Ai+ TRAINING

Advanced Fraud Modeling & Anomaly Detection

Part 1

Dr. Aric LaBarr

Associate Professor of Analytics

www.ariclabarr.com

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
- Supervised Modeling
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
  - Who am I?
  - What is Fraud?
  - Fraud Detection Analytical Framework
- Data Preparation
- Supervised Modeling
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns
- Supervised Modeling
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
- Supervised Modeling
    - Interpretable Models
    - Naïve Bayes Models
    - More Advanced Models
    - Model Evaluation
    - **NOT**-fraud Model
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
- Supervised Modeling
- Implementation Deployment
    - Clustering Revisited
    - Interpretability
    - Long-term Fraud Strategy
    - Chance & Loss Models
- Conclusion

Coding in Action

Example

Introduction

Introduction

- Introduction
  - Who am I?
  - What is Fraud?
  - Fraud Detection Analytical Framework

Introduction

Who Am I?

- Introduction
    - Who am I?
    - What is Fraud?
    - Fraud Detection Analytical Framework

Who Am I?

- 4-time North Carolina State University graduate:
  - BS in Statistics
  - BS in Economics
  - MS in Statistics
  - PhD in Statistics with minor in Economics

Who Am I?

- 4-time North Carolina State University graduate
- Former Senior Data Scientist and Director at Elder Research Inc.
    - Passionate about helping people solve challenges using their data.
    - Mentored a team of data scientists to work closely with clients and partners to solve problems in predictive modeling, advanced analytics, forecasting, and risk management.

Who Am I?

- 4-time North Carolina State University graduate
- Former Senior Data Scientist and Director at Elder Research Inc.
- Associate Professor of Analytics at Institute for Advanced Analytics at NC State University
    - Nation's first master of science in analytics degree program
    - Helped design the innovative program to prepare a modern work force to wisely communicate and handle a data-driven future.
    - Developed and taught courses in statistics, mathematics, finance, risk management, and operations research.

Who Am I?

- 4-time North Carolina State University graduate
- Former Senior Data Scientist and Director at Elder Research Inc.
- Associate Professor of Analytics at Institute for Advanced Analytics at NC State University
- Find me online:
  - https://www.linkedin.com/in/ariclabarr/
  - https://www.youtube.com/c/AricLaBarr/
  - https://www.ariclabarr.com/

Introduction

What is Fraud?

- Introduction
  - Who am I?
  - What is Fraud?
  - Fraud Detection Analytical Framework

What is an Anomaly?

anomaly
noun

/əˈnäməlē/

something that **deviates** from what is **standard**, **normal**, or **expected**

Why Detect Anomalies?

- Anomalies in data can lead to incorrect or out of date decisions to be made.
- Need to find these **outliers** before they become too much of a problem.
- Anomaly detection techniques used in variety of areas:
  - Cleaning data
  - Monitoring health of computer systems
  - Cybersecurity threats
  - Fraudulent claims or transactions

Why Detect Anomalies?

- Anomalies in data can lead to incorrect or out of date decisions to be made.
- Need to find these **outliers** before they become too much of a problem.
- Anomaly detection techniques used in variety of areas:
    - Cleaning data
    - Monitoring health of computer systems
    - Cybersecurity threats
    - Fraudulent claims or transactions

What is Fraud?

fraud

noun

/frôd/

Wrongful or criminal **deception** intended to result in financial or personal **gain**

Fraud Characteristics

1. Uncommon
2. Concealed and trying to be avoided
3. Ever changing and adapting
4. Thought out and organized
5. Doesn’t all look the same

Fraud Problem – Uncommon

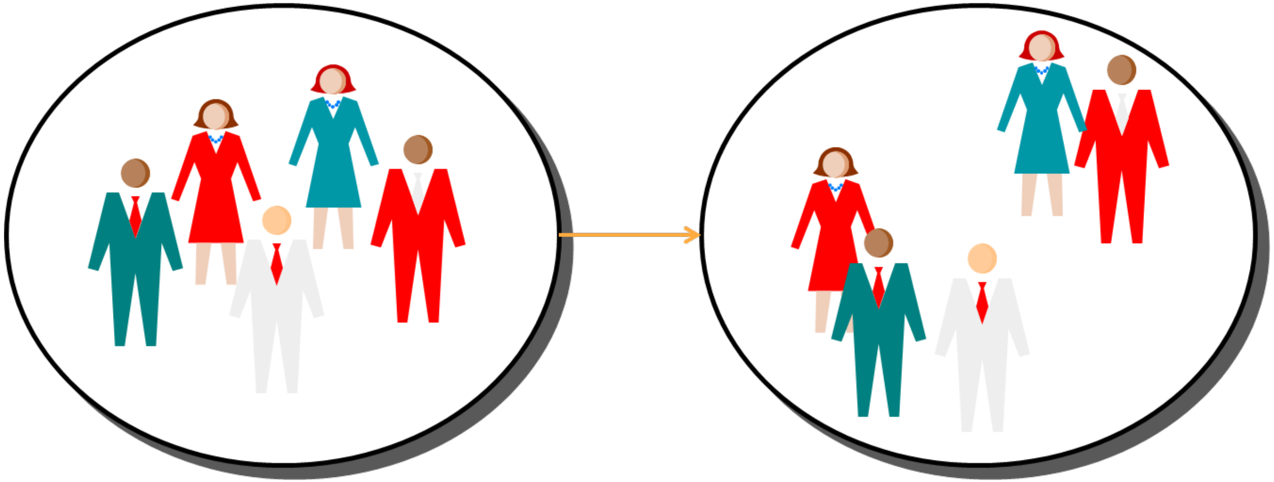
- In 2022, the ACFE (Association of Fraud Examiners) estimated that organizations lose approximately 5% of their revenues to fraud.
- Based on 2022 world GDP (IMF estimates) this would mean approximately $5.08 trillion is lost each year due to fraud.

Fraud Problem – Cat & Mouse Game

- In fraud data sets, observations are **trying to not be analyzed** or discovered – blending in.
  - Planned ahead of time – otherwise easier to detect in modeling.
  - Models have short shelf lives and are adapted often

Fraud Problem – Sociometry

- J L Moreno founded a social science called sociometry, where sociometrists believe that society is made up of individuals and their social, economic, or cultural ties.



Fraud Problem – Sociometry

- J L Moreno founded a social science called **sociometry**, where sociometrists believe that society is made up of individuals and their social, economic, or cultural ties.
- Fraud is often an organized crime.
    - No independence
    - Copycat
    - Homophily: “Birds of a feather flock together.”

Fraud Characteristics

1. Uncommon
2. Concealed and trying to be avoided
3. Ever changing and adapting
4. Thought out and organized
5. Doesn’t all look the same

- Because of these characteristics, fraud is a tough anomaly problem to solve.
- Data science can help aid in this problem!

Introduction

Fraud Detection Analytical Framework

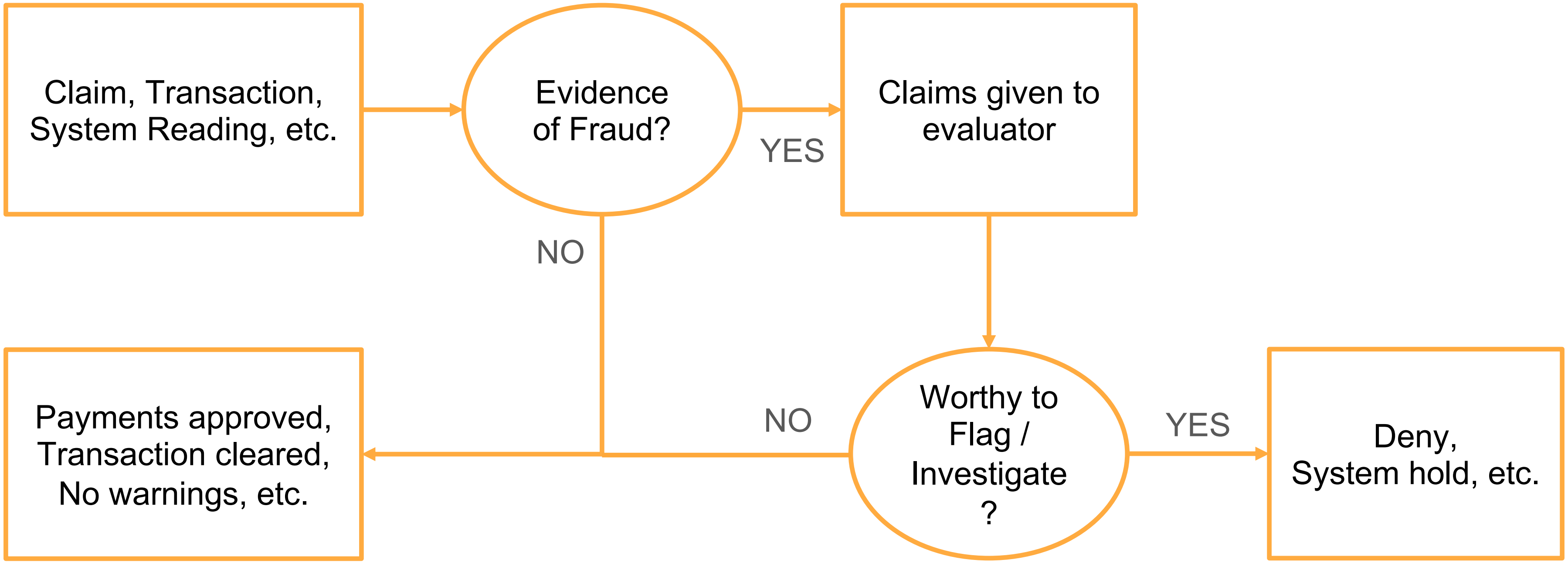
- Introduction
  - Who am I?
  - What is Fraud?
  - Fraud Detection Analytical Framework

Anomaly Detection Systems

- Regardless of the industry, two things are important for any anomaly detection solution or system:

  1. **DETECTION** – able to identify current anomalies in the system
  2. **PREVENTION** – able to flag potentially new anomalies in the system

Anomaly Detection Systems



Anomaly Detection Maturity – Card Transaction

- New / young anomaly detection solutions are based on **business rules**.
- Example:
    - IF:
        - Amount of transaction above threshold
    - THEN:
        - Flag as suspicious AND
        - Alert evaluator

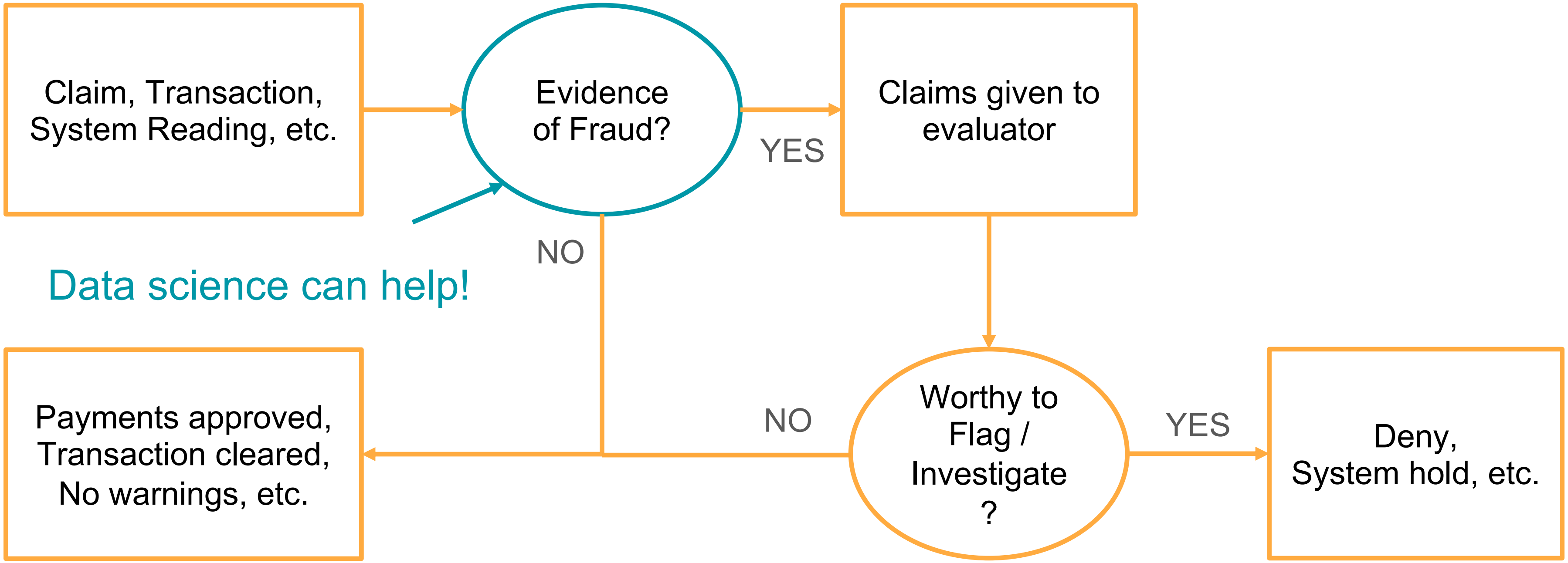
Anomaly Detection Maturity – Insurance Fraud

- New / young anomaly detection solutions are based on **business rules**.
- Example:
  - IF:
    - Severe injury but no doctor report
  - THEN:
    - Flag as suspicious AND
    - Alert evaluator

Business Rule Approach

- Advantages:
  - Simple
  - Easy to implement
- Disadvantages:
  - Expensive
  - Difficult to maintain and manage
  - Completely historical
  - Threats discover rules

Anomaly Detection Systems



Analytical Fraud Detection Framework

- Advantages
  1. **Precision**
     - Increased detection power
     - More information used in decisions
     - More anomalies evaluated

Analytical Fraud Detection Framework

- Advantages
  1. **Precision**
  2. **Efficiency in Operations**
     - Automated processing of claims
     - Ranked cases for evaluators

Analytical Fraud Detection Framework

- Advantages

    1. **Precision**
    2. **Efficiency in Operations**
    3. **Efficiency in Costs**
        - Cheaper to long-run maintain
        - Quicker evaluation
        - Higher return on evaluations

Introduction

Conclusion

- Introduction
  - Who am I?
  - What is Fraud?
  - Fraud Detection Analytical Framework

Data Preparation

Data Preparation

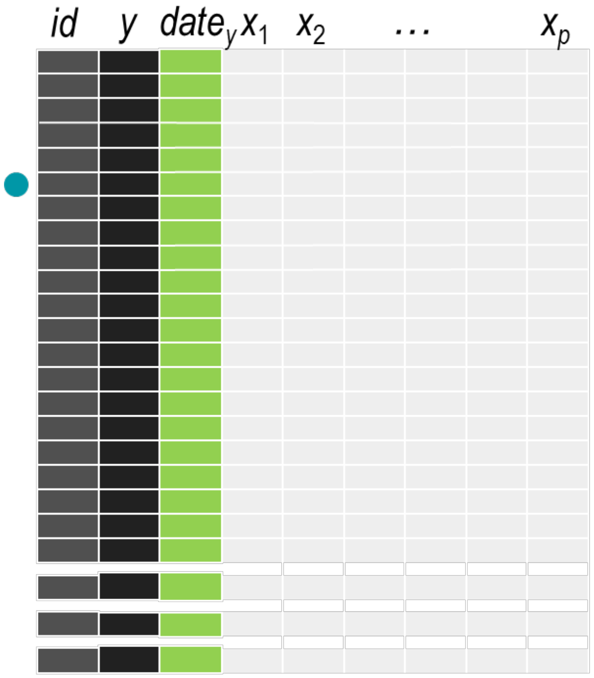
- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

Data Preparation

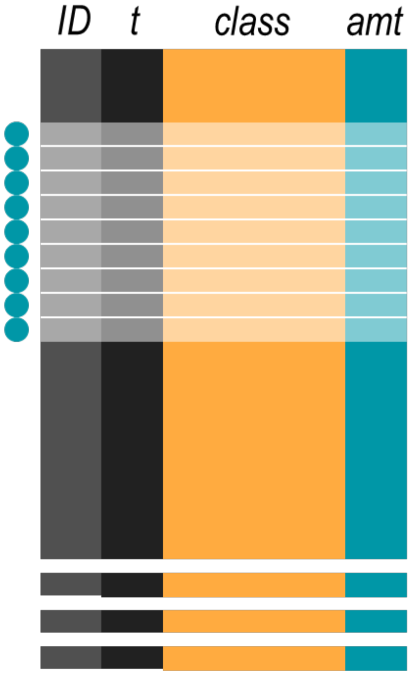
Feature Engineering

- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

Transaction Data



Model Development Data



Transaction Data

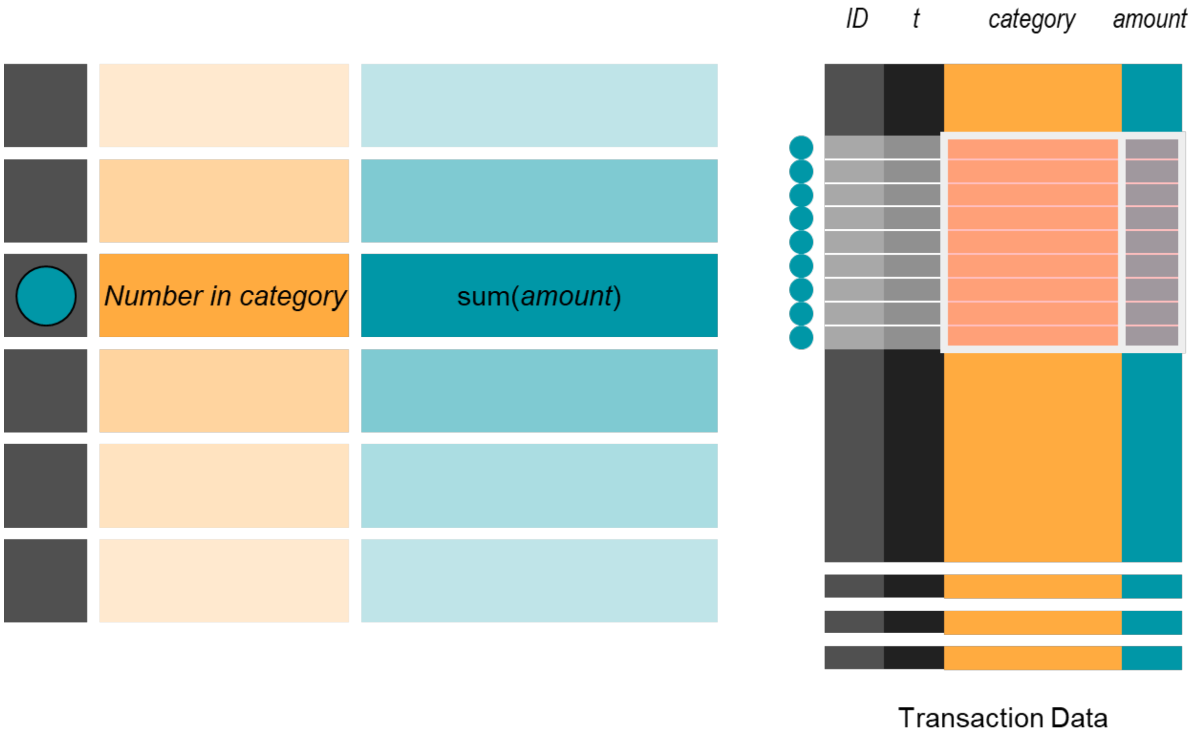
Transaction Data Examples

- There are many fields where transactional data plays an important role:
  - Credit card purchasing data
  - Medical / insurance claims data
  - Supply chain and logistics data
  - Censor / systems monitoring data
  - Etc.

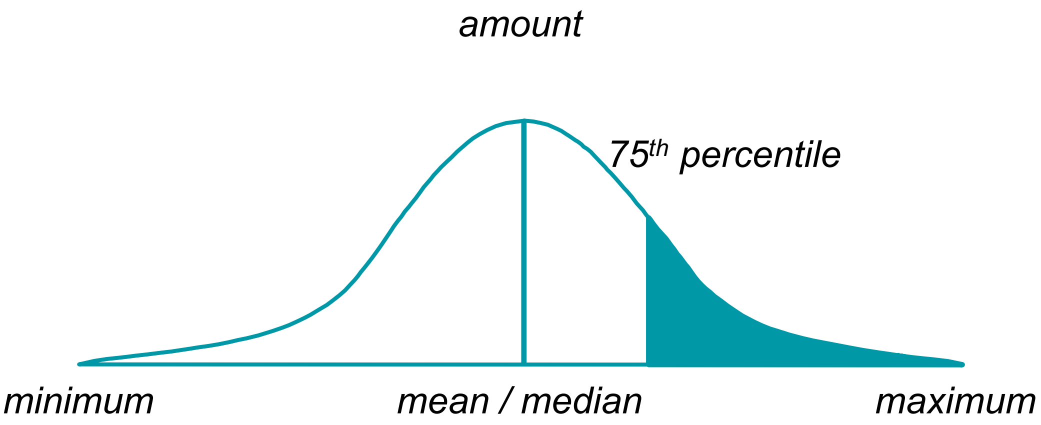
Transactions Data

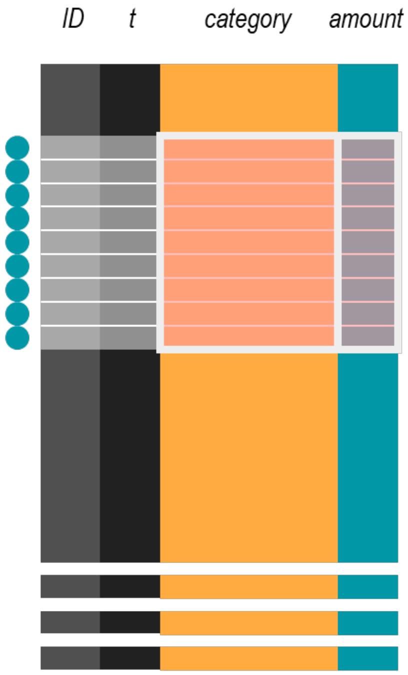
- Advantages
  - Highly detailed
  - Captures individual behavior
  - Strong prediction possible
- Challenges
  - Highly detailed
  - Difficult to obtain
  - Difficult to process

Input Possibilities: Tabulations



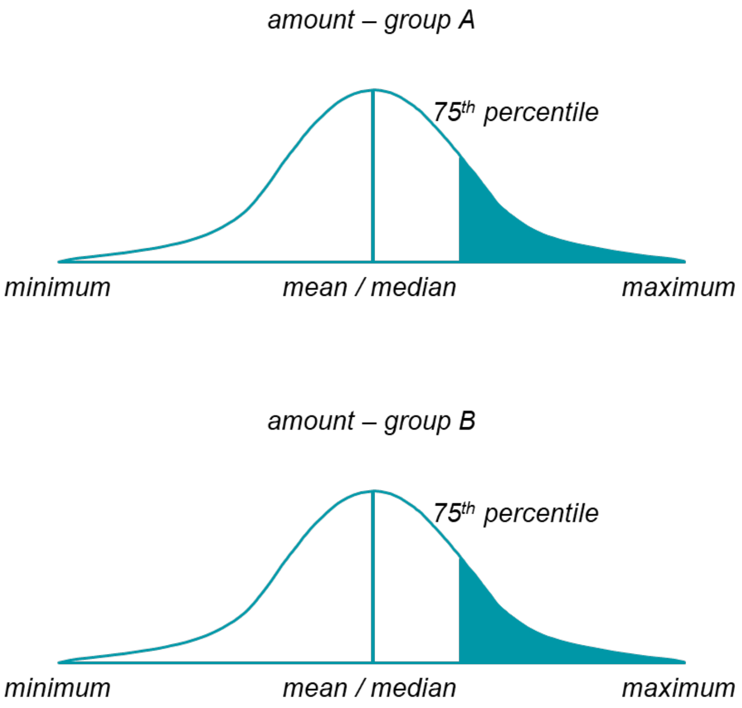
Input Possibilities: Tabulations

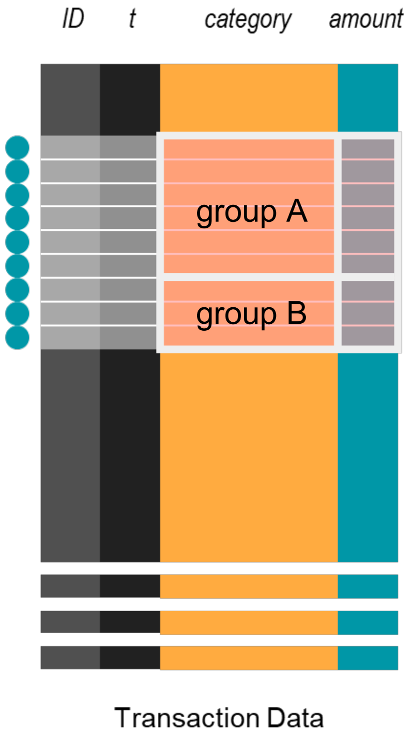




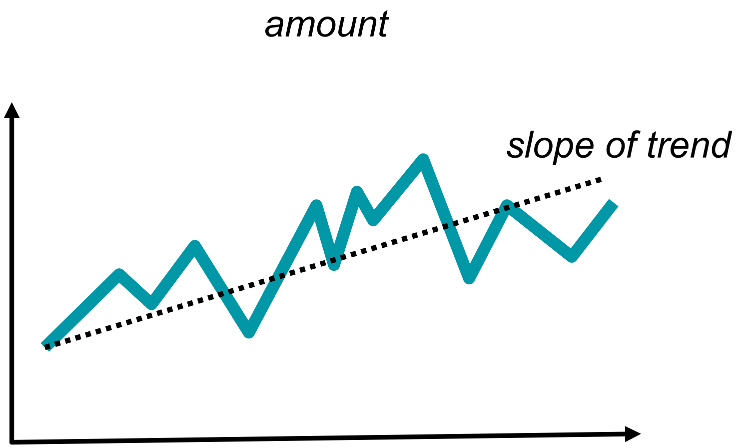
Transaction Data

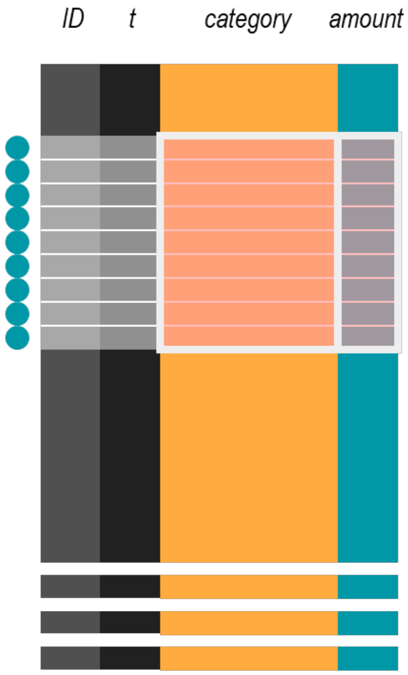
Input Possibilities: Tabulations





Input Possibilities: Tabulations





Transaction Data

Recency & Frequency

- Transactional data provides extensive information.
- Two of the most important things in fraud detection (as well as other fields) are **recency** and **frequency** of transaction.

- **Recency** – time in between transactions
- **Frequency** – how often transactions occur

Coding in Action

Data Preparation – Feature Engineering

Data Preparation

Fraud Data

- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

Fraud Data

- There are 3 common scenarios when it comes to fraud detection data sets:
    1. No previous data on fraudulent cases.

Fraud Data

- There are 3 common scenarios when it comes to fraud detection data sets:
    1. No previous data on fraudulent cases.
    2. Previous data on fraudulent cases, but can not use it.
        - Organizational structure prohibits exchange of information.
        - Retrieving data is too time consuming or expensive.
        - Fraudulent transactions can not be mapped to master database of important information.

Fraud Data

- There are 3 common scenarios when it comes to fraud detection data sets:
    1. No previous data on fraudulent cases.
    2. Previous data on fraudulent cases, but can not use it.
    3. Previous data on fraudulent cases that is fully integrated into company databases and structure.

Fraud Data

- There are 3 common scenarios when it comes to fraud detection data sets:
  1. No previous data on fraudulent cases.
  2. Previous data on fraudulent cases, but can not use it.
  3. Previous data on fraudulent cases that is fully integrated into company databases and structure.

How to handle these situations?

Anomaly Detection

- When no known fraud cases exist, we can find anomalous observations to serve as proxies.
- Anomaly detection techniques:
    - Probabilistic and Statistical Approaches
        - Benford’s Law, Z-scores, IQR Rule, Mahalanobis Distances
    - Machine Learning Approaches
        - k-NN, Local Outlier Factor, Isolation Forests, CADE, One-class SVM

Anomaly Detection

- When no known fraud cases exist, we can find anomalous observations to serve as proxies.
- 2 Paths from here:

  1. Wait for SIU to investigate anomalies and slowly gather data over time.
  2. Bring in subject matter experts (SME’s) to help with continuing modeling process.

Anomaly Detection

- Fraudulent cases will typically appear as anomalies.
- Here are the steps to take once you have your suspected anomalies:

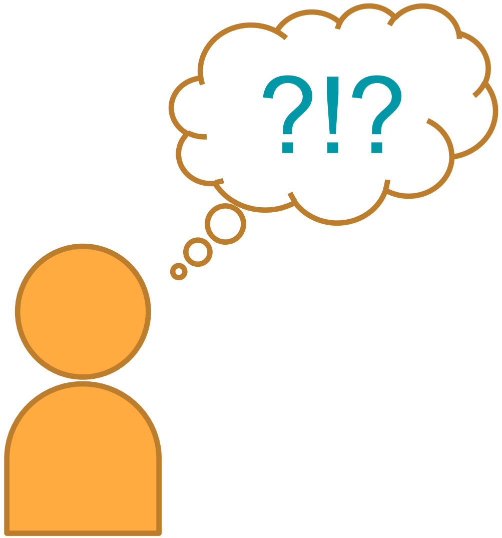
1. SME's look through the anomalies for possible fraud.
2. Tag possible fraud groups based on expert domain knowledge.
3. Treat these possible fraud cases as if they had committed fraud and other groups as if they have not.
4. Ideally, SME's also identify small set of legitimate claims.

Tagging Suspected Observations

- What are you modeling through these selection methods?

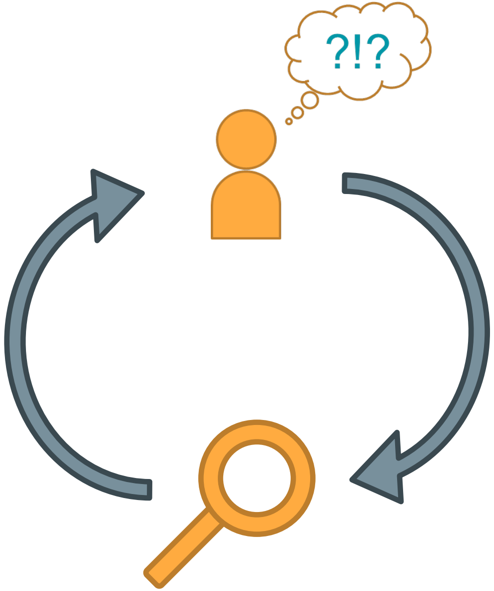
NOT FRAUD!

- Predicting domain expert classification instead of actual fraud.
- Depends on accuracy of SME's.



Tagging Suspected Observations

- This process of predicting classifications works for a limited time.
- As investigations occur and actual fraudulent claims are caught, these suspected fraud clusters are replaced with actual fraud data to help model future events.



Coding in Action

Data Preparation – Fraud Data

Data Preparation

Anomaly Detection with Statistical Techniques

- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

Anomaly Detection

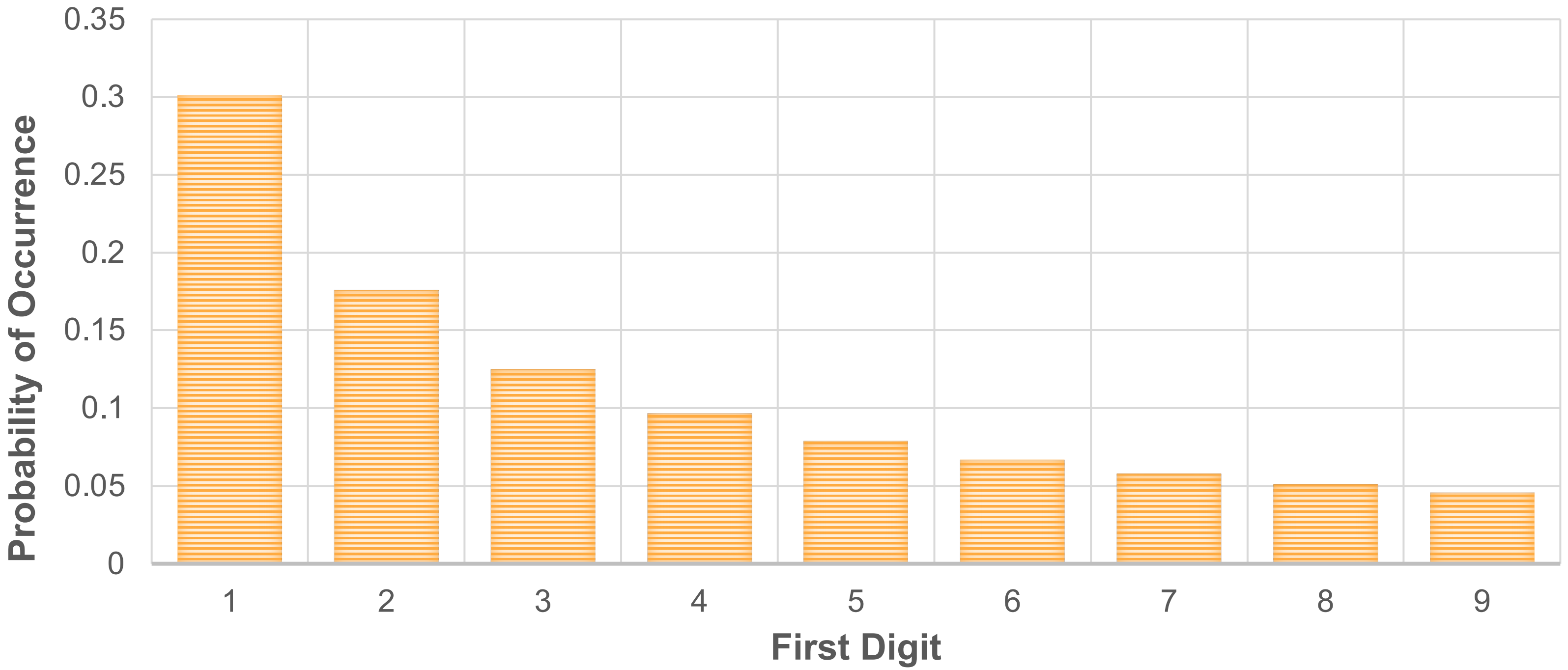
- When no known fraud cases exist, we can find anomalous observations to serve as proxies.
- Anomaly detection techniques:
  - Probabilistic and Statistical Approaches
    - Benford's Law, Z-scores, IQR Rule, Mahalanobis Distances
  - Machine Learning Approaches
    - k-NN, Local Outlier Factor, Isolation Forests, CADE, One-class SVM

Benford's Law

- Certain numbers do not occur uniformly despite what we might think.
- Digits of certain numbers follow Benford's Law.

- Example:
  - First digit of house/building numbers in addresses.
  - First digit of transaction amounts.

Benford's Law



Benford’s Law

- This wasn’t mathematically proven until the mid-90’s.
- http://testingbenfordslaw.com/
- Benford’s Law – First Digit

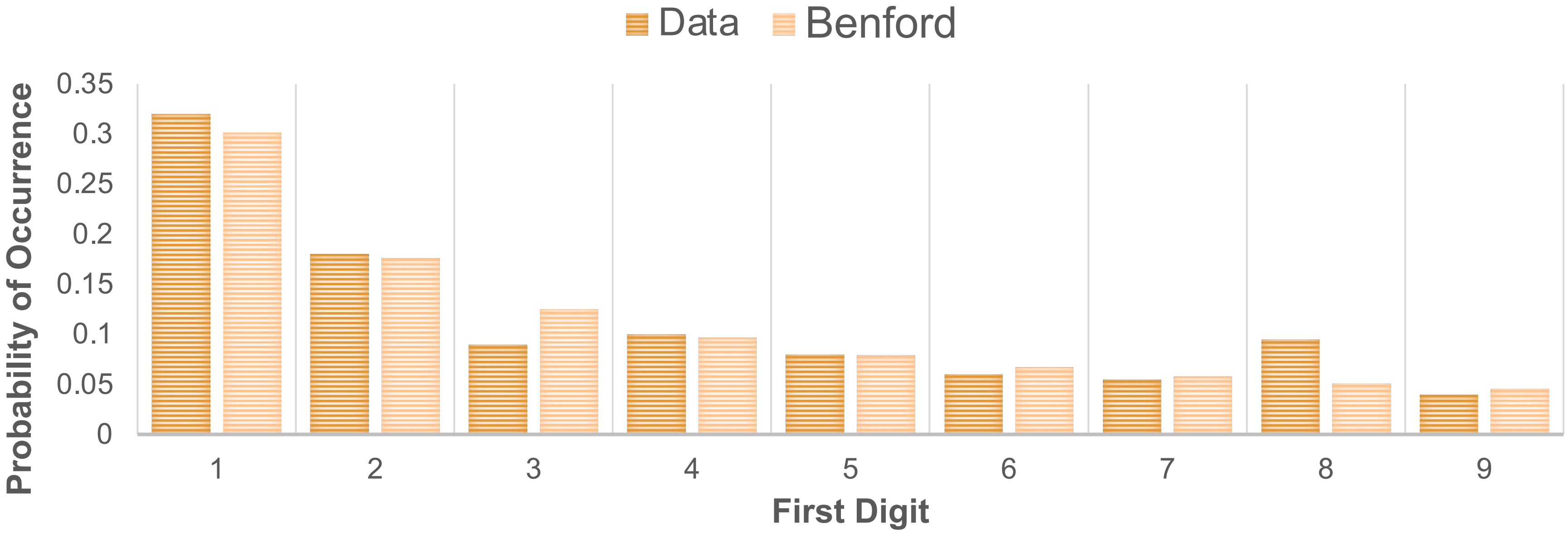
$$P(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right)$$

Benford's Law – Fraud Detection

- Fraud transactions typically involve inventing new numbers or changing real transactions into fraudulent ones.
- Legally admissible in Federal, State, and Local courts in United States as evidence.

Benford's Law – Fraud Detection

- Example transaction amounts submitted for reimbursement from scanned receipts



Benford's Law

- Fraud detection typically uses the first two digits in Benford's Law.
- Benford's Law – First Two Digits

$$P(d_1 d_2) = \log_{10}\left(1 + \frac{1}{d_1 d_2}\right)$$

$$d_1 d_2 \in [10, 11, 12, 13, \dots, 99]$$

Coding in Action

Data Preparation – Anomaly Detection with Statistical Techniques: Benford's Law

Statistical Methods

- Basic fraudulent systems look for abnormal observations from a statistical standpoint.
- Univariate analysis can help identify fraudulent **transactions** or **people** (aggregated transactions).

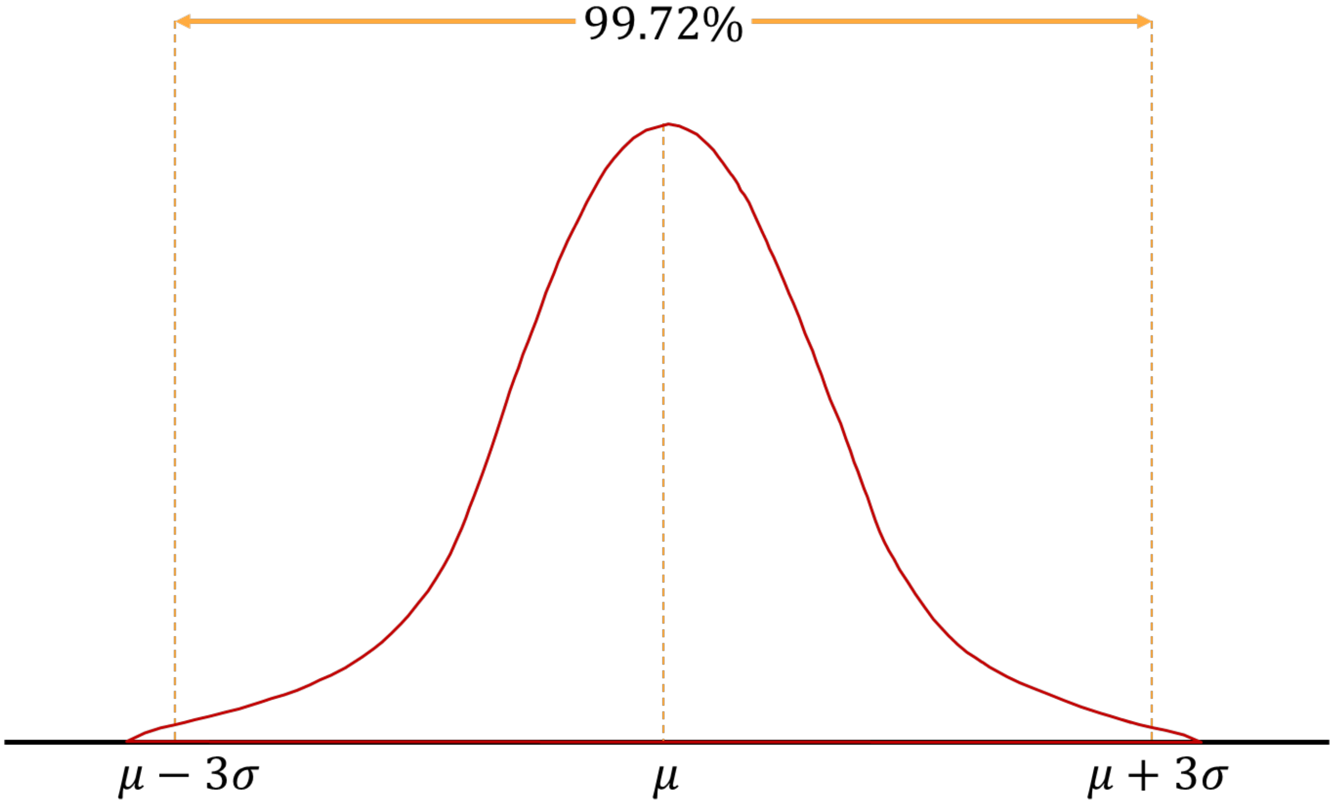
Z-Scores

- Typical with Normal distributions.

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Measures how many standard deviations away from mean each point is.
- Works best with **symmetric** distributions.

Empirical Rule



Z-Scores

- Typical with Normal distributions.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Bothered by outliers

- Measures how many standard deviations away from mean each point is.
- Works best with **symmetric** distributions.

Robust Statistics

- Outliers can greatly influence results.
- Robust techniques

1. Reliable when outliers present
2. Reliable when outliers **not** present (ideally)

Robust Z-Scores

- Robust adjustments to mean and standard deviation.

$$z_{R,i} = \frac{x_i - \mathrm{median(x)}}{\mathrm{MAD}(x)}$$

- Median Absolute Deviation (MAD):

$$\mathrm{MAD}(x) = k \times \mathrm{median}(|x_i - \mathrm{median}(x)|)$$

Robust Z-Scores

- Robust adjustments to mean and standard deviation.

$$z_{R,i} = \frac{x_i - \text{median(x)}}{\text{MAD}(x)}$$

- Median Absolute Deviation (MAD):

$$\text{MAD}(x) = k \times \text{median}(|x_i - \text{median}(x)|)$$

Adjustment factor per distribution

Robust Z-Scores

- Robust adjustments to mean and standard deviation.

$$z_{R,i} = \frac{x_i - \text{median(x)}}{\text{MAD}(x)}$$

- Median Absolute Deviation (MAD):

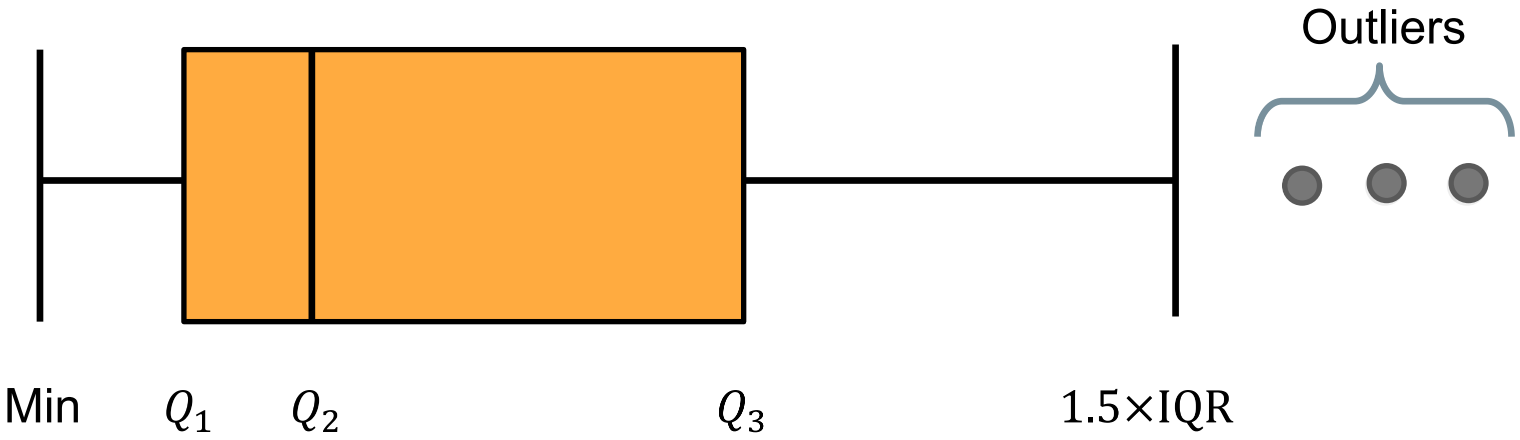
$$\text{MAD}(x) = k \times \text{median}(|x_i - \text{median}(x)|)$$

1.4826 for Normal distribution

Coding in Action

Data Preparation – Anomaly Detection with Statistical Techniques: Z-scores & Robust Z-scores

1.5 IQR Rule



1.5 IQR Rule

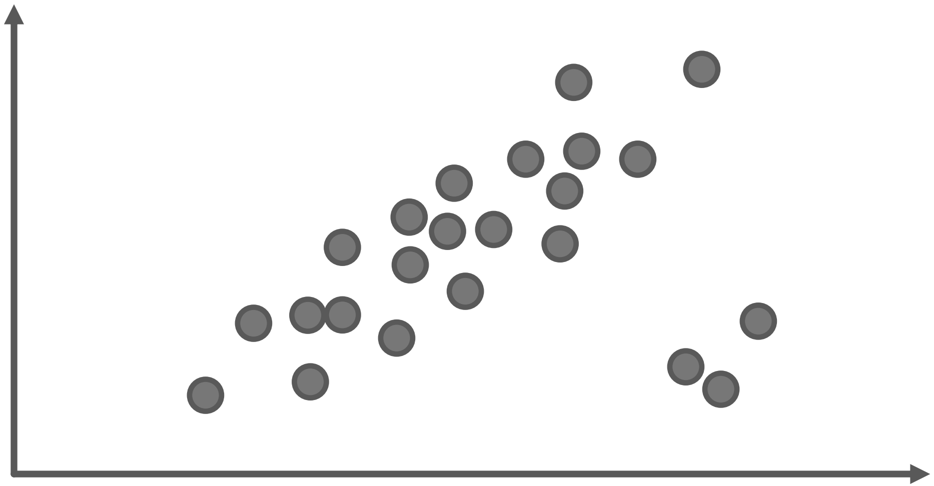
- Works best for **symmetric** distributions.
- Severely skewed distributions tend to report large number of outliers.
- Use **adjusted boxplot** instead – more robust to skewed distributions.

Coding in Action

Data Preparation – Anomaly Detection with Statistical Techniques:
IQR Rule and Its Adjustment

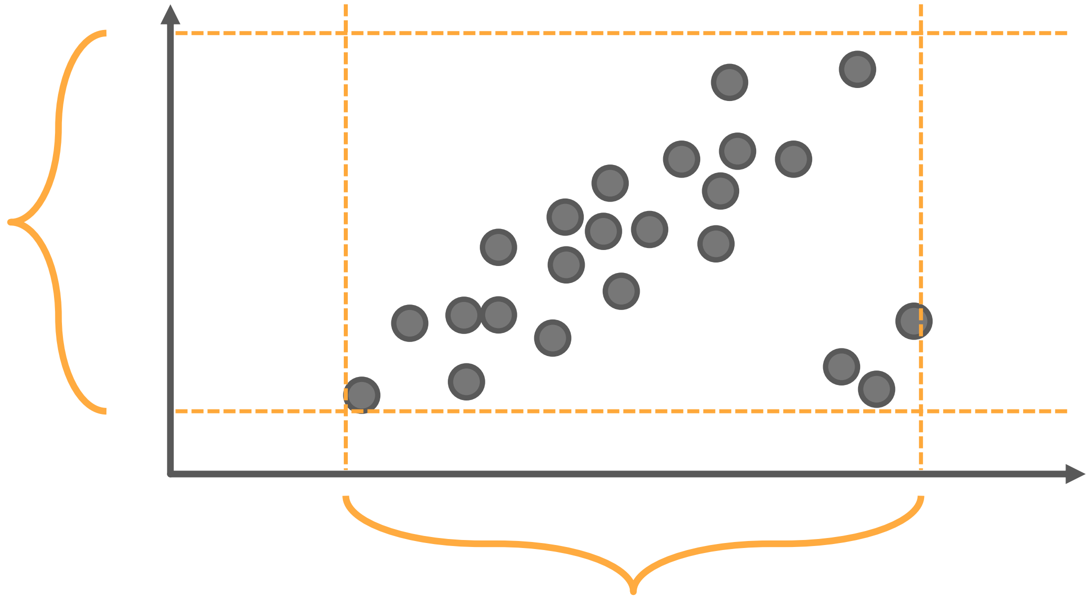
Multiple Dimensions

- Outliers in one dimension are possibly restrictive.



Multiple Dimensions

- Outliers in one dimension are possibly restrictive.



Mahalanobis Distances

- Generalization of z-scores to multi-dimensional space.
  - Replace univariate mean with **multivariate mean**
  - Replace standard deviation with **covariance matrix**

Mahalanobis Distances

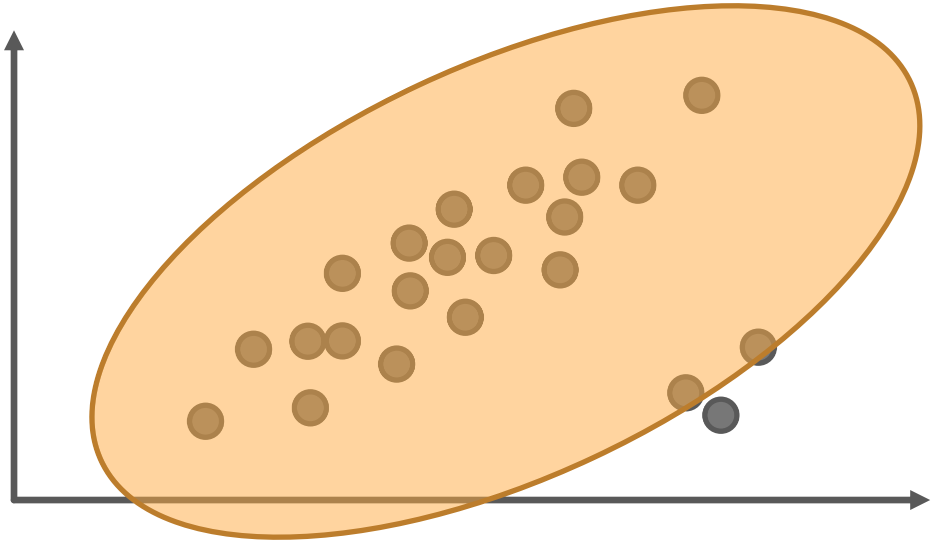
- Generalization of z-scores to multi-dimensional space.
  - Replace univariate mean with **multivariate mean**
  - Replace standard deviation with **covariance matrix**

- Euclidean Distance (L2): $D_{L2} = \sqrt{(x-\mu)^T(x-\mu)}$

- Mahalanobis Distance: $D_M = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$

Confidence Ellipsoids

- Still bothered by outliers since standard mean and covariance matrix used.



Robust Mahalanobis Distances

- Mahalanobis distances use mean and covariance matrix influenced by outliers.
- Use **robust** calculations of mean vector and covariance matrix instead:

$$D_M = \sqrt{(x - \mu_{MCD})^T \Sigma_{MCD}^{-1} (x - \mu_{MCD})}$$

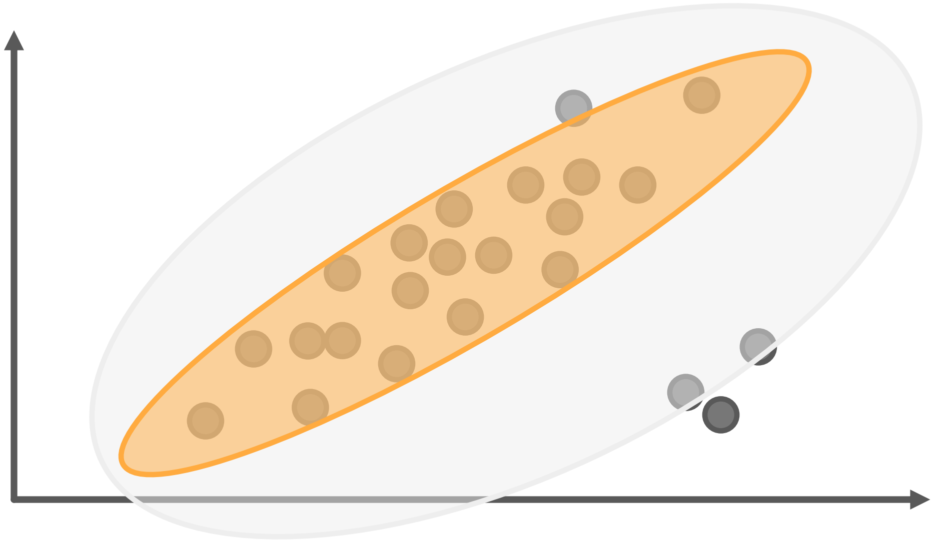
Robust Mahalanobis Distances

$$D_M = \sqrt{(x - \mu_{MCD})^T \Sigma_{MCD}^{-1} (x - \mu_{MCD})}$$

- MCD: Minimum Covariance Determinant
  - Find h ($< n$) observations that have MCD (essentially the tightest cloud)
  - Typically $h = 0.75 \times n$
  - Problem: How to find the right *h* observations?
  - Fast algorithms exist

Confidence Ellipsoids

- Robust version isn't impacted by outliers as drastically.



Coding in Action

Data Preparation – Anomaly Detection with Statistical Techniques: Mahalanobis Distances and Robust Mahalanobis

Data Preparation

Anomaly Detection with Machine Learning Techniques

- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

Anomaly Detection

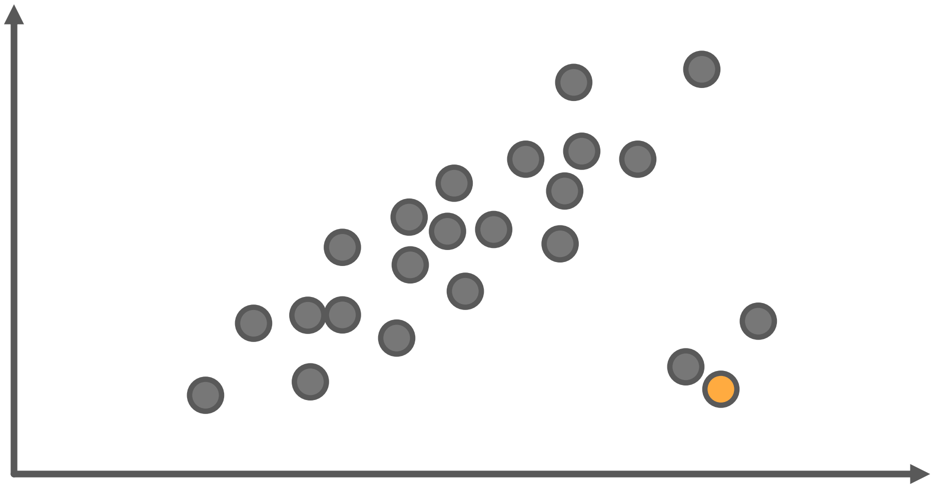
- When no known fraud cases exist, we can find anomalous observations to serve as proxies.
- Anomaly detection techniques:
    - Probabilistic and Statistical Approaches
        - Benford's Law, Z-scores, IQR Rule, Mahalanobis Distances
    - Machine Learning Approaches
        - k-NN, Local Outlier Factor, Isolation Forests, CADE, One-class SVM

k-Nearest Neighbors

- Want to discover points that are “not close” to the rest.
- Instead of distance from center of cloud, k-NN looks at distance from close points.
- Measure **average** distance from a point to each of the k-closest points.
    - Default: Euclidean distance

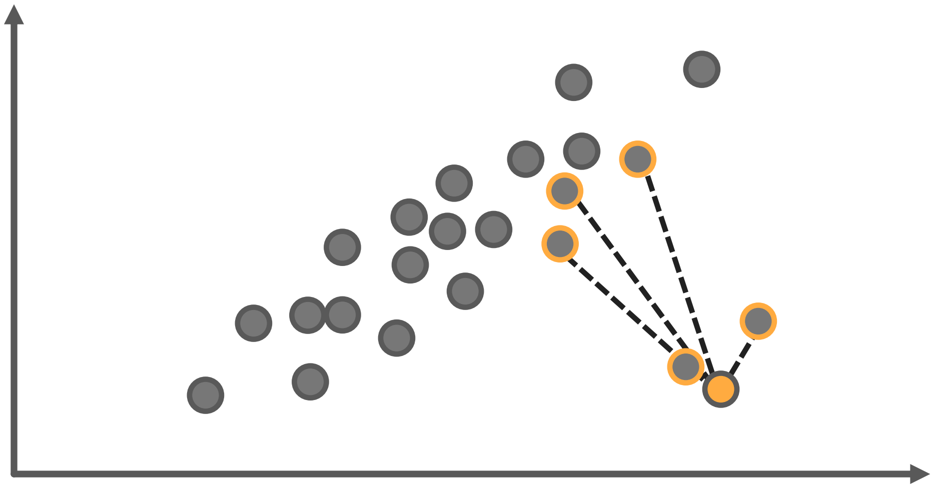
k-Nearest Neighbors

- Need to measure distances to *k* nearest observations.



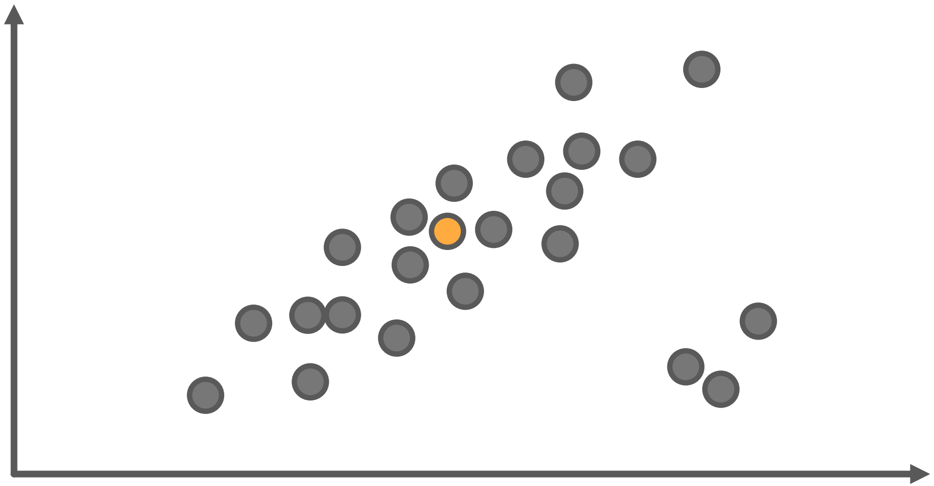
k-Nearest Neighbors

- Need to measure distances to 5 nearest observations.



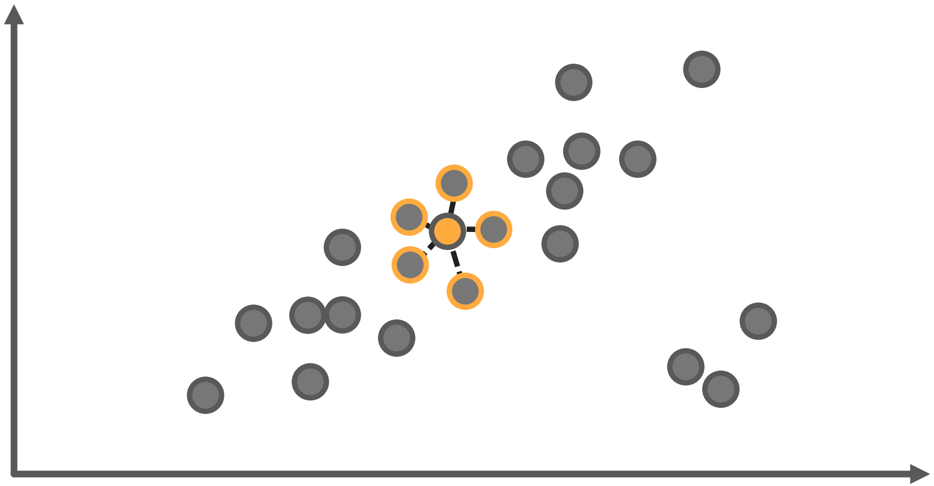
k-Nearest Neighbors

- Need to measure distances to *k* nearest observations.



k-Nearest Neighbors

- Need to measure distances to 5 nearest observations.

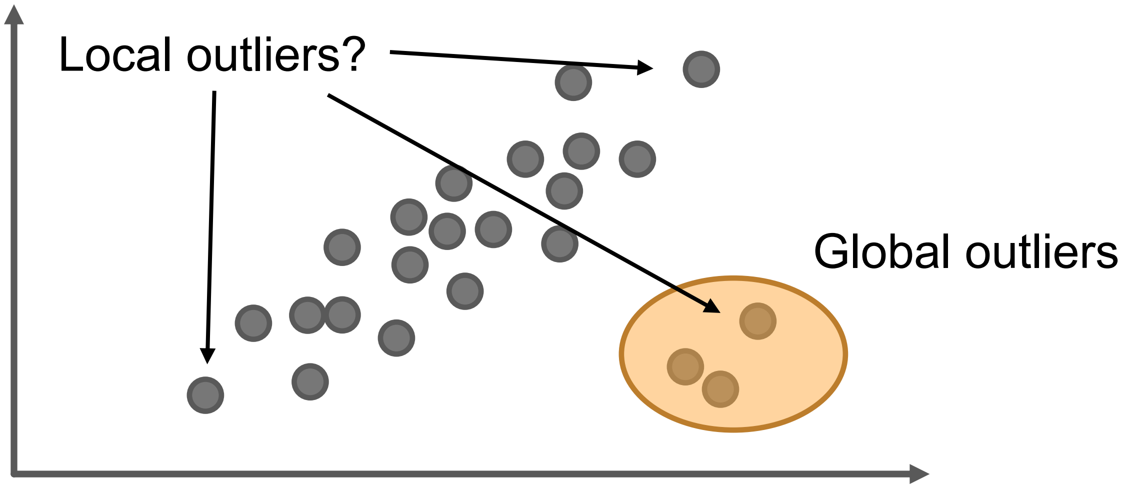


Coding in Action

Data Preparation – Anomaly Detection with Machine Learning Techniques:
k-Nearest Neighbors

Global vs. Local Outliers

- k-NN great at detecting **global** outliers, but not **local** outliers.



Local Outlier Factor (LOF)

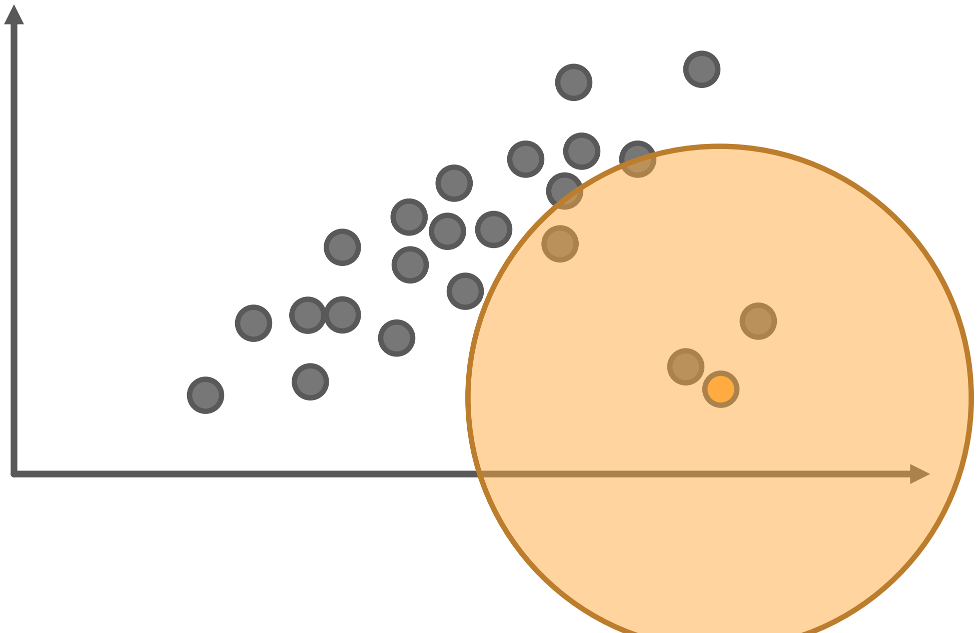
- LOF:
    - Ratio (comparison) of the average **density** of the k-NN of an observation to the **density** of the observation itself.
    - > 1 means more likely to be anomaly
    - < 1 means less likely to be anomaly

Local Outlier Factor (LOF)

- LOF:
    - Ratio (comparison) of the average **density** of the k-NN of an observation to the **density** of the observation itself.
- Density:
    - Inverse of the average **reachability** (distances) from observation to all of its k-NN.
    - Essentially, how far do we have to travel to nearest point, so less dense means farther travel.

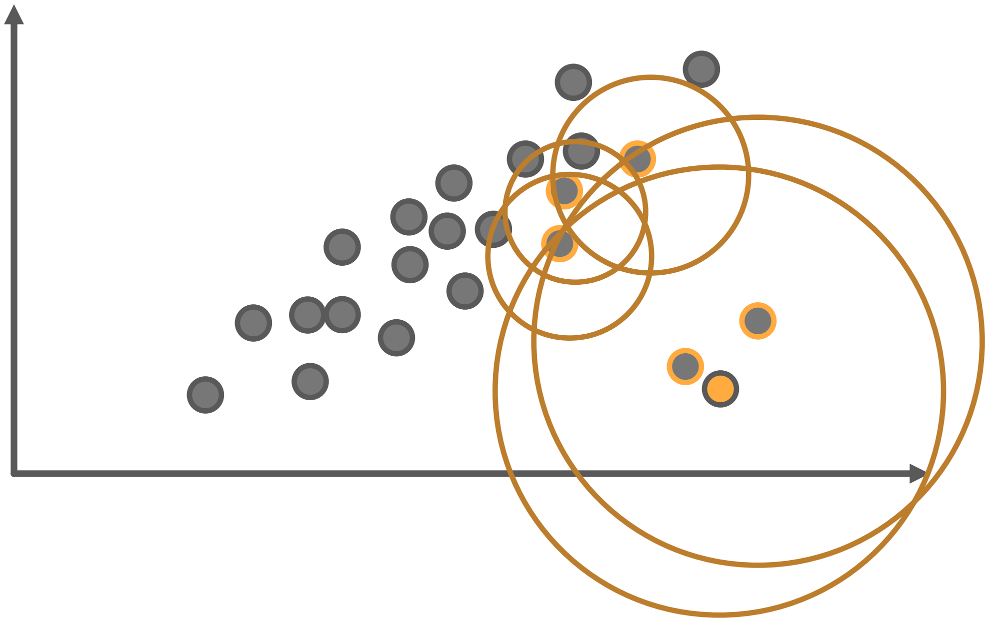
Local Outlier Factor (LOF)

- Density of observation of interest



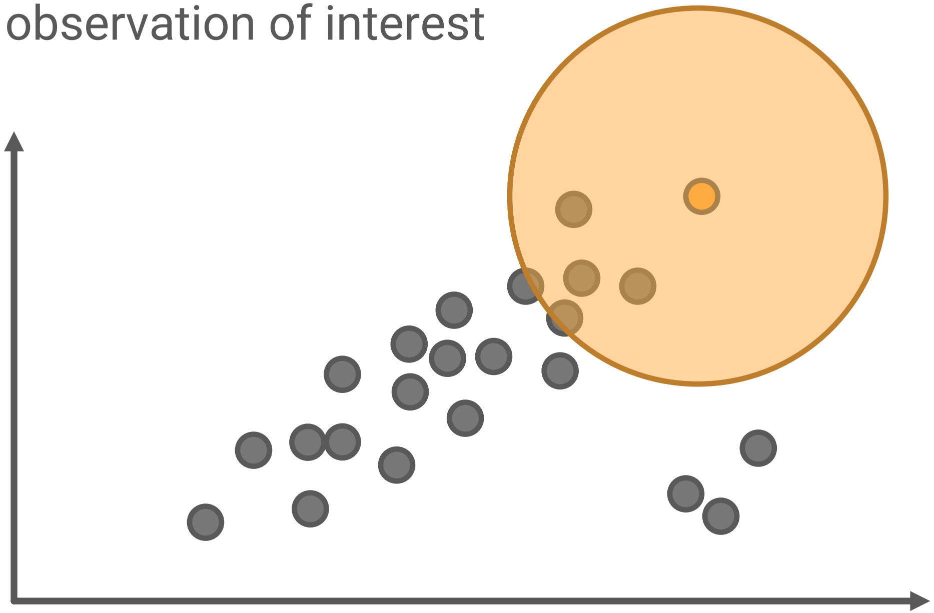
Local Outlier Factor (LOF)

- Need to average the densities of the k-NN observations.



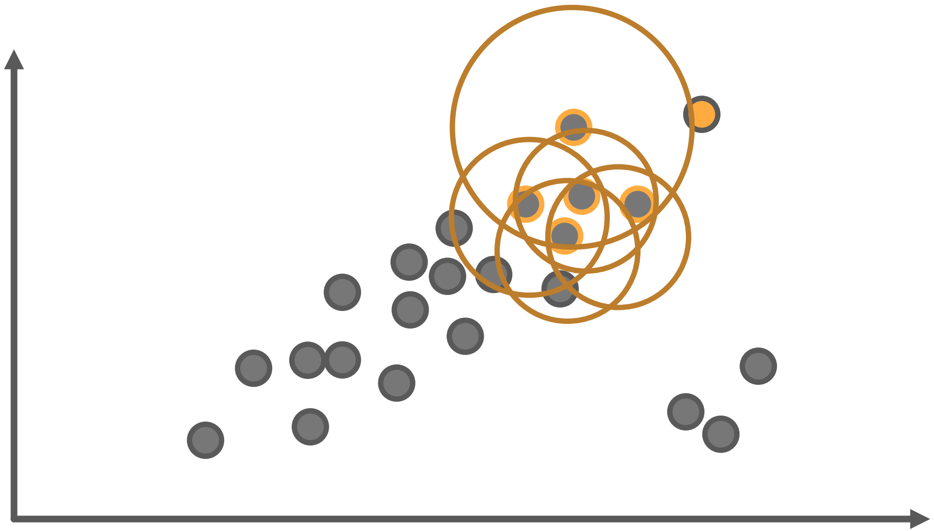
Local Outlier Factor (LOF)

- Density of observation of interest



Local Outlier Factor (LOF)

- Need to average the densities of the k-NN observations.

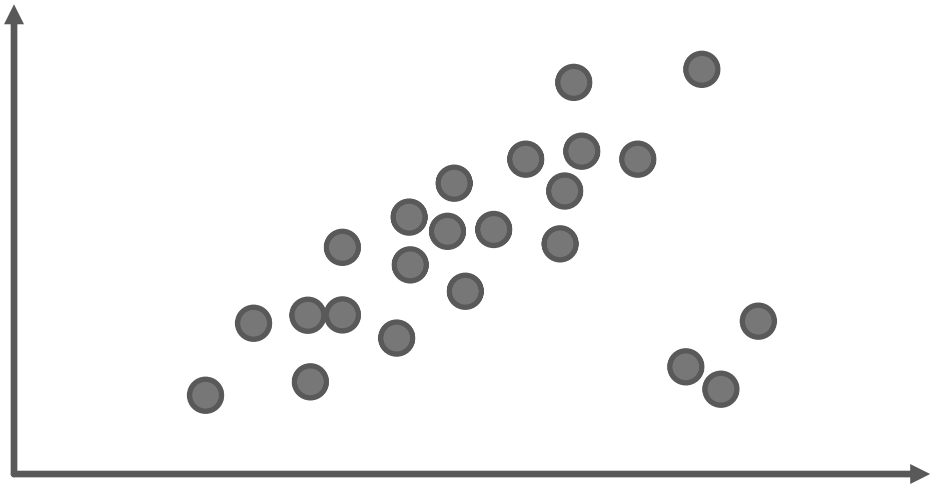


Coding in Action

Data Preparation – Anomaly Detection with Machine Learning Techniques: Local Outlier Factor

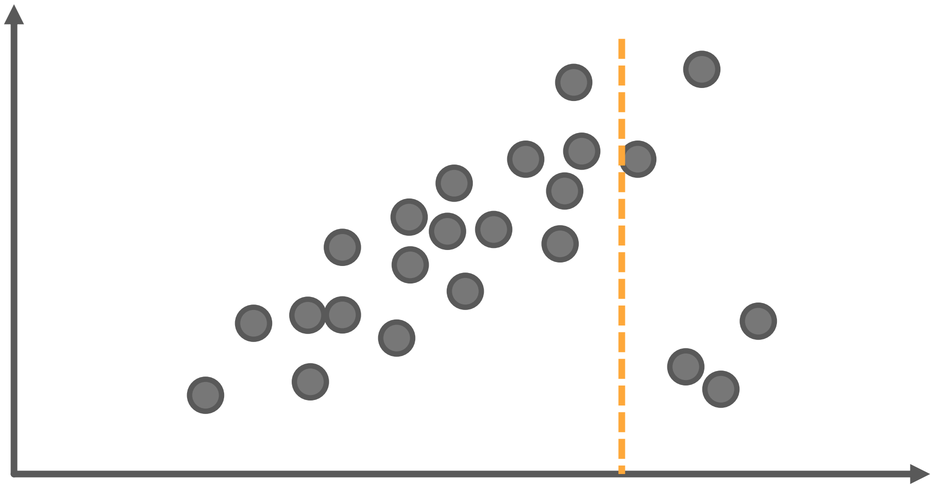
Isolation Tree

- Tree-based algorithm to isolate observations.
- Easier the isolation → More likely an anomaly!



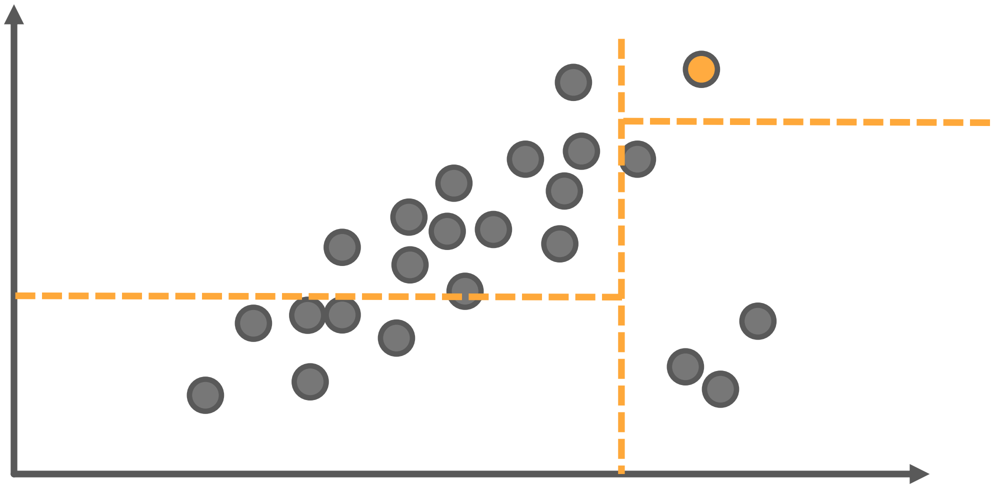
Isolation Tree

- Tree-based algorithm to isolate observations.
- Easier the isolation → More likely an anomaly!



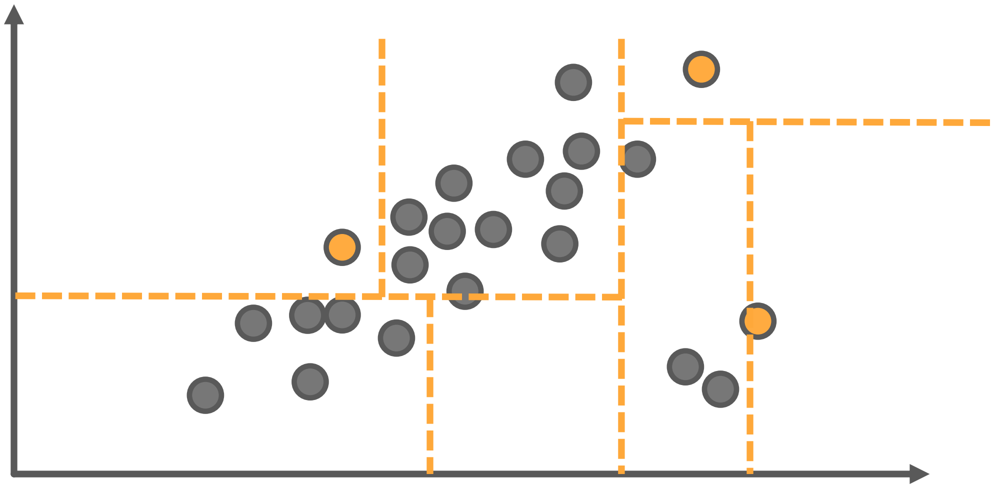
Isolation Tree

- Tree-based algorithm to isolate observations.
- Easier the isolation → More likely an anomaly!



Isolation Tree

- Tree-based algorithm to isolate observations.
- Easier the isolation → More likely an anomaly!



Isolation Tree

- Tree-based algorithm to isolate observations.
- Easier the isolation → More likely an anomaly!
- Isolation score is inversely related to number of needed splits to isolate observation.

  - Bounded between 0 and 1.

  - Closer to 1 → more likely an anomaly

  - Closer to 0 → less likely an anomaly

  - All observations ~ 0.5, no real anomalies

Isolation Forest

- Since the isolation trees are based on random splits on random dimensions, outlier might get lucky and survive longer than it really should.
- Isolation forest – combination of MANY isolation trees with averaged scores.
- Look for convergence of scores for optimal number of trees.

Coding in Action

Data Preparation – Anomaly Detection with Machine Learning Techniques: Isolation Forests

CADE

- Newer technique for density estimation.
- Value been found in anomaly detection and fraud applications.

CADE

- Process:

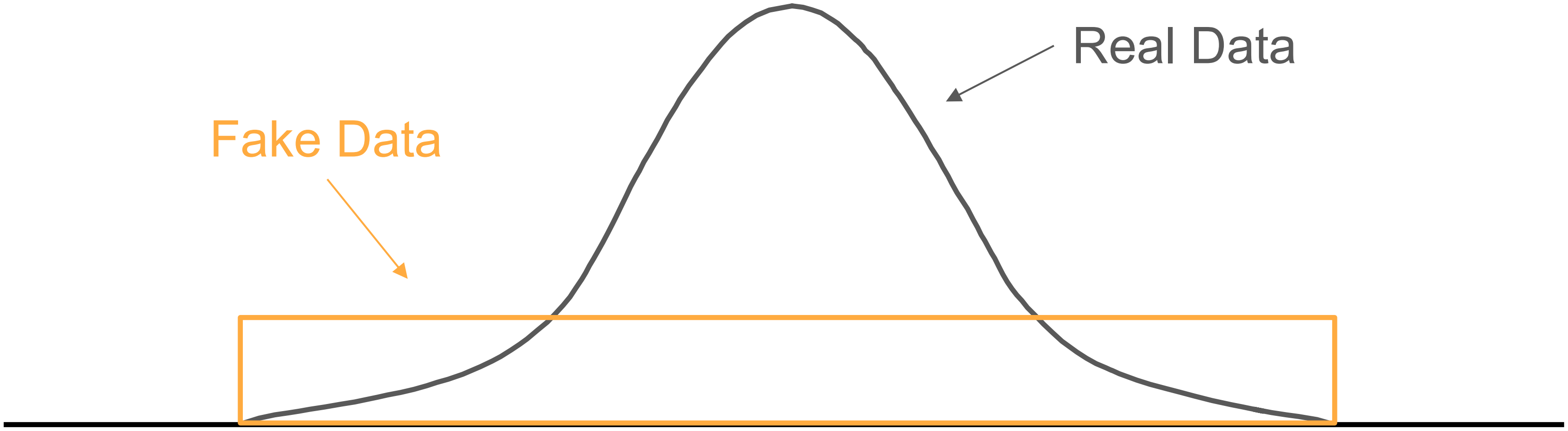
1. Label all original data as **not outliers**
2. Create new observations (same *n* as data) but variables are all uniformly distributed
3. Label all new data as **outliers**, merge old and new data
4. Use classification model to predict “outliers” (1’s).
5. Score original data

CADE

- High predicted probabilities → More likely an anomaly!
- Observation looks more like fake uniform data than actual distribution from which it came in multivariate space.

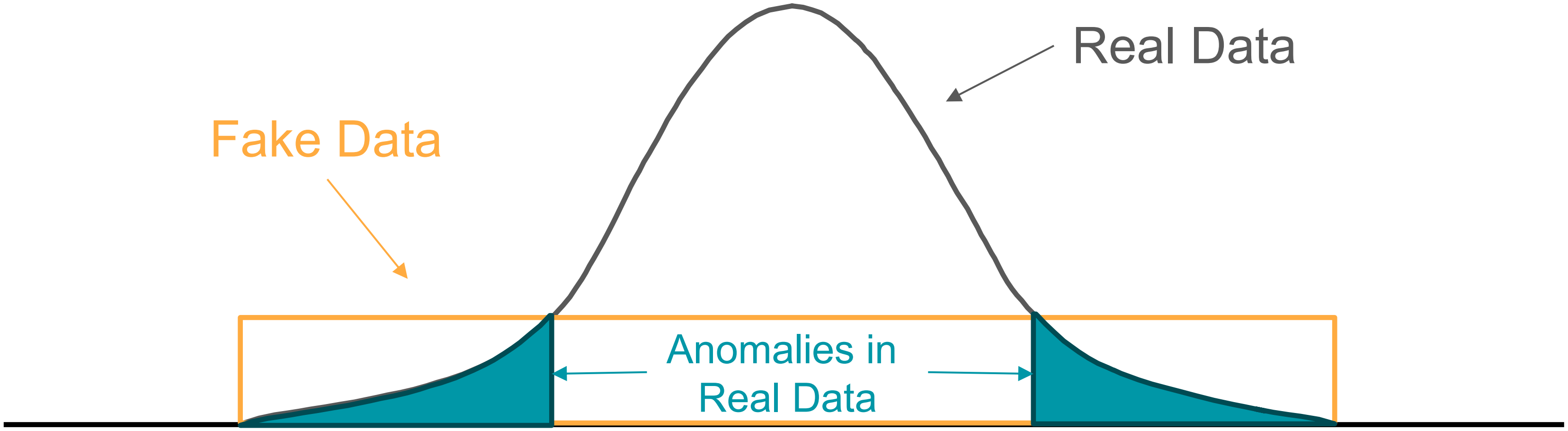
CADE

- High predicted probabilities → More likely an anomaly!
- Observation looks more like fake uniform data than actual distribution from which it came in multivariate space.



CADE

- High predicted probabilities → More likely an anomaly!
- Observation looks more like fake uniform data than actual distribution from which it came in multivariate space.

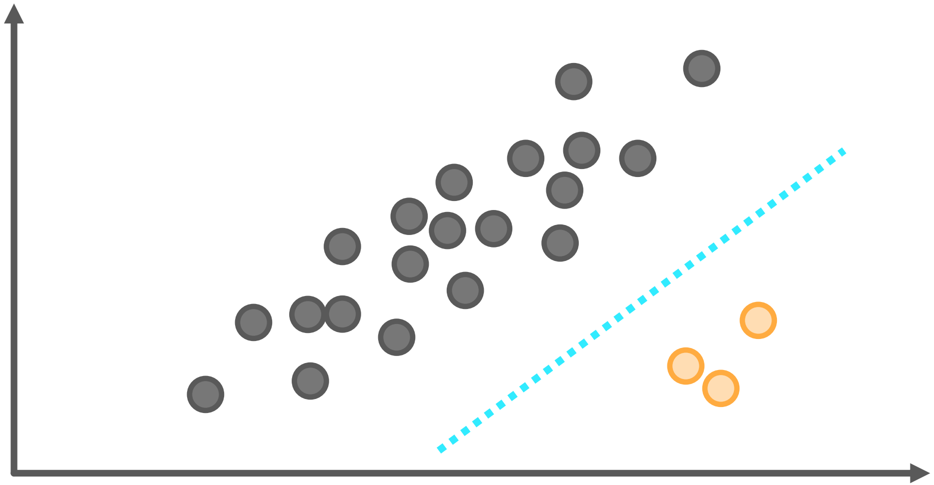


Coding in Action

Data Preparation – Anomaly Detection with Machine Learning Techniques: Classifier-Adjusted Density Estimation (CADE)

Support Vector Machines (SVM)

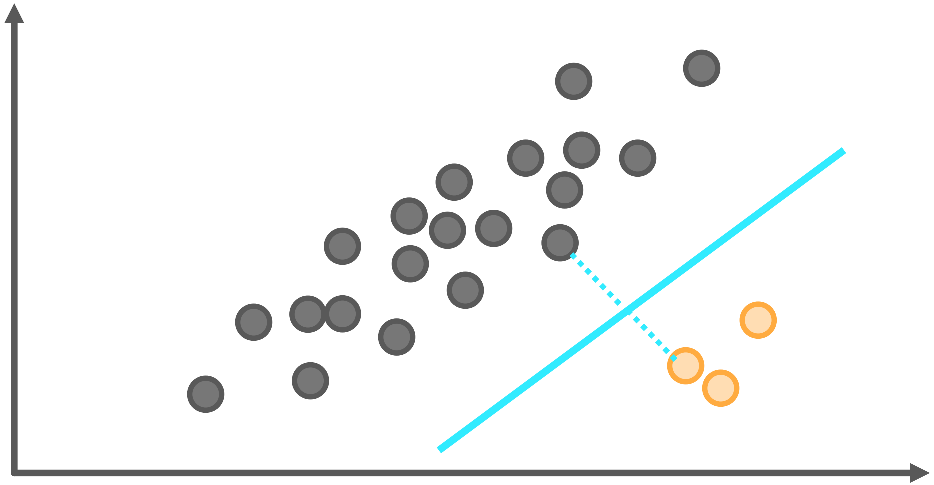
- A traditional two-class SVM is a classifier.
- It creates a hyperplane that “best” separates the two classes.



Ai+ TRAINING

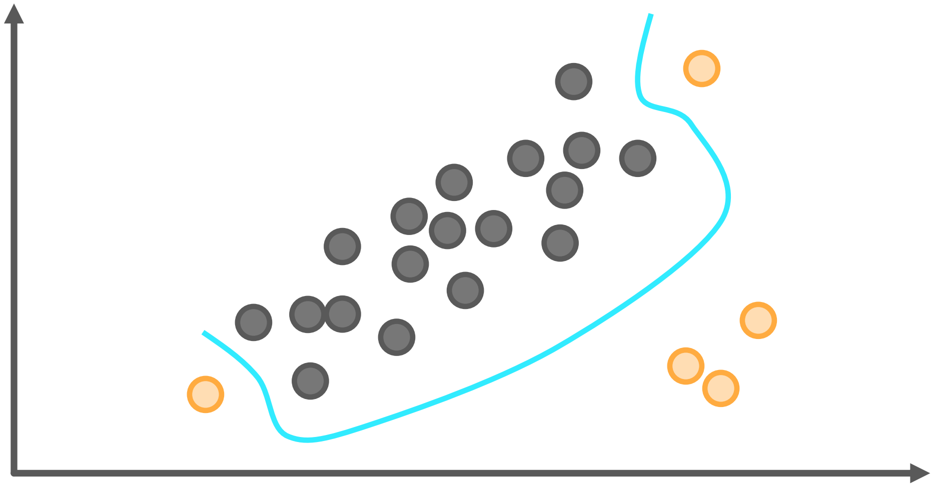
Support Vector Machines (SVM)

- A traditional two-class SVM is a classifier.
- It creates a hyperplane that “best” separates the two classes.
- “Best” is maximizing the distance (in every dimension) from the two classes.



Support Vector Machines (SVM)

- A traditional two-class SVM is a classifier.
- Not limited to linear separation! Kernels are used to make hyperplanes nonlinear.



One-Class Support Vector Machines

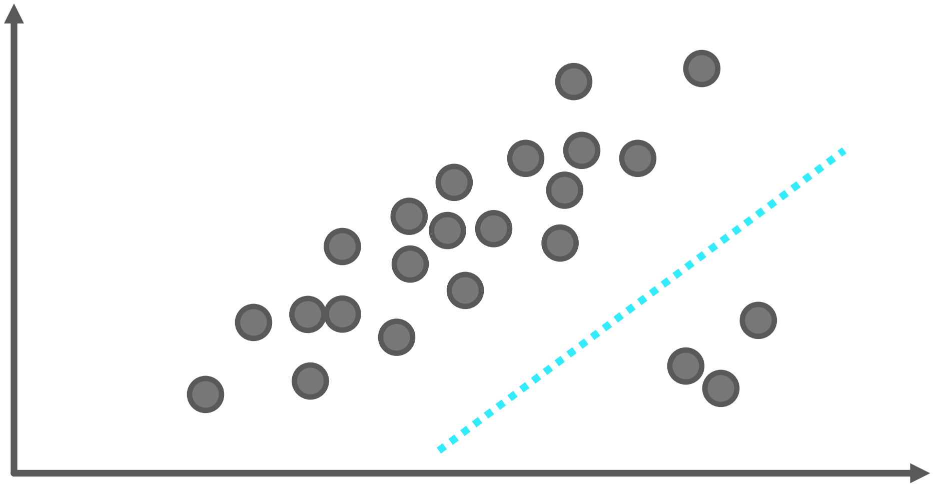
- SVM’s are also used in an unsupervised learning scenario as well.
- Instead of thinking about two classes, we can take one of two approaches:

1. Tell the SVM to isolate the X% of observations. If we think we have 5% anomalies, we tell the SVM to isolate the “most” anomalous 5% of observations.

2. Train the SVM on all data as normal and score new data to see if it falls “within normal”.

One-Class Support Vector Machines

- Tell the SVM to isolate the X% of observations. If we think we have 5% anomalies, we tell the SVM to isolate the “most” anomalous 5% of observations.



Coding in Action

Data Preparation – Anomaly Detection with Machine Learning Techniques: One-Class Support Vector Machines

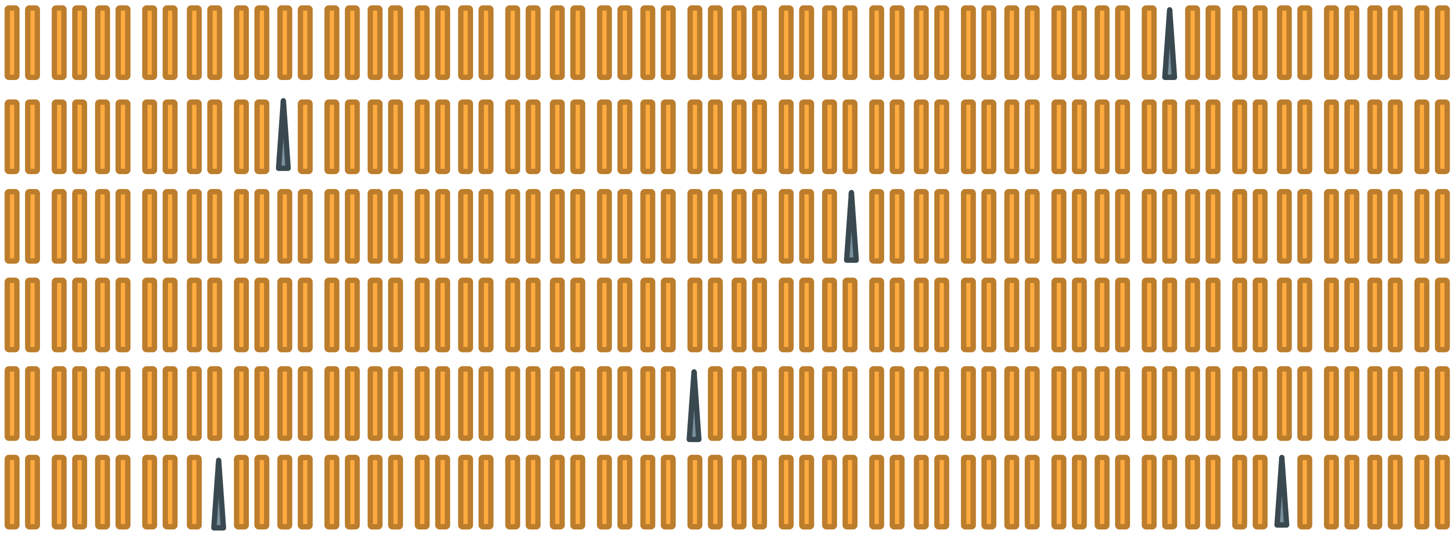
Data Preparation

Sampling Concerns

- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

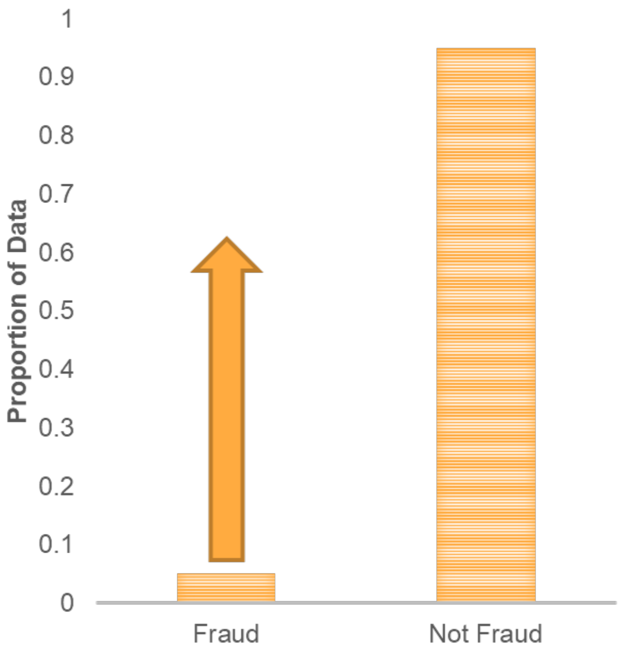
Rare Event Modeling

- Fraud modeling is difficult due to sampling concerns.
- 5% or smaller in a category can lead to classification problems.

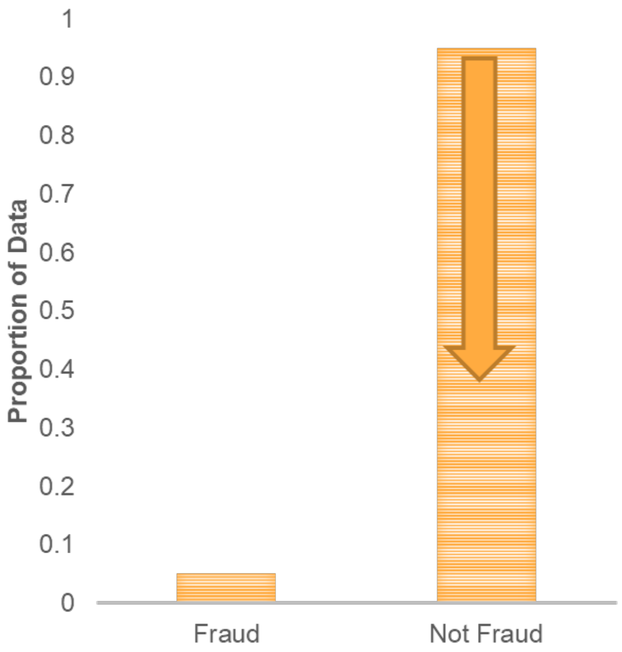


Rare Event Sampling Correction

Oversampling



Undersampling



Rare Event Sampling Correction

Oversampling

- Duplicate current fraud cases in training set to balance better with non-fraud cases.
- Keep test set as original population proportion.

Undersampling

- Randomly sample current non-fraud cases to keep in the training set to balance with fraud cases.
- Keep test set as original population proportion.

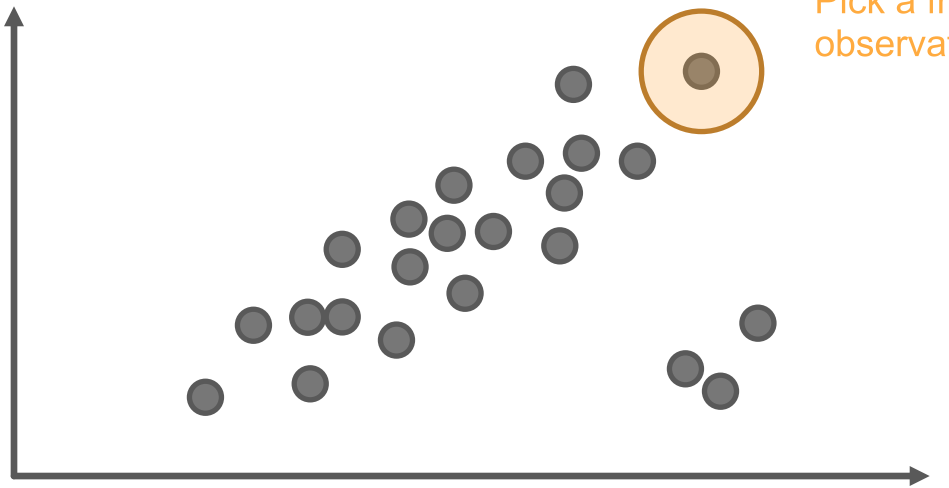
Coding in Action

Data Preparation – Sampling Concerns

Synthetic **M**inority **O**versampling **TE**chnique

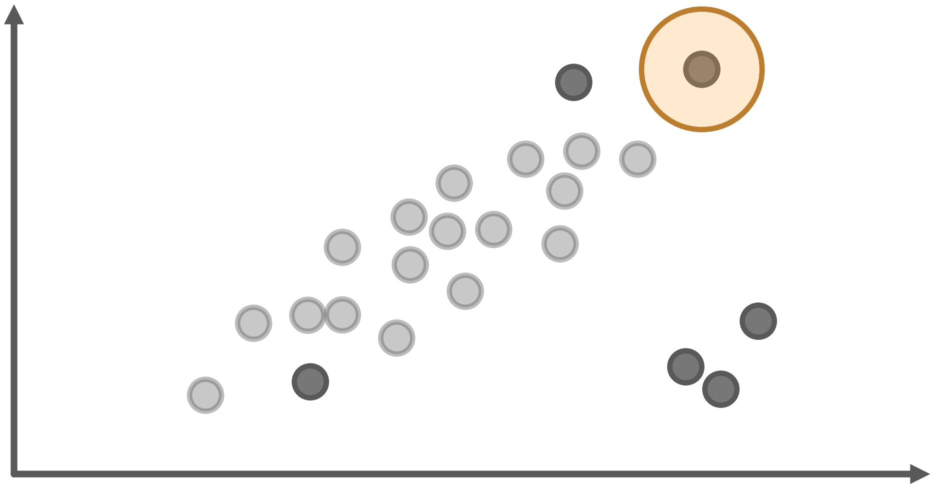
- SMOTE is a technique used to oversample rare data that creates synthetic observations that are close, but not exact replicates of your original data.
- SMOTE has shown great results in the fraud modeling space when adjusting for unbalanced samples.

SMOTE Process Example



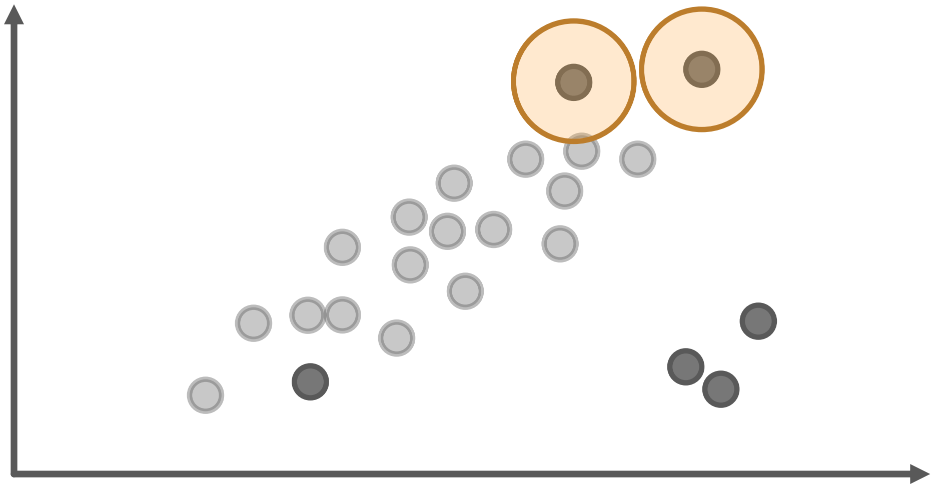
SMOTE Process

1. Isolate the other fraud cases



SMOTE Process

2. Randomly choose one of k-Nearest Neighbors.



SMOTE Process

3. Create the synthetic sample.

Data	Fraud Obs.	k-NN Fraud Obs.
X variable	8	6
Y variable	9	8.5

SMOTE Process

3. Create the synthetic sample.

Data	Fraud Obs.	k-NN Fraud Obs.
X variable	8	6
Y variable	9	8.5

Randomly select number between 0 and 1.

SMOTE Process

3. Create the synthetic sample.

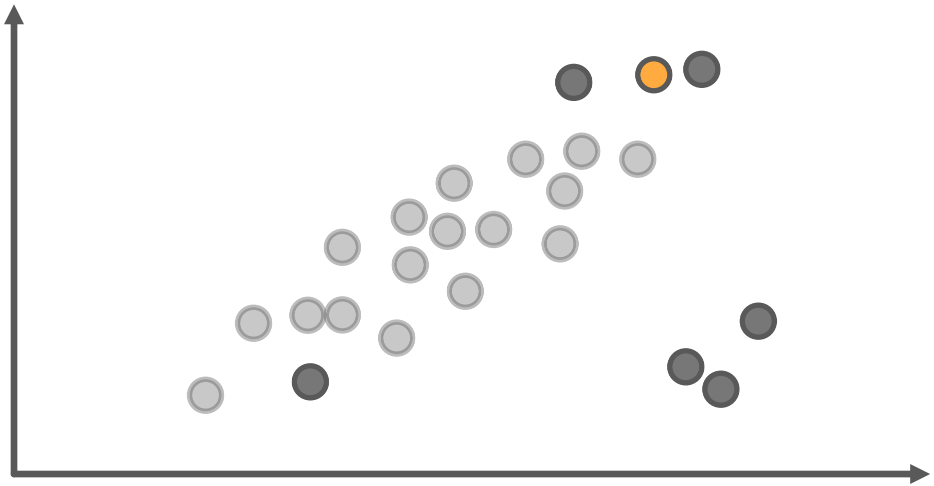
Data	Fraud Obs.	k-NN Fraud Obs.
X variable	8	6
Y variable	9	8.5

Randomly select number between 0 and 1: **0.3**

Data	Fraud Obs.	k-NN Fraud Obs.	Synthetic Obs.
X variable	8	6	$8 + (6 - 8) * 0.3 = 7.4$
Y variable	9	8.5	$9 + (8.5 - 9) * 0.3 = 8.85$

SMOTE Process

3. Create the synthetic sample.



SMOTE Process

4. Repeat for **every** fraud case a certain number of times to get balanced samples.

Coding in Action

Data Preparation – Sampling Concerns with SMOTE

Data Preparation

Conclusion

- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns

Conclusion

AI+ TRAINING

Course Outline

- Introduction
- Data Preparation
- Supervised Modeling
- Implementation / Deployment
- Conclusion

AI+ TRAINING

Course Outline – Part 1 & Part 2

- Introduction
  - Who am I?
  - What is Fraud?
  - Fraud Detection Analytical Framework
- Data Preparation
- Supervised Modeling
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
  - Feature Engineering
  - Fraud Data
  - Anomaly Detection with Statistical Techniques
  - Anomaly Detection with Machine Learning Techniques
  - Sampling Concerns
- Supervised Modeling
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
- Supervised Modeling
  - Interpretable Models
  - Naïve Bayes Models
  - More Advanced Models
  - Model Evaluation
  - **NOT**-fraud Model
- Implementation / Deployment
- Conclusion

Course Outline – Part 1 & Part 2

- Introduction
- Data Preparation
- Supervised Modeling
- Implementation Deployment
    - Clustering Revisited
    - Interpretability
    - Long-term Fraud Strategy
    - Chance & Loss Models
- Conclusion

Where Am I?

- Find me online:
    - https://www.linkedin.com/in/ariclabarr/
    - https://www.youtube.com/c/AricLaBarr/
    - https://www.ariclabarr.com/

Thank you

Ai+ TRAINING