

Electricity Demand Forecasting

Ayan Nandi, an2683

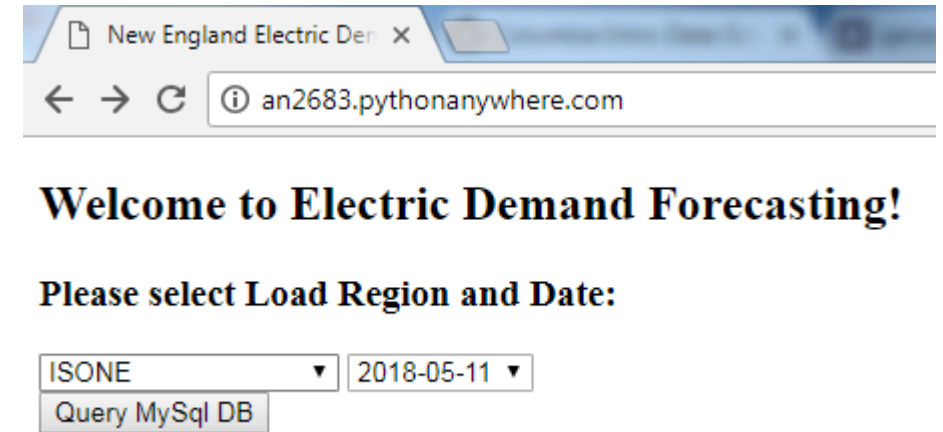
APMA E4990 Final Project

Problem Statement / Motivation

- The Price of Power in a particular region is highly related to the demand in that area – and as a residential power customer, you pay a fixed rate which compensates your provider for the risk of your usage and market Power Prices being highly correlated (embedded option).
- You could save money each month by becoming a Variable Rate Customer – but then you would bear the market risk yourself. One thing you can do to save money every month is time your usage around the Peak Demand for Electricity in your region. This means you avoid getting billed for large amounts of Power during the most expensive (High Demand Hours).
- Having a forecast of tomorrow's load can help you understand what times of the day Power will cost you the most.

Application

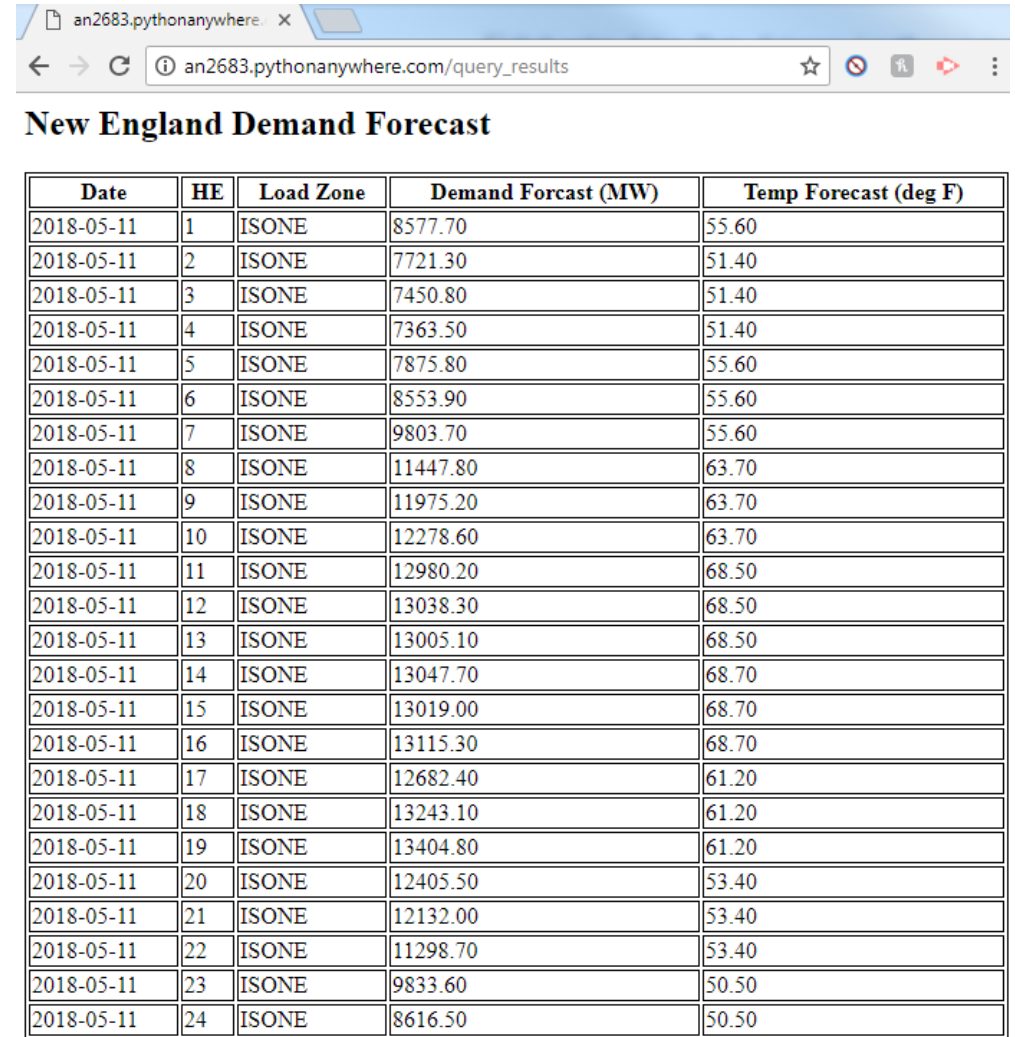
- Located at: <http://an2683.pythonanywhere.com/>
- For New England Residents - this simple application will allow you to have an idea of what the Power Demand in your region will be.
- Simply Select the Load Zone in which you reside (i.e Z.NEMASSBOS or Boston) or “ISONE” if you want total New England Load.
- Select the Date for which you wish to see the forecast and click “Query MySQL DB”



The screenshot shows a web browser window with the title "New England Electric Demand Forecasting". The address bar displays "an2683.pythonanywhere.com". The main content area features the heading "Welcome to Electric Demand Forecasting!" followed by the instruction "Please select Load Region and Date:". Below this, there are two dropdown menus: the first is set to "ISONE" and the second is set to "2018-05-11". A button labeled "Query MySQL DB" is positioned below the dropdowns.

Application

- The application will query a MySQL database and fetch Hour By Hour Demand Forecast for that Load Zone and Date (this is the output of my model) if available.
- For reference it will show the NWS provided forecast of Dry Bulb Temperature.
- You can see here that the highest demand throughout New England tomorrow is expected to occur at HE (Hour Ending) 19, or between 6:00-7:00 PM. This is around when people come home and turn their lights on have dinner, etc.

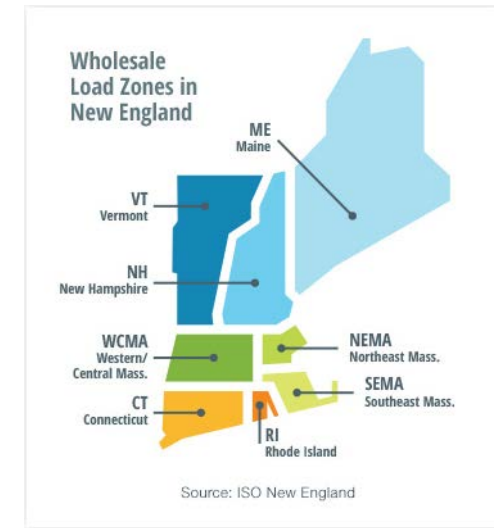


The screenshot shows a web browser window with the address bar displaying 'an2683.pythonanywhere.com/query_results'. The page title is 'New England Demand Forecast'. Below the title is a table with five columns: 'Date', 'HE', 'Load Zone', 'Demand Forecast (MW)', and 'Temp Forecast (deg F)'. The table contains 24 rows of data for the date 2018-05-11, showing demand forecasts for each hour of the day and corresponding temperature forecasts.

Date	HE	Load Zone	Demand Forecast (MW)	Temp Forecast (deg F)
2018-05-11	1	ISONE	8577.70	55.60
2018-05-11	2	ISONE	7721.30	51.40
2018-05-11	3	ISONE	7450.80	51.40
2018-05-11	4	ISONE	7363.50	51.40
2018-05-11	5	ISONE	7875.80	55.60
2018-05-11	6	ISONE	8553.90	55.60
2018-05-11	7	ISONE	9803.70	55.60
2018-05-11	8	ISONE	11447.80	63.70
2018-05-11	9	ISONE	11975.20	63.70
2018-05-11	10	ISONE	12278.60	63.70
2018-05-11	11	ISONE	12980.20	68.50
2018-05-11	12	ISONE	13038.30	68.50
2018-05-11	13	ISONE	13005.10	68.50
2018-05-11	14	ISONE	13047.70	68.70
2018-05-11	15	ISONE	13019.00	68.70
2018-05-11	16	ISONE	13115.30	68.70
2018-05-11	17	ISONE	12682.40	61.20
2018-05-11	18	ISONE	13243.10	61.20
2018-05-11	19	ISONE	13404.80	61.20
2018-05-11	20	ISONE	12405.50	53.40
2018-05-11	21	ISONE	12132.00	53.40
2018-05-11	22	ISONE	11298.70	53.40
2018-05-11	23	ISONE	9833.60	50.50
2018-05-11	24	ISONE	8616.50	50.50

Notebook Overview – General Modeling Framework

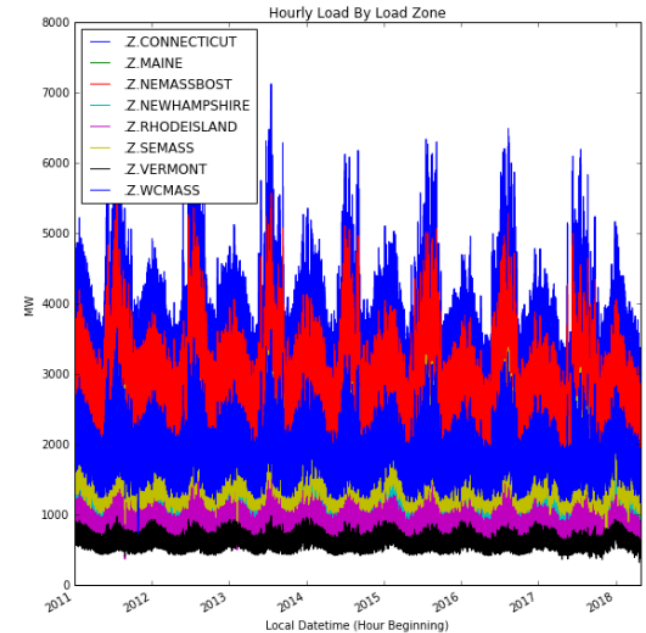
- [Link to iPython Notebook on GitHub](#)
- General idea is to produce an Hourly Forecast of Demand based on the relevant Locational Weather Factors for each Load Zone. Once this is accomplished, you can also aggregate hour by hour to create a Forecast for New England Demand.
- Problem is two-fold:
 - NWS can provide us with historical hourly weather data for many cities, airports in New England (22 locations with relatively clean data).
 - But which locations can influence the Demand in a particular Load Zone? Can more than one Weather Station drive Demand for say, Connecticut? Can this be influenced population density across the state? We need to figure this out.
 - Once we can figure out which weather stations can influence a particular Load Zone – what are the actual Factors which drive Demand?
 - Does it matter if it is a weekend versus weekday? The time of year (January vs July)?
 - Can apparent temperature (Heat Index / Wind Chill) influence Demand? What if it is Raining, Snowing, Cloudy?



Weather Station Name	State	County
Bangor Intl Arpt	ME	Penobscot
Bedford Hanscom Field	MA	Middlesex
Berlin Municipal Arpt	NH	Coos
Beverly Municipal Arpt	MA	Essex
Boston Logan Intl Arpt	MA	Suffolk
Bridgeport/Igor Sikorsky Memorial	CT	Fairfield
Burlington Intl Arpt	VT	Chittenden
Concord Municipal Arpt	NH	Merrimack
Danbury Municipal Arpt	CT	Fairfield
Hartford Brainard Arpt	CT	Hartford
Hartness State Springfield Arpt	VT	Windsor
Hyannis Barnstable Municipal Boardman Arpt	MA	Barnstable
Millinocket Municipal Arpt	ME	Penobscot
New Bedford Regional Arpt	MA	Bristol
Newport State Arpt	RI	Newport
Pittsfield Municipal Arpt	MA	Berkshire
Portland Intl Arpt	ME	Cumberland
Presque Isle	ME	Aroostook
Providence Green State Arpt	RI	Kent
St Johnsbury	NH	Grafton
Westfield Barnes Municipal Arpt	MA	Hampden
Windsor Locks Bradley Intl Arpt	CT	Hartford
Worcester Regional Arpt	MA	Worcester

Notebook Overview – Data Cleaning / Preparation

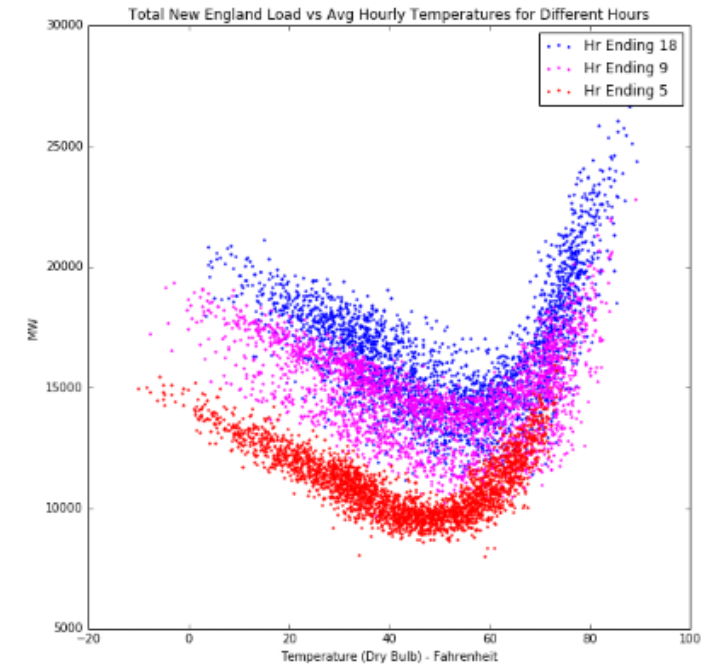
- Grab Load Zone Level Demand and Hourly Weather Data from Ventex since Jan 2011 (about 6.5 years of Hourly Data) ~ 2 GB.
 - Weather Data needs to be organized by Historical Date and Hour for Each Weather Station.
 - Load Data needs to be organized by Historical Date and Hour for Each load Zone.
- Parameterize as Dummies Seasonal Information: Calendar Month, Year, WeekDay vs Weekend, Hour,
- Parameterize as Dummies Qualitative Weather Variables: if it is raining, snowing, degree of cloud cover.
- In addition we have Heat Index Temp, Dry Bulb, Wind Chill Temp.
- Every hour we use in the model needs to have valid entries for all of these, if not, we cut it from the sample.
- Some Weather Stations were missing enough data that I cut them from the process – like Pittsfield Airport). Still ended up with ~60,000 samples for each Zone Level regression – still small relative to # of Tested Factors.



Notebook Overview – Modeling Approach

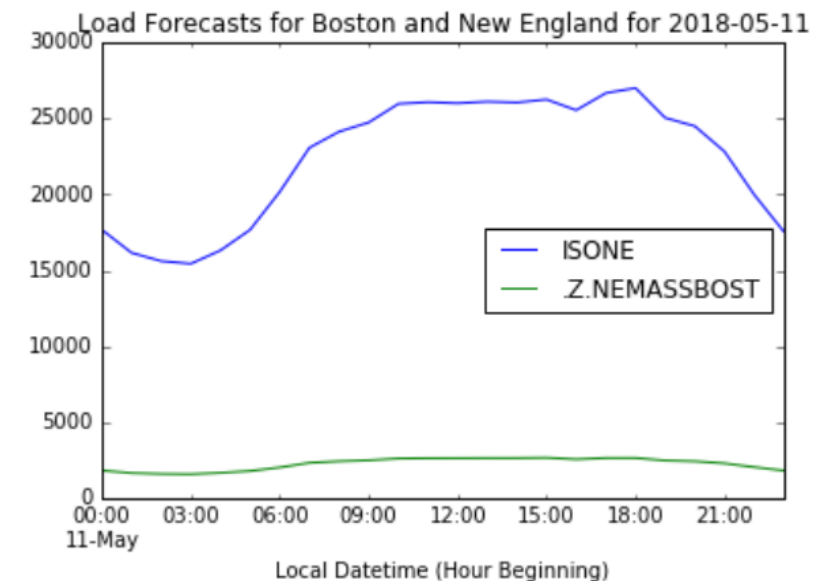
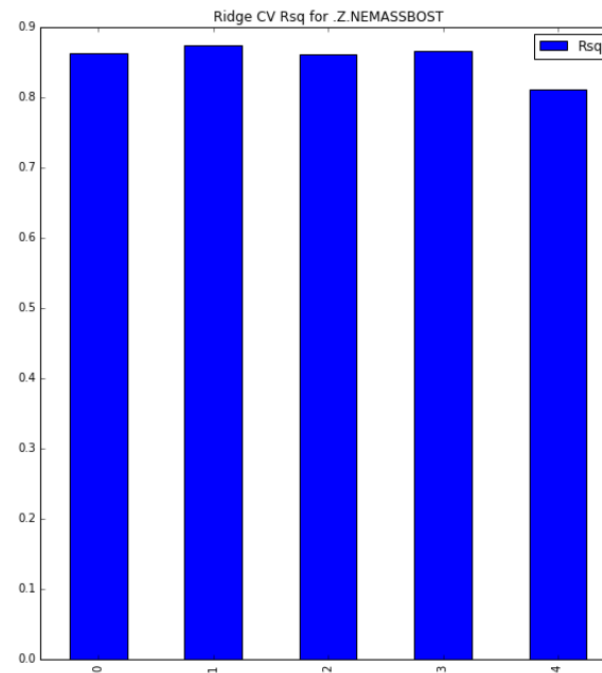
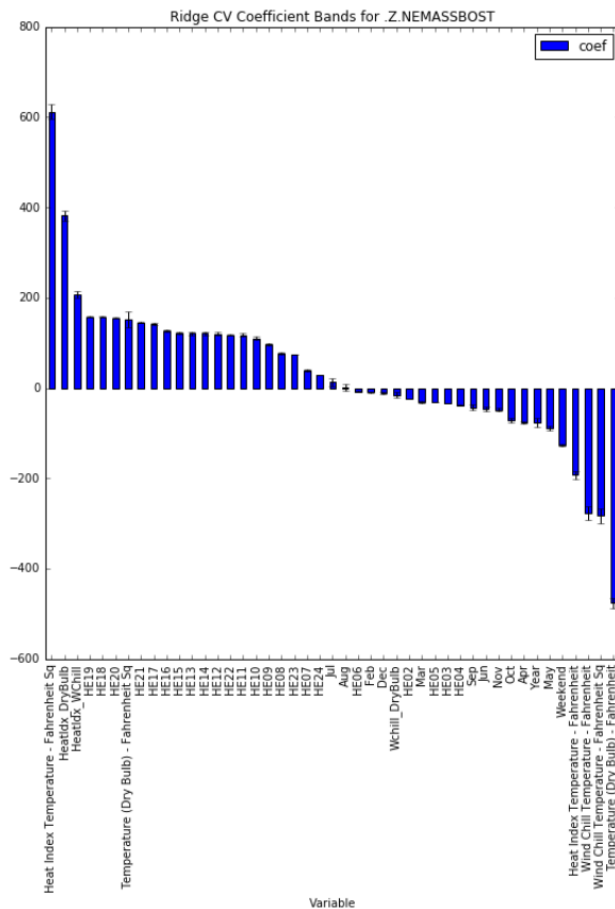
For Each Load Zone (There are Seven throughout New England):

1. Use a Lasso Regression against Dry-Bulb Temperature for the Summer to get weights for the 22 weather stations – we want most of them to be Zero. Why Summer? Look at the nonlinear relationship between temp and load, correlation flips signs in the Winter. We have to use a somewhat linear region to map this relationship.
2. Now, you have Load Zone Historical Demand as Y-Variable. For the X Variables
 - Take a weighted average of the Temperature Metrics across weather station. Need to be careful when not all stations have availability of data. Readjust the scaling proportionally.
 - Use the most prominent weather station to determine the Qualitative States (Rain/Snow/degree of Cloud Cover).
 - Static Seasonal Information (Hour, Jan-Dec, Year, Weekday vs Weekend)
 - Add Squared Terms of Temperature to help capture the Nonlinearities
 - You have correlation between the Temp Variables, Try Removing interaction with either a PCA representation of temperature or by using products of the Temperature Metrics: Wchill x Dry_Bulb, HeatIdx x Wchill, Dry_Bulb x HeatIdx
 - **Standardize All Data!**
3. Run a Ridge Regression and Estimate Rsq and Coefficient Ranges across 5 CV Folds.
4. Filter out the factors which aren't relevant, and rerun CV estimation of Coefficients and Rsq/
5. Compare this result to using Random Forest. Don't need to standardize for this.
6. Get NWS Weather Forecst for Next Two Days at 3 Hour Intervals and Project using GridSearchCV() Estimated Model.



Notebook Overview – Model Results

- Cross Validated Rsq ~ 85-90% with Stable Coefficient Ranges across the folds, and similar selection (ordering of positive to negative) across the Load Zones .



.Z.CONNECTICUT CV Rsq:

	Random Forest (No Std)	Ridge Regression
CV Split/Folds		
0	0.867167	0.848849
1	0.893919	0.883915
2	0.881647	0.872130
3	0.873431	0.881294
4	0.852941	0.849148

Conclusions

- Model is pretty effective in forecasting load. We could have done a few things differently:
 - Use harmonic regressions to estimate seasonality
 - For Higher Rsq , we could have tried to Model By Rate Class (Industrial, Residential with Electric Heating, Residential with No Electric Heating) – but this is excruciatingly cumbersome and unnecessary for our purposes.
 - Our goal is to give a regular guy an idea of when Power Demand / Prices might be high – not to bid Load Service in Competitive Bidding Process.
- Not shown in notebook, but the Random Forest Results didn't look as clean, qualitatively. Also, having a robust estimation of coefficients – and being able to see how coefficients are important based on Ridge Reg give me more confidence in the final forecast output.