Anomaly Detection
(time series data)

مقدمه

یکی از مسائلی که امروزه حائز اهمیت است تشخیص داده های ناهنجاری می باشد که جزو مسائل اصلی در یادگیری ماشین(Unsupervised) به حساب می آید. تشخیص ناهنجاری کاربرد های زیادی دارد از جمله تشخیص تقلب یا حملات سایبری، تشخیص علائم نادر در بیماری ها، تشخیص رویداد های نادر. برای این کار الگوریتم های مختلفی وجود دارد که می شود از آن ها بهره برد. ما از الگوریتم Isolation Forest و روش های مبتنی بر چگالی(Density Based) مانند Multivariate Normal و Gaussian استفاده می کنیم. ضمناً این الگوریتم و روش ها را بر روی نوع داده ای سری زمانی که مربوط به داده های AWS می باشد استفاده می کنیم.

بررسی و نمایش داده ها

	timestamp	value
0	2014-04-10 00:04:00	94.0
1	2014-04-10 00:09:00	56.0
2	2014-04-10 00:14:00	187.0
3	2014-04-10 00:19:00	95.0
4	2014-04-10 00:24:00	51.0
...	...	...
4027	2014-04-24 00:19:00	32.0
4028	2014-04-24 00:24:00	57.0
4029	2014-04-24 00:29:00	10.0
4030	2014-04-24 00:34:00	18.0
4031	2014-04-24 00:39:00	60.0

4032 rows × 2 columns

1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4032 entries, 0 to 4031
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   timestamp  4032 non-null   object
 1   value      4032 non-null   float64
dtypes: float64(1), object(1)
memory usage: 63.1+ KB
```

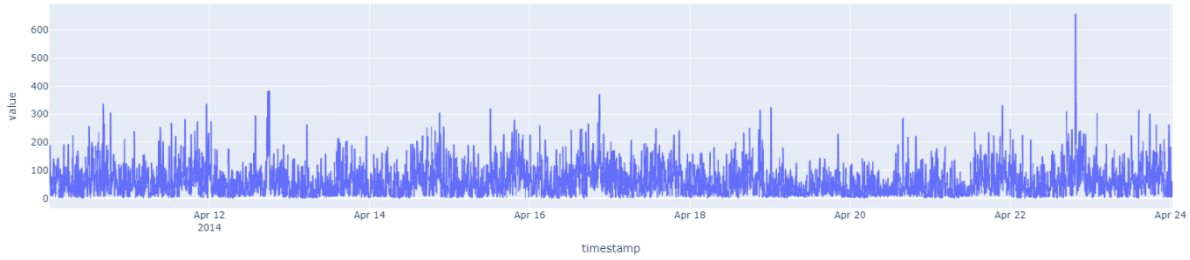
2

در شکل 1 داده های ما شامل 2 ستون است که یکی مربوط به زمان نمونه و دیگری مربوط به مقدار آن است. زمان نمونه ها با فاصله زمانی پنج دقیقه از هم فاصله دارند. تعداد داده ها 4032 عدد می باشد که بایستی بر روی آن ها پردازش های لازم را انجام دهیم.

در شکل 2 با بررسی اطلاعات اولیه دیتاست متوجه می شویم که نوع داده ای ستون timestamp از نوع سری زمانی نمی باشد که باید آن را تغییر دهیم.

	value
count	4032.000000
mean	61.837054
std	56.664703
min	1.000000
25%	15.000000
50%	48.000000
75%	89.000000
max	656.000000

تصویر روبه به رو نشان می دهد که انحراف معیار ستون value بالا است که ممکن است وجود ناهنجاری در میان این داده ها سبب این اتفاق شده باشد.



تصویر بالا نمایشی از دیتاست موجود می باشد که از تاریخ 10-04-2014 تا 24-10-2014 می باشد و در آن شاهد نقاطی با تغییرات زیاد هستیم که احتمالاً ناهنجاری ها هستند.

	timestamp	value	prev_1	prev_2	prev_3	prev_4	prev_5
5	2014-04-10 00:29:00	10.0	51.0	95.0	187.0	56.0	94.0
6	2014-04-10 00:34:00	49.0	10.0	51.0	95.0	187.0	56.0
7	2014-04-10 00:39:00	79.0	49.0	10.0	51.0	95.0	187.0
8	2014-04-10 00:44:00	24.0	79.0	49.0	10.0	51.0	95.0
9	2014-04-10 00:49:00	73.0	24.0	79.0	49.0	10.0	51.0
...	...	...	...	...	...	...	...
4027	2014-04-24 00:19:00	32.0	4.0	12.0	29.0	182.0	13.0
4028	2014-04-24 00:24:00	57.0	32.0	4.0	12.0	29.0	182.0
4029	2014-04-24 00:29:00	10.0	57.0	32.0	4.0	12.0	29.0
4030	2014-04-24 00:34:00	18.0	10.0	57.0	32.0	4.0	12.0
4031	2014-04-24 00:39:00	60.0	18.0	10.0	57.0	32.0	4.0

4027 rows × 7 columns

برای تحلیل داده ها و بررسی رفتار آنها نمی شود تنها با بررسی مقدار یک نمونه تشخیص دهیم که ناهنجار است یا خیر، پس رفتار آن را نسبت به 5 نمونه قبلی خود بررسی می کنیم که برای این کار 5 ستون دیگر اضافه می کنیم که نشان دهنده 5 نمونه قبلی هر داده هستند. پس از انجام این کار برخی ستون های 5 نمونه اول خالی می ماند(نبود داده قبل از آن ها) که برای برخورد با آن ها با توجه به تعداد پایین موارد می شود آن ها را حذف کرد.

حال داده ها را درون ماتریسی از اعداد قرار می دهیم تا به عنوان ورودی الگوریتم آماده شوند.

آموزش مدل

Isolation Forest

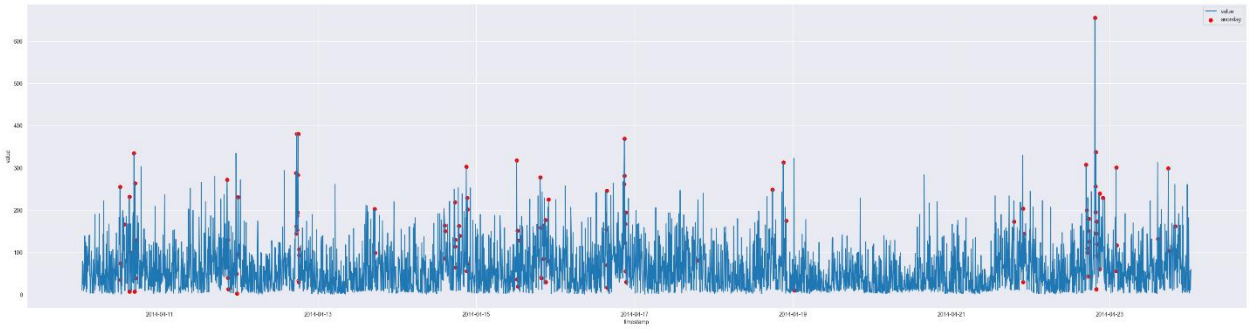
الگوریتم های از نوع "جنگل" به این منظور هستند که از تعدادی درخت تشکیل می شوند تا در نهایت به کمک همه ی آنها به یک جمع بندی کلی رسید. Isolation Forest هم از این نوع است که روش کار آن به این صورت است که یک ویژگی را به صورت تصادفی انتخاب می کند، همچنین یک مقدار تصادفی بین کمینه و بیشینه ی آن انتخاب می کند تا ویژگی را جدا کند. به این ترتیب با تقسیم کردن ویژگی به قسمت های زیاد، داده هایی که ناهنجار باشند مشخص می شوند.

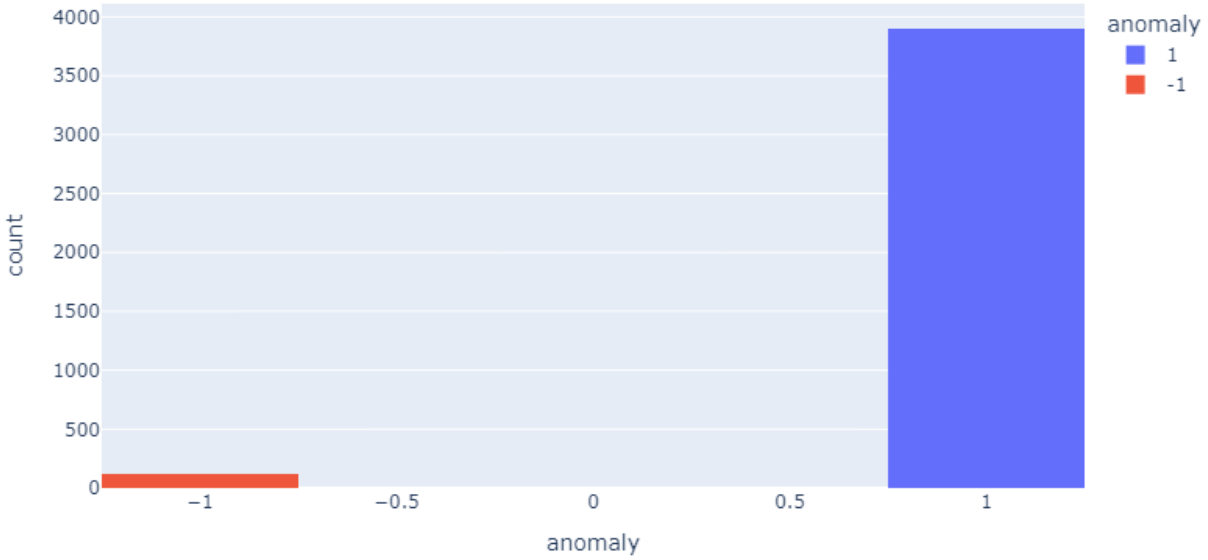
مدل را روی داده های ورودی با نرخ آلودگی 3 درصد (نسبت ناهنجاری ها به کل داده ها) آموزش می دهیم و در نتیجه به هر نمونه یک بر چسب ناهنجار(-1) یا نرمال(1) می دهد(تصویر زیر)

	timestamp	value	prev_1	prev_2	prev_3	prev_4	prev_5	anomaly
5	2014-04-10 00:29:00	10.0	51.0	95.0	187.0	56.0	94.0	1
6	2014-04-10 00:34:00	49.0	10.0	51.0	95.0	187.0	56.0	1
7	2014-04-10 00:39:00	79.0	49.0	10.0	51.0	95.0	187.0	1
8	2014-04-10 00:44:00	24.0	79.0	49.0	10.0	51.0	95.0	1
9	2014-04-10 00:49:00	73.0	24.0	79.0	49.0	10.0	51.0	1
...	...	...	...	...	...	...	...	...
4027	2014-04-24 00:19:00	32.0	4.0	12.0	29.0	182.0	13.0	1
4028	2014-04-24 00:24:00	57.0	32.0	4.0	12.0	29.0	182.0	1
4029	2014-04-24 00:29:00	10.0	57.0	32.0	4.0	12.0	29.0	1
4030	2014-04-24 00:34:00	18.0	10.0	57.0	32.0	4.0	12.0	1
4031	2014-04-24 00:39:00	60.0	18.0	10.0	57.0	32.0	4.0	1

4027 rows × 8 columns

حال با مشخص کردن ناهنجاری ها در میان کل داده ها می توان نمایش خوبی از آن ها را دید.





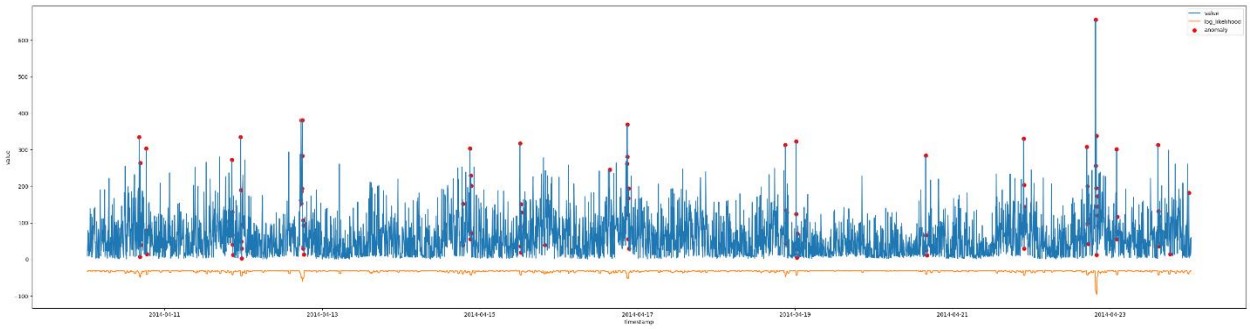
همچنین با رسم نمودار هیستوگرام داده ها، به نسبت ناهنجاری ها و داده های نرمال دست می یابیم که نشان می دهد ناهنجاری های کمی وجود دارند.

Density Based (Normal Distribution)

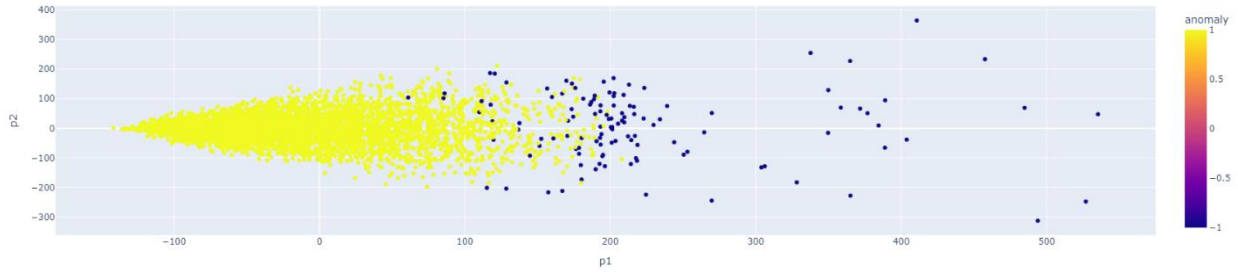
با استفاده از توزیع های آماری می توان ناهنجاری ها را تشخیص داد؛ بدین صورت که داده ها را بر روی توزیع نرمال می بریم و با توجه اینکه در این توزیع بیشتر داده ها حول میانگین هستند، هر داده ای که بسیار دور تر از میانگین بود(دور تر از مجموع میانگین و 2 یا 3 برابر انحراف معیار)، ناهنجاری محسوب می شود.

ابتدا میانگین هر سطر از ماتریس ورودی را به دست می آوریم، سپس ماتریس کوواریانس ستون ها را حساب می کنیم(ارتباط میان ستون ها) و در آخر توزیعی نرمال از این مقادیر به دست آمده را می سازیم؛ البته برای محاسبات عددی معمولاً لگاریتم این توزیع را حساب می کنند. به این توزیع به دست آمده like_lihood گفته می شود که نشان دهنده احتمال وجود داده بر روی توزیع مربوطه است که در اینجا ما لگاریتم آن را به دلایل گفته شده، محاسبه کردیم.

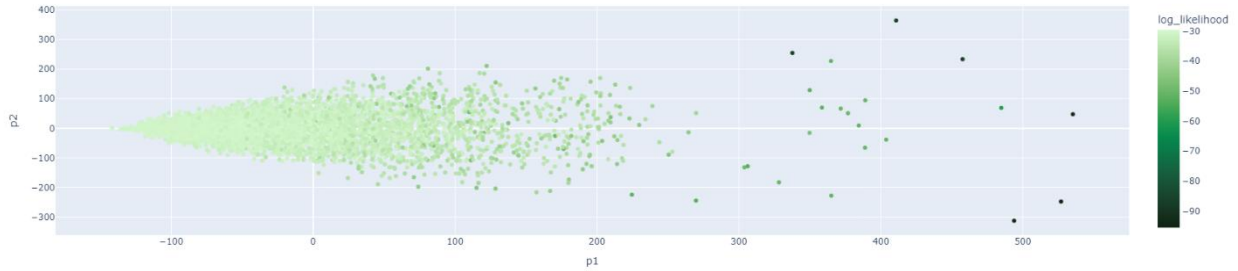
در تصویر زیر همه ی داده ها با مشخص بودن ناهنجاری ها و لگاریتم like_lihood هر داده را مشاهده می کنیم. داده هایی که لگاریتم like_lihood کمتر از 40- دارند را به عنوان ناهنجاری در نظر گرفته ایم(اختلاف های زیاد در نمودار نارنجی رنگ)



همچنین می توان با کاهش ابعاد داده ها، ناهنجاری ها را تشخیص داد که این کار را با روش Principle Component Analysis انجام می دهیم، در واقع تمام آن ویژگی ها را به دو ویژگی تبدیل می کنیم با دارا بودن بیشترین اطلاعات و کمترین پِرتی اطلاعات.

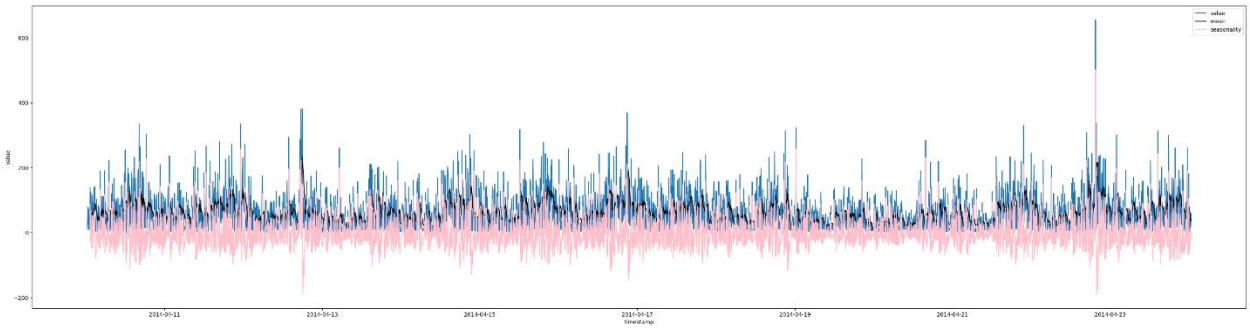


تشخیص داده های پرت از روی برچسب ناهنجاری در 2 بعد



تشخیص داده های پرت از روی لگاریتم like_lihood در 2 بعد

حتی می توان با محاسبه اختلاف هر داده با میانگین، ناهنجاری ها را تشخیص داد. این روش به نام seasonality شناخته می شود که نشان دهنده فصلی بودن(نادر بودن) رویداد ها(می تواند ناهنجاری برداشت شود)



نمودار صورتی رنگ seasonality را نشان می دهد که اختلافات زیاد نسبت به داده های قبلی همان ناهنجاری ها هستند

Density Based (Gaussian Distribution)

این روش نیز روشی مبتنی بر توزیع های آماری و چگالی است.

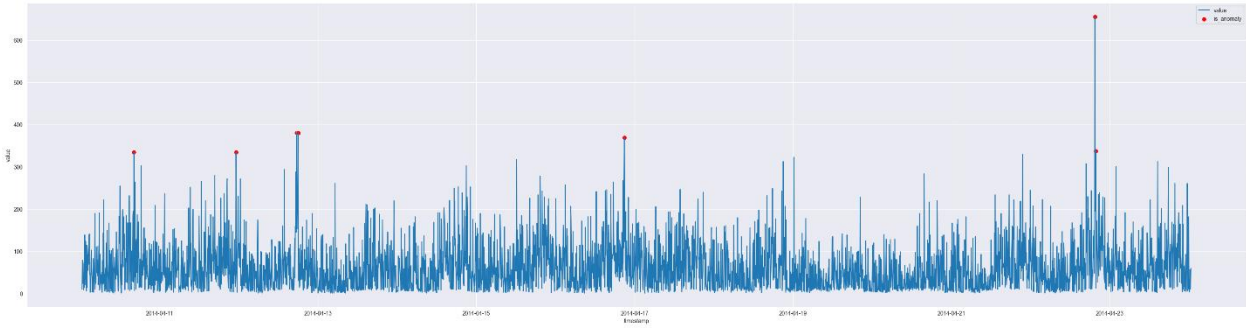
توزیع گوسی از داده ها را ایجاد می کند. با استفاده از توزیع گوسی داده ها را به دو جزء تقسیم می کنیم. سپس احتمال وجود هر نمونه را در این توزیع به دست می آوریم.

سپس با آستانه گذاری 15٪ (15 درصد کوچک ترین مقادیر احتمال) برای تشخیص داده ها به عنوان ناهنجاری، داده های پرت را به دو دسته ناهنجار و نرمال تقسیم می کنیم.

	timestamp	value	prev_1	prev_2	prev_3	prev_4	prev_5	anomaly	log_likelihood	p1	p2	mean	seasonality	probability	is_anomaly	is_not_anomaly
193	2014-04-10 16:14:00	335.0	48.0	27.0	195.0	186.0	83.0	-1	-45.897683	200.910812	-1.279429	116.3	218.7	-12.473326	True	False
564	2014-04-11 23:09:00	335.0	31.0	16.0	55.0	45.0	128.0	1	-44.045512	88.643060	72.090089	79.1	255.9	-12.473326	True	False
785	2014-04-12 17:34:00	381.0	162.0	145.0	288.0	156.0	159.0	-1	-50.615333	371.619877	66.598034	158.3	222.7	-15.434840	True	False
790	2014-04-12 17:59:00	381.0	283.0	194.0	187.0	153.0	381.0	-1	-60.147687	484.718533	69.176679	233.0	148.0	-15.434840	True	False
1973	2014-04-16 20:54:00	369.0	262.0	108.0	169.0	71.0	268.0	-1	-52.304852	349.751997	129.024453	157.2	211.8	-14.607687	True	False
3682	2014-04-22 19:34:00	656.0	175.0	48.0	150.0	84.0	116.0	-1	-86.478961	337.555929	254.686858	154.8	501.2	-44.949446	True	False
3685	2014-04-22 19:49:00	338.0	195.0	256.0	656.0	175.0	48.0	-1	-92.792283	535.165456	47.487105	215.4	122.6	-12.649210	True	False

تصویر بالا ناهنجاری ها را نشان می دهد که از روش توزیع گوسی به دست آمده اند. در واقع احتمال وجود آنها در توزیع کمتر از 15 درصد کوچک ترین مقادیر است.

با تنظیم آستانه گذاری می توان داده های بیشتر یا کمتری را به عنوان ناهنجاری در نظر گرفت؛ هر چقدر عدد آستانه گذاری پایین تر باشد، سخت گیرانه تر عمل می کنیم.



نتیجه گیری

برای تشخیص ناهنجاری در داده ها روش های متعددی وجود دارد که ما از Isolation Forest و Density Based(Normal&Gaussian Distribution) استفاده کردیم. نتیجه هایی که هر دو روش دادند تقریباً یکی است و بستگی به تحلیل ما دارد که چگونه با این ناهنجاری ها رفتار کنیم.