Solar Energy Generation

Team 1-10

Agenda

- Business Problem
- Data Source
- Data Cleansing
- Unsupervised Learning Models
    - Clustering
- Supervised Learning Models
    - Tree + Linear Regression, Lasso, Ridge, Logistic, GAMs, KNN
- Best model
    - Performance on test data (two new sites!)

Business Problem

- Background
  - Energy and weather data for sites around the United States
- Goal
  - Understand significant variables
  - Predict solar energy generation
    - based on weather & irradiance
    - Important for utilities to **know loads on the grid in advance**
      - E.g. Predict next-day-loads the night beforehand

Data Source

- Source: SunDance dataset, part of SMART project
- 100 sites across US
- 1 csv for each site with hourly weather data for a year (2015)
    - Date, Time, Location, Temperature, Humidity, Wind Speed, Wind Direction, Pressure, Wind chill, Heat Index, Conditions (clear, hazy, ect.), Fog, Rain, Snow, Hail, Thunder, Tornado, ...
- 1 csv for each site with hourly energy data for a year
    - Date, Time, Energy Usage, Solar Energy Generated
- Irradiance calculations for the locations in our training data
    - 1 csv for Denver, 1 for LA

Data Cleansing

- We chose 10 sites (in LA and Denver)
- Merged energy and weather data for each site
- Also merged in irradiance data from separate files using date, time, and location
- Merged all sites together
- Removed NAs, changed negative energy generated to zero
- Training data - 79497 observations of 24 variables

	Hour	datetzname	tempm	tempi	dewptm	dewpti	hum	wspdm	wspdi	wdird	wdire	vism	visi	pressurem	pressurei	icon	fog	rain	snow	hail	thunder	tornado	Radiance	EnergyGenerated
1	1	LA	4.4	39.9	-6.1	21.0	47	5.6	3.5	300	WNW	16.1	10	1019.3	30.10	clear	0	0	0	0	0	0	0.0000	0.0000000
2	11	LA	10.0	50.0	-6.1	21.0	32	0.0	0.0	0	North	16.1	10	1020.8	30.15	clear	0	0	0	0	0	0	737.8477	0.2185778
3	13	LA	11.7	53.1	-4.4	24.1	32	9.3	5.8	0	Variable	16.1	10	1019.9	30.12	clear	0	0	0	0	0	0	729.6409	1.9646015
4	15	LA	12.2	54.0	-3.9	25.0	33	9.3	5.8	360	North	16.1	10	1020.4	30.14	clear	0	0	0	0	0	0	443.9369	3.6580735
5	17	LA	8.9	48.0	-3.9	25.0	41	0.0	0.0	0	North	16.1	10	1021.0	30.15	clear	0	0	0	0	0	0	0.0000	2.5067029
6	19	LA	7.2	45.0	-3.3	26.1	48	9.3	5.8	260	West	16.1	10	1021.5	30.17	clear	0	0	0	0	0	0	0.0000	0.0082800

- Caret - dummy, near zero variance, and correlated variables examined
- Variable selection for various models

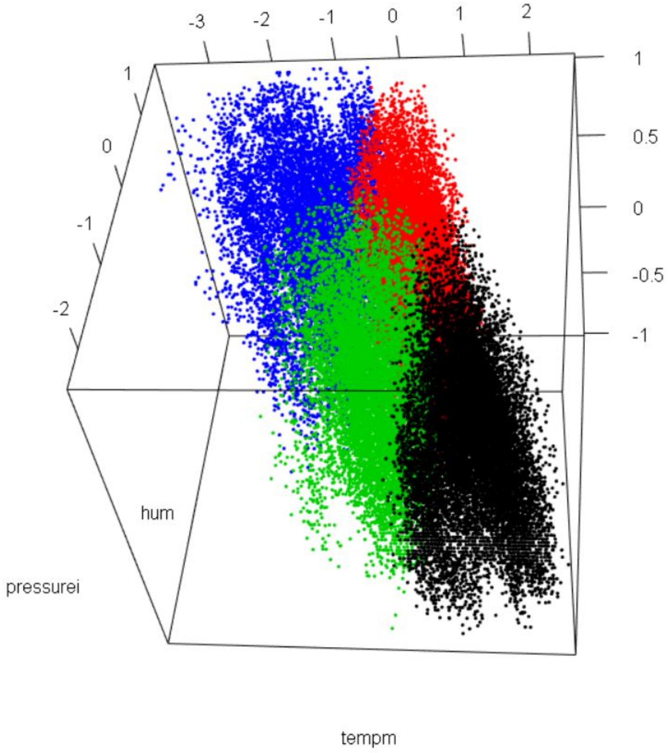
Unsupervised Learning Model

- Clustering
  - Wanted to understand the different types of days and combinations of conditions that occur
  - 3D plotting using the rgl package

Clustering

- Pressure was not significant
- Low humidity and high temps -> more solar energy

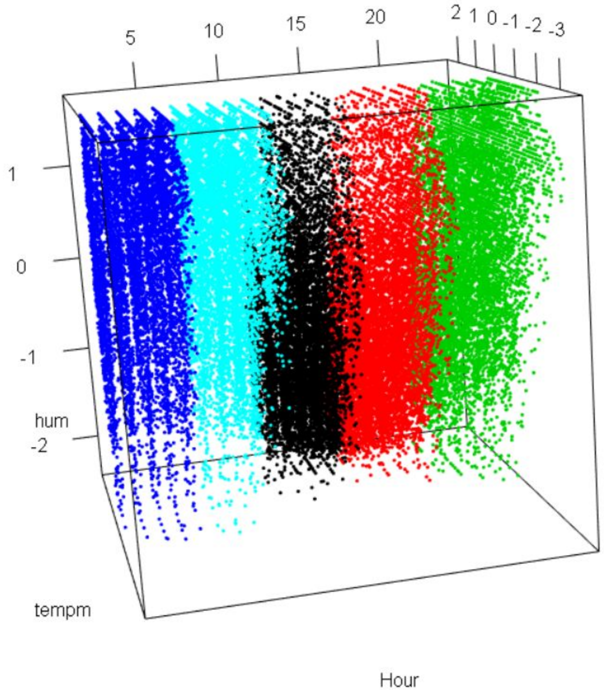
Cluster	Conditions	Avg Solar Energy Generated
1 (Green)	Med temp, low humidity	0.641
2 (Red)	Med temp, high humidity	0.000155
3 (Black)	High temp, low humidity	2.64
4 (Blue)	Low temp, high humidity	0.365



Clustering

- Most energy generated between 3 PM and 7 PM

Cluster	Time	Average Solar Energy Generated
1	11 AM - 3 PM	1.43
2	3 PM - 7 PM	2.68
3	8 PM - Midnight	0.365
4	Midnight - 5 AM	0.0016
5	5 AM - 11 AM	0.0122



Supervised Learning Models

- Logistic
- KNN
- Linear Regression
- Boosting
- Ridge & Lasso
- GAMs
- Random Forest

Logistic Regression

Tuning Method

- Added binary variable--1 if any energy was generated, 0 otherwise

Final Parameters

- Ran one model with all predictors, one with just three (irradiance, hour, location)

Training Accuracy

- First model predicted 92% of cases accurately
- Second model predicted 89% of cases accurately

KNN

Tuning Method

- For loop over K values (and select min MSE)

Final Parameters

- K = 19

Training Accuracy: MSE=0.420

Linear Regression

Tuning Method

- Ran a basic linear regression on all variables of a training set

Final Parameters

- lm.fit<- lm(EnergyGenerated ~ ., data=trainset)

Training Accuracy

- MSE of 0.7

GBM Boosting Model

Tuning Method

- Tuned model by hand

Final Parameters

```
boostmodel = gbm(EnergyGenerated~., data = trainset,
        distribution = "gaussian", n.trees = 100,
        interaction.depth = 15, shrinkage = 0.01, bag.fraction = 0.5,
        cv.folds = 10)
```

Training Accuracy

- MSE of 1.11

Ridge and Lasso

Tuning Method

- best lambda (feature penalty) chosen by 10-fold CV
- Best alpha (between ridge & lasso selection) chosen using a grid search
  - 0 to 1 in steps of 0.05

Final Parameters

- Alpha = 0.05 (close to ridge)
- Lambda = 0.005395222 (small penalty for including additional variables)
- thresh = 1e-12

Training Accuracy: 1.161

GAMs

Tuning Method - By hand

Natural and smoothing of different degrees and local splines

Final Parameters -

```
gam6 <- gam(EnergyGenerated ~ s(Hour, 23) + datetznameLA  +
        ns(dewpti,20) + s(hum, 20) + ns(wspdm,20) + ns(wdird, 20) +
        ns(pressurem, 15) + s(tempi, 30) +
        s(Radiance, 30) + iconclear + iconcloudy +
        iconmostlycloudy + iconpartlycloudy + datetznameLA, data=train)
```

Training Accuracy: 0.692

Random Forests

Tuning Method

- Looped over mtry = 1 to 10 (out of 13 predictors)
- Best ntrees (from 1 to 500) chosen by lowest OOB error

Final Parameters

- Best mtry = 3
- Best ntree = 496

Training Accuracy 0.2941536

Best Model

Compare MSE for each model on the new test set

Model	MSE	MAE	MAAPE*
Linear Regression	0.805	0.708	1.167
GAMs	1.090	0.752	1.167
KNN	0.621	0.411	0.795
Random Forest	0.743	0.551	0.997

*MAAPE is the mean arctangent absolute percentage error

Conclusions

- Best Model: KNN
- We can predict how much solar energy a site will generate with an average error of 0.411 kW
- We understand the important factors that influence solar generation
  - Time of Day, Temperature, Humidity, Location, and Irradiance

Next Steps

- Include more of the sites
    - especially those in **other locations**
- Obtain additional data
    - **Size** of array
    - **Tilt**-capability of array
        - E collected very different if tilting towards the sun
    - Solar Panel Physical Characteristics
        - **Material** (efficiency)

Links to Data

Energy/Weather Data (Sundance) - http://traces.cs.umass.edu/index.php/Smart/Smart

Irradiance Calculator - https://midcdmz.nrel.gov/solpos/solpos.html