Sushil Shah, Data Solution Architect

Anomaly Detection / Fraud Analytics

DISCLAIMER

- I am a private citizen
- All views and opinions expressed here are mine and does not represent of my employer or clients.
- All works cited here is available from public references and I acknowledge their great work in this area.

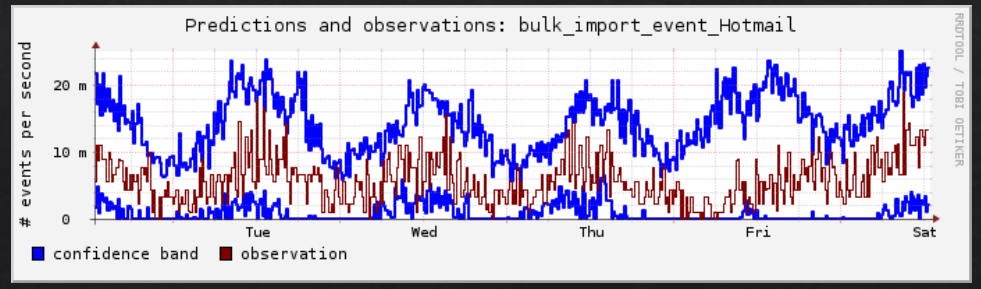
Agenda

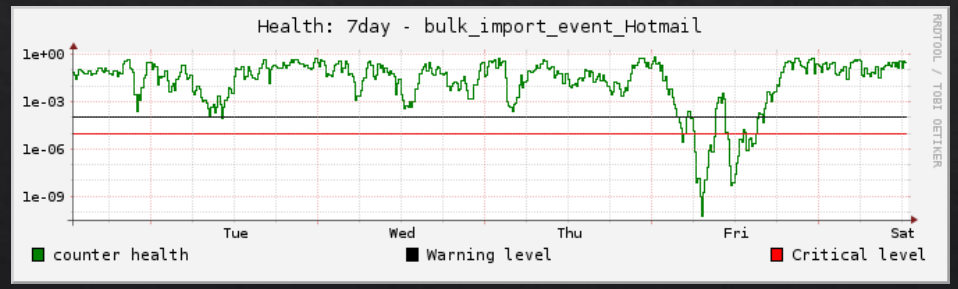
- What is Anomaly Detection
- Fraud Analytics
- Methods Used in Detecting Fraud
  - Isolation Forest
  - Autoencoders
- Architecture & Process Flow
- Q&A

Anomaly Detection

Anomaly detection is a set of techniques and systems to find unusual behaviours and/or states in systems and their observable signals.

- Detection of unusual or unexpected event or value
- Is a way to help find signal in noisy metrics.
- Detection of Outliers and Inliers





At approximately 5 AM Friday, it first detects a problem [in the number of IMVU users who invited their Hotmail contacts to open an account], which persists most of the day. In fact, an external service provider had changed an interface early Friday morning, affecting some but not all of our users.

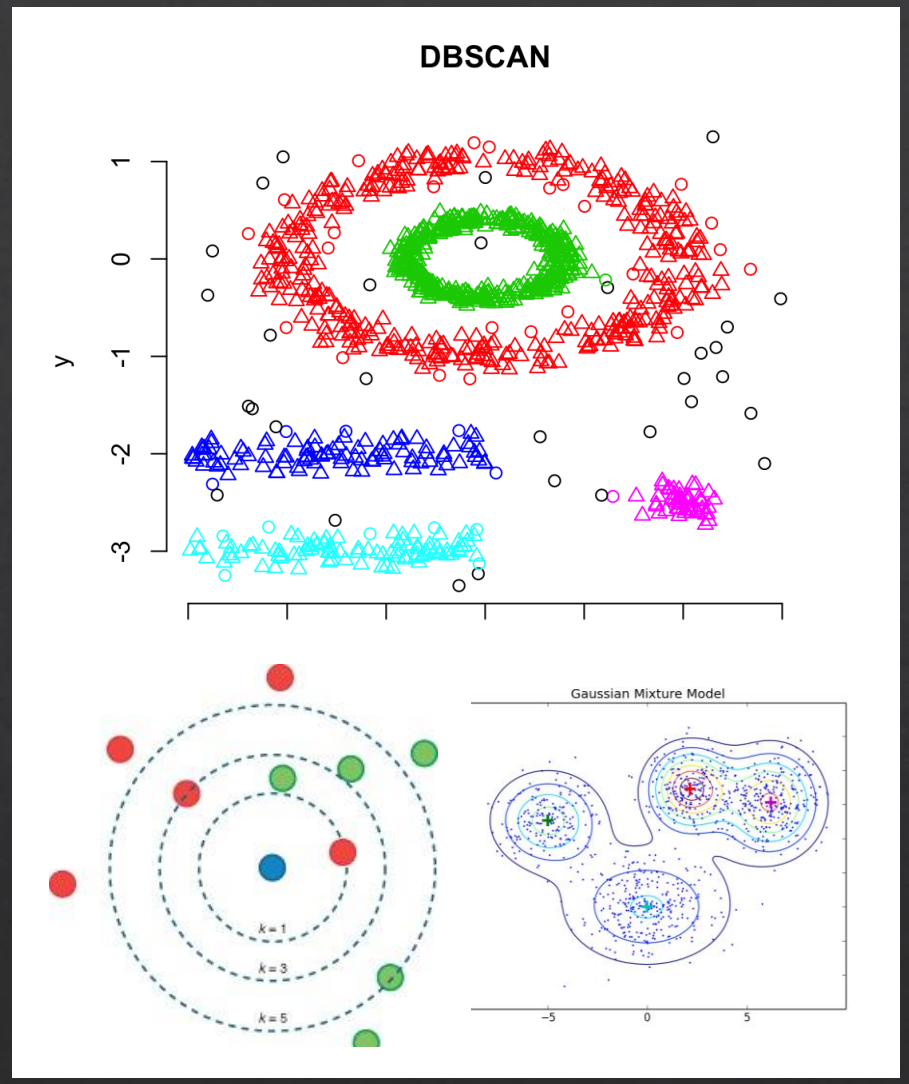
Evan Miller, 2008 paper describing real time anomaly detection in operation at IMVU
http://www.imvu.com/technology/anomalous-behavior.pdf

Anomalies Detection

- Understanding Outliers and Inliers
- Two sides of Anomalies
- Outliers
  - Points having very high values
  - Points having very low values
  - Easy to identify using their p-values
- Inliers
  - Points within main body of the distribution
  - Their neighbours tend to be far, in general
  - Harder to identify (their p-values appear normal)
- Understanding the patterns

Anomaly Detection

- Density Based (Spatial Proximity)
  - DBSCAN, LOF
- Distance Based (Clustering techniques)
  - K-Nearest Neighbours (K –NN)
  - K – MEANS
  - Regression Hyperplane Distance
- Parametric (Form based methods)
  - Gaussian Mixture Model (GMM)
  - Single Class SVMs
  - Extreme Value Theory
- Non-parametric
  - One Class SVM with RBF kernel



Anomaly Detection

- Supervised anomaly detection
  - Labels available for both normal data and anomalies
  - Similar to rare class mining/imbalanced classification
- Unsupervised anomaly detection (outlier detection):
  - No labels; *training set* = *normal* + *abnormal data*,
  - Assumption: anomalies are very rare
- Semi-supervised anomaly detection (novelty detection):
  - Only normal data available to train
  - The algorithm learns on normal data only

Fraud Analytics

Used in Data Cleaning and Labelling

- NLP Related use cases
- Labelling Data Sets

Known Red Flags

- Credit Card Fraud
- Market Abuse

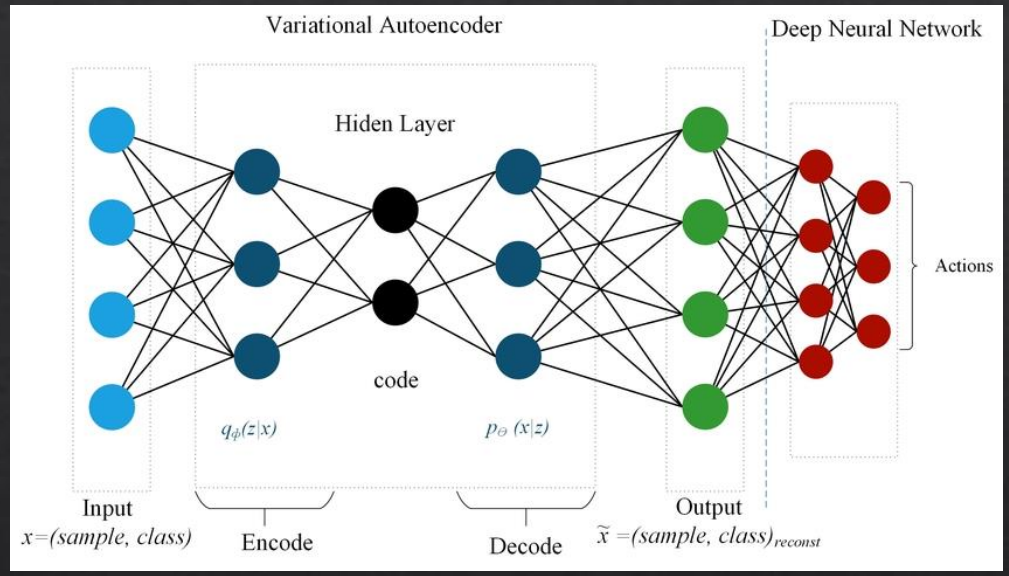
Complex Layering Anomalies

- Anti Money Laundering (AML)
  - Money Mules
  - Scams
  - Sleeper accounts
- Intrusion Detection
- Cyber Security

Fraud Analytics

(Var)Autoencoders

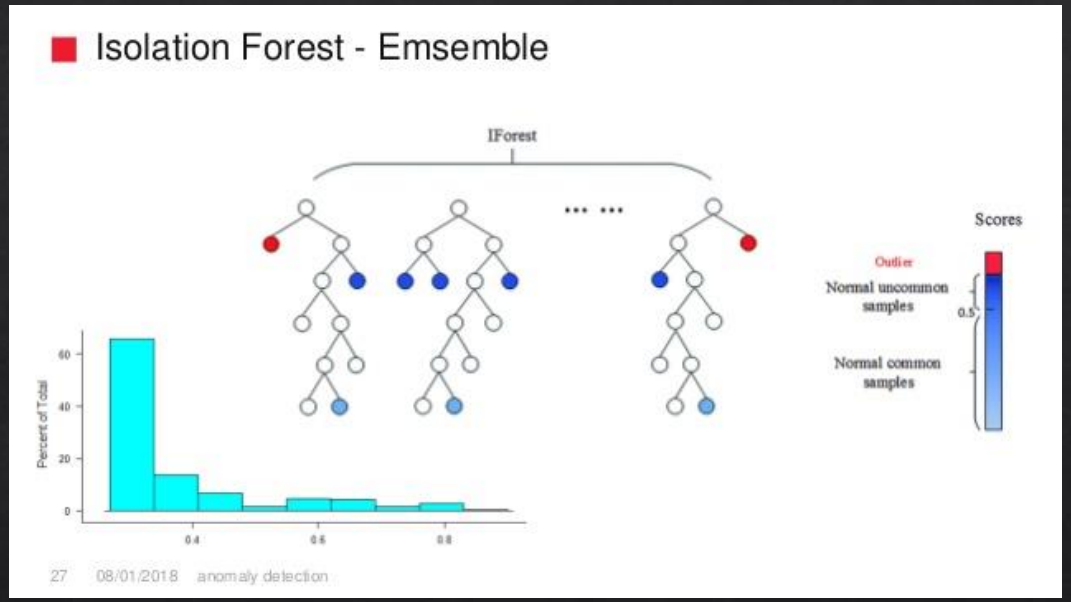
- Powerful Feature Detector
- Unsupervised Neural Network ML Technique
- Neural Networks try to reconstruct the output exactly what is input by encoding neurons..
- Adding constraints (bottleneck) to NN forces to learn the important features of Input
- Tries to reproduce the input with reconstruction error based on complexity of the NN layer.
- Anomalies are detected by evaluating the magnitude of *reconstruction error*
- The quality of the anomaly detection rely on the Mean Absolute Error (MAE) and on the Root Mean Squared Error (RSME)
- Directional Probabilistic Graphical Model – Reconstruction Probability derived from latent variable distribution.



Fraud Analytics

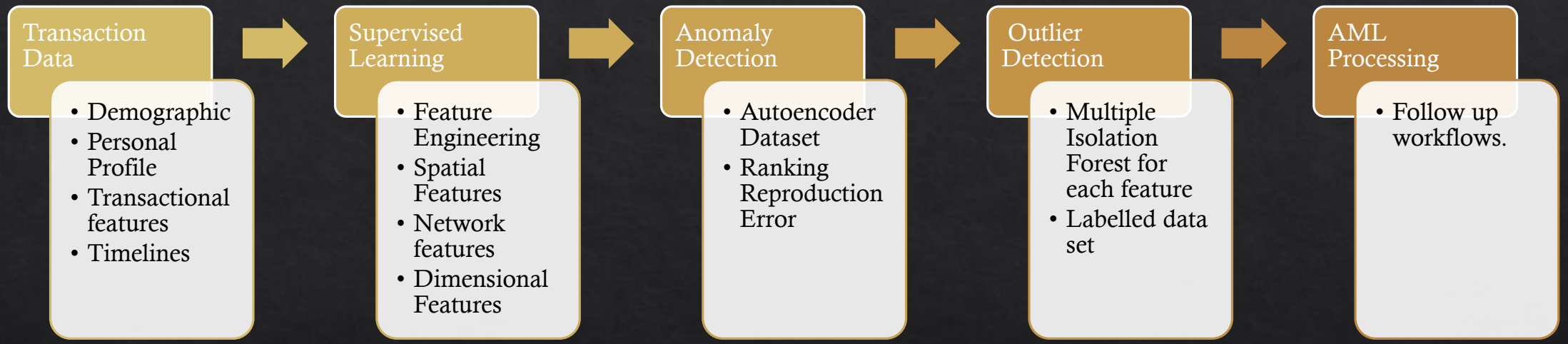
- The **isolation forest** isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.
- No profiling of normal instances, its builds ensemble of random trees of given data set and anomalies are points with shortest average path length.
- Based on random partition value for a given data point, calculates the observation distance.
- Shortest distance is considered anomalies: the anomalous data points are easier to explain than the normal data points.
- Can be used in Supervised and Unsupervised Scenario.
- Multiple patterns of Normal Data sets
- *In Money Laundering case, legitimate transactions differ very little from the ML transactions. This results in high dimension data.*

Isolation Forest

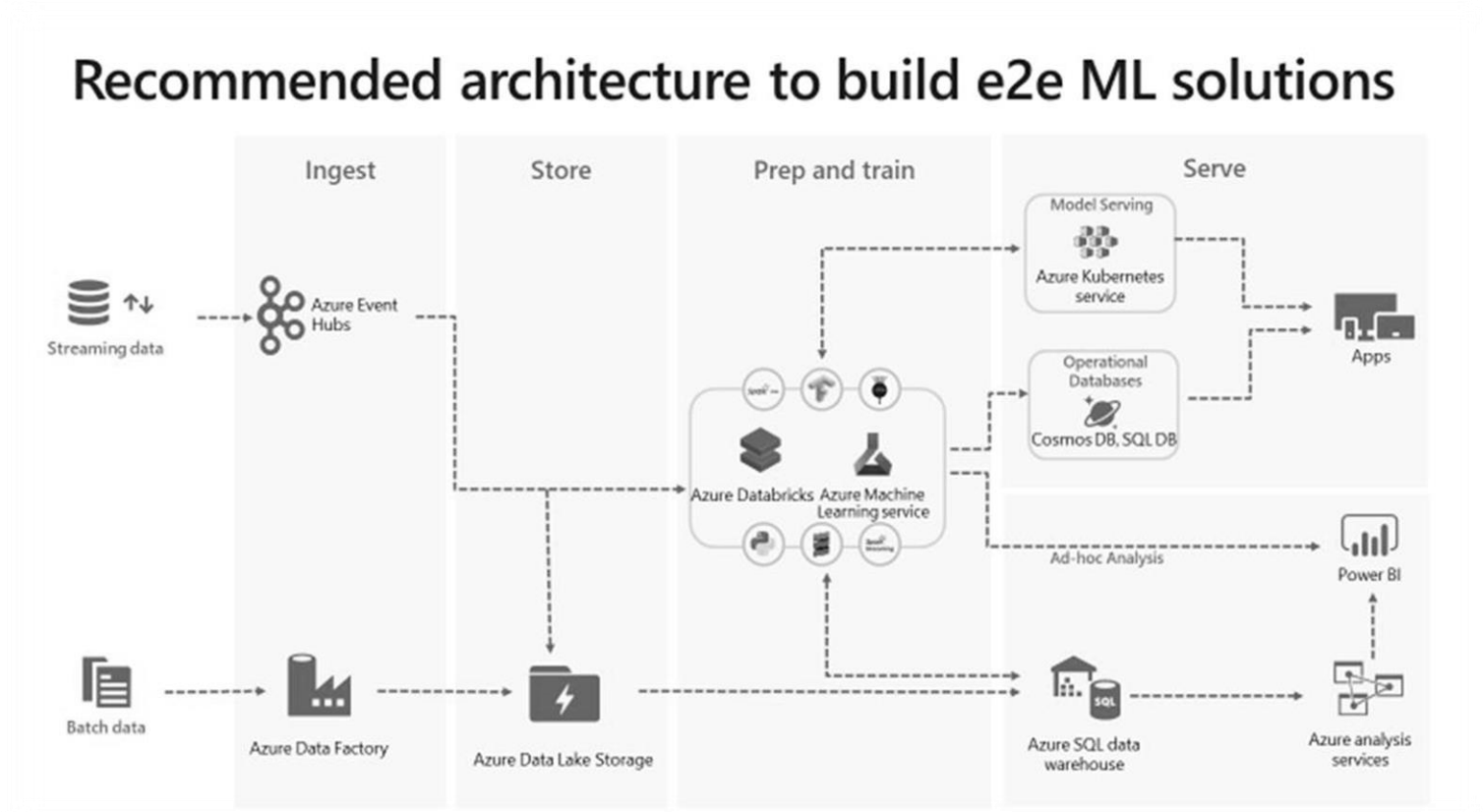


Isolation Forest Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou

Architecture & Process Flow



Architecture & Process Flow



Lookout

- Anomaly detection is probabilistic exercise reducing False Positives rather than determining the exact result.
- Multiple anomalies sets represents multiple normal data patterns.
- Data Engineering of large scale and disparate data is no mean task.
- Large feature set that requires parallel processing of the models.
- Incompatibility of version of libraries.

Thank You