In this project we wanted to capture the anomalies in the Expedia Hotel Search dataset.

We searched for anomalies in the time series of hotel room prices with unsupervised learning (no labeled data).

The data we used can be gathered from: https://www.kaggle.com/c/expedia-personalized-sort/data

(You may need to sign up to fetch the data)

Train.csv file size is 2.5gb so it is not included in the project instead we filtered based on the most data pointed hotel and for visitors coing from USA for the integrity of the price column. You can find the data after being filtered in the project named “TimeSeriesExpedia.csv”

Since we are looking for anomalies, we used some different anomaly detection techniques.

**-) Clustering-Based Anomaly Detection (K-means)**

The underline assumption in the clustering approach is that if we cluster the data, normal data will belong to clusters while anomalies will not belong to any clusters or belong to small clusters.

**-) Isolation Forest for anomaly detection**

The Isolation Forest "isolates" observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

**-) Support Vector Machine-Based Anomaly Detection**

A support vector machine is another effective technique for detecting anomalies. A SVM is typically associated with supervised learning, but OneClassSVM can be used to identify anomalies as an unsupervised problems.

**-) Anomaly Detection using Gaussian Distribution**

We’ll assume that our data are normally distributed. This is an assumption that cannot hold true for all data sets, yet when it does, it proves an effective method for spotting outliers.

**-) Markov Chain**

Markov chains can measure the probability of a sequence of events happening. This approach builds a Markov chain for the underline process, and when a sequence of events has happened, we can use the Markov Chain to measure the probability of that sequence occurring and use that to detect any rare sequences.

In all of the above methodologies we used, we were able to capture only the high priced anomalies, this was interesting.

It is also noteworthy that we used unsupervised algorithms and the only way to test the results is by checking with some supervision before using it in real life scenarios.

Our plots are attached too to provide some results.

Ebru Aksoy

Öykü Zümrütdal

Cenk Erim Özsevim

Gülbay Başar