'Actual':y test, \
                                                    Predictions':pred},
          index=y test.index
In [61]:
          results MLP
Out[61]:
                        Actual Predictions
            startdate
          2016-01-01
                      8.568278
                                 7.288152
          2016-01-01 10.872447
                                10.716764
          2016-01-01
                     -7.024219
                                -7.287463
          2016-01-01
                     -6.184810
                                -5.734422
          2016-01-01
                                 8.020726
                      9.002110
         2016-08-31 19.772009
                                12.712725
         2016-08-31 19.998930
                                12.922713
         2016-08-31 20.392469
                                12.926745
         2016-08-31
                    10.406187
                                 6.491517
         2016-08-31 15.910995
                                11.701444
         125244 rows × 2 columns
In [62]:
          print('R-squared =
           :.3f}'.format(r2 score(results MLP['Actual'],results MLP['Predictions
          print('RMSE =
```

#### **Initial MLPRegressor Evaluation**

Alright we are getting reasonable results when scaling the data. These metrics are not better than our Random Forest Regressor in the last notebook. We will try to optimize the

'.format(sqrt(mean squared error(results MLP['Actual'], results

MLPRegressor with a grid search and hope to get better scoring metrics. Initial MLPRegressor using Google Collab took 31min 43s. With Jupyter Notebooks on my cpu, the run time was 13 min 24s.

#### **Optimized RFR**

- R-squared = 0.909
- RMSE = 1.694

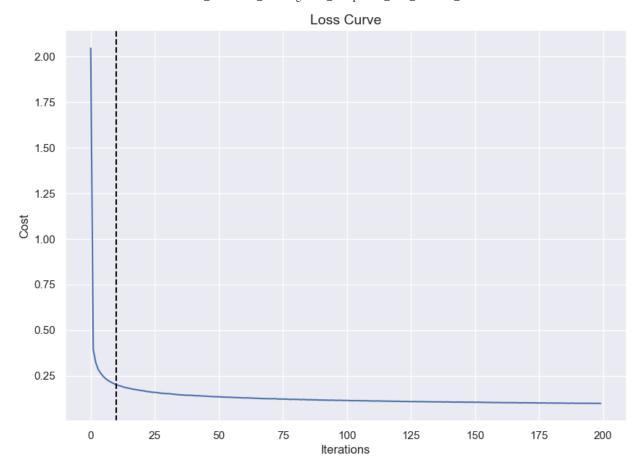
**MLPR - Initial attempt** -- 5 splits on Times Series split - With smaller test set in May-Aug 2016

- R-squared = 0.828
- RMSE = 2.327

MLPR - Initial attempt -- 2 splits on Times Series split - With larger test set in Jan-

- R-squared = 0.851
- RMSE = 3.662

```
In [60]:
    plt.figure()
    sns.set(rc={'figure.figsize':(10,7)})
    plt.plot(MLPmodel.loss_curve_)
    plt.title("Loss Curve", fontsize=14)
    plt.xlabel('Iterations')
    plt.ylabel('Cost')
    plt.axvline(x=10, color = 'black', linestyle = '---')
    plt.show()
```



#### Loss vs. Iteration Analysis

 Ther is a significant reduction in loss within the first 5-10 interations of our MLPRegressor. Then, there is a incremental improvement in loss until iteration 100 where the model loss plateaus more or less.

```
In [22]: #Set up Grid Search to tune hyper parameters of MLPR
MLPmodel = MLPRegressor(verbose = 1, random_state = 32)

param_grid = {
    'hidden_layer_sizes': [(150,100,50), (120,80,40), (100,50,30)],
    'max_iter': [50, 100],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant','adaptive'],
}
```

```
In [23]:
%%time
from sklearn.model_selection import GridSearchCV
```

```
tscv = TimeSeriesSplit(n_splits=2)

MLPgrid = GridSearchCV(MLPmodel, param_grid, cv=tscv)

MLPgrid.fit(X_trainscaled, y_train)
```

```
teration 47, loss = 0.09492709
```

```
teration 47, loss = 0.08647433
```

```
teration 17, loss = 0.09821414
teration 47, loss = 0.04470001
```

```
teration 47, loss = 0.04801227
```

```
teration 47, loss = 0.09492709
```

```
teration 47, loss = 0.08647433
```

```
teration 17, loss = 0.09821414
teration 47, loss = 0.04470001
```

```
teration 47, loss = 0.04801227
```

```
teration 47, loss = 0.09492709
```

```
teration 47, loss = 0.08647433
```

```
teration 17, loss = 0.09821414
teration 47, loss = 0.04470001
```

```
teration 47, loss = 0.04801227
```

```
teration 47, loss = 0.09492709
```

```
teration 47, loss = 0.08647433
```

```
teration 17, loss = 0.09821414
teration 47, loss = 0.04470001
```

```
teration 47, loss = 0.04801227
```

```
teration 47, loss = 0.09858791
```

```
teration 47, loss = 0.09337258
```

```
teration 47, loss = 0.05409210
```

```
teration 47, loss = 0.05435184
teration 49, loss = 0.05358814
```

```
teration 47, loss = 0.09858791
```

```
teration 47, loss = 0.09337258
```

```
teration 47, loss = 0.05409210
```

```
teration 47, loss = 0.05435184
```

```
teration 47, loss = 0.09858791
```

```
teration 47, loss = 0.09337258
```

```
teration 47, loss = 0.05409210
```

```
teration 5, loss = 0.20885924
teration 47, loss = 0.05435184
teration 49, loss = 0.05358814
```

```
teration 47, loss = 0.09858791
```

```
teration 47, loss = 0.09337258
```

```
teration 47, loss = 0.05409210
```

```
teration 47, loss = 0.05435184
```

```
teration 47, loss = 0.10542086
```

```
teration 47, loss = 0.10461887
```

```
teration 47, loss = 0.06177700
```

```
teration 47, loss = 0.06965836
```

```
teration 47, loss = 0.10542086
```

```
teration 45, loss = 0.10673274
teration 47, loss = 0.10461887
```

```
teration 47, loss = 0.06177700
```

```
teration 47, loss = 0.06965836
```

```
teration 47, loss = 0.10542086
```

```
teration 47, loss = 0.10461887
```

```
teration 47, loss = 0.06177700
```

```
teration 47, loss = 0.06965836
```

```
teration 47, loss = 0.10542086
```

```
teration 45, loss = 0.10673274
teration 47, loss = 0.10461887
```

```
teration 47, loss = 0.06177700
```

```
teration 47, loss = 0.06965836
```

```
teration 47, loss = 0.14828642
```

```
teration 47, loss = 0.14079204
```

```
teration 47, loss = 0.11497279
```

```
teration 5, loss = 0.26071124
teration 47, loss = 0.12761791
```

```
teration 47, loss = 0.14828642
```

```
teration 47, loss = 0.14079204
```

```
teration 47, loss = 0.11497279
```

```
teration 47, loss = 0.12761791
```

```
teration 47, loss = 0.14828642
```

```
teration 47, loss = 0.14079204
```

```
teration 47, loss = 0.11497279
```

```
teration 5, loss = 0.26071124
teration 47, loss = 0.12761791
```

```
teration 47, loss = 0.14828642
```

```
teration 47, loss = 0.14079204
```

```
teration 47, loss = 0.11497279
```

```
teration 47, loss = 0.12761791
```

```
teration 47, loss = 0.14442351
```

```
teration 47, loss = 0.14031432
teration 49, loss = 0.13822514
```

```
teration 3, loss = 0.47832817
teration 47, loss = 0.11942022
```

```
teration 47, loss = 0.12584254
```

```
teration 47, loss = 0.14442351
```

```
teration 47, loss = 0.14031432
```

```
teration 47, loss = 0.11942022
```

```
teration 47, loss = 0.12584254
```

```
teration 47, loss = 0.14442351
```

```
teration 47, loss = 0.14031432
teration 49, loss = 0.13822514
```

```
teration 3, loss = 0.47832817
teration 47, loss = 0.11942022
```

```
teration 47, loss = 0.12584254
```

```
teration 47, loss = 0.14442351
```

```
teration 47, loss = 0.14031432
```

```
teration 47, loss = 0.11942022
```

```
teration 47, loss = 0.12584254
```

```
teration 8, loss = 0.25844814
teration 47, loss = 0.14372020
```

```
teration 5, loss = 0.30550774
teration 47, loss = 0.14492434
```

```
teration 47, loss = 0.11684709
```

```
teration 47, loss = 0.13194850
```

```
teration 8, loss = 0.25844814
teration 47, loss = 0.14372020
```

```
teration 5, loss = 0.30550774
teration 47, loss = 0.14492434
```

```
teration 47, loss = 0.11684709
```

```
teration 47, loss = 0.13194850
```

```
Iteration 92, loss = 0.11932074
```

```
teration 47, loss = 0.14372020
```

```
teration 5, loss = 0.30550774
teration 47, loss = 0.14492434
```

```
teration 47, loss = 0.11684709
```

```
teration 47, loss = 0.13194850
```

```
teration 8, loss = 0.25844814
teration 47, loss = 0.14372020
```

```
teration 5, loss = 0.30550774
teration 47, loss = 0.14492434
```

```
teration 47, loss = 0.11684709
```

```
teration 47, loss = 0.13194850
```

```
Iteration 92, loss = 0.11932074
```

```
teration 47, loss = 0.09919038
```

```
teration 47, loss = 0.09424295
```

```
teration 47, loss = 0.05173414
```

```
teration 47, loss = 0.04941696
```

```
teration 47, loss = 0.09919038
```

```
teration 47, loss = 0.09424295
```

```
teration 47, loss = 0.05173414
```

```
teration 47, loss = 0.04941696
```

```
teration 47, loss = 0.09919038
```

```
teration 47, loss = 0.09424295
```

```
teration 47, loss = 0.05173414
```

```
teration 47, loss = 0.04941696
```

```
teration 47, loss = 0.09919038
```

```
teration 47, loss = 0.09424295
```

```
teration 47, loss = 0.05173414
```

```
teration 47, loss = 0.04941696
```

```
teration 47, loss = 0.10673537
```

```
teration 47, loss = 0.10405984
```

```
teration 47, loss = 0.05972790
```

```
teration 47, loss = 0.05761878
```

```
teration 47, loss = 0.10673537
```

```
teration 47, loss = 0.10405984
```

```
teration 38, loss = 0.06918874
teration 47, loss = 0.05972790
```

```
teration 47, loss = 0.05761878
```

```
teration 47, loss = 0.10673537
```

```
teration 47, loss = 0.10405984
```

```
teration 47, loss = 0.05972790
```

```
teration 47, loss = 0.05761878
```

```
teration 47, loss = 0.10673537
```

```
teration 47, loss = 0.10405984
```

```
teration 38, loss = 0.06918874
teration 47, loss = 0.05972790
```

```
teration 47, loss = 0.05761878
```

```
teration 47, loss = 0.12119400
```

```
teration 47, loss = 0.12153941
```

```
teration 47, loss = 0.07080609
```

```
teration 47, loss = 0.07433728
```

```
teration 47, loss = 0.12119400
```

```
teration 47, loss = 0.12153941
```

```
teration 47, loss = 0.07080609
```

```
teration 47, loss = 0.07433728
```

```
teration 47, loss = 0.12119400
```

```
teration 47, loss = 0.12153941
```

```
teration 47, loss = 0.07080609
```

```
teration 47, loss = 0.07433728
```

```
teration 47, loss = 0.12119400
```

```
teration 47, loss = 0.12153941
```

```
teration 47, loss = 0.07080609
```

```
teration 47, loss = 0.07433728
```

```
teration 47, loss = 0.14762990
```

```
teration 47, loss = 0.14026808
```

```
teration 47, loss = 0.10986600
```

```
teration 47, loss = 0.11866679
```

```
teration 47, loss = 0.14762990
```

```
teration 47, loss = 0.14026808
```

```
teration 47, loss = 0.10986600
```

```
teration 93, loss = 0.09533974
```

```
teration 47, loss = 0.11866679
```

```
teration 47, loss = 0.14762990
```

```
teration 47, loss = 0.14026808
```

```
teration 47, loss = 0.10986600
```

```
teration 47, loss = 0.11866679
```

```
teration 47, loss = 0.14762990
```

```
teration 47, loss = 0.14026808
```

```
teration 47, loss = 0.10986600
```

```
teration 93, loss = 0.09533974
```

```
teration 47, loss = 0.11866679
```

```
teration 47, loss = 0.14788922
```

```
teration 5, loss = 0.31537574
teration 47, loss = 0.14169314
```

```
teration 47, loss = 0.11454541
```

```
teration 47, loss = 0.12111008
```

```
teration 47, loss = 0.14788922
```

```
Steration 5, loss = 0.31537574
teration 47, loss = 0.14169314
```

```
teration 47, loss = 0.11454541
```

```
teration 47, loss = 0.12111008
```

```
teration 47, loss = 0.14788922
```

```
teration 5, loss = 0.31537574
teration 47, loss = 0.14169314
```

```
teration 47, loss = 0.11454541
```

```
teration 47, loss = 0.12111008
```

```
teration 47, loss = 0.14788922
```

```
Steration 5, loss = 0.31537574
teration 47, loss = 0.14169314
```

```
teration 47, loss = 0.11454541
```

```
teration 47, loss = 0.12111008
```

```
teration 47, loss = 0.15398052
```

```
teration 47, loss = 0.15015645
```

```
teration 47, loss = 0.11934114
```

```
teration 47, loss = 0.12936939
```

```
teration 47, loss = 0.15398052
```

```
teration 47, loss = 0.15015645
```

```
teration 47, loss = 0.11934114
```

```
teration 47, loss = 0.12936939
```

```
teration 47, loss = 0.15398052
```

```
teration 47, loss = 0.15015645
```

```
teration 47, loss = 0.11934114
```

```
teration 47, loss = 0.12936939
```

```
teration 47, loss = 0.15398052
```

```
teration 47, loss = 0.15015645
```

```
teration 47, loss = 0.11934114
```

```
teration 47, loss = 0.12936939
```

/Users/kileymack/opt/anaconda3/lib/python3.9/site-packages/sklearn/neural\_network/\_multilayer\_perceptron.py:692: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (100) reached and the optimization hasn't converged yet.

warnings.warn(

```
teration 47, loss = 0.09370193
```

```
Out[23]:
In [24]:
        MLPgrid.best params
Out[24]:
In [26]:
         MLPmodel2 = MLPRegressor(verbose = 1, random state = 32, activation
           'relu', alpha = 0.0001, \
                                hidden layer sizes = (150, 100, 50),
         learning rate = 'constant'\
                                 max iter = 200, solver = 'sgd')
         MLPmodel2.fit(X trainscaled, y train)
         print(MLPmodel2)
```

# MLPRegressor Grid Search - Best Parameters and Run Time

- Best Parameters {'activation': 'relu', 'alpha': 0.0001, 'hidden\_layer\_sizes': (150, 100, 50), 'learning\_rate': 'constant', 'max\_iter': 100, 'solver': 'sgd'}
- Run Time: 8h 37min 2s

```
In [28]: # Gather grid search predictions
grid_pred = MLPgrid.predict(X_testscaled)
```

# In [31]: results\_MLP\_grid

Out[31]:		Actual	Predictions
	startdate		
	2016-05-02	21.080032	21.303924
	2016-05-02	12.921798	13.245692
	2016-05-02	11.742004	11.297715

**2016-05-02** 18.386656

**2016-08-31** 19.998930

2016-05-02	10.771266	10.771513	
•••			
2016-08-31	19.772009	16.822504	

17.989621

17.165045

**2016-08-31** 20.392469 17.283563 **2016-08-31** 10.406187 11.432712

**2016-08-31** 15.910995 14.843682

62622 rows × 2 columns

```
In [32]: # Error Motrics for MIR Cris
```

```
# Error Metrics for MLP Grid Search
print('R-squared =
{:.3f}'.format(r2_score(results_MLP_grid['Actual'],results_MLP_grid['
print('RMSE =
{:.3f}'.format(sqrt(mean_squared_error(results_MLP_grid['Actual'],res
```

```
R-squared = 0.754
RMSE = 2.783
```

### MLP Grid Search Evaluation

R-squared = 0.754 RMSE = 2.783

The grid search proved to have poorer scores than the default MLPRegresssor. We will try out the optimized grid search parameters with 200 iterrations and see if the metrics improve.

print(MLPmodel2)

```
teration 47, loss = 0.09370193
```

```
CPU times: user 1h 47min 50s, sys: 3min 5s, total: 1h 50min 55s
```

/Users/kileymack/opt/anaconda3/lib/python3.9/site-packages/sklearn/neural\_network/\_multilayer\_perceptron.py:692: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.

In [36]:

```
pred2 = MLPmodel2.predict(X testscaled
In [37]:
         results MLP2 = pd.DataFrame(data = {'Actual':y test, \
                                                  Predictions':pred2},
          index=y test.index
In [38]:
         results MLP2
Out[38]:
                       Actual Predictions
            startdate
         2016-05-02 21.080032
                               21.276929
         2016-05-02 12.921798
                               13.328666
         2016-05-02 11.742004
                               11.743602
         2016-05-02 18.386656
                               17.927709
         2016-05-02 10.771266
                               10.915823
         2016-08-31 19.772009
                               16.797451
         2016-08-31 19.998930
                               17.183268
         2016-08-31 20.392469
                               17.279610
         2016-08-31 10.406187
                               10.420368
         2016-08-31 15.910995
                               14.193742
        62622 rows × 2 columns
In [39]:
         # Error Metrics for MLP2
         print('R-squared =
           :.3f}'.format(r2 score(results MLP2['Actual'],results MLP2['Prediction
          print('RMSE =
                '.format(sqrt(mean squared error(results MLP2['Actual'],results
```

# MLPRegressor - 200 Iterations with "Optimized Parameters"

• R-squared = 0.720

• RMSE = 2.966

Our metrics did not improve, lending more evidence that the default MLPRegressor to be superior with this dataset.

We also ran some different length testing sets to see how our error scores would be affected. With this small sample size, we see our R-squared score increase with the longer test set but the RMSE score decrease. The test set for the datathon submission is only two months long, and our RSME could improve upon submission with the shorter testing window.

The MLPRegressor is scoring lower than our RandomForestRegressor, and we did not have to scale the data. Our next step will be to compose a submission for the Datathon with the RFR model since it has consistently performed better than the MLP Regressor. This submission will give use a baseline to improve from.

We will then have to look to feature engineering and/or other regression models to further improve upon our modeling error metrics in subsequent notebooks.

### **RFR - Submission**

- We will fit our RFR with a train/test split of the WiDS training\_data with the optimized parameters from the grid search in Notebook 1.
- We evaluate on the test split to make sure things transferred well to this notebook.
- Finally, we will then fit the RFR on the complete training\_data and make predictions on the WiDS test\_data and reformat for a submission to the datathon.

In [76]:

#Check X\_test global variable -- Look good X\_test Out[76]:

		index	lat	lon	contest- pevpr-sfc- gauss- 14d_pevpr	nmme0- tmp2m- 34wcancm30	nmme0- tmp2m- 34wcancm40	nm tm 34wcc
S	tartdate							
	2016- 05-02	95639	0.409091	0.266667	374.589996	13.440000	14.210000	11.84
	2016- 05-02	212599	0.636364	0.800000	293.670013	10.670000	12.210000	10.26
	2016- 05-02	184090	0.590909	0.466667	197.000000	3.580000	6.840000	2.62
	2016- 05-02	116838	0.454545	0.300000	395.859985	10.430000	11.370000	9.08
	2016- 05-02	369764	1.000000	0.600000	221.509995	4.290000	6.690000	6.38
	•••	•••						
	2016- 08-31	255118	0.727273	0.833333	302.059998	24.760000	33.759998	25.90
	2016- 08-31	255849	0.727273	0.866667	324.470001	25.049999	33.889999	26.74
	2016- 08-31	256580	0.727273	0.900000	326.140015	25.139999	33.590000	27.46
	2016- 08-31	187135	0.590909	0.600000	494.890015	21.930000	22.770000	21.43
	2016- 08-31	375733	1.000000	0.866667	295.290009	23.129999	27.200001	20.2

62622 rows × 260 columns

```
Out[79]:
In [80]:
          pred = RFRmodel.predict(X test)
           Parallel(n jobs=1)]: Done 100 out of 100 | elapsed: 4.6s finished
In [81]:
         results RFR = pd.DataFrame(data = {'Actual':y test, \
                                                  'Predictions':pred},
          index=y test.index
In [82]:
         results RFR
Out[82]:
                       Actual Predictions
           startdate
         2016-05-02 21.080032
                              20.437980
         2016-05-02 12.921798
                              13.407156
         2016-05-02 11.742004
                              11.323933
         2016-05-02 18.386656
                              17.999544
         2016-05-02 10.771266
                              11.368966
         2016-08-31 19.772009
                              19.208867
         2016-08-31 19.998930
                              19.483500
         2016-08-31 20.392469
                               19.578225
         2016-08-31 10.406187
                               11.230049
         2016-08-31 15.910995
                               16.410757
        62622 rows × 2 columns
In [83]:
         print('R-squared =
            .3f}'.format(r2 score(results RFR['Actual'], results RFR['Predictions
         print
                 '.format(sqrt(mean_squared_error(results_RFR['Actual'],results
```

```
R-squared = 0.908
RMSE = 1.696
```

The RFR is performing at pretty much the same level as in our last notebook after optimized. We will proceed.

Notebook 1: R-squared = 0.909 RMSE = 1.694

Notebook 2: R-squared = 0.908 RMSE = 1.696

## Transform Test\_Data

- Set up WiDS test\_data with a stardate as index as we did for the training\_data allowing our models fit on our cleaned and transformed data to run the WiDS wiDS test data.
- The test\_data from WiDS has no target variable. Therefore, we will not split it into X and y.
- We will predict the target and then upon submission to the WiDS Kaggle site we will know our RMSE metric.
- This will be our baseline score to improve upon.

```
In [84]: # Read in clean training set with start date as datetime index
test_data = pd.read_csv('data/test_data_clean.csv')
```

In [85]: # Let's view the indices and shape of the original test set. S
 can make sure the index will realign
 # when we make our submission.
test data.head()

Out[85]:	indev	lat	lon	startdate	contest- pevpr-sfc-	nmme0-	nmme0-	nmme
	test_da	ta.head	l()					
	# when	we make	our	submissi	on.			

	index	lat	lon	startdate	pevpr-sfc- gauss- 14d_pevpr	nmme0- tmp2m- 34wcancm30	nmme0- tmp2m- 34wcancm40	nmme tmp2ı 34wccsm
0	375734	0.0	0.833333	2022-11- 01	339.88	30.88	30.92	29
1	375735	0.0	0.833333	2022-11- 02	334.63	30.88	30.92	29
2	375736	0.0	0.833333	2022-11- 03	337.83	30.88	30.92	29
3	375737	0.0	0.833333	2022-11- 04	345.81	30.88	30.92	29
4	375738	0.0	0.833333	2022-11- 05	357.39	30.88	30.92	29

5 rows × 261 columns

contact-

```
In [86]: test_data.tail()
```

Out[86]:

	index	lat	lon	startdate	pevpr-sfc- gauss- 14dpevpr	nmme0- tmp2m- 34wcancm30	nmme0- tmp2m- 34wcancm40	n t 34wc
31349	407083	1.0	0.866667	2022-12- 27	62.72	4.6	8.71	
31350	407084	1.0	0.866667	2022-12- 28	73.41	4.6	8.71	
31351	407085	1.0	0.866667	2022-12- 29	70.00	4.6	8.71	
31352	407086	1.0	0.866667	2022-12- 30	79.81	4.6	8.71	
31353	407087	1.0	0.866667	2022-12- 31	86.17	4.6	8.71	

5 rows × 261 columns

```
In [87]: test_data.shape
Out[87]: (31354, 261)
```

In [89]:

```
# Check that start date is still in date time data type
with pd.option_context('display.max_rows', None,
'display.max_columns', None):
    print(test_data.dtypes)
```

WiDs_Notebook2_MLPRegressor_Com	
index	int64
lat	float64
lon	float64
startdate	object
contest-pevpr-sfc-gauss-14d_pevpr	float64
nmme0-tmp2m-34wcancm30	float64
nmme0-tmp2m-34wcancm40	float64
nmme0-tmp2m-34wccsm30	float64
nmme0-tmp2m-34wccsm40	float64
nmme0-tmp2m-34wcfsv20	float64
nmme0-tmp2m-34wgfdlflora0	float64
nmme0-tmp2m-34wgfdlflorb0	float64
nmme0-tmp2m-34wgfd10	float64
nmme0-tmp2m-34wnasa0	float64
nmme0-tmp2m-34wnmme0mean	float64
contest-wind-h10-14d_wind-hgt-10	float64
nmme-tmp2m-56wcancm3	float64
nmme-tmp2m-56wcancm4	float64
nmme-tmp2m-56wccsm3	float64
nmme-tmp2m-56wccsm4	float64
nmme-tmp2m-56wcfsv2	float64
nmme-tmp2m-56wgfdl	float64
nmme-tmp2m-56wgfdlflora	float64
nmme-tmp2m-56wgfdlflorb	float64
nmme-tmp2m-56wnasa	float64
nmme-tmp2m-56wnmmemean	float64
contest-rhum-sig995-14drhum	float64
nmme-prate-34wcancm3	float64
nmme-prate-34wcancm4	float64
nmme-prate-34wccsm3	float64
nmme-prate-34wccsm4	float64
nmme-prate-34wcfsv2	float64
nmme-prate-34wgfdl	float64
nmme-prate-34wgfdlflora	float64
nmme-prate-34wgfdlflorb	float64
nmme-prate-34wnasa	float64
nmme-prate-34wnmmemean	float64
contest-wind-h100-14dwind-hgt-100	float64
nmme0-prate-56wcancm30	float64
nmme0-prate-56wcancm40	float64
nmme0-prate-56wccsm30	float64
nmme0-prate-56wccsm40	float64
nmme0-prate-56wcfsv20	float64
nmme0-prate-56wgfdlflora0	float64
nmme0-prate-56wgfdlflorb0	float64
nmme0-prate-56wgfdl0	float64
nmme0-prate-56wnasa0	float64
nmme0-prate-56wnmme0mean	float64
nmme0-prate-34w_cancm30	float64
nmme0-prate-34w cancm40	float64
nmme0-prate-34w_ccsm30	float64
nmme0-prate-34w ccsm40	float64

WiDs_Notebook2_MLPRegressor_Com	•
nmme0-prate-34wcfsv20	float64
nmme0-prate-34wgfdlflora0	float64
nmme0-prate-34wgfdlflorb0	float64
nmme0-prate-34wgfd10	float64
nmme0-prate-34wnasa0	float64
nmme0-prate-34wnmme0mean	float64
contest-slp-14d_slp	float64
contest-wind-vwnd-925-14dwind-vwnd-925	float64
nmme-prate-56wcancm3	float64
nmme-prate-56wcancm4	float64
nmme-prate-56wccsm3	float64
nmme-prate-56wccsm4	float64
nmme-prate-56wcfsv2	float64
nmme-prate-56wgfdl	float64
nmme-prate-56wgfdlflora	float64
nmme-prate-56wgfdlflorb	float64
nmme-prate-56wnasa	float64
nmme-prate-56wnmmemean	float64
contest-pres-sfc-gauss-14dpres	float64
contest-wind-uwnd-250-14dwind-uwnd-250	float64
nmme-tmp2m-34wcancm3	float64
nmme-tmp2m-34wcancm4	float64
nmme-tmp2m-34wccsm3	float64
nmme-tmp2m-34wccsm4	float64
nmme-tmp2m-34wcfsv2	float64
nmme-tmp2m-34wgfdl	float64
nmme-tmp2m-34wgfdlflora	float64
nmme-tmp2m-34wgfdlflorb	float64
nmme-tmp2m-34wnasa	float64
nmme-tmp2m-34wnmmemean	float64
contest-prwtr-eatm-14dprwtr	float64
contest-wind-vwnd-250-14dwind-vwnd-250	float64
contest-precip-14dprecip	float64
contest-wind-h850-14dwind-hgt-850	float64
contest-wind-uwnd-925-14d_wind-uwnd-925	float64
contest-wind-h500-14dwind-hgt-500	float64
cancm30	float64
cancm40	float64
ccsm30	float64
ccsm40	float64
cfsv20	float64
gfdlflora0	float64
gfdlflorb0	float64
gfdl0	float64
nasa0	float64
nmme0mean	float64
elevationelevation	int64
wind-vwnd-250-2010-1	float64
wind-vwnd-250-2010-2	float64
wind-vwnd-250-2010-3	float64
wind-vwnd-250-2010-4	float64
wind-vwnd-250-2010-5	float64

WiDs_Notebook2_MLPRegressor_Co	mparison_KFK_Datathon_Submission
wind-vwnd-250-2010-6	float64
wind-vwnd-250-2010-7	float64
wind-vwnd-250-2010-8	float64
wind-vwnd-250-2010-9	float64
wind-vwnd-250-2010-10	float64
wind-vwnd-250-2010-11	float64
wind-vwnd-250-2010-12	float64
wind-vwnd-250-2010-13	float64
wind-vwnd-250-2010-14	float64
wind-vwnd-250-2010-15	float64
wind-vwnd-250-2010-16	float64
wind-vwnd-250-2010-17	float64
wind-vwnd-250-2010-18	float64
wind-vwnd-250-2010-19	float64
wind-vwnd-250-2010-20	float64
wind-uwnd-250-2010-1	float64
wind-uwnd-250-2010-2	float64
wind-uwnd-250-2010-3	float64
wind-uwnd-250-2010-4	float64
wind-uwnd-250-2010-5	float64
wind-uwnd-250-2010-6	float64
wind-uwnd-250-2010-7	float64
wind-uwnd-250-2010-8	float64
wind-uwnd-250-2010-9	float64
wind-uwnd-250-2010-10	float64
wind-uwnd-250-2010-11	float64
wind-uwnd-250-2010-12	float64
wind-uwnd-250-2010-13	float64
wind-uwnd-250-2010-14	float64
wind-uwnd-250-2010-15	float64
wind-uwnd-250-2010-16	float64
wind-uwnd-250-2010-17	float64
wind-uwnd-250-2010-18	float64
wind-uwnd-250-2010-19	float64
wind-uwnd-250-2010-20	float64
mjo1dphase	float64
mjoldamplitude	float64
meimei	float64
meimeirank	float64
meinip	float64
wind-hgt-850-2010-1	float64
wind-hgt-850-2010-2	float64
wind-hgt-850-2010-3	float64
wind-hgt-850-2010-4	float64
wind-hgt-850-2010-5	float64
wind-hgt-850-2010-6	float64
wind-hgt-850-2010-7	float64
wind-hgt-850-2010-8	float64
wind-hgt-850-2010-9	float64
wind-hgt-850-2010-10	float64
sst-2010-1	float64
sst-2010-2	float64

WiDs_Notebook2_MLPRegressor_Co	mparison_RFR_Datathon_Submission
sst-2010-3	float64
sst-2010-4	float64
sst-2010-5	float64
sst-2010-6	float64
sst-2010-7	float64
sst-2010-8	float64
sst-2010-9	float64
sst-2010-10	float64
wind-hgt-500-2010-1	float64
wind-hgt-500-2010-2	float64
wind-hgt-500-2010-3	float64
wind-hgt-500-2010-4	float64
wind-hgt-500-2010-5	float64
wind-hgt-500-2010-6	float64
wind-hgt-500-2010-7	float64
wind-hgt-500-2010-8	float64
wind-hgt-500-2010-9	float64
wind-hgt-500-2010-10	float64
icec-2010-1	float64
icec-2010-2	float64
icec-2010-3	float64
icec-2010-4	float64
icec-2010-5	float64
icec-2010-6	float64
icec-2010-7	float64
icec-2010-8	float64
icec-2010-9	float64
icec-2010-10	float64
wind-uwnd-925-2010-1	float64
wind-uwnd-925-2010-2	float64
wind-uwnd-925-2010-3	float64
wind-uwnd-925-2010-4	float64
wind-uwnd-925-2010-5	float64
wind-uwnd-925-2010-6	float64
wind-uwnd-925-2010-7	float64
wind-uwnd-925-2010-8	float64
wind-uwnd-925-2010-9	float64
wind-uwnd-925-2010-10	float64
wind-uwnd-925-2010-11	float64
wind-uwnd-925-2010-12	float64
wind-uwnd-925-2010-13	float64
wind-uwnd-925-2010-14	float64
wind-uwnd-925-2010-15	float64
wind-uwnd-925-2010-16	float64
wind-uwnd-925-2010-17	float64
wind-uwnd-925-2010-18	float64
wind-uwnd-925-2010-19	float64
wind-uwnd-925-2010-20	float64
wind-hgt-10-2010-1	float64
wind-hgt-10-2010-2	float64
wind-hgt-10-2010-3	float64
wind-hgt-10-2010-4	float64

	Regressor_Comparison_RFR_Datathon_Submission
wind-hgt-10-2010-5	float64
wind-hgt-10-2010-6	float64
wind-hgt-10-2010-7	float64
wind-hgt-10-2010-8	float64
wind-hgt-10-2010-9	float64
wind-hgt-10-2010-10	float64
wind-hgt-100-2010-1	float64
wind-hgt-100-2010-2	float64
wind-hgt-100-2010-3	float64
wind-hgt-100-2010-4	float64
wind-hgt-100-2010-5	float64
wind-hgt-100-2010-6	float64
wind-hgt-100-2010-7	float64
wind-hgt-100-2010-8	float64
wind-hgt-100-2010-9	float64
wind-hgt-100-2010-10	float64
wind-vwnd-925-2010-1	float64
wind-vwnd-925-2010-2	float64
wind-vwnd-925-2010-3	float64
wind-vwnd-925-2010-4	float64
wind-vwnd-925-2010-5	float64
wind-vwnd-925-2010-6	float64
wind-vwnd-925-2010-7	float64
wind-vwnd-925-2010-8	float64
wind-vwnd-925-2010-9	float64
wind-vwnd-925-2010-10	float64
wind-vwnd-925-2010-11	float64 float64
wind-vwnd-925-2010-12	
wind-vwnd-925-2010-13 wind-vwnd-925-2010-14	float64 float64
wind-vwnd-925-2010-14 wind-vwnd-925-2010-15	float64
wind-vwnd-925-2010-15 wind-vwnd-925-2010-16	float64
wind-vwnd-925-2010-16 wind-vwnd-925-2010-17	float64
wind-vwnd-925-2010-17 wind-vwnd-925-2010-18	float64
wind-vwnd-925-2010-18 wind-vwnd-925-2010-19	float64
wind-vwnd-925-2010-19 wind-vwnd-925-2010-20	float64
BSh	int64
BSk	int64
BWh	int64
BWk	int64
Cfa	int64
Cfb	int64
Csa	int64
Csb	int64
Dfa	int64
Dfb	int64
Dfc	int64
Dsb	int64
Dsc	int64
Dwa	int64
Dwb	int64
month number	int64
_	

```
In [90]:
         test data['startdate'] = pd.to datetime(test data['startdate'])
In [91]:
          rith pd.option_context('display.max_rows', None,
             print(test data.dtypes
```

	or_Comparison_RFR_Datathon_Submission
index	int64
lat	float64
lon	float64
startdate	datetime64[ns]
contest-pevpr-sfc-gauss-14d_pevpr	float64
nmme0-tmp2m-34wcancm30	float64
nmme0-tmp2m-34wcancm40	float64
nmme0-tmp2m-34wccsm30	float64
nmme0-tmp2m-34wccsm40	float64
nmme0-tmp2m-34wcfsv20	float64
nmme0-tmp2m-34wgfdlflora0	float64
nmme0-tmp2m-34wgfdlflorb0	float64
nmme0-tmp2m-34wgfd10	float64
nmme0-tmp2m-34wnasa0	float64
nmme0-tmp2m-34wnmme0mean	float64
contest-wind-h10-14dwind-hgt-10	float64
nmme-tmp2m-56wcancm3	float64
nmme-tmp2m-56wcancm4	float64
nmme-tmp2m-56wccsm3	float64
nmme-tmp2m-56wccsm4	float64
nmme-tmp2m-56wcfsv2	float64
nmme-tmp2m-56wgfdl	float64
nmme-tmp2m-56wgfdlflora	float64
nmme-tmp2m-56wgfdlflorb	float64
nmme-tmp2m-56wnasa	float64
nmme-tmp2m-56wnmmemean	float64
contest-rhum-sig995-14drhum	float64
nmme-prate-34wcancm3	float64
nmme-prate-34wcancm4	float64
nmme-prate-34wccsm3	float64
nmme-prate-34wccsm4	float64
nmme-prate-34wcfsv2	float64
nmme-prate-34wgfdl	float64
nmme-prate-34wgfdlflora	float64
nmme-prate-34wgfdlflorb	float64
nmme-prate-34wnasa	float64
nmme-prate-34wnmmemean	float64
contest-wind-h100-14dwind-hgt-100	float64
nmme0-prate-56wcancm30	float64
nmme0-prate-56wcancm40	float64
nmme0-prate-56wccsm30	float64
nmme0-prate-56wccsm40	float64
nmme0-prate-56wcfsv20	float64
nmme0-prate-56wgfdlflora0	float64
nmme0-prate-56wgfdlflorb0	float64
nmme0-prate-56wgfdl0	float64
nmme0-prate-56wnasa0	float64
nmme0-prate-56wnmme0mean	float64
nmme0-prate-34wcancm30	float64
nmme0-prate-34wcancm40	float64
nmme0-prate-34wccsm30	float64
nmme0-prate-34wccsm40	float64

WiDs_Notebook2_MLPRegressor_Comparison	_RFR_Datathon_Submission
nmme0-prate-34wcfsv20	float64
nmme0-prate-34wgfdlflora0	float64
nmme0-prate-34wgfdlflorb0	float64
nmme0-prate-34wgfd10	float64
nmme0-prate-34wnasa0	float64
nmme0-prate-34wnmme0mean	float64
contest-slp-14d_slp	float64
contest-wind-vwnd-925-14dwind-vwnd-925	float64
nmme-prate-56wcancm3	float64
nmme-prate-56wcancm4	float64
nmme-prate-56wccsm3	float64
nmme-prate-56wccsm4	float64
nmme-prate-56wcfsv2	float64
nmme-prate-56w_gfdl	float64
nmme-prate-56w gfdlflora	float64
nmme-prate-56w_gfdlflorb	float64
nmme-prate-56w nasa	float64
nmme-prate-56w nmmemean	float64
contest-pres-sfc-gauss-14d pres	float64
contest-wind-uwnd-250-14d wind-uwnd-250	float64
nmme-tmp2m-34w cancm3	float64
nmme-tmp2m-34w cancm4	float64
nmme-tmp2m-34w ccsm3	float64
nmme-tmp2m-34w ccsm4	float64
nmme-tmp2m-34w cfsv2	float64
nmme-tmp2m-34wgfdl	float64
nmme-tmp2m-34w gfdlflora	float64
nmme-tmp2m-34w_gfdlflorb	float64
nmme-tmp2m-34w nasa	float64
nmme-tmp2m-34w nmmemean	float64
contest-prwtr-eatm-14d prwtr	float64
	float64
contest-wind-vwnd-250-14dwind-vwnd-250 contest-precip-14d precip	float64
contest-wind-h850-14d wind-hgt-850	float64
contest-wind-uwnd-925-14d wind-uwnd-925	float64
contest-wind-h500-14d wind-hgt-500	float64
cancm30	float64
cancm40	float64
ccsm30	float64
ccsm30	float64
cfsv20	float64
gfdlflora0	float64
gfdlflorb0	float64
	float64
gfdl0	
nasa0 nmme0mean	float64 float64
elevationelevation	int64
wind-vwnd-250-2010-1	float64
wind-vwnd-250-2010-2	float64
wind-vwnd-250-2010-3	float64
wind-vwnd-250-2010-4	float64
wind-vwnd-250-2010-5	float64

	WiDs_Notebook2_MLPRegressor_Comparison_RFR_D	
wind-vwnd-250-2010-6	f	loat64
wind-vwnd-250-2010-7	f	loat64
wind-vwnd-250-2010-8	f	loat64
wind-vwnd-250-2010-9		loat64
wind-vwnd-250-2010-1	0 f	loat64
wind-vwnd-250-2010-1	1 f	loat64
wind-vwnd-250-2010-1	2 f	loat64
wind-vwnd-250-2010-1	3 f	loat64
wind-vwnd-250-2010-1	4 f	loat64
wind-vwnd-250-2010-1	5 f	loat64
wind-vwnd-250-2010-1	6 f	loat64
wind-vwnd-250-2010-1	7 f	loat64
wind-vwnd-250-2010-1	8 f	loat64
wind-vwnd-250-2010-1	9 f	loat64
wind-vwnd-250-2010-2	0 f	loat64
wind-uwnd-250-2010-1	f	loat64
wind-uwnd-250-2010-2	f	loat64
wind-uwnd-250-2010-3	f	loat64
wind-uwnd-250-2010-4	f	loat64
wind-uwnd-250-2010-5	f	loat64
wind-uwnd-250-2010-6	f	loat64
wind-uwnd-250-2010-7	f	loat64
wind-uwnd-250-2010-8	f	loat64
wind-uwnd-250-2010-9	f	loat64
wind-uwnd-250-2010-1	0 f	loat64
wind-uwnd-250-2010-1	1 f	loat64
wind-uwnd-250-2010-1	2 f	loat64
wind-uwnd-250-2010-1	3 f	loat64
wind-uwnd-250-2010-1	4 f	loat64
wind-uwnd-250-2010-1	5 f	loat64
wind-uwnd-250-2010-1	6 <b>f</b>	loat64
wind-uwnd-250-2010-1	7 f	loat64
wind-uwnd-250-2010-1	8 f	loat64
wind-uwnd-250-2010-1	g f	loat64
wind-uwnd-250-2010-2	0 <b>f</b>	loat64
mjoldphase	f	loat64
mjoldamplitude	f	loat64
meimei	f	loat64
meimeirank	f	loat64
meinip	f	loat64
wind-hgt-850-2010-1	f	loat64
wind-hgt-850-2010-2	f	loat64
wind-hgt-850-2010-3	f	loat64
wind-hgt-850-2010-4	f	loat64
wind-hgt-850-2010-5	f	loat64
wind-hgt-850-2010-6	f	loat64
wind-hgt-850-2010-7	f	loat64
wind-hgt-850-2010-8	f	loat64
wind-hgt-850-2010-9	f	loat64
wind-hgt-850-2010-10	f	loat64
sst-2010-1	f	loat64
sst-2010-2	f	loat64

	WiDs_Notebook2_MLPRegressor_Comparison_RFR_Datathon_Subn
sst-2010-3	float64
sst-2010-4	float64
sst-2010-5	float64
sst-2010-6	float64
sst-2010-7	float64
sst-2010-8	float64
sst-2010-9	float64
sst-2010-10	float64
wind-hgt-500-2010-1	float64
wind-hgt-500-2010-2	float64
wind-hgt-500-2010-3	float64
wind-hgt-500-2010-4	float64
wind-hgt-500-2010-5	float64
wind-hgt-500-2010-6	float64
wind-hgt-500-2010-7	float64
wind-hgt-500-2010-8	float64
wind-hgt-500-2010-9	float64
wind-hgt-500-2010-10	float64
icec-2010-1	float64
icec-2010-2	float64
icec-2010-3	float64
icec-2010-4	float64
icec-2010-5	float64
icec-2010-6	float64
icec-2010-7	float64
icec-2010-8	float64
icec-2010-9	float64
icec-2010-10	float64
wind-uwnd-925-2010-1	float64
wind-uwnd-925-2010-2	float64
wind-uwnd-925-2010-3	float64
wind-uwnd-925-2010-4	float64
wind-uwnd-925-2010-5	float64
wind-uwnd-925-2010-6	float64
wind-uwnd-925-2010-7	float64
wind-uwnd-925-2010-8	float64
wind-uwnd-925-2010-9	float64
wind-uwnd-925-2010-10	
wind-uwnd-925-2010-11	
wind-uwnd-925-2010-12	
wind-uwnd-925-2010-13	float64
wind-uwnd-925-2010-14	
wind-uwnd-925-2010-15	
wind-uwnd-925-2010-16	float64
wind-uwnd-925-2010-17	
wind-uwnd-925-2010-18	
wind-uwnd-925-2010-19	
wind-uwnd-925-2010-20	
wind-hgt-10-2010-1	float64
wind-hgt-10-2010-2	float64
wind-hgt-10-2010-3	float64
wind-hgt-10-2010-4	float64

	WiDs_Notebook2_MLPRegressor_Comparison_RFR_Datathon_Submission
wind-hgt-10-2010-5	float64
wind-hgt-10-2010-6	float64
wind-hgt-10-2010-7	float64
wind-hgt-10-2010-8	float64
wind-hgt-10-2010-9	float64
wind-hgt-10-2010-10	float64
wind-hgt-100-2010-1	float64
wind-hgt-100-2010-2	float64
wind-hgt-100-2010-3	float64
wind-hgt-100-2010-4	float64
wind-hgt-100-2010-5	float64
wind-hgt-100-2010-6	float64
wind-hgt-100-2010-7	float64
wind-hgt-100-2010-8	float64
wind-hgt-100-2010-9	float64
wind-hgt-100-2010-10	float64
wind-vwnd-925-2010-1	float64
wind-vwnd-925-2010-2	float64
wind-vwnd-925-2010-3	float64
wind-vwnd-925-2010-4	float64
wind-vwnd-925-2010-5	float64
wind-vwnd-925-2010-6	float64
wind-vwnd-925-2010-7	float64
wind-vwnd-925-2010-8	float64
wind-vwnd-925-2010-9	float64
wind-vwnd-925-2010-1	float64
wind-vwnd-925-2010-1	1 float64
wind-vwnd-925-2010-12	2 float64
wind-vwnd-925-2010-13	float64
wind-vwnd-925-2010-1	float64
wind-vwnd-925-2010-1	
wind-vwnd-925-2010-1	float64
wind-vwnd-925-2010-1	float64
wind-vwnd-925-2010-18	
wind-vwnd-925-2010-19	
wind-vwnd-925-2010-20	float64
BSh	int64
BSk	int64
BWh	int64
BWk	int64
Cfa	int64
Cfb	int64
Csa	int64
Csb	int64
Dfa	int64
Dfb	int64
Dfc	int64
Dsb	int64
Dsc	int64
Dwa	int64
Dwb	int64
month_number	int64

Out[95]:

```
In [92]:
          time test data = test data.copy()
          time test data.set index('startdate', inplace=Ecue)
          time test data.sort index(inplace=Brue)
In [93]:
          time test data.head(
Out [93]:
                                               contest-
                                                             nmme0-
                                                                           nmme0-
                                                                                        nm
                                             pevpr-sfc-
                    index
                               lat
                                        Ion
                                                             tmp2m-
                                                                           tmp2m-
                                                                                         tm
                                                gauss-
                                                       14d__pevpr
         startdate
          2022-11-
                   375734 0.000000 0.833333
                                                339.88
                                                               30.88
                                                                              30.92
               01
          2022-11-
                  404099 0.954545
                                   0.100000
                                                224.64
                                                                16.43
                                                                              17.98
               01
          2022-11-
                                                 417.74
                   394827
                          0.681818 0.600000
                                                                22.07
                                                                              24.84
               01
          2022-11-
                   386287 0.454545 0.766667
                                                               28.95
                                                                              33.25
                                                411.23
               01
          2022-11-
                   390679 0.590909 0.266667
                                                433.60
                                                               23.09
                                                                              23.65
         5 rows × 260 columns
In [94]:
          time test data.shape
Out[94]:
In [95]:
          X.shape
```

# Fit RFR on Full training\_data from WiDS

• We can use the full X and y dataset to train our model.

```
ouilding tree 4 of 100
ouilding tree 41 of 100
```

Out[100]:

```
In [102...
          pred = RFRmodel final.predict(time test data)
In [192...
          submission = pd.DataFrame(data =
               edictions':pred,'Index':time test data['index']})
In [193...
          submission=submission.sort values(by='Index', ascending=Exue)
In [194...
          submission
Out[194]:
                     Predictions
                                 Index
            startdate
           2022-11-01
                      28.385288 375734
           2022-11-02
                      28.325583 375735
          2022-11-03
                      28.530644 375736
          2022-11-04
                      28.483178 375737
          2022-11-05
                      28.648197 375738
          2022-12-27
                      3.478709 407083
          2022-12-28
                       3.660469 407084
          2022-12-29
                      3.361483 407085
          2022-12-30
                       2.882080 407086
          2022-12-31
                       2.885870 407087
         31354 rows × 2 columns
In [195...
          submission = submission.reset index()
```

In [196...

### submission

### Out[196]:

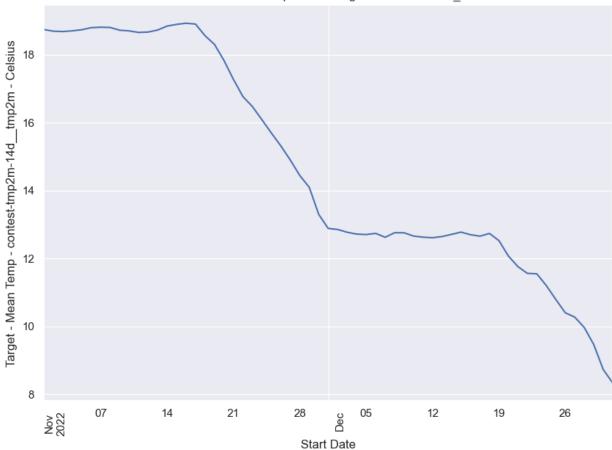
	startdate	Predictions	Index
0	2022-11-01	28.385288	375734
1	2022-11-02	28.325583	375735
2	2022-11-03	28.530644	375736
3	2022-11-04	28.483178	375737
4	2022-11-05	28.648197	375738
•••			•••
31349	2022-12-27	3.478709	407083
31350	2022-12-28	3.660469	407084
31351	2022-12-29	3.361483	407085
31352	2022-12-30	2.882080	407086
31353	2022-12-31	2.885870	407087

31354 rows × 3 columns

### In [211...

```
#Visualize mean daily prediction temps over startdate
plt.figure()
submission.groupby(by=['startdate']).mean()['Predictions'].plot()
plt.title('Mean Predicted Temperature/Target vs. Date of Test_Data')
plt.ylabel('Target - Mean Temp - contest-tmp2m-14d__tmp2m -
Celsius')
plt.xlabel('Start Date')
plt.xticks(rotation=90)
plt.show()
```

Mean Predicted Temperature/Target vs. Date of Test\_Data



These temps look reasonable and we will proceed with the submission.

```
In [212... # Rename predictions to target variable as in sample solution from
    WiDS
    submission = submission.rename(columns = {'Predictions':'contest-
        tmp2m-14d_tmp2m'})
In [215... # Drop start date to match sample solution
    submission = submission.drop(['startdate'], axis = 1)
In [216... # Check
    submission
```

Out[216]:		contest-tmp2m-14dtmp2m	Index
	0	28.385288	375734
2 3 4  31349 31350 31351	1	28.325583	375735
	2	28.530644	375736
	3	28.483178	375737
	4	28.648197	375738
	•••		•••
	31349	3.478709	407083
	31350	3.660469	407084
	31351	3.361483	407085
	31352	2.882080	407086
	31353	2.885870	407087

31354 rows × 2 columns

```
In [173... # Read in sample solution to compare format
sample_solution = pd.read_csv('data/sample_solution.csv')
```

In [174... # Check - Our submission is in the same format as the sample solution.

sample\_solution

Out[174]:		contest-tmp2m-14dtmp2m	
	0	27.073876	375734
	1	25.109308	375735
	2	22.557390	375736
	3	25.572875	375737
	4	20.781073	375738
	•••		•••
	31349	28.303967	407083
	31350	26.635933	407084
	31351	27.057762	407085
	31352	26.871066	407086
	31353	21.253714	407087

31354 rows × 2 columns

```
In [218... #Export Submission
submission.to_csv('data/submission_mack.csv', index=False)
In [219... #Export time_test_data
time test data.to csv('data/time test data clean.csv', index=False)
```

# **Notebook 2 Conclusion**

Workflow Initial MLPRegressor --> MLPRegressor Optimization --> MLPR vs. RFR Comparison--> RFR Baseline WiDS Submission

We have refined our ability to prepare and evaluate models in this notebook. We adopted a data reduction strategy to reduce each parameter's data type to minimize memory load and run times. We determined that the regressor models have quicker run times on my local machine over Google Collab. We can look into other methods of running cloud-based models such as AWS if run times becoming a major in inhibition to advancement.

My task as part of my WiDS team was to evaluate the Random Forest Regressor and Multilayer Perceptron Regressor models. Both models have proven effective in predictive regression time series modeling.

After optimizing the MLPRegressor, we see that the Random Forest Regressor optimized in Notebook 1 achieved better scoring metrics. We wanted to make sure to get in a submission to WiDS to establish our baseline and know that we could properly format a submission.

After fitting our optimized RFR with the full training\_data from WiDS, we ran the WiDS test\_data through the RFR model to gain predictions for our baseline submission. We successfully made the submission to the WiDS Kaggle site. WiDS evaluated our predictions on the actual target mean temperatures and we scored a root mean squared error of 1.91 degrees celsius. This gives us a good baseline, and we will feel confident formatting submissions in the future.

We will need to carry out a variety of steps in successive notebooks to achieve better results.

### **Next Steps**

- Feature engineering
  - We will research steps to boost the predictive power of our features.
- More comprehensive model evaluation
  - We will need to extend our grid searches to other regression models.