# TimeSeries Analysis:

Seasonal Decomposition: Break down login data into trend, seasonal, and residual components to detect anomalies.

Moving Average: Calculate moving averages to identify deviations from normal login patterns.

ARIMA Models: Use Autoregressive Integrated Moving Average models to forecast expected login times and identify outliers.

# Statistical Methods:

ZScore Analysis: Calculate the zscore of login counts per user to identify logins that are significantly different from the mean login count.

Gaussian Mixture Models (GMM): Model the login times using a mixture of Gaussian distributions and detect logins that have a low probability of belonging to any of the modeled distributions.

# Machine Learning Methods:

Clustering (e.g., Kmeans, DBSCAN): Cluster the login data and identify logins that fall into lowdensity regions as anomalies.

Isolation Forest: A treebased method that isolates anomalies by partitioning the data.

Autoencoders: Train an autoencoder neural network to learn normal login patterns and detect logins that result in high reconstruction error as anomalies.

# RuleBased Methods:

ThresholdBased Rules: Define rules based on business knowledge, such as maximum number of logins per hour or allowable login hours, and flag violations as anomalies.

# Hybrid Methods:

Combining Statistical and Machine Learning: Use statistical methods to preprocess data and machine learning models to detect anomalies.

Ensemble Methods: Combine the results of multiple models to improve detection accuracy.

ایجاده داده ترینینگ:

گروهها های کاری 10 عدد شامل 150 کاربر

بازه اطلاعات 6 هفته

در مجموع 3% کاربرها ادمین، 10% کارمند پشتیبانی، 30% کارمندهای عادی شیفتی و 53% کارمند عادی روزکار هستند. 5% تلورانس در این توزیع

~~گروه های با درجه اهمیت بیشتر ساعت کاری بیشتری دارند~~

شیفت صبح: 8 صبح تا 16 عصر، شیفت عصر 16 تا 24 شیفت شب 24 شب تا 8 صبح فردا

80% از همه کارکنان در 80% مواقع بین نیم ساعت قبل از آغاز ساعت کاری تا 45 دقیقه بعد از آغاز ساعت کاریشان حداقل به یک ایستگاه کاری لاگین می کنند و این 80% در 20% مواقع در زمان دیگری در داخل ساعت کاری به یک ایستگاه کاری لاگین می کنند و در کل این 80% میانگین در زمان آغاز ساعت کاری

انواع گروه کاری:

کارمندهای عادی شیفتی: 50% شیفت صبح کارمند عادی، 25% شیفت عصر کارمند عادی، 25% شیفت شب کارمند عادی.

5% کاربرها در هر هفته شیفت کاری شان را جابجا می کنند.

1% کاربرها در هر هفته بین گروه های کاری اصلی جابجا می شوند.

کارمندهای پشتیبانی: 70% شیفت روز کارمند پشتیبانی، 15% شیفت عصر کارمند پشتیبانی، 15% شیفت شب کارمند پشتیبانی.

کارمند عادی روزکار: ساعت کاری 8 صبح تا 17

ادمین روز ساعت کاری: 90% ساعت کاری 8 صبح تا 17 ادمین شب: 10% ساعت کاری 17 تا 8 صبح فردا.

کارمند عادی به دستگاه خودش و یکی دو دستگاه دیگر لاگین می نمایند. کارمند عادی در هر شیفت کاری اگر غایب نباشد (80% شیفت های کاری) به 1 تا 3 ایستگاه کاری لاگین می نماید.

کارمند پشتیبانی اگر غایب نباشد (80% شیفت های کاری) به 1 تا 10 ایستگاه کاری لاگین می نماید و به 0 تا 3 سرور غیر حساس لاگین می نماید.

ادمین اگر غایب نباشد در یک شیفت کاری به 1 تا 10 ایستگاه کاری لاگین می نماید و 0 تا 3 سرور غیر حساس و 0 تا 2 سرور حساس لاگین می نماید.

دستگاه های:

ایستگاه های کاری 100 عدد، کلا متوسط روزانه یک نفر به آنها لاگین می نماید.

سرورهای غیر حساس 5 عدد

سرورهای حساس 2 عدد

Training Data Generation:

# User role Distribution:

* 10 work groups including 150 users in total
* 3% Admins
* 10% Support Staff Shift Employees, at least 2 work groups
* 30% Regular Shift Employees, at least 2 work groups
* 53% Regular Day Employees, at least 2 work groups
* 5% Tolerance in the number of employees distribution across roles

# Distribution of Shift Employees between shifts:

## General rule

* Shift Changes: 5% of users change their shifts each 5-8 days.
* Group Changes: 1% of users switch between main workgroups each week.

## Shift employees working time:

* Morning Shift: 8 AM to 4 PM
* Afternoon Shift: 4 PM to 12 AM
* Night Shift: 12 AM to 8 AM the next day

## Regular Shift Employees:

* 50% Morning Shift
* 25% Afternoon Shift
* 25% Night Shift

## Support Staff Shift Employees:

* 70% Morning Shift
* 15% Afternoon Shift
* 15% Night Shift

## Admins:

* 90% Day Shift (8 AM to 5 PM)
* 10% Night Shift (5 PM to 8 AM the next day)

## Regular Day Employees:

* Working hours: 8 AM to 5 PM

# Login Behavior:

* Employees normally log in to at least one workstation between 30mins and 90mins after the start of their shift~~.~~
* Duration Range of Data: 6 weeks
* Presents/Absents rate is 90%

## Employees:

* Every present employee login to his/her workstation at least once in 95% of shifts.
* 50% of count of shifts employee login to their workstations more than once.
* Every employee takes at least 10 minutes to login to a new device
* Regular Employees:
  + 20% times Login to max of 2 other workstations and 80% time max of 2 non-sensitive servers and 0 sensitive servers.
  + Min time before new login is 30 mins
* Support Staff:
  + login to 50% max of 5 workstations and 50% max of 4 non-sensitive servers and 50% 1 sensitive servers.
* Admins:
  + log in to 10% max of 5 workstations, and 80% max of 4 non-sensitive servers, and 50% max of 2 sensitive servers per shift.

## Devices:

* Workstations: 100 in total, normally (95% of days) used by one person per day.
* Non-sensitive Servers: 5 in total, normally (95% of days) used by 10 users of any kind
* Sensitive Servers: 2 in total, normally (95% of days) used by 2 users of any kind

~~Higher Importance Groups: Have more working hours~~

ساعت کاری در هر روز هفته نرمال گروه را از زمان لاگین یک ماه قبل گروه استخراج می کنیم.

مرحله اول: دسته بندی کارکنان به 5 دسته بند

تشخیص گروه های زمان:

گروه های زمانی نرمال بر اساس دسته بندی زمان های فعالیت اتوماتیک تشخیص داده شود.

تشخیص زمان کاری نرمال کاربر:

1. دسته بندی کاربرها بر اساس طول مدت کاری:
   1. کاربرها 80% در شیفت کاری 8 ساعته و 20% در شیفت کاری 16 ساعته کار می کنند.
   2. 80% کاربرها زمان آغاز و پایان شیفت کاری آنها بر اساس روز هفته ثابت است و 20% در شیفت چرخشی کار می کنند.
2. دسته بندی کاربرها بر اساس تعداد دستگاهی که با آنها کار می کنند.
   1. 80% کاربرها در 95% مواقع یک دستگاه
   2. 20% در 70% مواقع با یک دستگاه و 95% با کمتر از 5 دستگاه کار می کنند.
3. دسته بندی دستگاه ها:
4. هر چند روز یک بار و از چه نوع کاربری با دستگاه های کار می کند؟
   1. 80% دستگاه ها در 95% مواقع فقط یک نفر در یک روز با آنها کار می کند.
5. آیا تعداد دفعات کارکردن با دستگاه بر اساس ساعت و روز هفته فرکانس مشخص دارد؟

انومالی یک کاربر:

تعداد لاگین در روزش از حد نرمال فاصله داشته باشد.

خارج از تعداد نرمال در خارج از بازه زمانی نرمال به ایستگاه های کاری لاگین بنماید.

به تعداد ایستگاه های کاری بیشتر از حد نرمال لاگین کرده باشد.

اگر دستگاه هایی که به آنها لاگین می کرده است غیر حساس باشند و ناگهان به یک سرور حساس لاگین بنماید.

زمانی بین جابجایی:

بازه نرمال فاصله بین لاگین های متوالی به دستگاه های مختلف توسط یک کاربر معیار دسته بندی کاربرها در گروه های مختلف باشد.

بازه زمانی نرمال فاصله بین اتصال دوکاربر متفاوت به یک دستگاه معیار دسته بندی دستگاه ها باشد.

بازه نرمال لاگین کاربر در 7 روز گذشته: میانه، میانگین، انحراف معیار، تیزی و خمیدگی

تغییر شیفت کاری چقدر انحراف ایجاد می نماید؟

تعداد لاگین نرمال کاربر:

انومالی:

تعداد دستگاهی که کاربر در یک ساعت استفاده کرده خارج از 95% نرمال باشد.

When you do not know the number of clusters beforehand, you should consider using clustering algorithms that are capable of determining the number of clusters automatically. Here are some algorithms that are suitable for this scenario:

1. \*\*DBSCAN (Density-Based Spatial Clustering of Applications with Noise)\*\*:

- DBSCAN does not require specifying the number of clusters in advance.

- It groups together closely packed points based on a specified distance measure and identifies noise points as outliers.

- It's effective for datasets with varying cluster densities and irregularly shaped clusters.

2. \*\*Mean Shift\*\*:

- Mean Shift is a non-parametric clustering technique that doesn't require specifying the number of clusters.

- It iteratively shifts centroids to the mode of the data distribution, converging towards the densest regions of the data.

- It's suitable for datasets with complex and irregularly shaped clusters.

3. \*\*Agglomerative Clustering\*\*:

- Agglomerative clustering starts with each data point as a separate cluster and merges the closest clusters recursively.

- It results in a dendrogram that can be cut at different levels to obtain clusters.

- By inspecting the dendrogram or using criteria such as the silhouette score, you can determine the optimal number of clusters.

4. \*\*Gaussian Mixture Models (GMM)\*\*:

- GMM is a probabilistic model that represents data as a mixture of several Gaussian distributions.

- It allows for flexibility in the number of clusters and can estimate the number of components based on statistical criteria such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC).

5. \*\*Spectral Clustering\*\*:

- Spectral clustering does not require specifying the number of clusters in advance.

- It transforms data into a lower-dimensional space using the graph Laplacian matrix and then performs K-means clustering.

- It's effective for datasets with complex structures and when traditional distance-based methods may not work well.

These algorithms are more suitable when you don't have prior knowledge about the number of clusters in your data. Experimentation with different algorithms and evaluation techniques can help you determine which one works best for your specific dataset and clustering objectives.