FRAUD SUPERVISED MODELS

Dr. Aric LaBarr

Institute for Advanced Analytics

Course Layout

Data Preparation
- Transactional Data
- Recency vs. Frequency
- Network Features

Anomaly Models
- Univariate Analysis
- Clustering
- Isolation Forests
- CADE

Fraud Supervised Models
- SMOTE
- Models
- Labeled vs. Unlabeled Bias
- Not Fraud Model
- Evaluation

Clusters of Not Goods
- Cluster Analysis
- Social Network Analysis

Implement
- Investigators
- Traffic Light Indicators
- Backtesting

Fraud Maturity

Components	New / Young	Emerging SIU	Fraud Scoring	Holistic Solution
Simple Rules	Yes	Yes	Yes	Yes
Unlabeled Data	Yes / No	Yes / No	Yes	Yes
Labeled Fraud Cases	No	Yes	Yes	Yes
Anomaly Models	No	Yes / No	Yes	Yes
Supervised Models	No	No	Yes	Yes
Non-Fraud Models	No	No	No	Yes
Clusters of not Good	No	No	No	Yes

OBTAINING LABELED DATA

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 not in electronic form.
  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 not in electronic form.
  3. Previous data on fraudulent cases that is fully integrated into company databases and structure.
     - SKIP to section on sampling concerns!

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 not in electronic form.
     - Anomaly models identify potential observations
     - Clustering of observations
  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.
     - Anomaly models identify potential observations
     - Clustering of observations
  2. Previous data on fraudulent cases, but not in electronic form.
  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 not in electronic form.

- 2 Paths from here:
  1. Wait for SIU to investigate anomalies and slowly gather data over time.
  2. Bring in SME's to help with continuing modeling process.

Anomaly Models

- Fraudulent cases will typically appear as anomalies in multiple dimensions, but not necessarily all.
- Here are the steps to take once you have your suspected anomalies:
  1. Subject matter experts will look through the suspected anomalies for cases that appear to be fraudulent.

Anomaly Models

- Fraudulent cases will typically appear as anomalies in multiple dimensions, but not necessarily all.
- Here are the steps to take once you have your suspected anomalies:
  1. Subject matter experts will look through the suspected anomalies for cases that appear to be fraudulent.
  2. Tag suspected fraud groups based on expert domain knowledge.

Anomaly Models

- Fraudulent cases will typically appear as anomalies in multiple dimensions, but not necessarily all.
- Here are the steps to take once you have your suspected anomalies:
  1. Subject matter experts will look through the suspected anomalies for cases that appear to be fraudulent.
  2. Tag suspected fraud groups based on expert domain knowledge.
  3. Treat these suspected fraud groups as if they had committed fraud and other groups as if they have not.

Anomaly Models

- Fraudulent cases will typically appear as anomalies in multiple dimensions, but not necessarily all.
- Here are the steps to take once you have your suspected anomalies:
  1. Subject matter experts will look through the suspected anomalies for cases that appear to be fraudulent.
  2. Tag suspected fraud groups based on expert domain knowledge.
  3. Treat these suspected fraud groups as if they had committed fraud and other groups as if they have not.
  4. Ideally, have subject matter experts also identify small set of legitimate claims in non-anomaly data.

Unsupervised Learning

- Patterns should exist between fraudulent transactions.
- These patterns will typically be unseen by simply looking through the data.
- Unsupervised learning techniques can help identify fraudulent transactions.
  - K-means clustering
  - Self Organizing Maps (SOM)
  - Kohonen Vector Quantization (KVQ)

Unsupervised Learning

- Fraudulent cases will typically be isolated small clusters.
- Here are the steps to take once you have your suspected clusters:
  1. Subject matter experts will look through the suspected clusters for observations that appear to be fraudulent.

Unsupervised Learning

- Fraudulent cases will typically be isolated small clusters.
- Here are the steps to take once you have your suspected clusters:
  1. Subject matter experts will look through the suspected clusters for observations that appear to be fraudulent.
  2. Tag suspected fraud groups based on expert domain knowledge.

Unsupervised Learning

- Fraudulent cases will typically be isolated small clusters.
- Here are the steps to take once you have your suspected clusters:
  1. Subject matter experts will look through the suspected clusters for observations that appear to be fraudulent.
  2. Tag suspected fraud groups based on expert domain knowledge.
  3. Treat these suspected fraud groups as if they had committed fraud and other groups as if they have not.

Unsupervised Learning

- Fraudulent cases will typically be isolated small clusters.
- Here are the steps to take once you have your suspected clusters:

1. Subject matter experts will look through the suspected clusters for observations that appear to be fraudulent.
2. Tag suspected fraud groups based on expert domain knowledge.
3. Treat these suspected fraud groups as if they had committed fraud and other groups as if they have not.
4. Ideally, have subject matter experts also identify small set of legitimate claims inside the suspected clusters.

Clustering Techniques

- How many clusters to calculate?
  - Too few a clusters and you won’t have any small isolated situations.
  - Too many clusters and you won’t know which groups are the small isolated groups.
- Approximately 2-3% of claims are fraudulent.
  - Don’t want clusters that are too big.

Tagging Suspected Observations

- What are you modeling through these selection methods?

Tagging Suspected Observations

- What are you modeling through these selection methods?

NOT FRAUD!

Tagging Suspected Observations

- What are you modeling through these selection methods?

NOT FRAUD!

- This process means your model is predicting domain expert classifications instead of actual fraud.
- If domain experts are knowledgeable, then these classifications will be highly associated with fraudulent cases.

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.

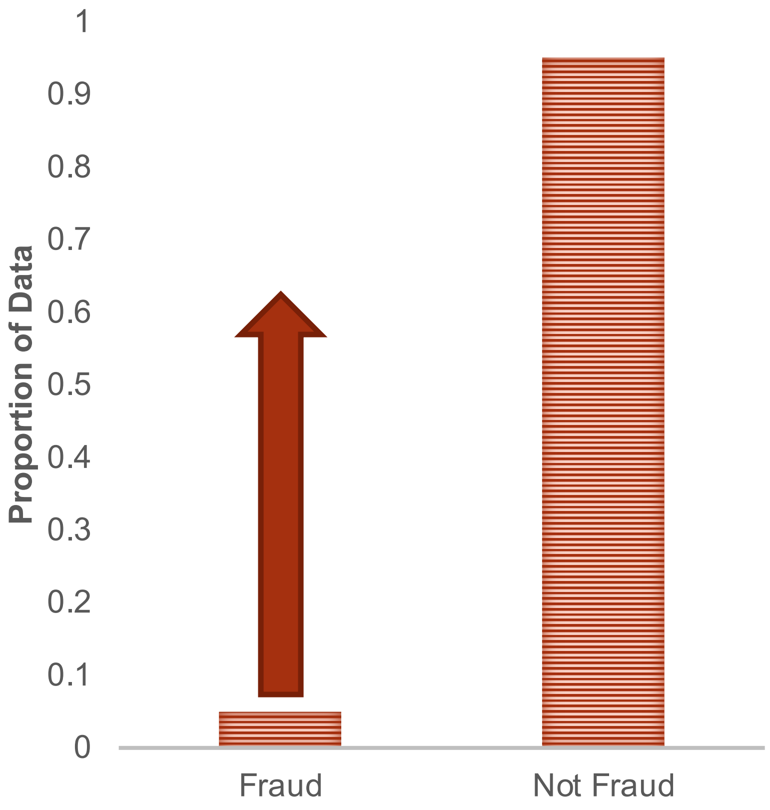
SAMPLING CONCERNS

Rare Event Modeling

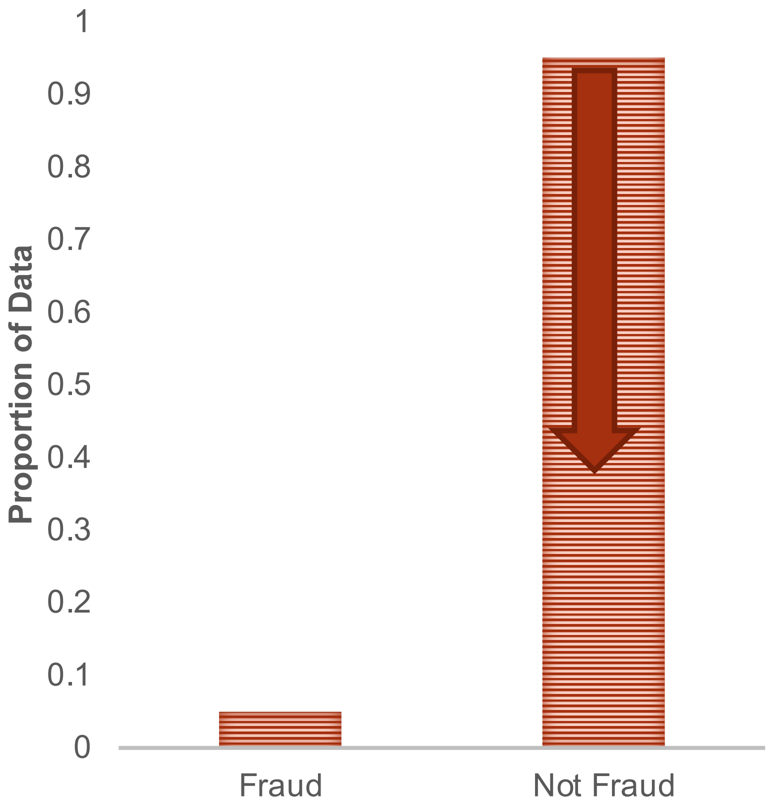
- Fraud modeling is difficult due to sampling concerns.

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.

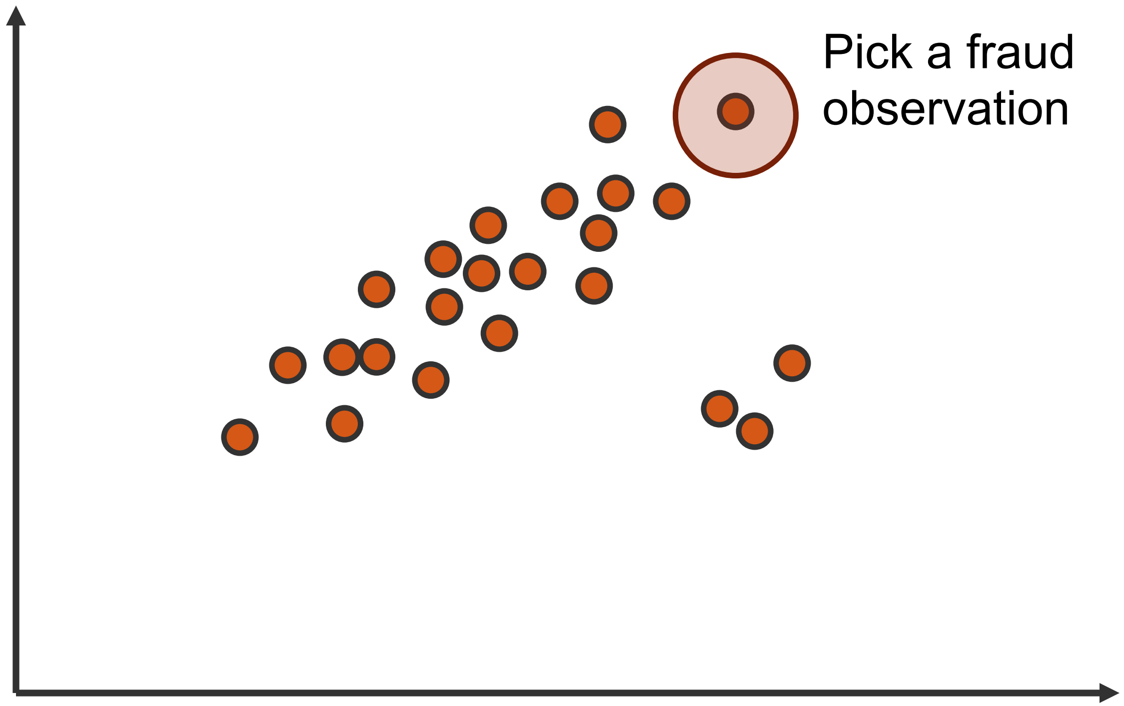
Undersampling

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

Synthetic **M**inority **O**versampling **TE**ch.

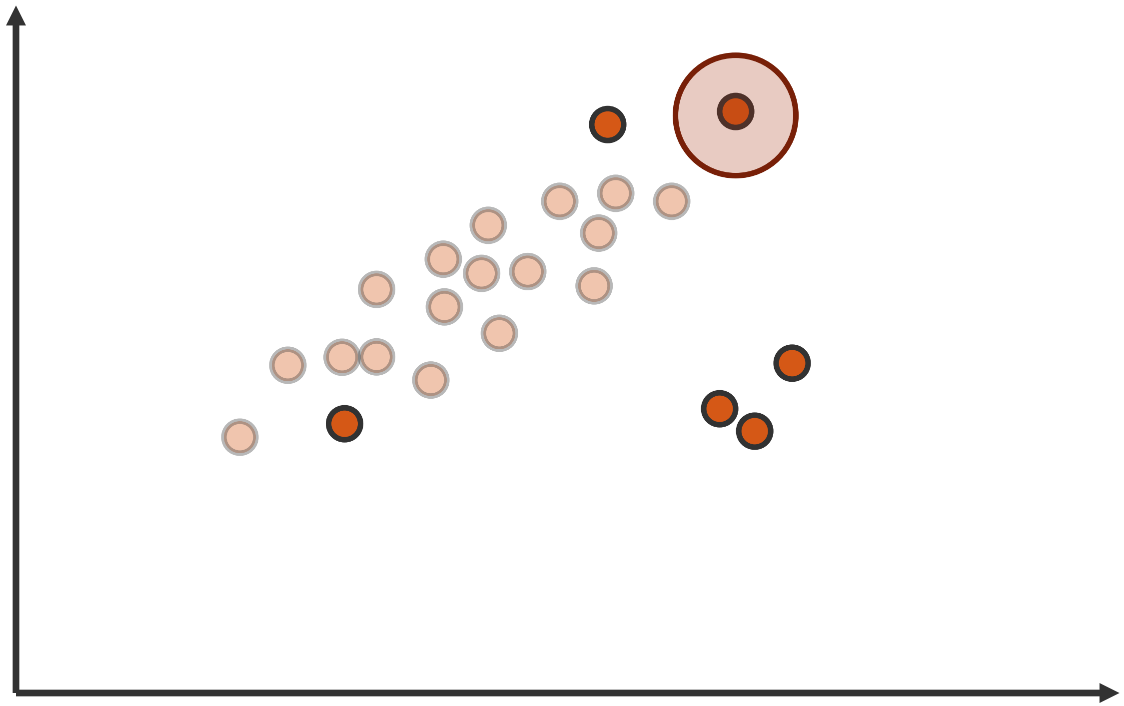
- SMOTE has shown great results in the fraud modeling space when adjusting for unbalanced samples.

SMOTE Process Example



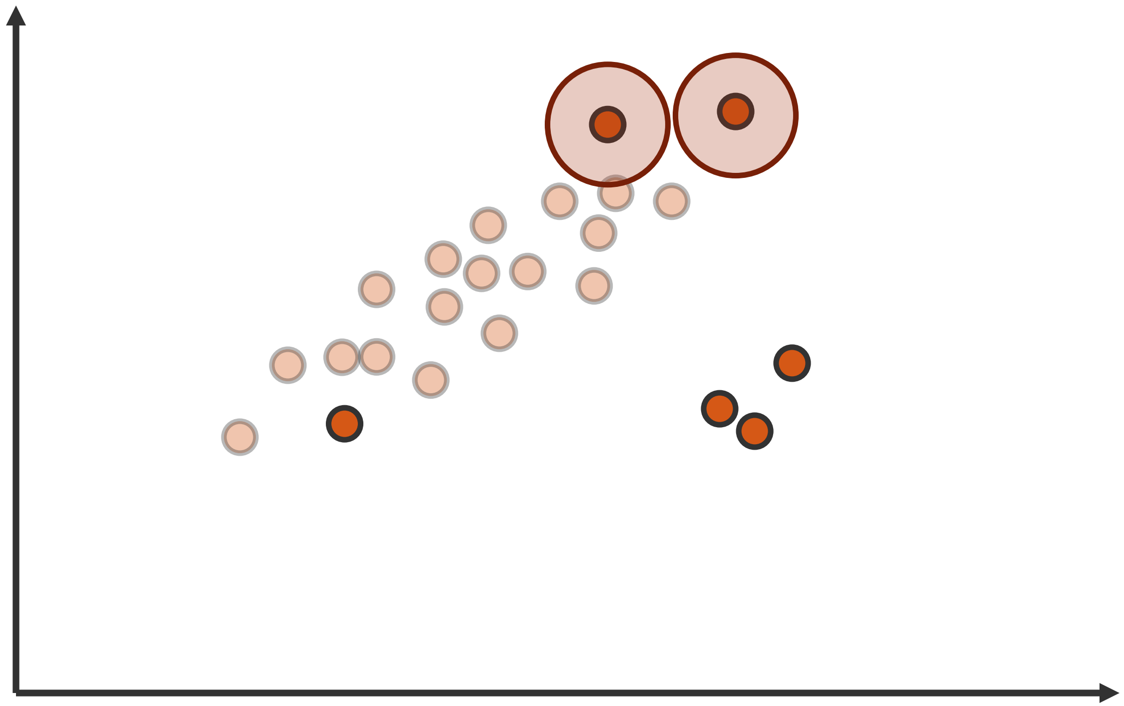
SMOTE Process

1. Isolate the other fraud cases.



SMOTE Process

2. Randomly choose one of k-Nearest Neighbors.



SMOTE Process

3. Create synthetic sample.

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

SMOTE Process

3. Create 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 synthetic sample.

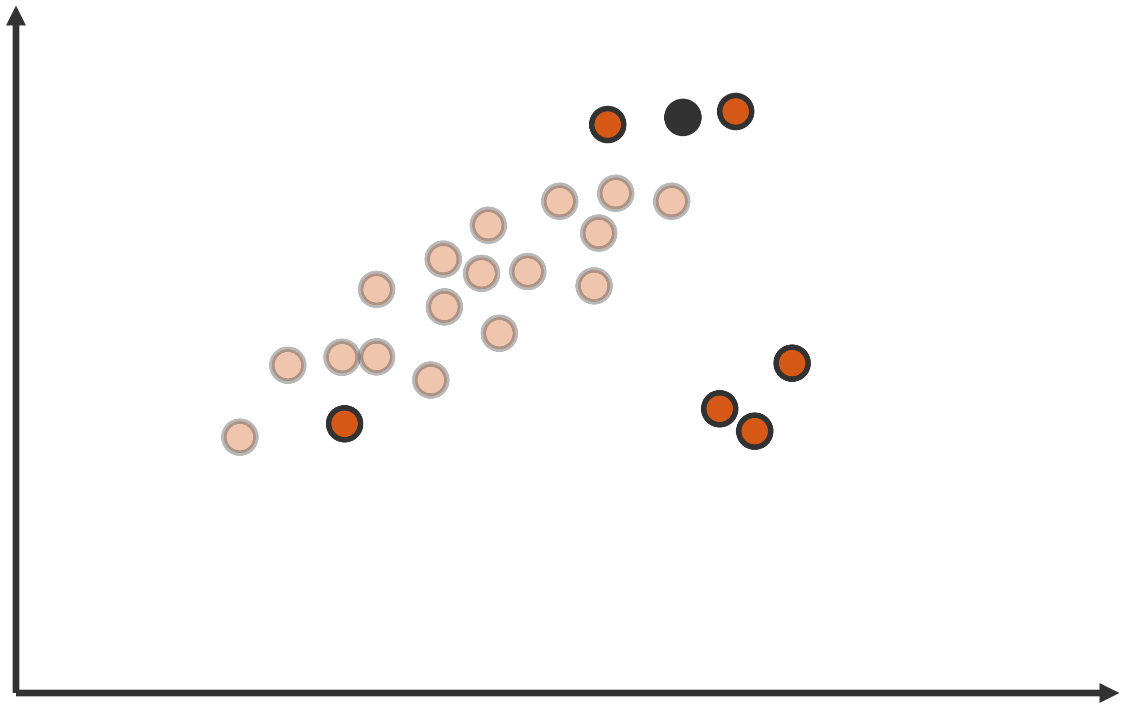
Data	Fraud Obs.	k-NN Fraud Obs.
X variable	8	6
Y variable	9	8.5

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 synthetic sample.



SMOTE Process

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

SMOTE – R

```
complete <- complete.cases(train)
num_names <- names(train)[sapply(train, is.numeric)]
inputs <- train[num_names]
inputs <- inputs[complete,]
target <- as.numeric(train[complete,120])

smote_sam <- SMOTE(X = inputs, target = target,
                   K = 5,
                   dup_size = 10)

train_s <- smote_sam$data
train_s$Fraud <- as.numeric(train_s$class) - 1
```

```
> table(train_s$Fraud)

    0     1 
12413  3707 
> prop.table((table(train_s$Fraud)))

        0         1 
0.7700372 0.2299628 
```

SUPERVISED FRAUD MODELS

Supervised Learning

- Supervised learning techniques are techniques where you know the values of the target value.
- The model will classify the individuals into one of two groups – suspected fraud or not.
- Models do this through scoring.

Scoring

- Models will produce a score for each individual between 0 and 1.
- A cut-off value is derived for the score where anything above the cut-off is suspected of fraud and anything below is not.
- Cut-off values are best determined through time and cost calculations.

Types of Models to Use

- There are many different supervised learning techniques.
  - Decision Trees
  - Logistic Regression
  - Neural Networks
  - Random Forests
  - Gradient Boosting
  - Etc.

Types of Models to Use

- There are many different supervised learning techniques.
  - Decision Trees
  - Logistic Regression
  - Neural Networks
  - Random Forests
  - Gradient Boosting
  - Etc.

Problem of repeating identified clusters.

Types of Models to Use

- There are many different supervised learning techniques.
  - Decision Trees
  - Logistic Regression
  - Neural Networks
  - Random Forests
  - Gradient Boosting
  - Etc.

Problems with certain interactions
causing quasi-complete separation.

Types of Models to Use

- There are many different supervised learning techniques.
  - Decision Trees
  - Logistic Regression
  - Neural Networks
  - Random Forests
  - Gradient Boosting
  - Etc.

Problems with certain interactions causing quasi-complete separation.

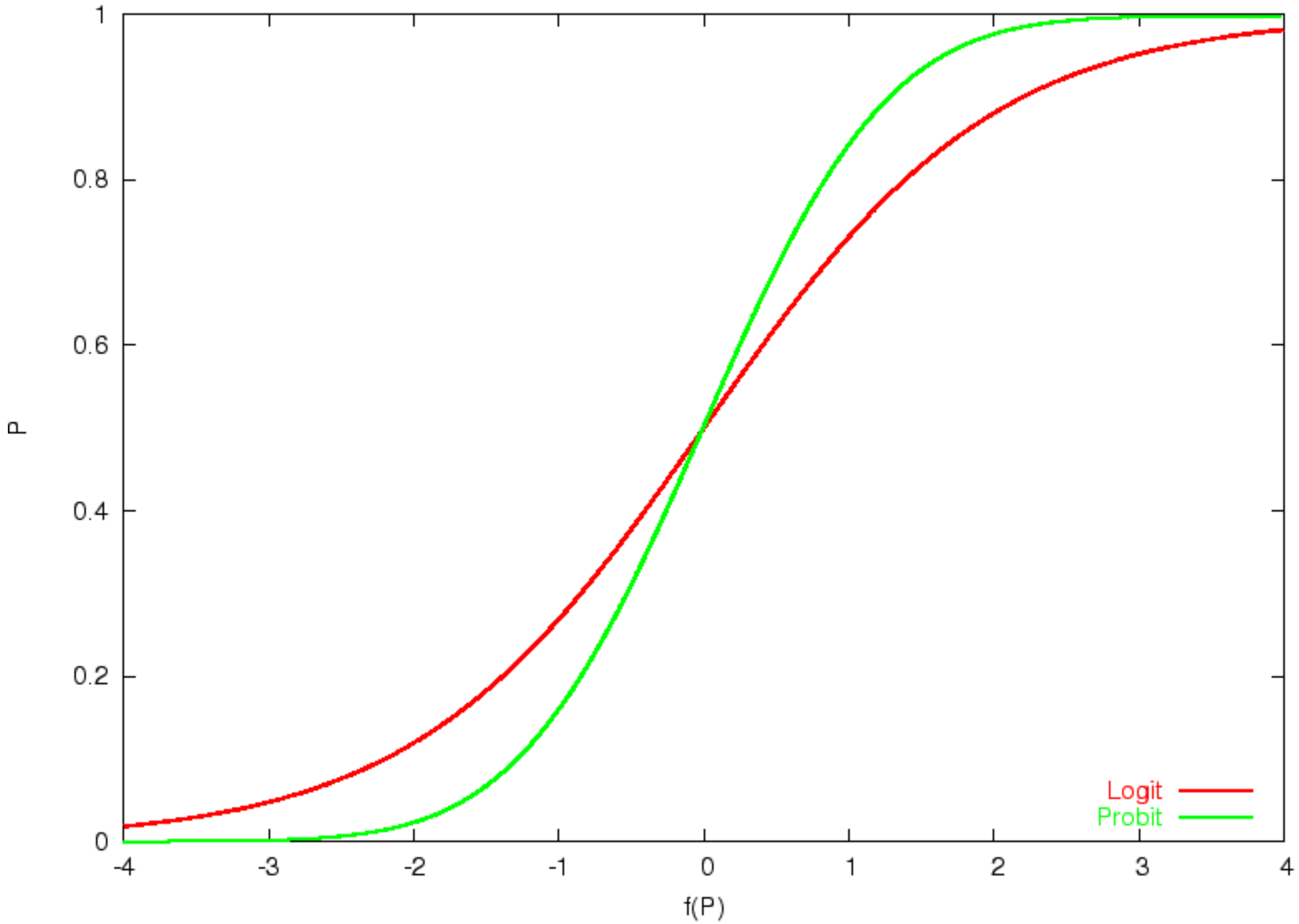
↓

Try to find main effects and then build interactions and nonlinearities off of those only.

Logistic / Probit Regression

- Both logistic and probit regressions are predicting the probability of an event occurring.
- They are based on different underlying distributions for the probability curve.

Logistic Regression



Logistic / Probit Regression

- Here is the equation for the logistic regression curve:

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}$$

- Here is the equation for the probit regression curve:

$$p = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$= \int_{-\infty}^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{t^2}{2}\right)} dt$$

Logistic / Probit Regression

- The assumptions for each of these models is essentially the same:
  - The transformation is the correct one.
- In other words, the transformation results in a linear relationship with the input variables.

Types of Models to Use

- There are many different supervised learning techniques.
  - Decision Trees
  - Logistic Regression / Probit Regression
  - Neural Networks
  - Random Forests
  - Gradient Boosting
  - Etc.

Problems with interpretability and
use by investigators.
Needs interpretable layer on top!

SUPERVISED **NOT**-FRAUD MODELS

The Fraud Solution

- Regardless of the industry, two things are important for any fraud detection solution:
  1. **DETECTION** – Observing **known** fraudulent observations to determine patterns that may assist in finding other fraudulent observations.

The Fraud Solution

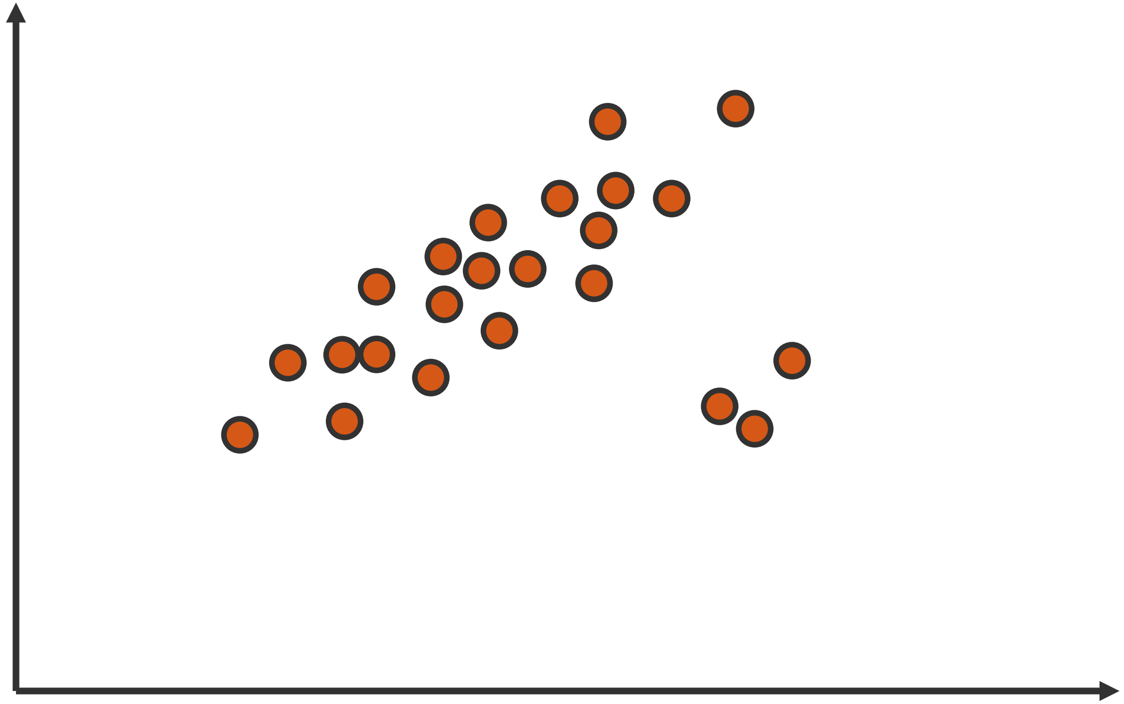
- Regardless of the industry, two things are important for any fraud detection solution:
  1. **DETECTION** – Observing **known** fraudulent observations to determine patterns that may assist in finding other fraudulent observations.
  2. **PREVENTION** – Observing behavior and identifying suspicious actions that might be fraudulent – lead to further investigation and identification of **new** fraudulent observations.

NOT-Fraud Supervised Model

- Predicting previous known cases of fraud works for fraud detection.
- Predicting previous known cases of **not**-fraud works for prevention of new fraud.

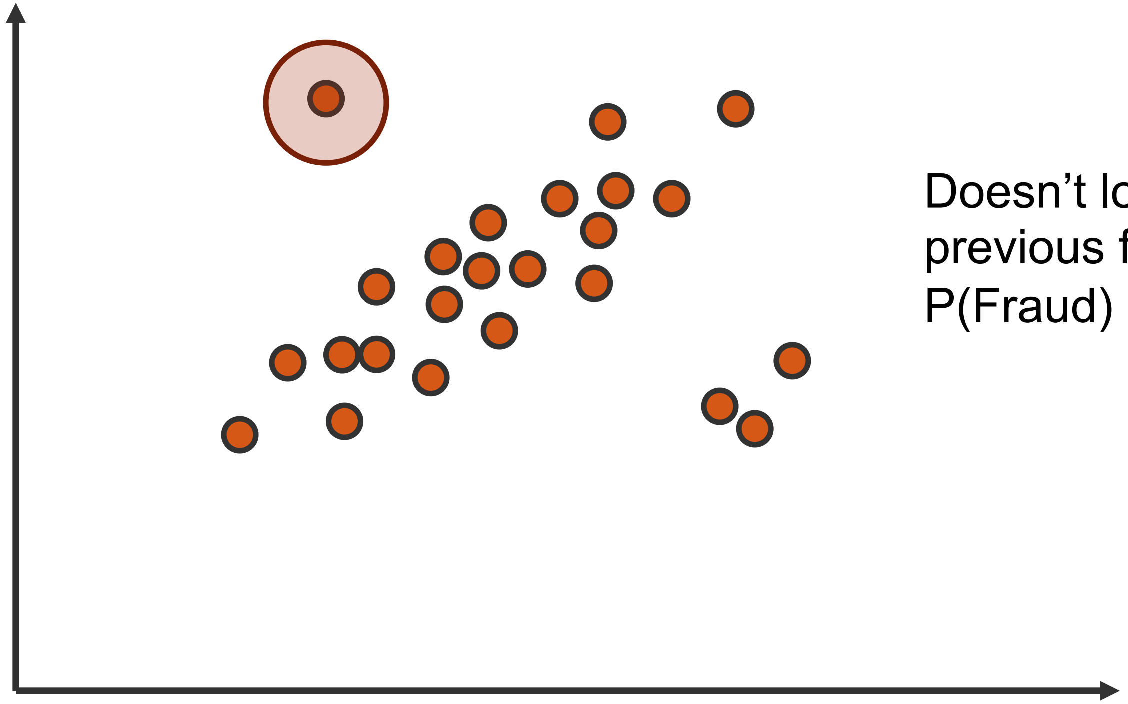
NOT-Fraud Supervised Model

- Predicting previous known cases of fraud works for fraud detection.
- Predicting previous known cases of **not**-fraud works for prevention of new fraud.



NOT-Fraud Supervised Model

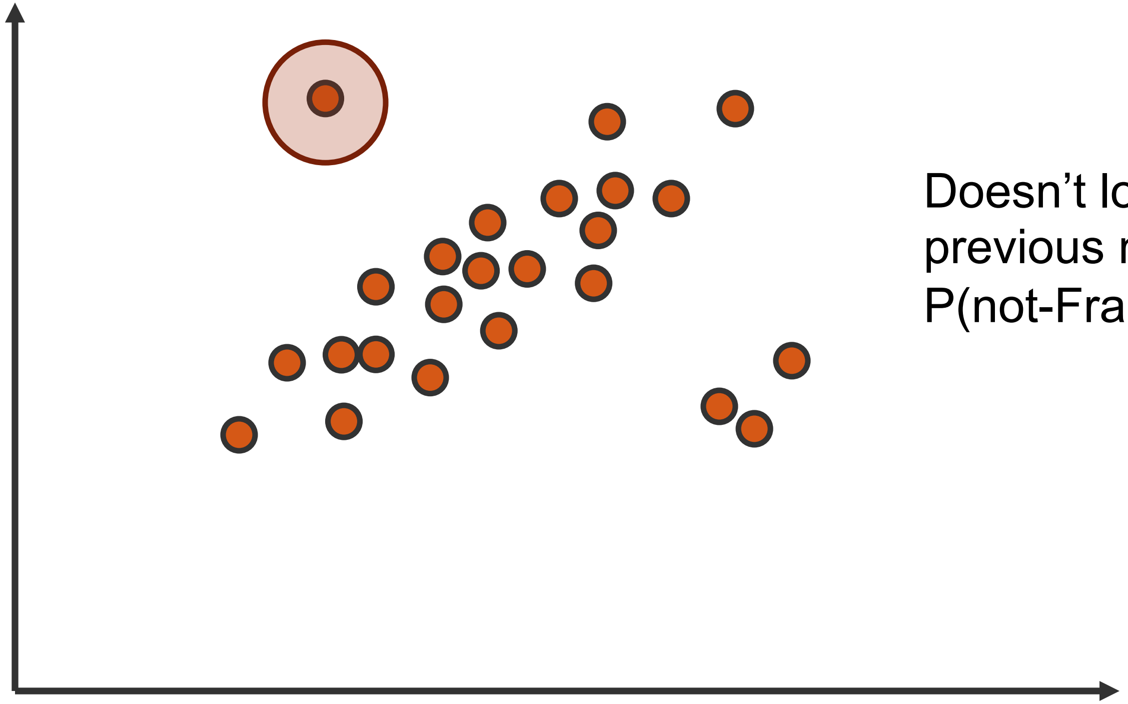
- Predicting previous known cases of fraud works for fraud detection.
- Predicting previous known cases of **not**-fraud works for prevention of new fraud.



Doesn’t look like
previous fraud.
P(Fraud) = **LOW**

NOT-Fraud Supervised Model

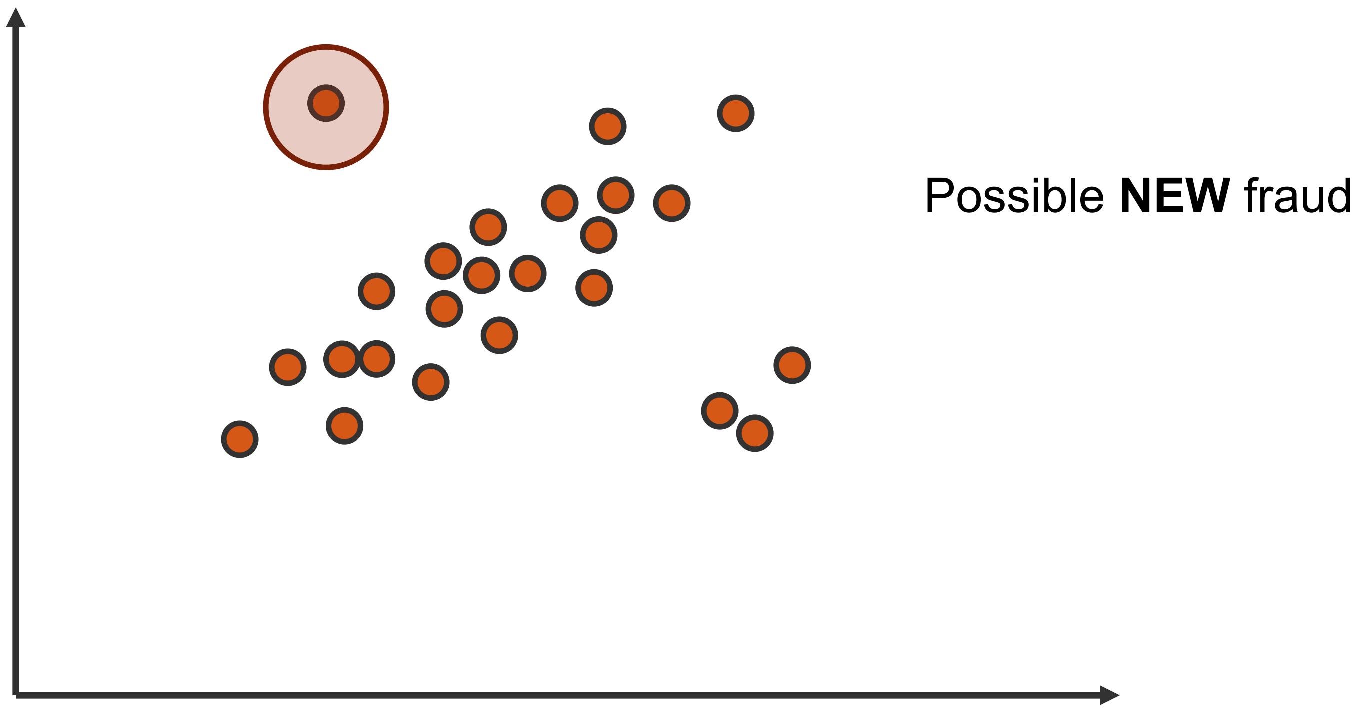
- Predicting previous known cases of fraud works for fraud detection.
- Predicting previous known cases of **not**-fraud works for prevention of new fraud.



Doesn't look like previous not-fraud.
P(not-Fraud) = **LOW**

NOT-Fraud Supervised Model

- Predicting previous known cases of fraud works for fraud detection.
- Predicting previous known cases of **not**-fraud works for prevention of new fraud.



MODEL EVALUATION

Balancing Unbalanced Costs

- Even the best fraud models catch about 25-35% of fraud initially.
- Models should be evaluated more on costs/savings than accuracy in fraud models.
  - May be **very** accurate due to correctly identifying non-fraud.

Balancing Unbalanced Costs

	True Non-Fraud	True Fraud
Predicted Non-Fraud	No Cost	Cost = Amount Paid
Predicted Fraud	Cost = Investigation	Cost = Investigation