PREDICTION OF PARTICULATE MATTER (PM2.5) IN LUCKNOW, INDIA

Shaunak Phatak

OUTLINE

- Problem Context and Definition
- Data Wrangling and Exploratory Data Analysis (EDA)
- Modeling Results
- Conclusion

PROBLEM CONTEXT

- Air pollution is a serious health hazard with 8.8 million deaths globally
- Particulate matter is very harmful due to its microscopic size entering lungs and bloodstream
- Serious problem in India with 0.98 out of 1.67 million deaths due to particulate matter in 2019
- Project aims at using publicly available pollution data to create a forecasting model for particulate matter (PM2.5)
- Model can be further extended into an app to warn people and provide recommendations to take necessary precaution

Reference: https://www.gettyimages.in/detail/illustration/deadly-smoke-pollution-from-industrial-smoke-royalty-free-illustration/485276208?adppopup=true

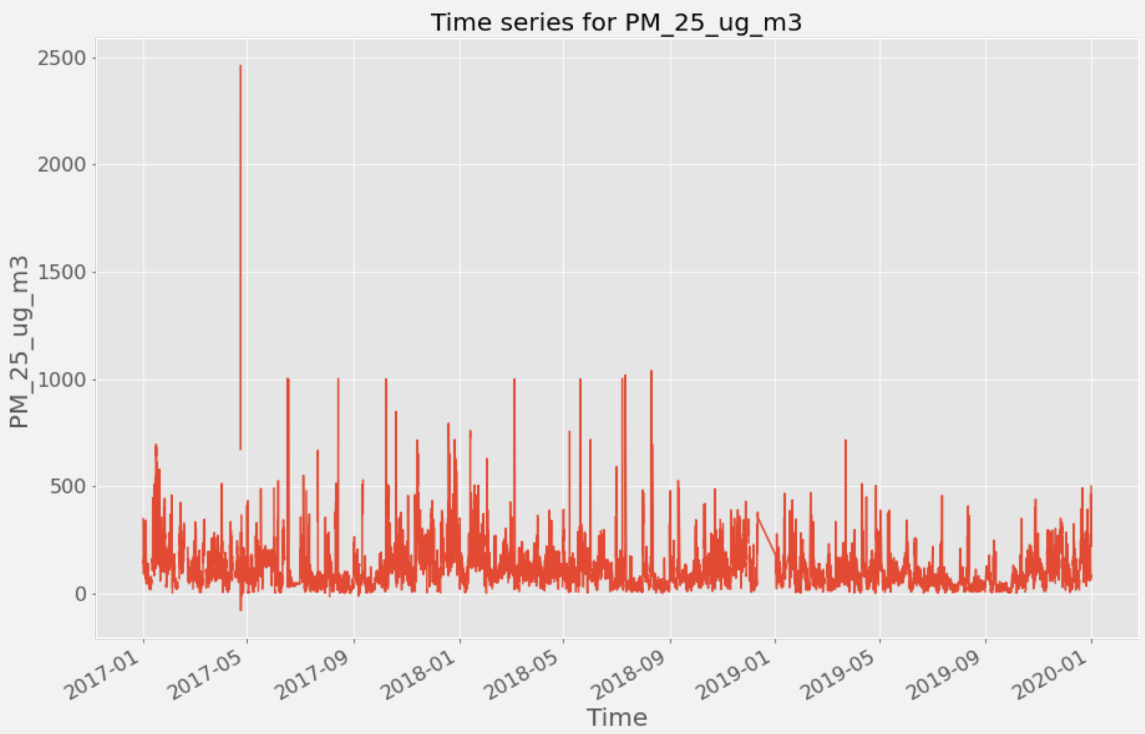
PROBLEM DEFINITION

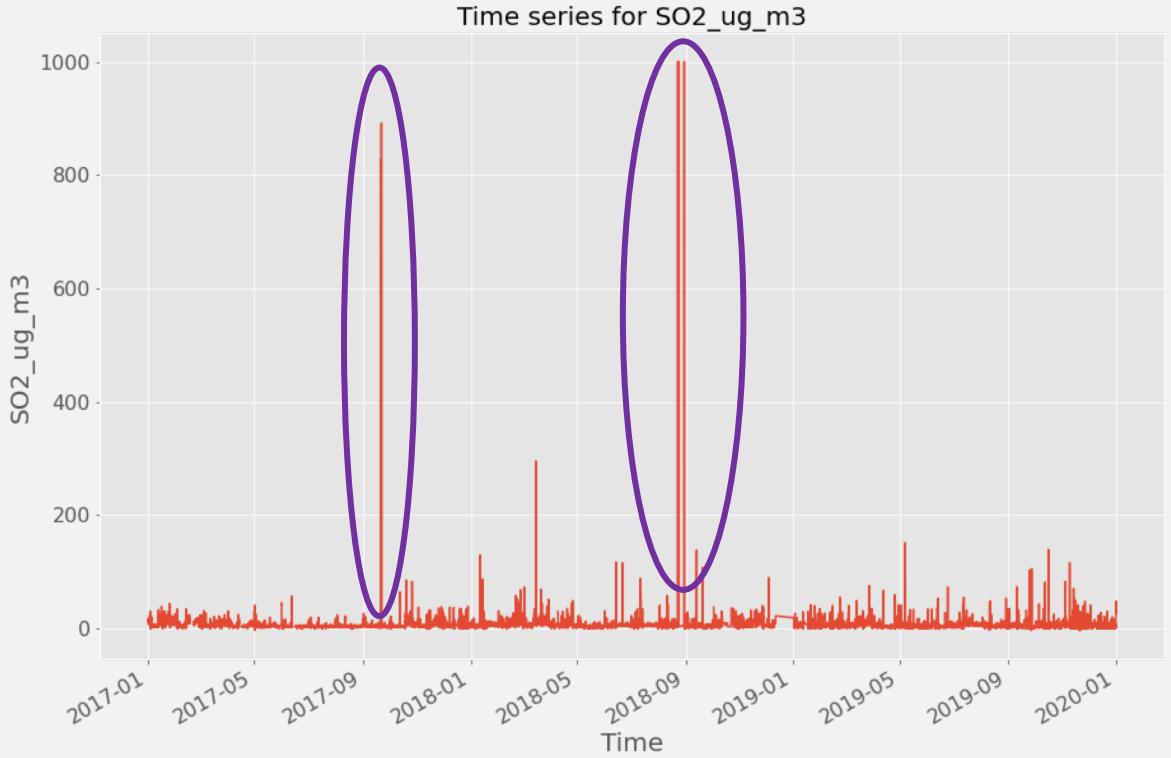
- Create an hourly forecasting model for particulate matter (PM2.5) in Lucknow, India applying supervised learning for a multivariate time series dataset
- Dataset Information:
  - Pollution data - Central Pollution Control Board, India
  - Temperature data - National Oceanic & Atmospheric Administration
- Independent Variables: Weather, other air pollutants and time based features
- Dependent Variable: PM2.5

DATA WRANGLING AND EDA

- Missing values for the dependent variable (PM2.5) were dropped
- Missing values for independent variables were imputed by their column means
- Outlier treatment methods tested:
  - Percentile Capping
  - Bounding by 3 standard deviations
- Did not use the above methods due to significant changes to the data distribution

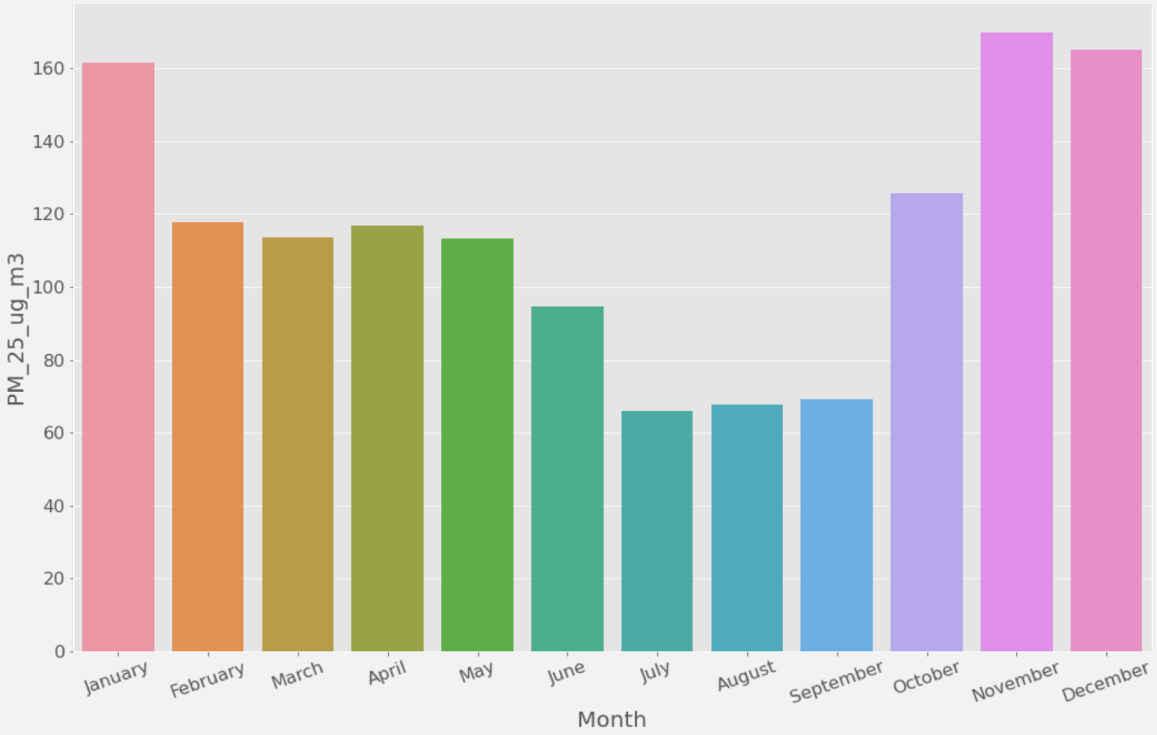
DATA WRANGLING AND EDA

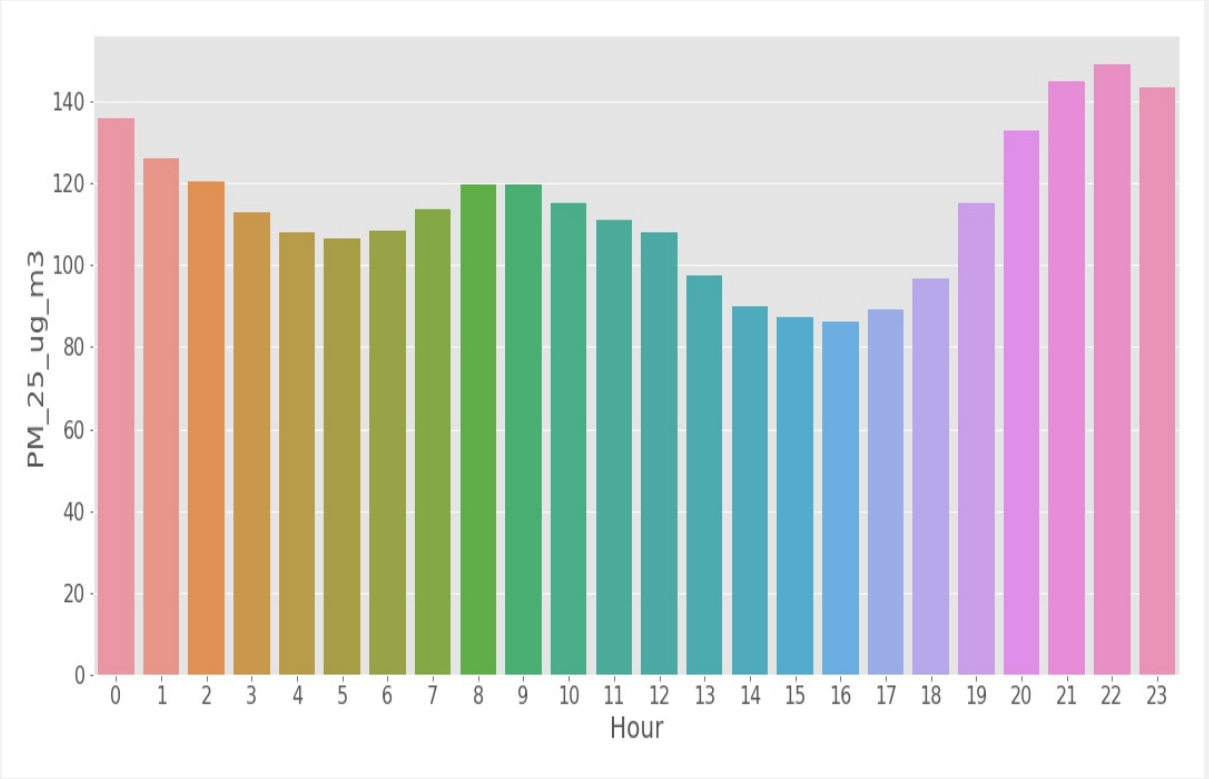




- Some outliers as marked on the figures were removed for PM2.5 and sulphur dioxide
- No other outlier changes were made

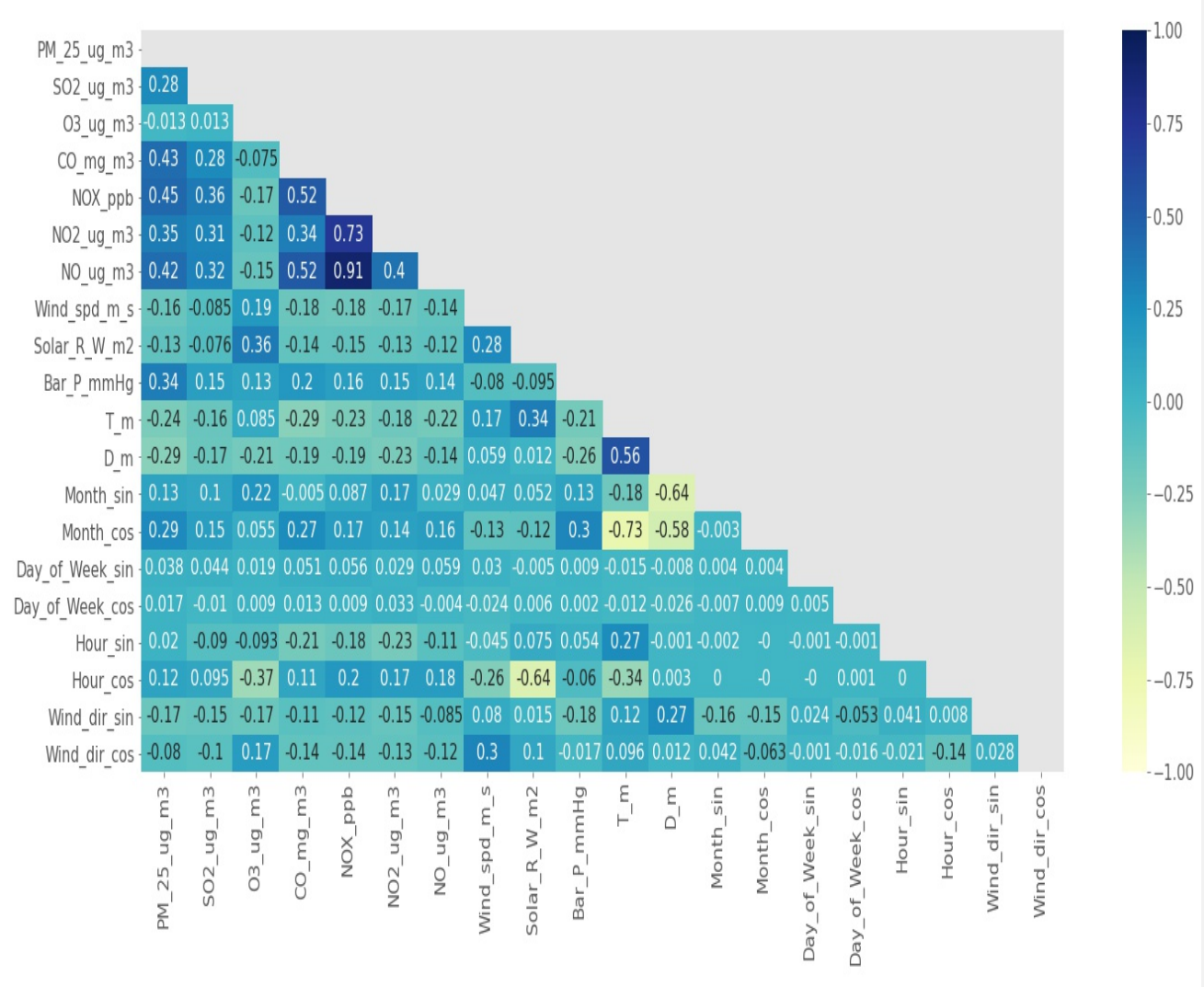
DATA WRANGLING AND EDA





- Extracted time based features from the date-time index
- Cyclic variations in PM2.5 concentration across month as well as hour of day

DATA WRANGLING AND EDA



- Negative correlation between PM2.5 and temperature and wind speed
- Moderate positive correlations between PM2.5 and other pollutant features
- Multicollinearity between other pollutant features
- Heat map also shows cyclical features converted to sine and cosine components to retain cyclical information

MODELING

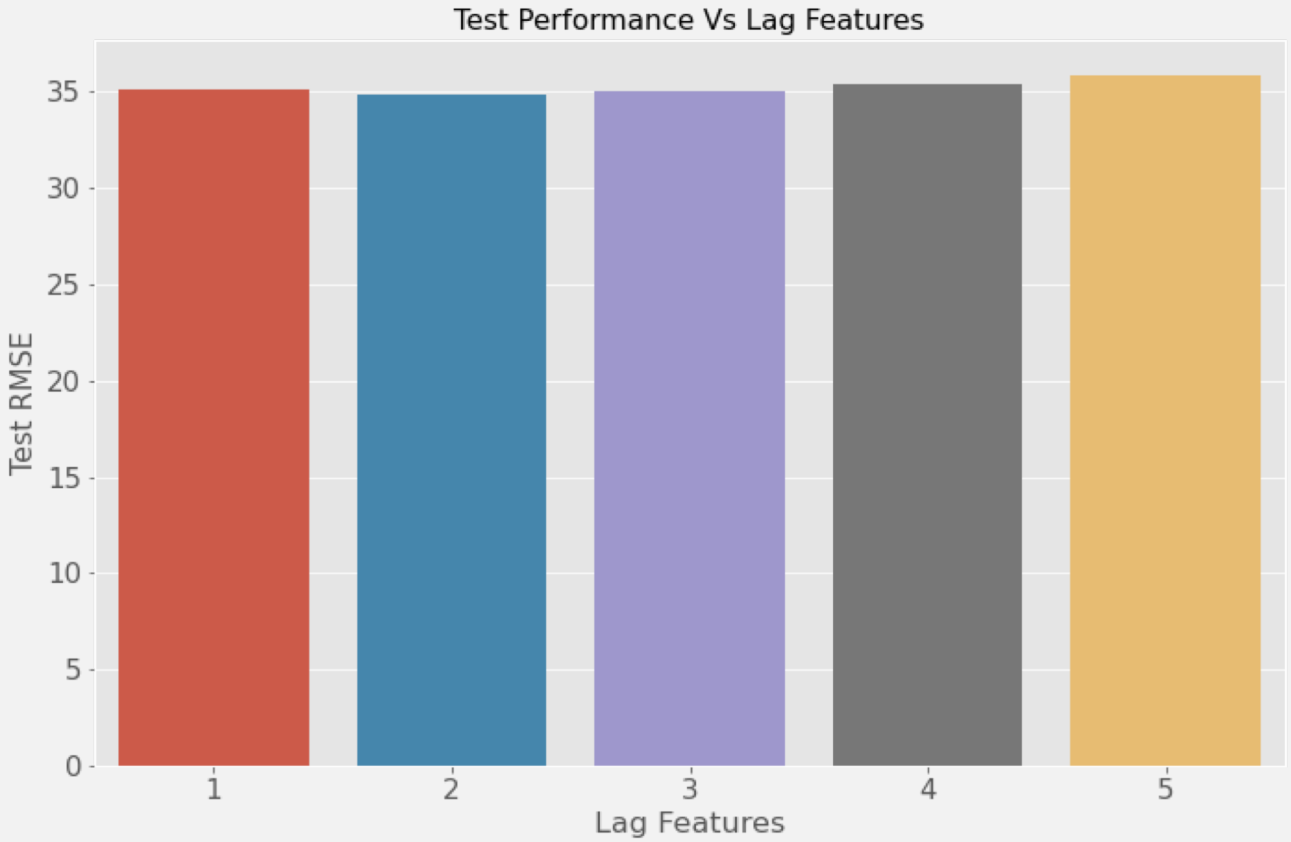
- Modeling studies involved two case studies:
  - Performance Comparison for randomly split train/test sets versus splits on temporal order
  - Comparing model performances while applying walk forward validation
- Nested cross validation to tune hyperparameters
- Tuning lag features for PM2.5 as a hyperparameter
- Check effectiveness of using some or all features

MODELING:
RANDOM VS TIME SERIES SPLIT

Split	Lag Features	Avg. Train RMSE	Avg. Test RMSE	Avg. Train MAE	Avg. Test MAE
Random Split	No	59.28	72.18	36.78	47.93
	Yes (3)	28.66	36.89	13.7	16.34
Time based Split	No	67.94	83.22	40.5	59.17
	Yes (3)	30.92	33.87	14.07	16.95

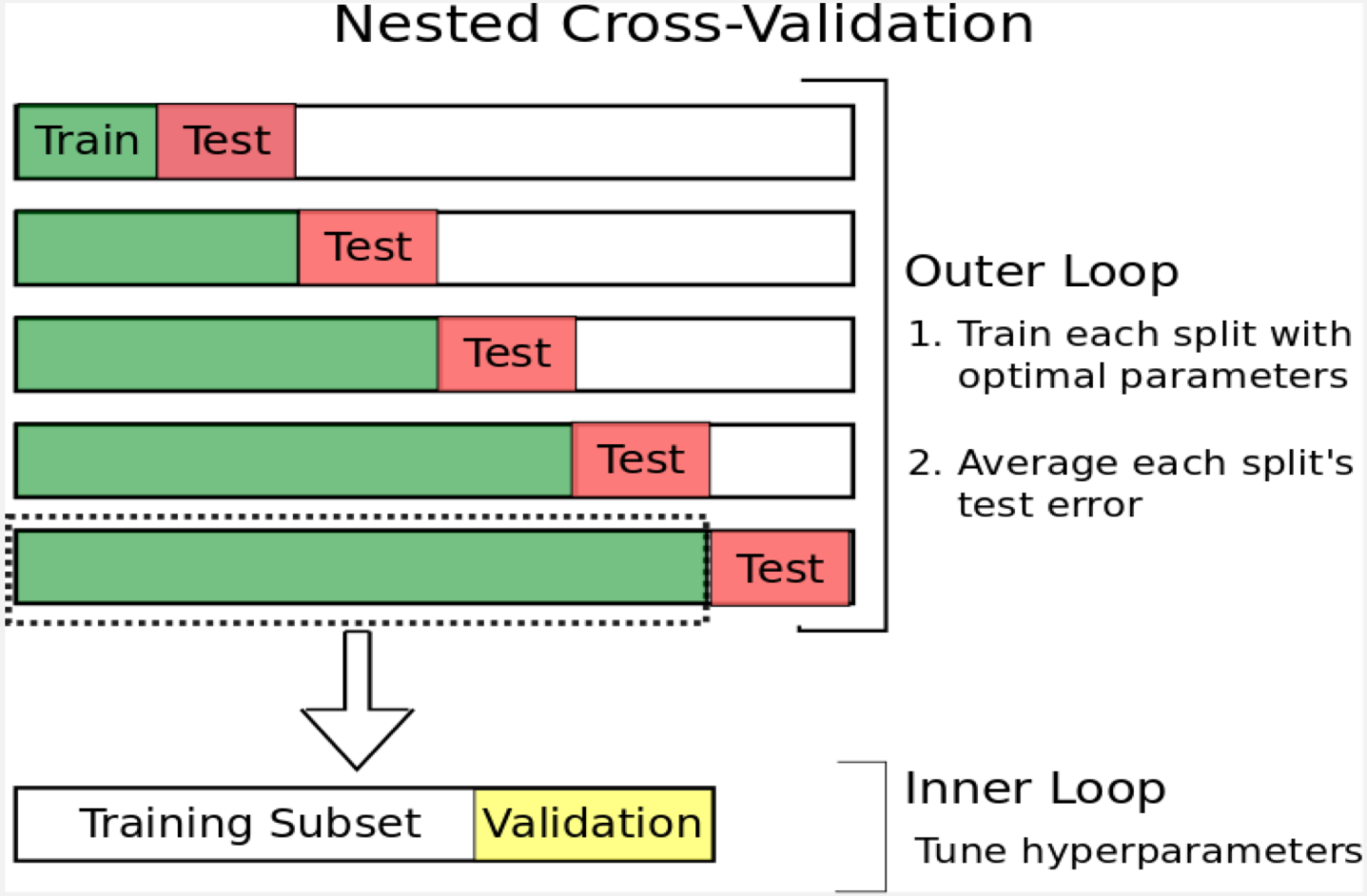
- Random split performed better than time series split without any lag features
- Similar performance when lag features were added
- All features were used for these studies without any feature selection process
- Results for a Light Gradient Boosting Model

MODELING
TUNING THE LAG FEATURE



- No performance improvements beyond using the 1st lag feature
- Similar behavior across all models tested

MODELING



- Time series split avoids data leakage that can happen for random splits
- Nested cross-validation avoids a problem of choosing an arbitrary test set to make predictions

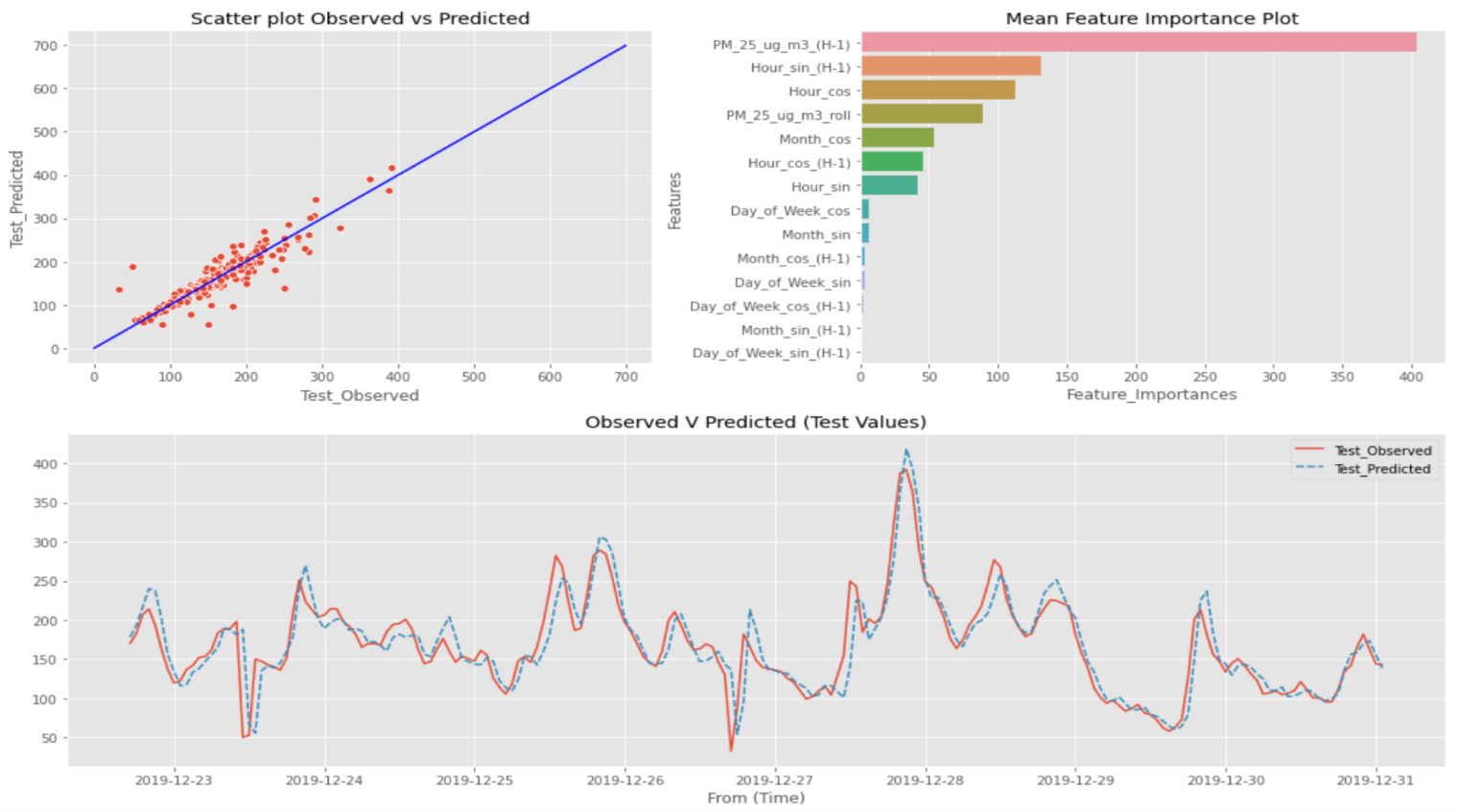
Reference: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

MODELING:
WALK FORWARD VALIDATION RESULTS

Model	Avg. Train RMSE	Avg. Test RMSE	Avg. Train MAE	Avg. Test MAE
Random Forest	37.49	33.56	14.82	15.66
Gradient Boost	38.62	34.35	17.39	16.62
Light Gradient Boost	38.6	33.85	16.88	16.22
XGBoost	38	34.05	16.68	16

- 1st lag and time features sufficient to maximize modeling performance
- Weather and pollutant variables were not useful to further improve performance
- Average test set metrics similar across all models
- Models based on a 3 splits for train/test as well as nested cross-validation

MODELING:
WALK FORWARD VALIDATION EXAMPLE



- Modeling results for 200 hourly walk forward predictions using Light Gradient Boost
- 1st lag feature had the highest importance

CONCLUSIONS

- Project aimed to predict hourly PM2.5 values using supervised learning for a multivariate time series
- Performance Comparison between random split and time series split
  - Random split fared better when no lags were added
  - Similar performance with lag features
  - Improved performance with lag features
- Model Comparison with Walk forward nested cross-validation
  - 1st lag and time features sufficient to maximize performance
  - Weather and pollution features not useful to improve model results
  - Similar performance across all models tested
  - Light Gradient Boost can be chosen as a final model due to its fast training time

REFERENCES

1. https://www.theguardian.com/environment/2019/mar/12/air-pollution-deaths-are-double-previous-estimates-finds-research
2. https://ourworldindata.org/air-pollution
3. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM
4. https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(20)30298-9/fulltext
5. https://www.researchgate.net/publication/341618027_Forecasting_Sales_of_Truck_Components_A_Machine_Learning_Approach
6. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9
7. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting
8. https://medium.com/mongolian-data-stories/ulaanbaatar-air-pollution-part-1-35e17c83f70b
9. https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/