THE UNIVERSITY OF CHICAGO

Predicting Daily Mean Temperature

Final Project Milestone
MSCA 31006 4 Time Series Analysis

Antonia Sanhueza,
Earnest Salgado,
Guillermo Trefogli

Outline

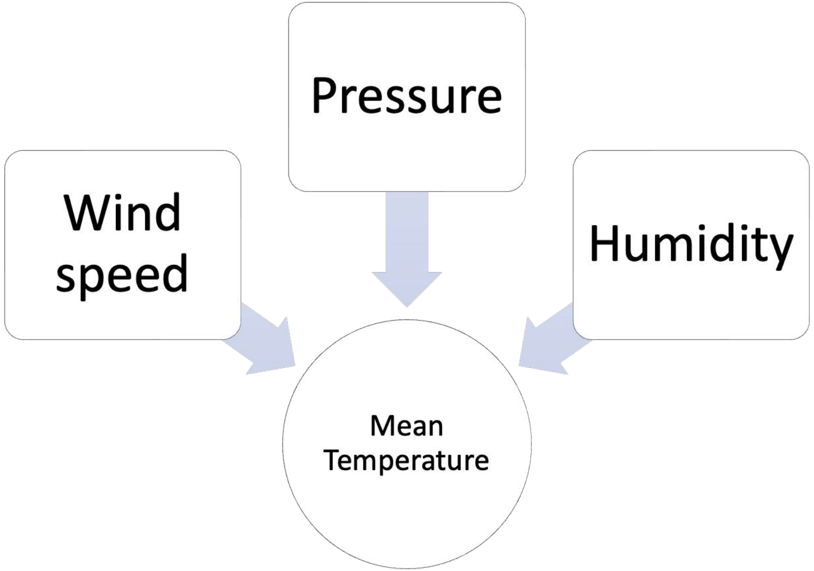
1. Business case and problem statement
2. Modeling hypothesis and assumptions
3. Data description
4. EDA and Feature Engineering
5. Proposed modeling approaches
6. Selected model results with justifications and tradeoffs
7. Insights/Recommendations & Future work

1. Business case and problem statement

- Changes in global temperature are becoming more abrupt and common every day (UN-IPPC).
- Predicting temperature is relevant for decision making in several business problems (i.e. Governmental investments for adaptation actions towards climate change)
- We make an attempt to predict the daily temperature based on past climate information.

2. Modeling hypothesis and assumptions

- We hypothesized that it is possible to fit a model using historical temperature levels to predict future mean temperature.
- We think that the other variables in the dataset will make our predictions more robust.



3. Data description

- Daily Climate time Series Data (source: Kaggle)
- Daily climate data for Delhi, India
- Period: 2013 to 2017
- Interest: Temperature

	meantemp	humidity	wind_speed	meanpressure
min	6.000000	13.428571	0.000000	-3.041667
max	38.714286	100.000000	42.220000	7679.333333
median	27.714286	62.625000	6.221667	1008.563492
mean	25.495521	60.771702	6.802209	1011.104548
std	7.348103	16.769652	4.561602	180.231668

4. Exploratory Data Analysis (EDA) and Feature Engineering

4. EDA and Feature Engineering

✔ No missing values

```
df.isnull().sum()
```

```
date            0
meantemp        0
humidity        0
wind_speed      0
meanpressure    0
dtype: int64
```

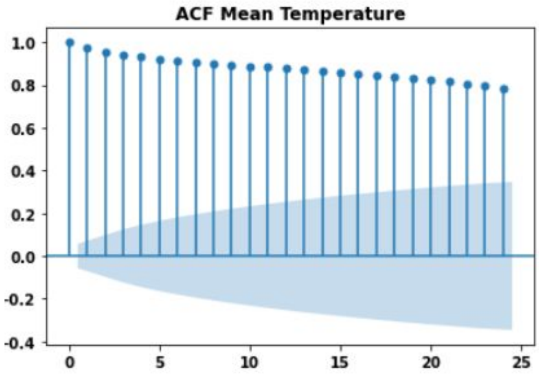
4. EDA and Feature Engineering

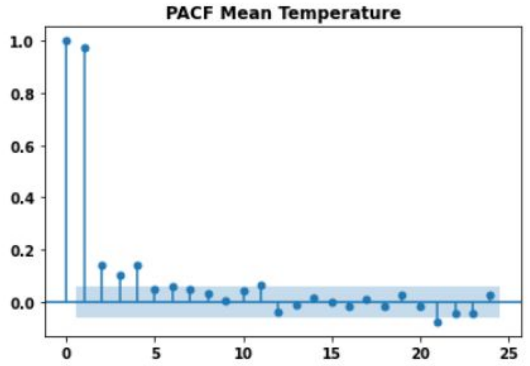
✔ Seasonality in the time series (peaks in summer, lows in winters)

4. EDA and Feature Engineering

- ✔ Non-stationarity based on ADF and KPSS tests, but difference stationarity

- ✔ ADF test: p-value 0.19 (False)
  "Non-stationarity cannot be rejected"
- ✔ KPSS test: p-value 0.1 (True)
  "Stationarity cannot be rejected"





4. EDA and Feature Engineering

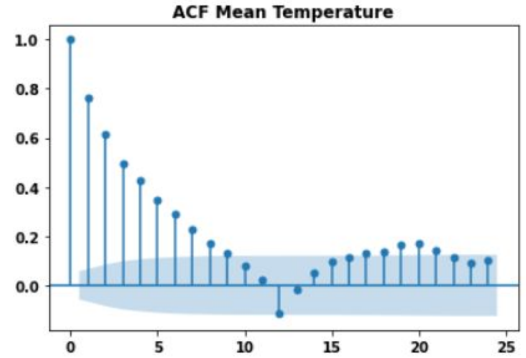
- ✔ Time series turns stationary after differentiating both by seasonal (m=12) and nonseasonal (P=1) patterns

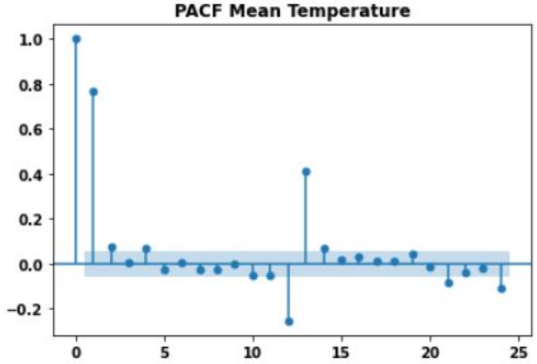
Non-seasonal component:

- ✔ ADF test: p-value 1.39702e-26 (True)
- ✔ KPSS test: p-value 0.1 (True)

Seasonal component:

- ✔ ADF test: p-value 4.49425e-08 (True)
  “Non-stationarity can be rejected”
- ✔ KPSS test: p-value 0.1 (True)
  “Stationarity cannot be rejected”





5. Proposed modelling approaches

We started exploring different models from more simple, to more complex. We developed 4 models:

1) Seasonal ARIMA/ AutoARIMA
2) Granger Causality and VAR/VARMA
3) Prophet
4) Random Forest Regression

Evaluation metric: Root Mean Squared Error

6. Selected model results with justifications and tradeoffs

6. 1 Seasonal ARIMA/ AutoARIMA

ARIMA:

- Model order: (1,0,0) (2, 1, 1, 12)
- AIC ~ 4,503
- Ljung-Box Test: fail to reject autocorrelation of residuals

Auto ARIMA:

- Model order (2, 0, 1)(3, 1, 0, 12)
- AIC ~ 4,542
- Ljung-Box Test: no autocorrelation of residuals
- RMSE: 12

6. 1 Seasonal ARIMA/ AutoARIMA

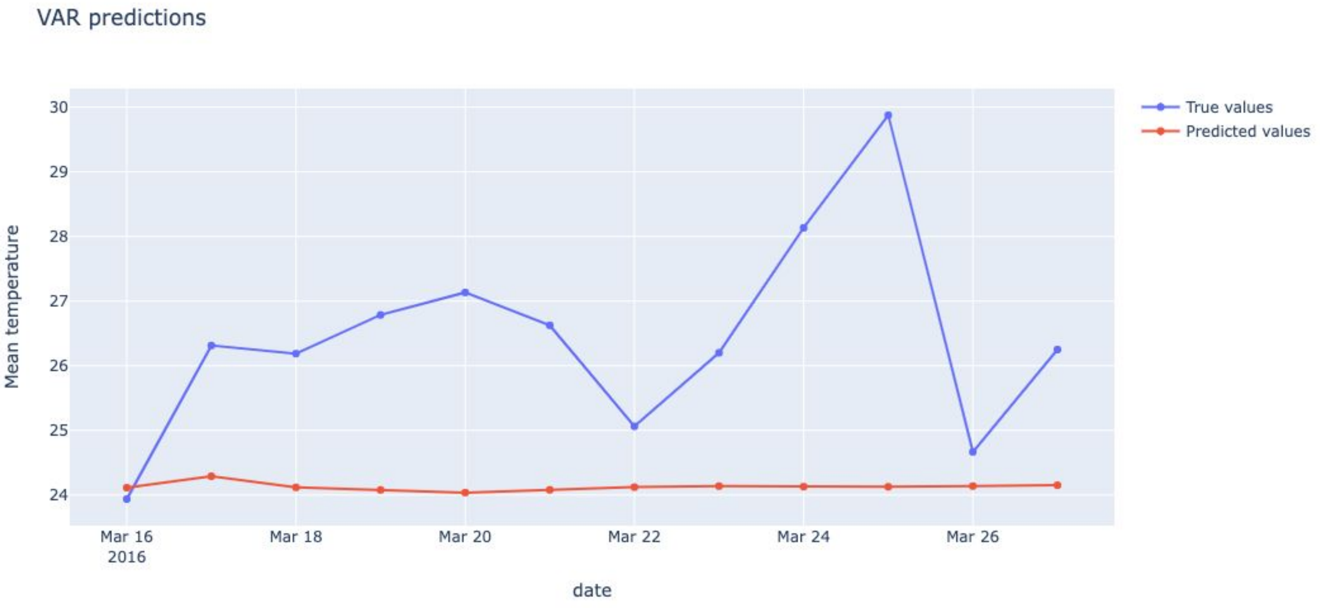
p = 1

6. 1 Seasonal ARIMA/ AutoARIMA

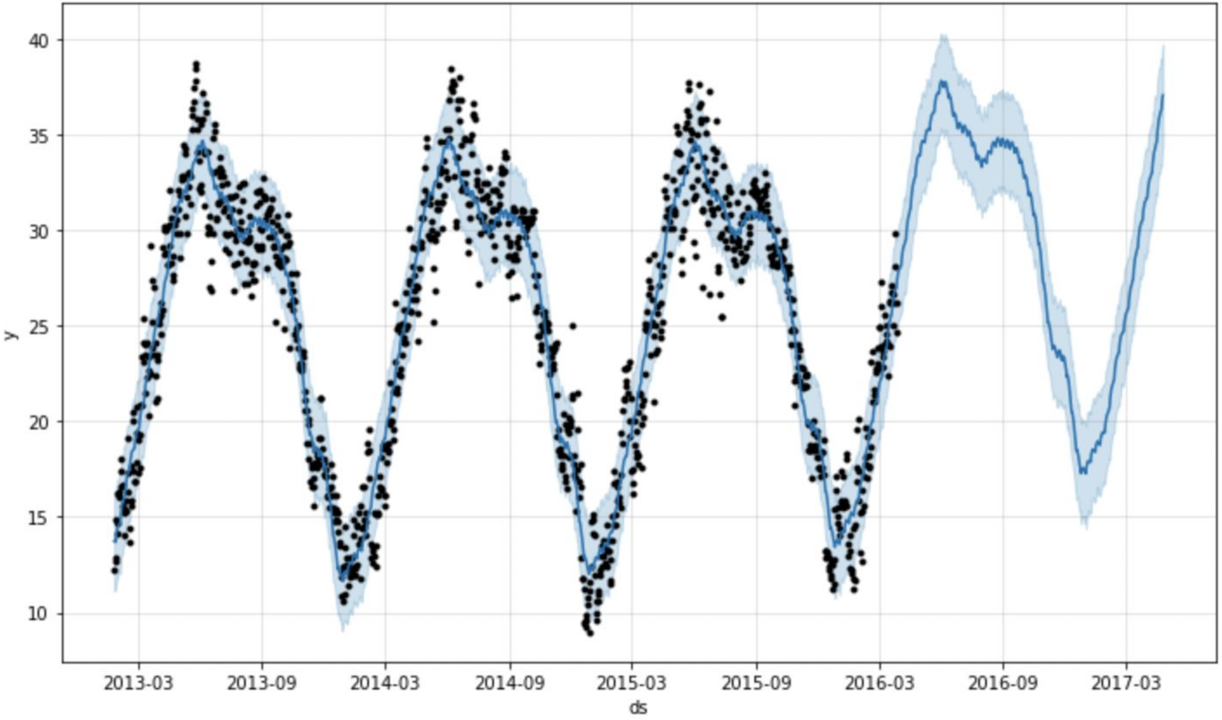
p = 9

6. 2 Granger Causality and VAR/VARMA

- Significant lags for humidity and wind speed
- Model order (3, 0)
- AIC: 7.37
- RMSE: 2.7



6. 3 Prophet



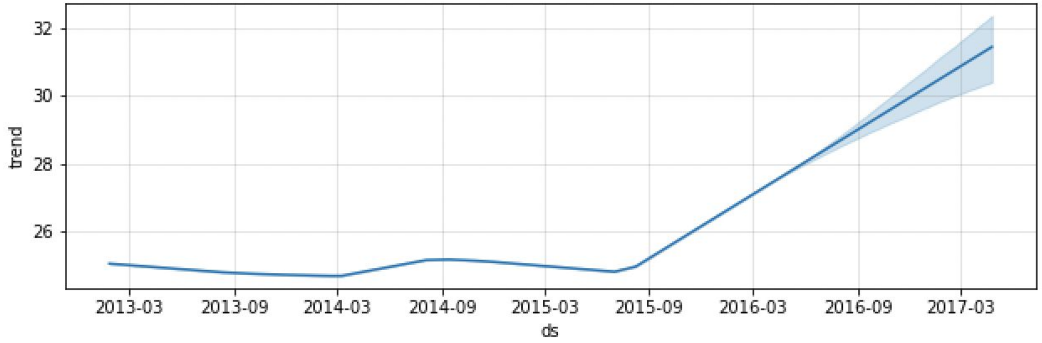
40
35
30
25
20
15
10
y
2013-03
2013-09
2014-03
2014-09
2015-03
2015-09
2016-03
2016-09
2017-03
ds

6. 3 Prophet

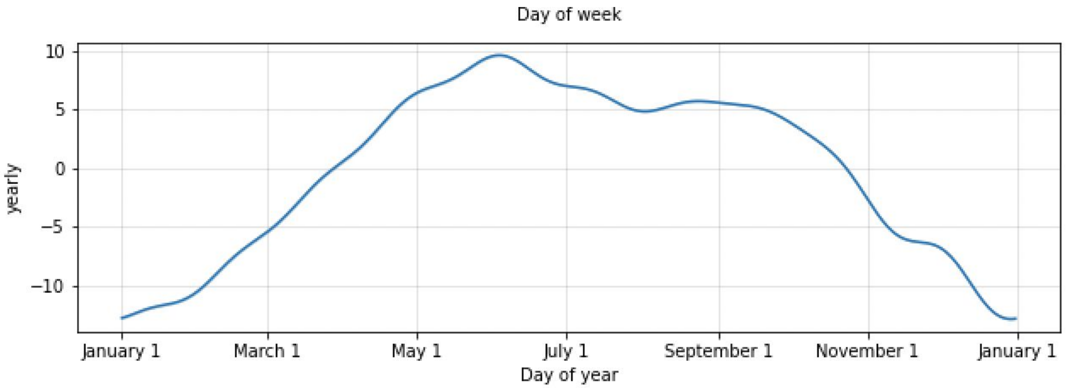
RMSE: 3.49

THE UNIVERSITY OF CHICAGO

6. 3 Prophet



32
30
28
26
trend
2013-03
2013-09
2014-03
2014-09
2015-03
2015-09
2016-03
2016-09
2017-03
ds



Day of week
10
5
0
−5
−10
yearly
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year

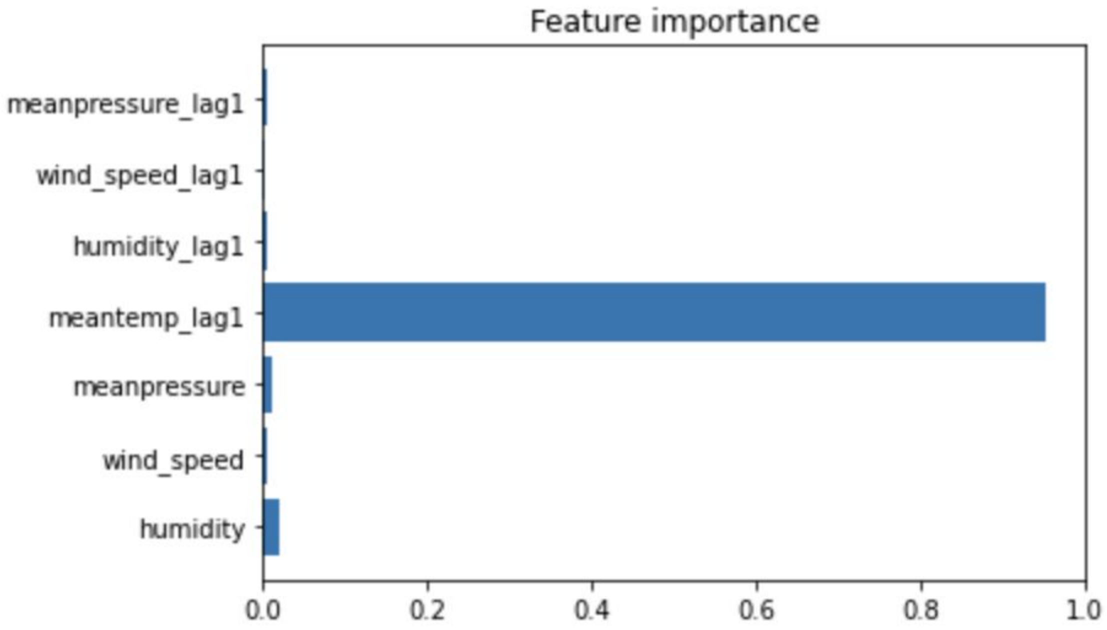
6. 4 Random Forest Regression - Sklearn

Model parameters:

- num_estimators: 1000
- min_interval: 3

RMSE: 0.4

6. 4 Random Forest Regression - Sklearn



6. 5 Random Forest Regression - Sktime

Model parameters:

- num_estimators: 1000
- min_interval: 3

RMSE: 1.77e-13

7. Insights/Recommendations & Future work

- From these can conclude Model choice is dependent on purpose
  - In classroom/theoretical settings, may choose ARIMA/Prophet for deeper understanding of measurements (e.g., p, q, # of lags)
  - In professional settings, it may be most efficient to opt for Machine Learning techniques
- RMSE is a robust comparison metric when observing performance across all models (SARIMA, AutoARIMA, Prophet, Granger, VAR, ML)
- In future works, it is insightful to apply these modeling approaches on other locations on Earth!