PSDS 4900 Capstone

JD Davis
11 Aug 2021

Goal

Determine if any one, or a combination of several, weather data sources can accurately predict the wind speed at my personal weather station in my backyard.

Importance

Currently, the Homeowners Association where I live does not have any architectural standards for personal wind-electric generation equipment.

I hope to produce a model to retroactively predict an approximation of wind speed in my backyard.

I intend to then use the prediction to estimate roughly how much electricity I might have produced in a given past time range.

Eventually I hope to develop and propose architectural guidelines for wind-electric generation equipment for our community using my findings.

Data Sources: My Sensor

- Ambient Weather WS-2902C
  - solarRadiation: Float
  - uv: Float
  - winddir: Integer
  - humidity: Float
  - temp:Float
  - windSpeed: Float
  - pressure: Float
  - precipRate: Float
- Data exported in Ambient format and/or Weather Underground format

Data Sources: Weather Underground

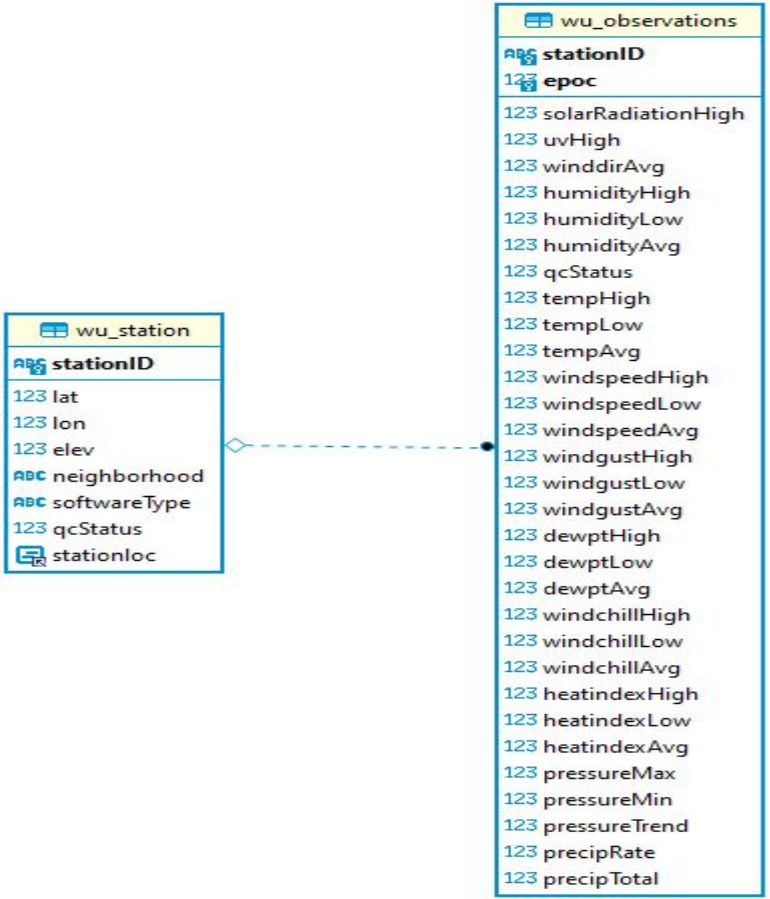
- Robust API
- Ability to query for historic data
- Weather station attributes available (lat/lon, station identifiers, etc)
- Data format identical for all stations (including mine)

Local DB For Storage

Local storage was necessary due to Weather Underground API limitations

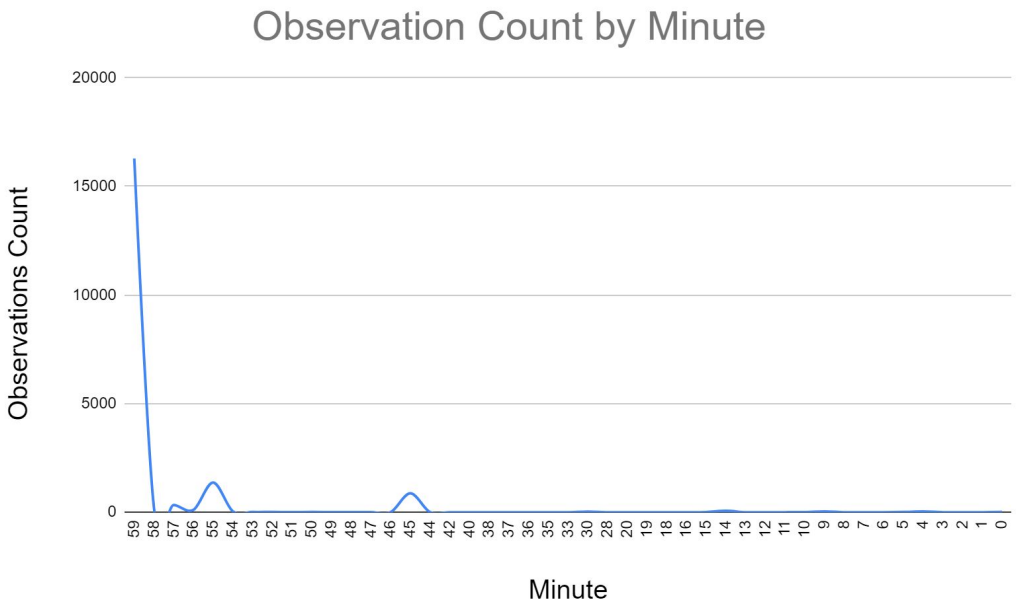
Raspberry Pi-based MariaDB

- 16 March - 3 July 2021
- 114 Stations
- > 246,000 Observations



Data Alignment

- Stations self-report measurements on their own schedule
- Majority of stations report at the bottom of the hour, but not all
- For use in prediction models, data must be aligned by time



Maximizing Target Station Representation

	5min	10min	15min	20min	30min	45min	60min
My Station Present	266	265	265	264	263	263	263
Total Groups	880	590	577	540	314	351	263
Percent Groups Containing My Station	30.23%	44.92%	45.93%	48.89%	83.76%	74.93%	100.00%

Deduplication of Data

- Alignment introduced/compounded the problem of stations reporting multiple observations within one grouping
- Multi-step process to remove duplicates
  - Keep target station observation closest to the bottom of the hour
  - Compute the time delta between every observation in the group and target
  - Keep the observation with the smallest absolute value of delta

Choosing Model Input - Step 1

- Create correlation matrix of data from my weather station
- Include all variables correlated to windspeedAvg at > 0.8
- Use most highly correlated stations for those variables

windspeedAvg	1
windgustAvg	0.995071
windspeedHigh	0.852027
windgustHigh	0.838374
windgustLow	0.523414
windspeedLow	0.459209
solarRadiationHigh	0.370677
uvHigh	0.364316
tempHigh	0.286263
windchillHigh	0.28625

Choosing Model Input - Step 2

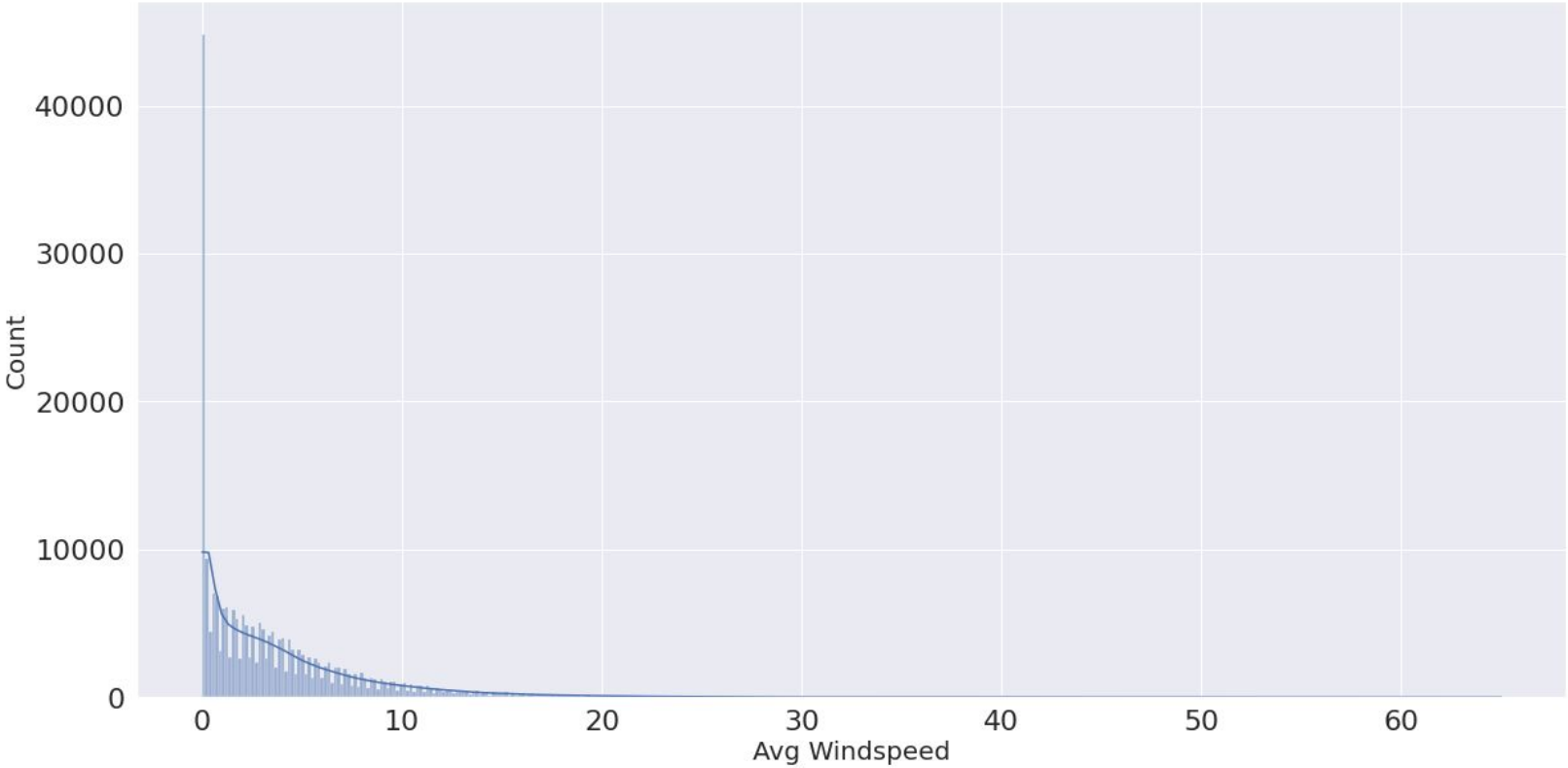
- Create correlation matrices for all stations compared to my station
- Choose stations with highest correlation for each of the variables from Step 1

stationID	windspeedAvg_corr	windgustAvg_corr	windspeedHigh_corr	windgustHigh_corr
KCOCASTL148	0.448414	0.547012	0.675833	0.730436
KCOCASTL161	0.765005	0.744230	0.788360	0.801225
KCOCASTL167	0.708132	0.727253	0.750650	0.758954
KCOCASTL195	0.638610	0.673448	0.747853	0.765771
KCOCASTL200	0.678979	0.721673	0.761377	0.782530
KCOCASTL204	0.716807	0.731671	0.749011	0.759554
KCOCASTL205	0.648407	0.692907	0.756221	0.777309
KCOCASTL208	0.746687	0.745058	0.758229	0.763326

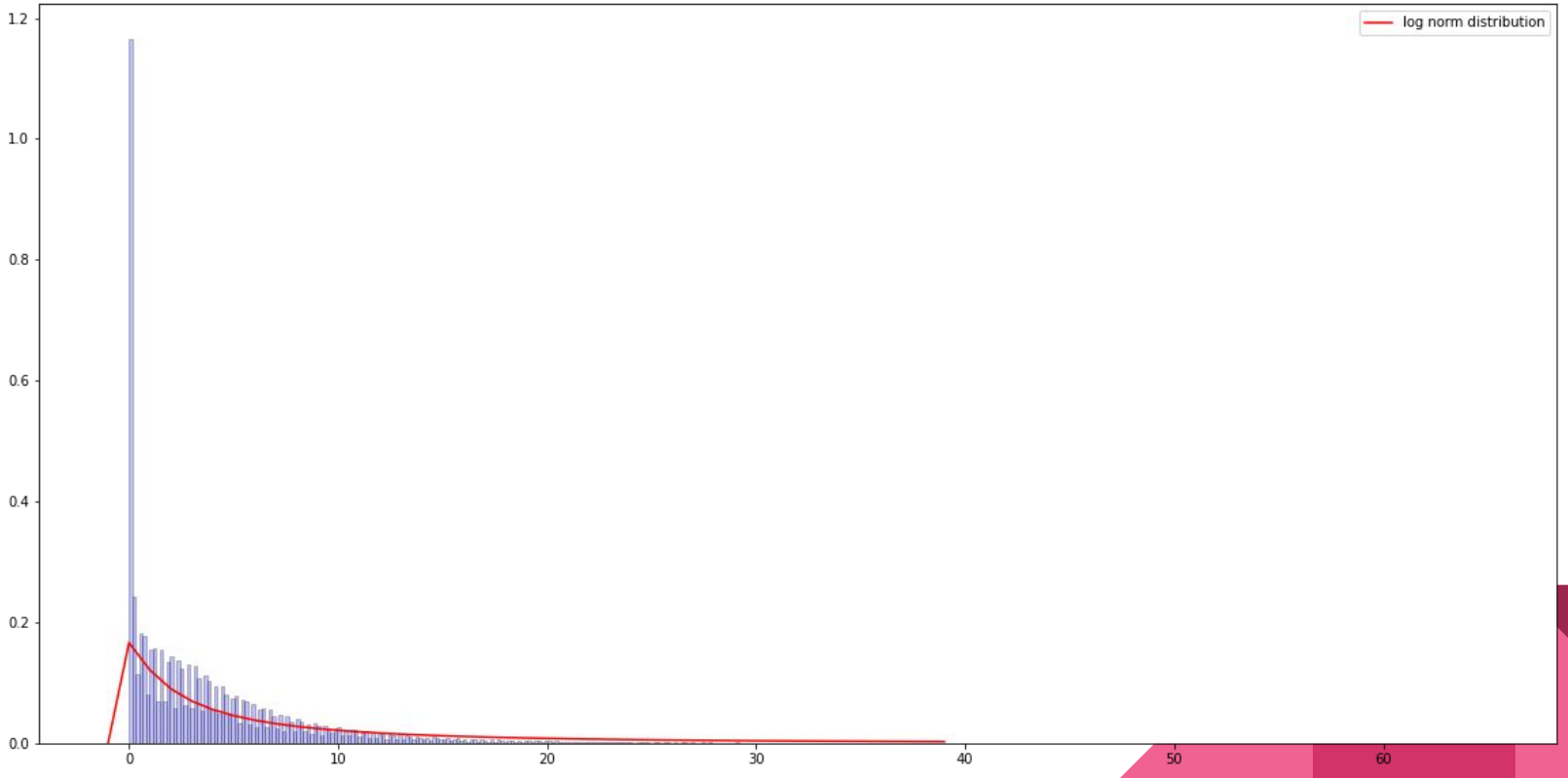
Choosing Model Input - Step 3

- Identify N most highly correlated stations for each of the variables from Step 1 using correlation matrix from Step 2
  - Pull observations of variables (Step 1) for N number of stations
  - Perform Linear Regression using N stations and score
  - Store number of stations with best score
- Use N stations for training and testing other models

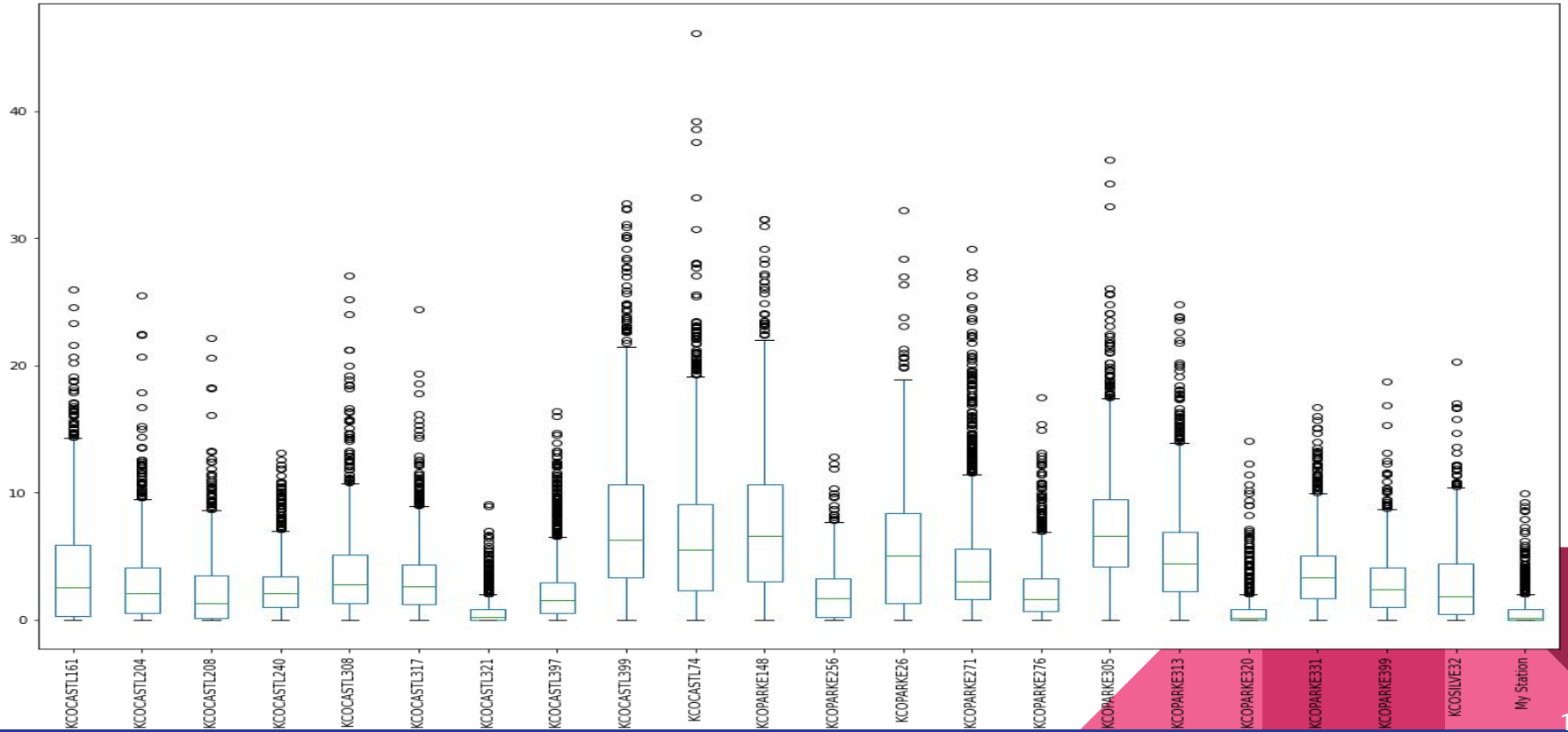
Data Distribution



Data Distribution: PDF With Log Norm Overlay



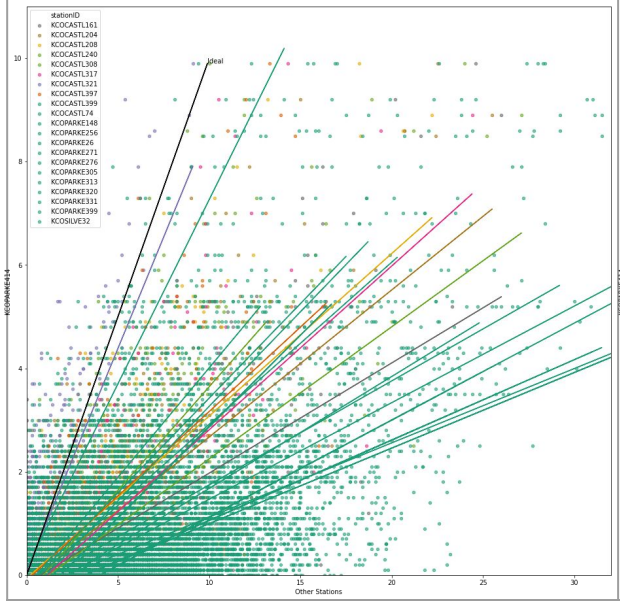
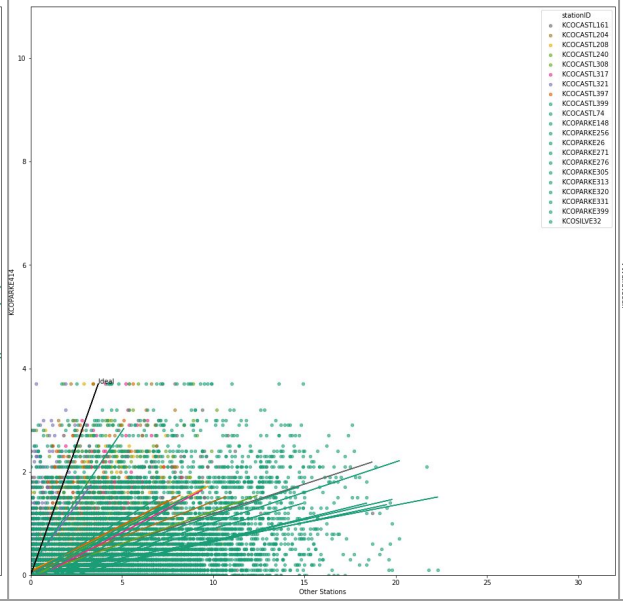
windspeedAvg For Most Correlated Stations



Impact of Removing Outliers

	Entire Dataset	Outliers Removed via Isolation Forest	Outliers Removed via LocalOutlierFactor
Rows Dropped	0	159 (13.27%)	31 (2.59%)
Linear Regression Accuracy	0.873127	0.638218 (-27%)	0.680896 (-22%)
Ridge Regression Accuracy	0.766903	0.620214 (-19%)	0.739091 (-4%)
Lasso Regression Accuracy	0.768726	0.610656 (-20%)	0.70588 (-8%)
SVR (Poly degree 2) Accuracy	0.766903	0.697551 (-9%)	0.746221 (-3%)

Impact of Removing Outliers

Entire Dataset	Outliers Removed via Isolation Forest	Outliers Removed via LocalOutlierFactor
		

Model Selection, Tuning and Training

- Experimented with
  - 10 different regression methods
  - Two different methods for outlier elimination
  - With and without input scaling
- Hyperparameter Tuning via Pipeline/GridSearchCV
- 85/15 Train Test Split
  - 1198 observations of
  - 61 variables from
  - 27 different stations
- Scoring is mean of 5-fold cross validation scores

Regressors, Best Parameters and Accuracy Scores

Model	Best Hyperparameters	Best Avg Accuracy
LinearRegression	N/A	0.873127
ElasticNet	'alpha': 0.01, 'l1_ratio': 1.0, 'max_iter': 2000, 'normalize': False	0.83345688
HuberRegressor with StandardScaler	'epsilon': 2.3, 'max_iter':201	0.803334
Lasso	'alpha': 0.01	0.768726
SVR poly kernel degree 2	N/A	0.766903
Ridge	'alpha': 0.01, 'normalize': True	0.765682
SVR poly kernel 2 localoutlierfactor removed	N/A	0.746221

Regressors, Best Parameters and Accuracy Scores

Model	Best Model Parameters	Best Avg Accuracy
HuberRegressor	'epsilon': 2.1, 'max_iter': 151	0.741381
Ridge localoutlierfactor removed	'alpha': 0.02, ‘normalize': True	0.739091
LinearRegression with StandardScaler	N/A	0.738264
BayesianRidge	'alpha_1': 0.09, 'alpha_2': 0.01, 'normalize': True	0.706051
Lasso localoutlierfactor removed	'alpha': 0.02	0.70588
SVR poly kernel 2 iso outliers removed	N/A	0.697551
Linear Regression localoutlierfactor removed	N/A	0.680896

Regressors, Best Parameters and Accuracy Scores

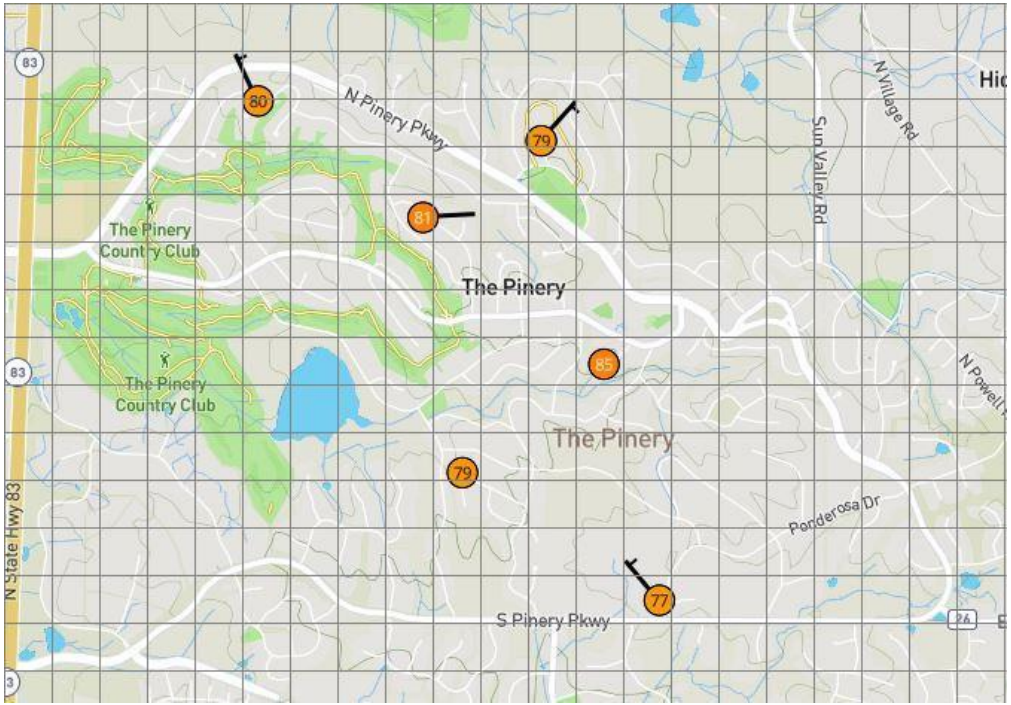
Model	Best Model Parameters	Best Avg Accuracy
Poisson	'alpha': 1.8	0.654925
Linear Regression iso outliers removed	N/A	0.638218
Ridge iso outliers removed	'alpha': 0.02, 'normalize': True	0.620214
Lasso iso outliers removed	'alpha': 0.01	0.610656
RandomForestRegressor	'max_depth': 100, 'max_features':'auto', 'n_estimators': 50	0.582931
Poisson localoutlierfactor removed	‘alpha': 1.8	0.476733
Poisson with StandardScaler	‘alpha': 0.2	0.438463
Tweedie	‘alpha': 1.9 ‘power”: 1	0.390051

Predicting January

- Best Model: LinearRegression
  - Average accuracy steady around 0.87
- Best Stations: 15 Most Correlated
- Retrieved Weather Underground data for target month: January 2021
  - After cleaning and grooming observations: 580 hour groups
- Assess Model Performance
  - Hourly average wind speed prediction: 0.439 mph
  - Expected hourly average wind speed between 0.5 and 0.38 (+/- .13)

Next Steps

- Use other personal weather stations to build new models
- Assess wind speed estimations for areas that do not have direct measurements
- Architectural Standards recommendations



Classified Applications & Discussion

Resources

GitHub: https://github.com/riverdogcabin/PSDS4900

Gustavsson, Sara. Sahlgrenska Academy at University of Gothenburg, “Evaluation of Regression Methods for Log-Normal Data Linear Models for Environmental Exposure and Biomarker Outcomes,” Occupational and Environmental Medicine Institute of Medicine, ISBN (e-publ.) 978-91-628-9295-1, 2015. [Online]. Available: https://gupea.ub.gu.se/bitstream/2077/37537/4/gupea_2077_37537_4.pdf. [Last Accessed: Jul. 28, 2021].