A Real Time Fraud Detection System Using Gradient Boosted Trees

Credit Scoring and Credit Control XVI
Edinburgh, August 2019
The UK Fraud Landscape

Fraud continues to increase: it is more important than ever to protect our customers

**UK Card Fraud Loss £s**

<table>
<thead>
<tr>
<th>Year</th>
<th>Card Present</th>
<th>Card Not Present</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>100m</td>
<td>5m</td>
<td>105m</td>
</tr>
<tr>
<td>2013</td>
<td>110m</td>
<td>5m</td>
<td>115m</td>
</tr>
<tr>
<td>2014</td>
<td>120m</td>
<td>5m</td>
<td>125m</td>
</tr>
<tr>
<td>2015</td>
<td>130m</td>
<td>5m</td>
<td>135m</td>
</tr>
<tr>
<td>2016</td>
<td>140m</td>
<td>5m</td>
<td>145m</td>
</tr>
<tr>
<td>2017</td>
<td>150m</td>
<td>5m</td>
<td>155m</td>
</tr>
<tr>
<td>2018</td>
<td>160m</td>
<td>5m</td>
<td>165m</td>
</tr>
</tbody>
</table>

19% increase from 2017 to 2018

**Monthly Stats**

- **220 – 300m** debit card transactions
- **£9b – 11b** value
- **CP** : 110m contactless, 20m ATM
- **CNP** : 5m e-commerce + telephony

- **10m unique customers** per month
- Transacting in 220 countries
- 12m debit cards
- 1.9m different merchants

**Fraud Rates**

- Card Present : < 0.01%
- Card Not Present : < 0.15%

*Source: Fraud the Facts 2019 by UK Finance*
Barclays Fraud Machine Learning Programme

Multiyear strategic investment in data, platforms and models to reduce the total cost of fraud

Step change in fraud detection capability

- Reduce fraud losses by >10%
- Reduce customer inconvenience
- Reduce operation costs
- Reduce fraud strategy and model suite complexity

Leverage latest technology

- Open source software
- Distributed internal and cloud infrastructure
- Non-linear ML techniques
- New data sources
The Machine Learning Lifecycle

To reach beyond ML proof of concepts we have built an end-to-end ML operating model

- **Data Science Platform**
  - **Monitoring**: Monitoring model performance on agreed metrics, retraining triggered as needed
  - **Data Curation**: Curation and maintenance of all data pipelines required for model development and management
  - **Modelling**: Development of model per Model Owner requirements

- **Model Repository**

- **Model Owner**
  - Accountability for model usage and governance

- **Validation**
  - Independent validation per Model Risk Policy

- **Legend**
  - Technology
  - Modelling
  - Business
  - Risk

- **Authoritative Data Source**

- **Dev Data**

- **Model Owner**

- **Live Data**

- **Implementation**
  - Continuous integration, testing and deployment of models into the production platform using DevOps processes

- **Production Platform**

- **Production Output Data**
Gradient Boosted Tree models are a good choice for 1st generation non-linear models

~35 fraud models used currently, many from vendors

UK Transactions
- Debit Cards (CP / CNP Segments)
- Credit Cards (CP / CNP Segments)

US Applications
- 1st Party
- 3rd Party

Gradient Boosting Trees using LightGBM
- Outperforms Random Forests and other GBT implementations such as XGBoost
- Reasonable data and compute requirements to train (non-frauds down sampled to 3% to reduce size and class imbalance)
- Heavily dependent on feature engineering (~100 features per model)
- Not too sensitive to hyper-parameters (random search and hyperopt give similar results, ~5,000 trees per model)
- Usual validation approach (in sample and out of sample hold-outs)
- Can execute in ~1ms per case on horizontally scalable distributed cloud environment
- Explainability slightly harder than linear models, but not as hard as deep learning (though not needed for fraud use case)
Debit Card Not Present Model Performance

Detects twice as much fraud than the incumbent – reduces losses, genuine declines and operational costs

Performance is higher than incumbent model and rules combined: allows ruleset reduction by 60%
Feature Engineering

Performance is heavily dependent on good feature engineering using domain knowledge.

**Feature Engineering**

Features based on:
- Current Transaction details
- Aggregated historical transactions (time, type, metric)
- Customer, card and account details

Domain knowledge important, but use the kitchen sink approach!

**Feature Selection**

- 20,000 Feature Pool
- 10,000 Predictive Features

By Segment

- Predictive Power of Feature
- Predictive Power of Feature
- Feature Importance in Toy Model
- Feature Importance in Toy Model

700 Features Selected

Model on all transactions

Performance overall and on segments

Remove low importance features

Add segment level features if needed
Implementation Approach

Model implementation pipeline using DevOps principles is still at early stages but is maturing.

Nexus Repository

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 Implementation and UAT Validation
- Feature Implementation and UAT Validation
- Feature Maturation and Shadow Running
- Production Validation and Go-Live Decision

Aim for maximum automation to reduce risk and increase time to market.
Learnings and Improvements

Model development is the easiest bit: focus on data engineering and the execution platform

Learnings
- Build data pipelines before the execution platform
- Use modern tools (Python, Spark, Jira, Nexus, etc)
- More features leads to better stability and performance
- Don’t assume you need to retrain frequently

Improvements
- More data sources (device data, inbound payments)
- Customer view data
- Better feature engineering (e.g. networks)
- Less down sampling, bigger models
- New modelling techniques (e.g. RNNs)