(Machine)
Learning To Detect Fraudsters

Hany Elemary

Sarah LeBlanc
CREDIT CARD FRAUD

TRANSACTION

APPLICATION

CARD NOT FOUND
MINIMIZE LOSSES

Lost Profitability =

$\text{(Fraud Cost} \times FN) + \text{(Opportunity Cost} \times FP)$

Legend:

- **FN** (Fraud missed)
- **FP** (Mistaken fraud)
CURRENT STATE

Vendor

Fraud Detection

Rules
Model
Strategies

Application Service

Customer
PROPOSED STATE

Fraud Detection

Customer

Application Service

Vendor
- Fraud Detection
  - Rules
  - Strategies

CHALLENGER MODELS

CHAMPION
MODEL TRAINING

Supervised Learning

Historical Data → Training → Model → Classification

Fraud
Not Fraud
DATA PATTERNS

Filter
Transform
Impute
Features
DATA FILTERING

Low Cardinality
DATA FILTERING

High Cardinality
DATA FILTERING

Medium Cardinality
DATA FILTERING

Medium Cardinality
# DATA TRANSFORMATION

<table>
<thead>
<tr>
<th>Email</th>
<th>Fraud Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:jack.smith@gmail.com">jack.smith@gmail.com</a></td>
<td>✓</td>
</tr>
<tr>
<td><a href="mailto:annie.may@fraudster.com">annie.may@fraudster.com</a></td>
<td>!</td>
</tr>
<tr>
<td><a href="mailto:freddy.jr@gmail.com">freddy.jr@gmail.com</a></td>
<td>✓</td>
</tr>
<tr>
<td><a href="mailto:nicole.jack@fraudster.com">nicole.jack@fraudster.com</a></td>
<td>!</td>
</tr>
<tr>
<td><a href="mailto:jon.johnston@gmail.com">jon.johnston@gmail.com</a></td>
<td>✓</td>
</tr>
<tr>
<td><a href="mailto:claudia.penns@us.gov">claudia.penns@us.gov</a></td>
<td>✓</td>
</tr>
<tr>
<td><a href="mailto:walter.carson@gmail.com">walter.carson@gmail.com</a></td>
<td>✓</td>
</tr>
<tr>
<td><a href="mailto:ben.benjamin@fraudster.com">ben.benjamin@fraudster.com</a></td>
<td>!</td>
</tr>
<tr>
<td>Domain name</td>
<td>Fraud Status</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td>gmail.com</td>
<td>![Green Check]</td>
</tr>
<tr>
<td>fraudster.com</td>
<td>![Orange Exclamation]</td>
</tr>
<tr>
<td>gmail.com</td>
<td>![Green Check]</td>
</tr>
<tr>
<td>fraudster.com</td>
<td>![Orange Exclamation]</td>
</tr>
<tr>
<td>gmail.com</td>
<td>![Green Check]</td>
</tr>
<tr>
<td>us.gov</td>
<td>![Green Check]</td>
</tr>
<tr>
<td>gmail.com</td>
<td>![Green Check]</td>
</tr>
<tr>
<td>fraudster.com</td>
<td>![Orange Exclamation]</td>
</tr>
</tbody>
</table>
## DATA IMPUTATION

### Handling Missing Data

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
</tr>
</tbody>
</table>
## DATA IMPUTATION

### Handling Missing Data

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
</tbody>
</table>
FEATURE SELECTION

IP to Zip Proximity
ARCHITECTURE

DATA SCIENTIST WORKFLOW

Raw Data → Transformed Data → Trained Model

DEVELOPER WORKFLOW

Applications → Trained Model → Score
DATA SCIENTIST WORKFLOW

Raw Data ➔ Transformed Data ➔ Trained Model

- Historical Data Store
- Binary Repository
- Clean
- Transform
- Impute

[H2O]

[JAR]
ARCHITECTURE

DATA SCIENTIST WORKFLOW

DEVELOPER WORKFLOW
VALUE STREAM

Data Ingestion → Model Training → Governance

Shadow Mode → Governance Evaluation → Champion Model

Publish Service → Publish Model
THANK YOU

Sarah LeBlanc
sleblanc@thoughtworks.com
@sarah_g_leblanc

Hany Elemary
helemary@thoughtworks.com
@hanyelemary

Questions?