Predicting Real-Time Transaction Fraud

Sami Niemi, PhD
Barclays, Quantitative Analytics, Fraud Detection
Contents

1. Background
2. Raw Data
3. Data Processing
4. Development
5. Validation
6. Implementation
7. Summary
3rd Party Fraud

an individual, or group of people, create or use a third-party's identity in order to apply for products or take over an account without the consent or knowledge of the third-party.

Card Transaction Fraud

Card Present (CP)
- e.g. lost, stolen, counterfeit/clone

Card not Present (CnP)
- e.g. identity theft, hacking, fake online shops
Background – Motivation (global view)

Figure 8: Fraud Incidence Rate and Dollar Amount of Losses, 2011-2017

Source: Javelin Strategy & Research, 2018

Consumers Who Had a Card Misused for POS and CNP Transactions

Source: Javelin Strategy and Research, 2018

#StrataData - Predicting real-time transaction fraud using supervised learning
# Background – Motivation (UK view)

<table>
<thead>
<tr>
<th>CARD FRAUD TYPE ON UK-ISSUED CREDIT AND DEBIT CARDS</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote Purchase (CNP)</td>
<td>752,450</td>
<td>951,998</td>
<td>1,019,146</td>
<td>1,113,084</td>
<td>1,437,832</td>
<td>1,399,031</td>
</tr>
<tr>
<td>Counterfeit (skimmed/clone)</td>
<td>98,555</td>
<td>101,109</td>
<td>99,279</td>
<td>86,021</td>
<td>108,597</td>
<td>84,861</td>
</tr>
<tr>
<td>Fraud on lost or stolen cards</td>
<td>113,162</td>
<td>138,967</td>
<td>133,943</td>
<td>143,802</td>
<td>231,164</td>
<td>350,066</td>
</tr>
<tr>
<td>Card ID theft</td>
<td>24,287</td>
<td>30,718</td>
<td>26,542</td>
<td>33,566</td>
<td>31,756</td>
<td>29,139</td>
</tr>
<tr>
<td>Card non-receipt</td>
<td>9,053</td>
<td>9,125</td>
<td>9,302</td>
<td>10,719</td>
<td>11,377</td>
<td>10,905</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>997,507</strong></td>
<td><strong>1,231,917</strong></td>
<td><strong>1,288,212</strong></td>
<td><strong>1,387,192</strong></td>
<td><strong>1,820,726</strong></td>
<td><strong>1,874,002</strong></td>
</tr>
</tbody>
</table>

Source: Fraud the Facts 2018 by UK Finance
Background – Challenges

Fraudsters Adapt and Invent New MOs

Real-Time Runtime Requirements

Front Page News Material
Aim of the project was to develop and implement new Debit CP and CnP real-time fraud detection models, which can reduce fraud losses and protect genuine customers.
Raw Data – Sources

Confirmed Frauds

Payment Instrument Info

Non-Mon Events

Other Cards and Accounts

Customer Info

Account Info

Debit Card Transactions
Data Processing – Quality Assurance and Data Exploration

Data Quality
- Reconciliation, Volumes, and Amounts
- Daily and Monthly Summary Statistics
- Anomaly and Outlier Detection

Exploration
- Trend Analysis and Anomaly Detection
- Distributions (PDFs and bar charts for Fraud / Non)
- Correlations (covariance, correlation w/ target, etc.)

Report
- Thresholding
- Issue Generation and Resolution
- Documentation and Governance
Data – High Level Statistics

Volumes
- Total: 220 – 300M debit card transactions with total value of £9 – 11B per month
  - CP: 110M contactless, 20M ATM
  - CnP: 85M e-commerce + telephony

Customers
- 10M unique customers per month
  - transacting in 220 countries
  - using 12M debit cards
  - with 1.9M different merchants

Frauds
- Fraud Rates
  - CP: less than 0.01%, depending on segment
  - CnP: less than 0.15%, depending on segment
Development – Datasets

Debit
- 14 months

CP
- Train: 12 months
- OOT: 2 recent months
- Sample: 45M transactions

CnP
- Train: 12 months
- OOT: 2 recent months
- Sample: 55M transactions

#StrataData - Predicting real-time transaction fraud using supervised learning
and many more (e.g. merchant)... finally, ratios between values and current transaction.
Development – Feature Selection

**Univariate**
- Remove zero or extremely low variance
- Remove if all or extremely high level of missing values
- Very low Information Value or Spearman rank correction

**Model**
- Lasso co-efficient importance
- Random Forest feature importance

**Wrapper**
- Recursive Feature Elimination

**Domain**
- Business Review
- Implementation Considerations

Counts:
- 20k
- 10k
- 1k
- 500
Development – Feature Selection & Business Review

Debit CP model feature:
ratio of current transaction amount and maximum contactless in last X days

![Box plot comparing genuine and fraud transactions](image-url)
Development – Model Development Cycle

1. Select Features
2. Pick Model and Train
3. Evaluate
4. Optimize
5. Review

StrataData - Predicting real-time transaction fraud using supervised learning
Development – Hyper-parameter Optimization

Example of Bayesian hyper-parameter optimization using hyper-opt
Validation – CP Model Performance

Precision Recall Curve: AUC \approx 0.23

ROC Curve: AUC \approx 0.95
Validation – CP Model Performance

Transaction Detection Rate

Value Detection Rate

False Positive Rate [bps]

New model

Incumbent model

False Positive Rate [bps]

New model

Incumbent model

Transaction Detection Rate

Value Detection Rate

False Positive Rate [bps]

New model

Incumbent model
Validation – CnP Model Performance

**Transaction Detection Rate**

- New model
- Incumbent model

**Value Detection Rate**

- New model
- Incumbent model
Validation – CP Model Interrogation

- Time since a new card was issued

- Fraud Risk

- Chip used
- Chip not-used
Implementation – Development Artefacts

Model Artefacts
- Model Specification (JSON)
- Model File (txt)
- Validation Data (parquet)

Feature Artefacts
- Feature Specification (JSON)
- Validation Data (parquet)
Artefacts from Nexus using Jenkins

Model File and Implementation Validation

Feature Code Gen and Validation

Feature Maturation and Shadow Operations

Production Validation and Go-Live
Summary

• Increasing number of customers become victims of fraud, especially remote purchase (e.g. e-commerce).

• To improve fraud prevention and customer experience, we undertook
  – Development of generation 1 models for Debit CP and CnP using tree ensemble algorithms
  – 12 months of training data were converted to about 20k features to develop the best possible models
  – Both models are in implementation, shadow operations due in May with go-live during summer

• R&D for generation 2 models (e.g. RNNs, autoencoders) on-going, promising results, but implementation requires more work…
Rate today’s session

Cyberconflict: A new era of war, sabotage, and fear

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our interconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the New York Times as well as a national security and political contributor for CNN and a frequent guest on CBS This Morning, Face the Nation, and many PBS shows.

O’Reilly Events App

Session page on conference website