

LEON, AMANDA, ANTONIO, JOHNY

Goal of the project...

...was to take a normal set of the U.S electrical power grid data and be able to differentiate between anomalous and normal electrical power usage when given another set



Overview

- Data Exploration
- Approach 1: Point Anomaly Detection
- Approach 2: Training and Testing of Hidden Markov Models

Data Exploration



Data Exploration – Exploring Each Feature

• 3 different time frames:

- 1. Weekends (Sat and Sun) 7am-3pm,
- 2. Mon, Wed, Fri 4pm-8pm
- 3. Monday 12pm-8pm.

-All in 2007-2008

Data Exploration – Training & Testing Dataset

• 80 % Training Dataset ----- 20% Training Dataset -----Testing Dataset

```
Global_active_power Global_active_power
                                         Global_active_power
       :0.0960
                   Min.
                           :0.098
Min.
                                                :0.0980
                                         Min.
                   1st Qu.: 0.550
1st Qu.: 0.4280
                                         1st Qu.: 0.4558
Median : 0.6604
                   Median :1.331
                                         Median : 0.7737
       :1.0495
Mean
                           :1.449
                                                :1.2215
                   Mean
                                         Mean
3rd Qu.:1.5040
                                         3rd Qu.:1.6880
                    3rd Qu.:1.930
     :7.5660
Max.
                                                :9.5900
                           :8.166
                                         Max.
                    Max.
Global_active_power Global_active_power
                                         Global_active_power
       :0.0780
Min.
                   Min.
                           :0.0780
                                         Min.
                                                :0.0780
1st Qu.: 0.4160
                   1st Qu.: 0.4655
                                         1st Qu.: 0.4360
Median : 0.6623
                   Median :1.2060
                                         Median : 0.7797
Mean : 0.9543
                           :1.3358
                   Mean
                                         Mean
                                                :1.1293
3rd Ou.:1.4248
                   3rd Qu.:1.8860
                                         3rd Qu.:1.6682
      :5.6880
Max.
                           :7.6600
                   Max.
                                                :6.7340
                                         Max.
Global_active_power
                   Global_active_power
                                        Global_active_power
Min.
       :-1.818
                   Min.
                          :-2.2776
                                        Min.
                                               :-1.8262
1st Qu.: 0.635
                   1st Qu.: 0.7631
                                        1st Qu.: 0.6823
Median : 1.408
                                        Median: 1.4140
                   Median: 1.5920
       : 1.585
Mean
                                        Mean : 1.6161
                           : 1.7738
                   Mean
3rd Qu.: 2.156
                                        3rd Qu.: 2.2255
                   3rd Qu.: 2.4326
       : 8.363
Max.
                                                :10.4176
                                        Max.
                           : 9.9552
                   Max.
```

Data Exploration – Training & Testing Dataset

- Global active power: consistent min, mean & SD, inconsistent max, testing values generally higher
- Voltage: consistent min, mean, max, max of testing dataset quite a big larger, inconsistent SD

Data Exploration – Training & Testing Dataset

- Global reactive power: consistent min, mean, & SD, inconsistent max
- Global intensity: consistent min w/ 1st 80% generally bigger, inconsistent mean & SD, does not look like training & testing are following similar trend
- Sub Metering 1, 2, & 3: showed similar results
 - Consistent min
 - Inconsistent max, mean, & SD

Data Exploration – Correlation Coefficient

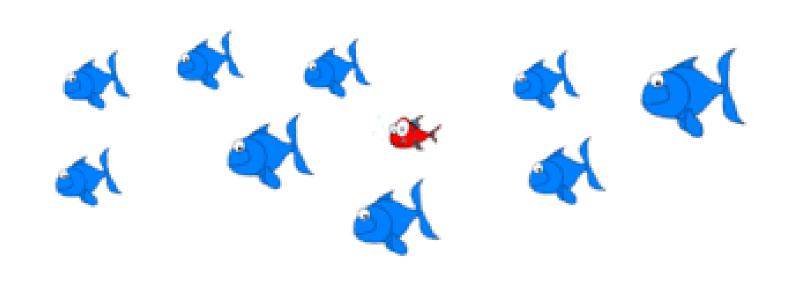
Wed 4-8pm Global active power vs:	Fri 9-12pm Global active power vs:
global_reactive_power =0.18201	global_reactive_power = - 0.2746126
global_intensity = 0.9962178	global_intensity = 0.9967875
voltage= -0.2721188	voltage = -0.055234488
sub_metering_1 =0.107927	sub_metering_1 = 0.4424803
sub_metering_2= 0.7268048	sub_metering_2 = -0.2026798
sub_metering_3 =0.449147	sub_metering_3 = 0.9468868
sub_metering_(1,2,3) =0.708217	sub_metering_(1,2,3) =0.8191631

Why Univariate?

- Criteria:
 - 1. Range
 - 2. % change in values from 1 SD from mean
 - 3. Exclude those with strong positive correlation
- Time Frame: 2007-2008



Approach 1: Point Anomaly Detection



Approach 1: Point Anomaly Detection

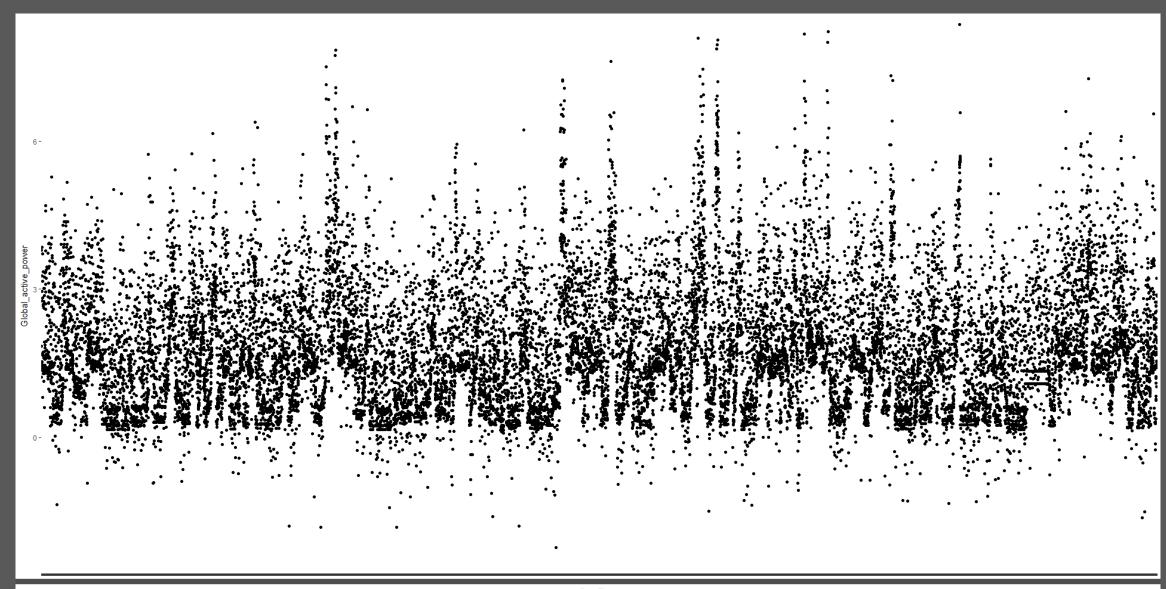
- Point Anomaly
 - A single instance of data is anomalous if it's too far off from the rest

- Mondays 12-8PM, 2007 & 2008
 - Global Active Power vs Date/Time

Anomaly Detection Techniques

- Out of Range
 - min/max of training
 - compare with test data

- Moving Average
 - Average specified time window of n observations

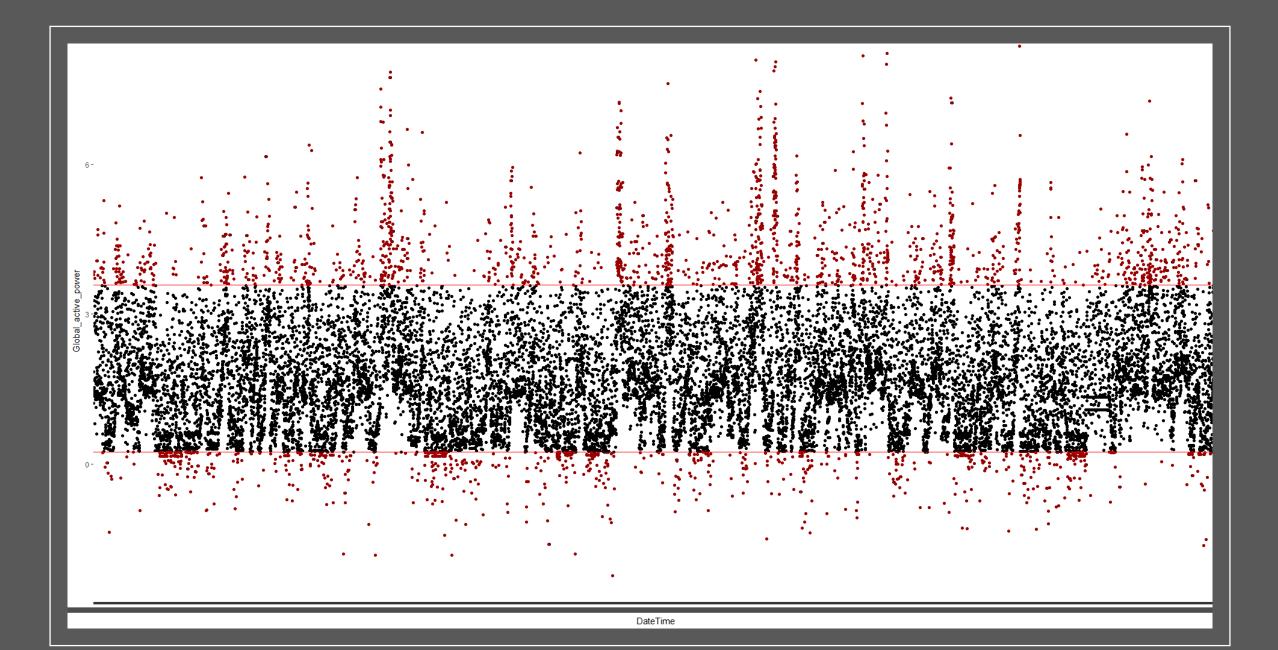


Out of Range

- I. Find minimum / maximum global active power for each Monday (training)
- II. Average out the total minimum / maximum across all Mondays

Average Min: 0.2411429

Average Max: 3.580115



Moving Average

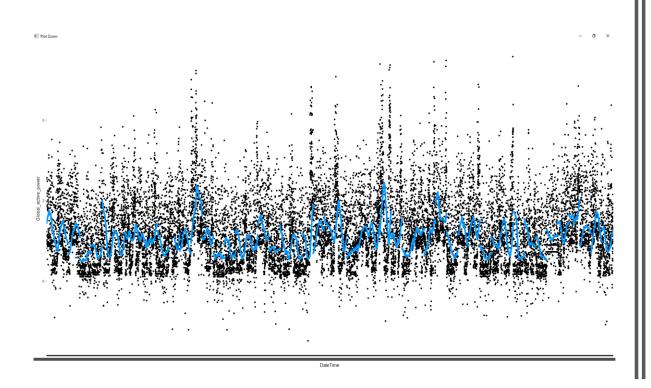
Specify a time window of N observations

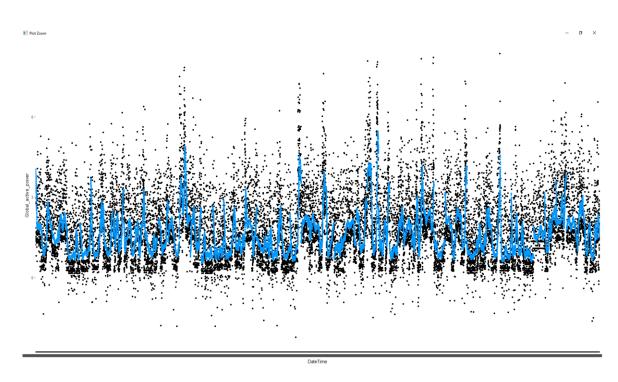
Define a threshold from the moving average

Points outside of this threshold is considered anomalous

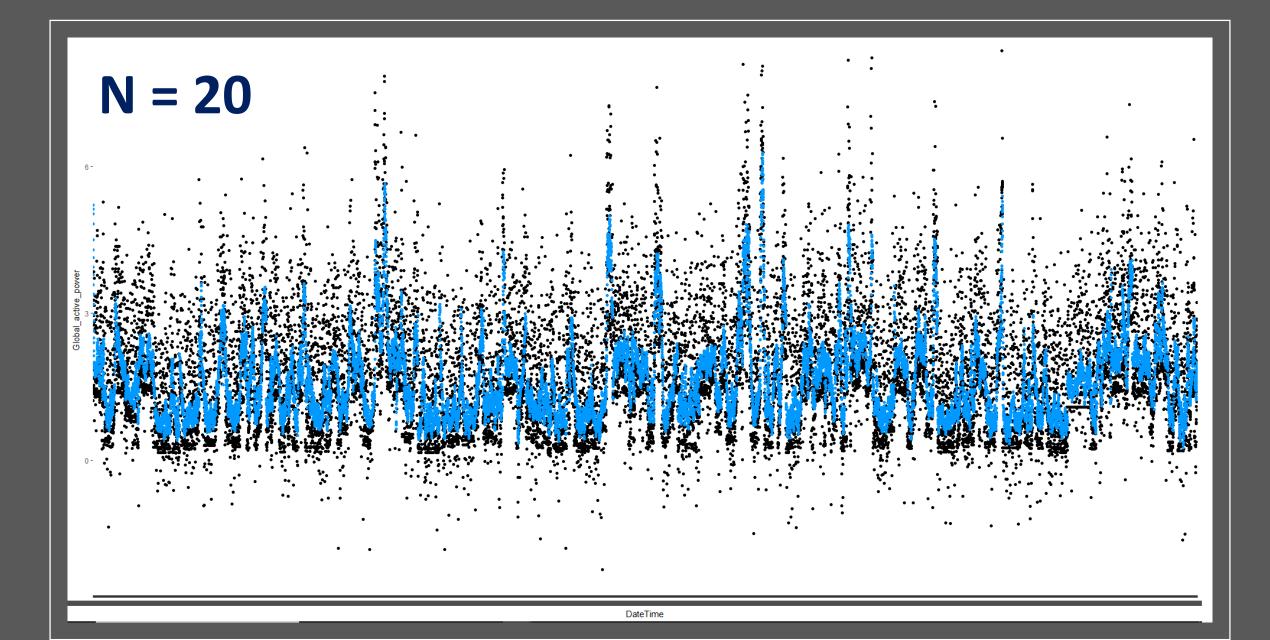
N = 480

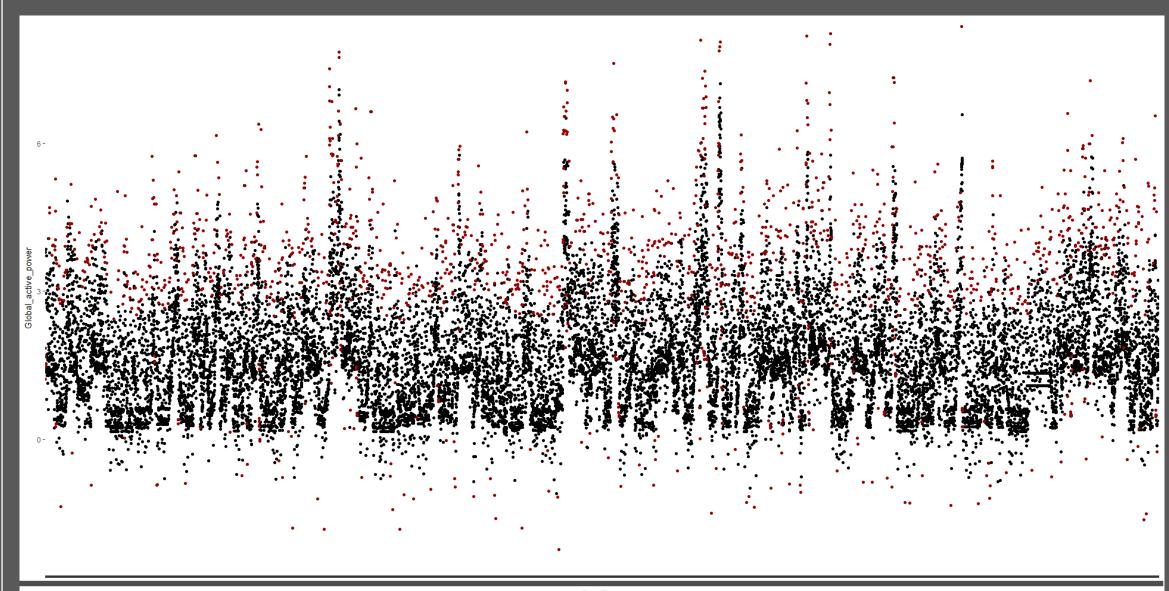
D-4-Time



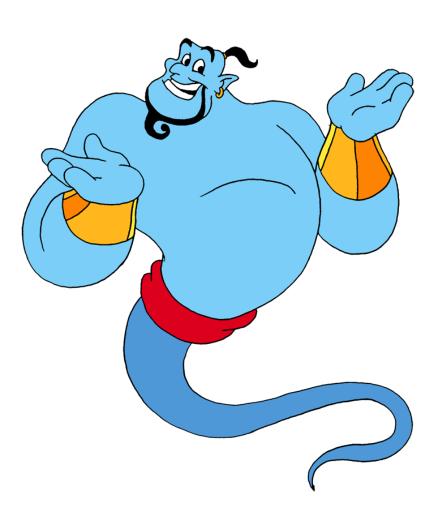


N = 240 & N = 60

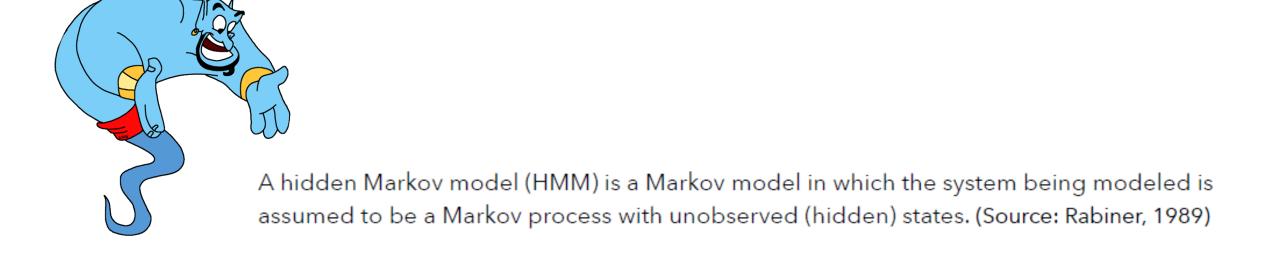




Approach 2: Hidden Markov Models



Hidden Markov Model



Training HMM



Values to look out for...

• Log-likelihood

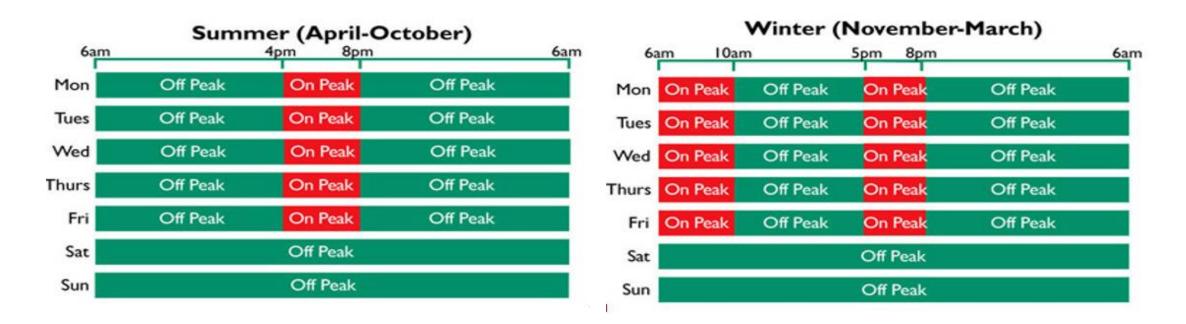


• BIC Value



- Log-likelihood ratio
 - training(80%) vs. validation(20%)
 - ranges between 0 and 1

Early models were based on intuition as well as statistics...



From Pacific Power

Wed 16	:00	0-19:59 08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	3	-3539.423	-84.27197619	7207.22	-1916.57	-174.2336364	3940.63	0.483672257
	4	-2795.2	-66.55238095	5801	-1547.535	-140.685	3271.66	0.473059537
	5	-2330.897	-55.49754762	4973.557	-1384.092	-125.8265455	3029.232	0.441063906
	8	-1001.864	-23.85390476	2728.121	-980.0264	-89.09330909	2566.604	0.267740698
	9	-428.547	-10.2035	1755.707	-885.6255	-80.51140909	2523.682	0.126733591

Wed 16	:00	0-19:59 08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	3	-3539.423	-84.27197619	7207.22	-1916.57	-174.2336364	3940.6	0.483672257
	4	-2795.2	-66.55238095	5801	-1547.535	-140.685	3271.66	0.473059537
	5	-2330.897	-55.49754762	4973.557	-1384.092	-125.8265455	3029.232	0.441063906
	8	-1001.864	-23.85390476	2728.121	-980.0264	-89.09330909	2566.60	0.267740698
	9	-428.547	-10.2035	1755.707	-885.6255	-80.51140909	2523.682	0.126733591

Wed 16:0	0-19:59 08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
3	-3539.423	-84.27197619	7207.22	-1916.57	-174.2336364	3940.63	0.483672257
	-2795.2	-66.55238095	5801	-1547.535	-140.685	3271.6	0.473059537
5	-2330.897	-55.49754762	4973.557	-1384.092	-125.8265455	3029.232	0.441063906
8	-1001.864	-23.85390476	2728.121	-980.0264	-89.09330909	2566.604	0.267740698
9	-428.547	-10.2035	1755.707	-885.6255	-80.51140909	2523.682	0.126733591

Wed 16:00	0-19:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
12	-685.9017	-8.263875904	3025.016	-505.4596	-22.97543636	2442.39	0.359683088
13	-117.517	-1.41586747	2155.533	-636.5176	-28.93261818	2935.941	0.048936721
14	positive						

Wed 16:	:00	0-19:59 08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	3	-3539.423	-84.27197619	7207.22	-1916.57	-174.2336364	3940.6	0.483672257
	4	-2795.2	-66.55238095	5801	-1547.535	-140.685	3271.6	0.473059537
	5	-2330.897	-55.49754762	4973.557	-1384.092	-125.8265455	3029.232	0.441063906
	8	-1001.864	-23.85390476	2728.121	-980.0264	-89.09330909	2566.604	2267740698
	9	-428.547	-10.2035	1755.707	-885.6255	-80.51140909	2523.6%2	0.126733591

Wed 16:00	0-19:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
12	-685.9017	-8.263875904	3025.016	-505.4596	-22.97543636	2442.39	0.359683088
13	-117.517	-1.41586747	2155.533	-636.5176	-28.93261818	2935.941	0.048936721
14	positive						

...poor ratios

Mon, Wed, Fri 10am-2pm

MWF 10:0	00-13:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
7	-4623.019	-18.64120565	9370.039	-576.8778	-8.740572727	1753.314	0.468884518
8	-3954.22	-15.94443548	8776.971	-1723.065	-26.10704545	4210.082	0.610733049
9	-11.00976	-0.044394194	1099.438	106.6644	1.616127273	734.36	-0.027469491

Mon, Wed, Fri 10am-2pm

MWF 10	0:0	0-13:59 07-08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	7	-4623.019	-18.64120565	9370.039	-576.8778	-8.740572727	1753.314	0.468884518
	8	-3954.22	-15.94443548	8776.971	-1723.065	-26.10704545	4210.082	0.610733049
	9	-11.00976	-0.044394194	1099.438	106.6644	1.616127273	734.36	-0.027469491

...inconsistent numbers

All Days 10am-2pm

ALL DAYS	ALL DAYS 10:00-13:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
9	-13374.24	-23.13882353	27908.81	-3097.357	-20.2441634	7224.799	0.874900289
10	-12898.98	-22.31657439	27206.94	-3553.303	-23.22420261	8357.425	0.960918864
13	-4345.402	-7.51799654	9078.804	-3410.649	-22.29182353	8860.446	0.337253546
14	-1322.774	-2.288536332	3091.549	-104.3524	-0.68204183	2552.675	0.298025345
15	1650.873	2.856181661	-294.3312	671.2543	4.387283007	1327.305	0.651013772

All Days 10am-2pm

ALL DAYS	10:00-13:59 07	'-08'					
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
9	-13374.24	-23.13882353	27908.81	-3097.357	-20.2441634	7224.799	0.874900289
10	-12898.98	-22.31657439	27206.94	-3553.303	-23.22420261	8357.425	0.960918864
13	-4345.402	-7.51799654	9078.804	-3410.649	-22.29182353	8860.446	0.337253546
14	-1322.774	-2.288536332	3091.549	-104.3524	-0.68204183	2552.675	0.298025345
15	1650.873	2.856181661	-294.3312	671.2543	4.387283007	1327.305	0.651013772

...ratio still not that good

Finally...

We wanted to try a time that encompassed both "mid-peak" and "on-peak" hours

...12pm-8pm

Trial and error on all 7 days of the week...

Mon 12:00	-19:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-19678	-237.0843373	39430.15	-4340.496	-197.2952727	8745.846	0.832173373
5	-2919.447	-35.17406024	6199.043	-339.0766	-15.41257273	993.1575	0.438180086
6	-319.0362	-3.843809639	1135.926	-114.9913	-5.226877273	665.4296	0.73539313
7	1589.869	19.15504819	-2522.995	456.81	20.76409091	-339.2006	0.922508396
Tues 12:00	-19:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-32773.64	-394.8631325	65621.42	-6821.28	-310.0581818	13707.4	0.785229504
6	-12533.65	-151.0078313	25565.15	-2210	-100.4545455	4855	0.66522739
8	-7483.172	-90.1586988	15803.16	-1278.711	-58.12322727	3289.34	0.644676865
12	-3838.78	-46.25036145	9446.529	-24.2274	-1.101245455	1595.68	0.023810526
14	-4862.457	-58.58381928	12087.07	-39.45315	-1.793325	2144.96	0.030611268
15	positive						
Wed 12:00)-19:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	 	-314.3253012	52252.14	-8673.365	-394.2438636		0.797286477
6	-8319.358	-100.2332289	17136.57	-3442.964	-156.4983636	7321.376	0.640474613
10	-1999.841	-24.09446988	5260.205	-2482.161	-112.8255	6066.837	0.213555179
11	-654.7305	-7.888319277	2813.614	-3074.441	-139.7473182	7464.487	0.056447017
12	positive						
Γhu 12:00-	19:59 07-08'						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-16099.9	-193.9746988	32273.94	-3823.409	-182.0670952	7711.35	0.938612594
4	-3739.995	-45.06018072	7723.621	-677.1251	-32.24405238	1566.27	0.71557752
5	-1534.507	-18.48803614	3429.163	-207.9034	-9.900161905	729.229	0.535490186
6	1187.508	14.3073253	-1877.163	241.5677	11.50322381	-49.875	0.804009385
ri 12:00-1	9:59 07-08'						
tates	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
6		-40.34956098	7067.353	-1078.645	-49.02931818	2592.737	0.822968021
8	-3073.006	-37.47568293	6902.436	-1296.645	-58.93840909	3325.212	0.635844834
10	-1796.535	-21.90896341	4732.128	-322.8599	-14.67545	1748.234	0.669837715
11	-1766.154	-21.53846341	4891.956	-396.7697	-18.03498636	2109.145	0.83733858
	2.00.20	22.000 700 12	.552.550	220.7037	25.55.50000		5.55788888
at 12:00-1	19:59 07-08'						
		Norm. Train LogLike	RIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
5		-267.850122	44287.15	-4252.27		8819.55	0.721616248
10		-191.8318293	32719.51	-2890.063	-131.3665		0.68480033
10			29917.35	-2576.033	-117.0924091		0.682162779
		-171.6487805					
15		-138.8619512	25460.82	-2484.556	-112.9343636		0.813285156
20	-9025.868	-110.071561	22696.58	-1448.214	-65.82790909	6963.69	0.598046475
	19:59 07-08'						
tates	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-46291.68	-557.7310843	92657.42	-11968.69	-544.0313636	24002.23	0.975436691
6	-24049.38	-289.7515663	48596.05	-4353.747	-197.8975909	9142.941	0.682990582
U					13710370303		

Trial and error on all 7 days of the week...

Mon 12	2:00	-19:59 07-08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	2	-19678	-237.0843373	39430.15	-4340.496	-197.2952727	8745.846	0.832173373
	5	-2919.447	-35.17406024	6199.043	-339.0766	-15.41257273	993.1575	0.438180086
	6	-319.0362	-3.843809639	1135.926	-114.9913	-5.226877273	665.4296	0.73539313
	7	1589.869	19.15504819	-2522.995	456.81	20.76409091	-339.2006	0.922508396
Tues 11		-19:59 07-08'						
States	2.00	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
States	2	-32773.64	-394.8631325	65621.42				0.785229504
	6	-12533.65	-394.8031323	25565.15			4855	0.783229304
	8	-7483.172						
	12	-3838.78	-90.1586988 -46.25036145	9446.529		-58.12322727 -1.101245455		0.644676865 0.023810526
	14	-4862.457	-58.58381928			-1.101243433		
		positive	-38,38381928	12007.07	-39.43313	-1.793323	2144.90	0.030611268
	10	posterve						
Wed 12	2:00	-19:59 07-08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	2	-26089	-314.3253012	52252.14	-8673.365	-394.2438636	17411.58	0.797286477
	6	-8319.358	-100.2332289	17136.57	-3442.964	-156.4983636	7321.376	0.640474613
	10	-1999.841	-24.09446988	5260.205	-2482.161	-112.8255	6066.837	0.213555179
	11	-654.7305	-7.888319277	2813.614	-3074.441	-139.7473182	7464.487	0.056447017
	12	positive						
	:00-	19:59 07-08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	2		-193.9746988		-			0.938612594
	4		-45.06018072	7723.621		-32.24405238		0.71557752
	5	-1534.507	-18.48803614					0.535490186
	6		14.3073253	-1877.163	241.5677	11.50322381	-49.875	0.804009385
	_	9:59 07-08'						
States			Norm. Train LogLike		Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	6	-3308.664	-40.34956098	7067.353	-1078.645	-49.02931818	2592.737	0.822968021
	8	-3073.006	-37.47568293	6902.436	-1296.645	-58.93840909	3325.212	0.635844834
	10	-1796.535	-21.90896341	4732.128	-322.8599	-14.67545	1748.234	0.669837715
	11	-1766.154	-21.53846341	4891.956	-396.7697	-18.03498636	2109.145	0.83733858
Sat 12:0	00-1	9:59 07-08'						
States			Norm. Train LogLike	RIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
States	5	-21963.71	-267.850122	44287.15			8819.55	0.721616248
	10	-15730.21	-191.8318293	32719.51	-2890.063	-131.3665		0.68480033
	12	-14075.2	-171.6487805	29917.35	-2576.033	-117.0924091	6699.29	0.682162779
	15	-11386.68	-138.8619512	25460.82	-2484.556	-112.9343636		0.813285156
	20	-9025.868	-110.071561	22696.58	-1448.214	-65.82790909	6963.69	0.598046475
Sun 12:	00-:	19:59 07-08'						
States		Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
	2	-46291.68	-557.7310843	92657.42	-11968.69	-544.0313636	24002.23	0.975436691
	6	-24049.38	-289.7515663	48596.05	-4353.747	-197.8975909	9142.941	0.682990582
	23	-12355.6	-148.8626506	30784.4	-1600.867	-72.76668182	8519.745	0.488817588
		12000.0	21010020000	2070-714	1000,307	. 217 0000102	352317-75	01-0001/500

...Monday 12pm-8pm was the best!

Mon 12:0	0-19:59 07-08						
States	Train LogLike	Norm. Train LogLike	BIC Train	Val LogLike	Norm. Val LogLike	BIC Val	Norm. LogLikelihood Ratio
2	-19678	-237.0843373	39430.15	-4340.496	-197.2952727	8745.85	0.832173373
5	-2919.447	-35.17406024	6199.043	-339.0766	-15.41257273	993.158	0.438180086
6	-319.0362	-3.843809639	1135.926	-114.9913	-5.226877273	665.43	0.73539313
7	1589.869	19.15504819	-2522.995	456.81	20.76409091	-339.201	0.922508396

Good balance of all 3 values

Testing



Ran the test data on the model...

Monday 1	2:00-19:59 200	7-2008 Model			
Month	Train LogLike	Norm. Train LogLike	Test LogLike	Norm. Test LogLike	Norm. LogLike Ratio
Dec 09'	-319.0362	-3.843809639	-2976.96	-744.24	0.005164745
Jan 09'	-319.0362	-3.843809639	-2844.231	-711.05775	0.005405763
Feb 10'	-319.0362	-3.843809639	-2822.686	-705.6715	0.005447024
Mar 10'	-319.0362	-3.843809639	-3530.595	-706.119	0.005443572
Apr 10'	-319.0362	-3.843809639	-2847.655	-711.91375	0.005399263
May 10'	-319.0362	-3.843809639	-3607.563	-721.5126	0.005327432
Jun 10'	-319.0362	-3.843809639	-3215.242	-803.8105	0.004781985
Jul 10'	-319.0362	-3.843809639	-3141.011	-785.25275	0.004894997
Aug 10'	-319.0362	-3.843809639	-3854.993	-770.9986	0.004985495
Sep 10'	-319.0362	-3.843809639	-2900.847	-725.21175	0.005300258
Oct 10'	-319.0362	-3.843809639	-3391.731	-847.93275	0.004533154
Nov 10'	-319.0362	-3.843809639	-2943.52	-735.88	0.005223419
ALL	-319.0362	-3.843809639	-38077.04	-746.6086275	0.00514836

Ran the test data on the model...

Monday 12:00-19:59 2007-2008 Model							
Month	Train LogLike	Norm. Train LogLike	Test LogLike	Norm. Test LogLike	Norm	. LogLike Ratio	
Dec 09'	-319.0362	-3.843809639	-2976.96	-744.24		0.005164745	
Jan 09'	-319.0362	-3.843809639	-2844.231	-711.05775		0.005405763	
Feb 10'	-319.0362	-3.843809639	-2822.686	-705.6715		0.005447024	
Mar 10'	-319.0362	-3.843809639	-3530.595	-706.119		0.005443572	
Apr 10'	-319.0362	-3.843809639	-2847.655	-711.91375		0.005399263	
May 10'	-319.0362	-3.843809639	-3607.563	-721.5126		0.005327432	
Jun 10'	-319.0362	-3.843809639	-3215.242	-803.8105		0.004781985	
Jul 10'	-319.0362	-3.843809639	-3141.011	-785.25275		0.004894997	
Aug 10'	-319.0362	-3.843809639	-3854.993	-770.9986		0.004985495	
Sep 10'	-319.0362	-3.843809639	-2900.847	-725.21175		0.005300258	
Oct 10'	-319.0362	-3.843809639	-3391.731	-847.93275		0.004533154	
Nov 10'	-319.0362	-3.843809639	-2943.52	-735.88		0.005223419	
ALL	-319.0362	-3.843809639	-38077.04	-746.6086275		0.00514835	

...ratios significantly lower which was expected

Conclusion



Conclusion

Applications

- Anomaly Detection is a challenge
 - Time and Effort
 - Trial and Error
 - Attention to Detail
 - Multiple approaches/solutions to a problem
- The Importance of the Industry

Thank you for listening!

Any Questions?

- Leon (Trung) Trieu
- **A**ntonio Leung
- Amanda (Fang Chi) Chang
- **J**ohny Kuang

CMPT 318

Cybersecurity

