Automated Data Quality Assurance with Machine Learning

4th European Conference On Artificial Intelligence in Finance and Industry
Winterthur, 05.09.2019

Martin Müller-Lennert
Senior Data Scientist
martin.mueller-lennert@incubegroup.com

Milica Petrović
Senior Data Scientist
milica.petrovic@incubegroup.com
Talk Outline

1. What’s Wrong with Data Quality?
2. Error Detection using ML
3. Demo & Features
4. Error Remediation using RPA
Talk Outline

1. What’s Wrong with Data Quality?
2. Error Detection using ML
3. Demo & Features
4. Error Remediation using RPA
Data Quality Today
What’s wrong with it?

Data Sources

Data Quality

End User

```
select ID, height, weight
from datatable
where height is not NULL
and weight is not NULL
and height > weight
and weight/(height*height) > 0
and weight/(height*height) < 100;
```
Data Quality Today
What’s wrong with it?

Data Sources

End User

Data Quality

select ID, height, weight from datatable
where height is not NULL
  and weight is not NULL
  and height > weight
  and weight/(height*height) > 0
  and weight/(height*height) < 100;
Data Quality Today
What’s wrong with it?

Data Sources

End User

Data Quality

```sql
select ID, height, weight
from datatable
where height is not NULL
  and weight is not NULL
  and height > weight
  and weight/(height*height) > 0
  and weight/(height*height) < 100;
```
Data Quality Today
What’s wrong with it?

Data Sources

End User

Data Quality

```sql
select ID, height, weight
from datatable
where height is not NULL
and weight is not NULL
and height > weight
and weight/(height*height) > 0
and weight/(height*height) < 100;
```
Data Quality Today
What’s wrong with it?

Data Sources

End User

Data Quality

Data Sources

select ID, height, weight
from datatable
where height is not NULL
and weight is not NULL
and height > weight
and weight/(height*height) > 0
and weight/(height*height) < 100;
Data Quality Today
What’s wrong with it?

Data Sources

End User

Data Quality

```
select ID, height, weight
from datatable
where height is not NULL
and weight is not NULL
and height > weight
and weight/(height^2) > 6
and weight/(height^2) < 100;
```
Data Quality Today
What’s wrong with it?

End User

Data Sources

Data Quality

```
select ID, height, weight
from datatable
where height is not NULL
and weight is not NULL
and height > weight
and weight/(height*height) > 0
and weight/(height*height) < 100;
```
Data Quality Today
What’s wrong with it?

Data Sources

End User

Data Quality

select ID, height, weight
from datatable
where height is not NULL
and weight is not NULL
and height > weight
and weight/(height*height) > 0
and weight/(height*height) < 100;
Data Quality Today
Our Take at a Solution

Data Quality Today
- Manually coded SQL rules
- Uni-/bi-variate checks

Challenges
- Too much data
- Too few rules
- Too narrow focus
- Too late

Solutions with Machine Learning

Automate: simultaneous error detection & faster process
Reusability: tailored ML algorithms reused for fields of similar type
Deep dive: discovery of new types of errors based on multivariate relationships

Unsupervised
- Capture multivariate relationships

Autoencoders
Talk Outline

1. What’s Wrong with Data Quality?
2. Error Detection using ML
3. Demo & Features
4. Error Remediation using RPA
Autoencoders for Data Quality
Architecture and Training

Target: Reconstruct input

Bottleneck: Enforced by architecture or regularization
Ensures network learns structure of input data

For good data only
Training on imperfect data: Requires large share of good data

Limits potency of network: More layers not always better
Discriminating Good and Bad Data Records
Clustering the Reconstruction Errors

Challenge: Many data points and potentially extreme class imbalance
Discriminating Good and Bad Data Records
Clustering the Reconstruction Errors

Challenge: Many data points and potentially extreme class imbalance
Discriminating Good and Bad Data Records
Sequence of Autoencoders

1st iteration
Remove Detected Anomalies
Keep Rest of Data

Challenge: Magnitude of reconstruction error varies across data error types
Discriminating Good and Bad Data Records
Sequence of Autoencoders
Discriminating Good and Bad Data Records
Sequence of Autoencoders
Discriminating Good and Bad Data Records
Sequence of Autoencoders

**Across iterations:** Increase model complexity

**Stopping:** When threshold separates large chunk of data
Talk Outline

1. What’s Wrong with Data Quality?
2. Error Detection using ML
3. Demo & Features
4. Error Remediation using RPA
Birth date

<table>
<thead>
<tr>
<th>Customer type</th>
<th>First name</th>
<th>Last name</th>
<th>Company name</th>
<th>Birth date</th>
<th>Income</th>
<th>Income class</th>
<th>Tax class</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>Mrs Georgia</td>
<td>Kingsley</td>
<td></td>
<td>11/19/1875</td>
<td>540000</td>
<td>high</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Demetria</td>
<td>Harmon</td>
<td></td>
<td>08/19/2020</td>
<td>270000</td>
<td>high</td>
<td>full</td>
<td>yes</td>
</tr>
<tr>
<td>private</td>
<td>Homer</td>
<td>Carr</td>
<td></td>
<td>10/12/1984</td>
<td>180000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>company</td>
<td>Littleton</td>
<td>Dean</td>
<td>FreeSeas Inc.</td>
<td>01/22/1957</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Allana</td>
<td>Snider</td>
<td></td>
<td>01/30/1976</td>
<td>112500</td>
<td>medium</td>
<td>reduced</td>
<td>yes</td>
</tr>
<tr>
<td>private</td>
<td>Ralph</td>
<td>Jones</td>
<td></td>
<td>06/28/1953</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Jimmy</td>
<td>Jordahl</td>
<td></td>
<td>04/21/1958</td>
<td>67500</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Amos</td>
<td>Miller</td>
<td></td>
<td>01/23/1982</td>
<td>126000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>company</td>
<td>Peregrine Pharmaceuticals Inc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>private</td>
<td>Elgin</td>
<td>Howell</td>
<td></td>
<td>02/24/1952</td>
<td>81000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Adriana</td>
<td>Bailey</td>
<td></td>
<td>07/09/1980</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Jamin</td>
<td>Sakaguchi</td>
<td></td>
<td>08/05/1971</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Charles</td>
<td>Partin</td>
<td></td>
<td>10/27/1977</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>yes</td>
</tr>
<tr>
<td>private</td>
<td>Jeanetta</td>
<td>Clark</td>
<td></td>
<td>08/03/1966</td>
<td>292500</td>
<td>high</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>Customer type</td>
<td>First name</td>
<td>Last name</td>
<td>Company name</td>
<td>Birth date</td>
<td>Income</td>
<td>Income class</td>
<td>Tax class</td>
<td>Children</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>------------</td>
<td>------------------------------------------</td>
<td>--------------</td>
<td>--------</td>
<td>--------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>1</td>
<td>private</td>
<td>Mrs Georgia Kingsley</td>
<td></td>
<td>11/19/1875</td>
<td>540000</td>
<td>high</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>private</td>
<td>Demetria Harmon</td>
<td></td>
<td>08/19/2020</td>
<td>270000</td>
<td>high</td>
<td>full</td>
<td>yes</td>
</tr>
<tr>
<td>70</td>
<td>company</td>
<td>Palmetto Bancshares, Inc. (SC)</td>
<td></td>
<td>02/15/1943</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>78</td>
<td>private</td>
<td>Rupert Carty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Demo

### First name

<table>
<thead>
<tr>
<th>Customer type</th>
<th>First name</th>
<th>Last name</th>
<th>Company name</th>
<th>Birth date</th>
<th>Income</th>
<th>Income class</th>
<th>Tax class</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>Mrs Georgia</td>
<td>Kingsley</td>
<td></td>
<td>11/19/1875</td>
<td>540000</td>
<td>high</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Demetria</td>
<td>Harmon</td>
<td></td>
<td>08/19/2020</td>
<td>270000</td>
<td>high</td>
<td>full</td>
<td>yes</td>
</tr>
<tr>
<td>private</td>
<td>Homer</td>
<td>Carr</td>
<td>FreeSeas Inc.</td>
<td>10/12/1984</td>
<td>180000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Littleton</td>
<td>Dean</td>
<td></td>
<td>01/22/1957</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Aliana</td>
<td>Snider</td>
<td></td>
<td>01/30/1976</td>
<td>112500</td>
<td>medium</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Ralph</td>
<td>Jones</td>
<td></td>
<td>06/28/1953</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Jimmy</td>
<td>Jordahl</td>
<td></td>
<td>04/21/1958</td>
<td>67500</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Amos</td>
<td>Miller</td>
<td></td>
<td>01/23/1982</td>
<td>126000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
</tr>
<tr>
<td>company</td>
<td>Peregrine Pharmaceuticals Inc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>private</td>
<td>Elgin</td>
<td>Howell</td>
<td></td>
<td>02/24/1952</td>
<td>81000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Adriana</td>
<td>Bailey</td>
<td></td>
<td>07/09/1980</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Jamin</td>
<td>Sakaguchi</td>
<td></td>
<td>08/05/1971</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
</tr>
<tr>
<td>private</td>
<td>Charles</td>
<td>Partin</td>
<td></td>
<td>10/27/1977</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>yes</td>
</tr>
<tr>
<td>private</td>
<td>Jeanetta</td>
<td>Clark</td>
<td></td>
<td>08/03/1966</td>
<td>292500</td>
<td>high</td>
<td>full</td>
<td>no</td>
</tr>
</tbody>
</table>

**KPIs for selected field**

- **True Positives**: 2
- **True Negatives**: 98
- **False Positives**: 0
- **False Negatives**: 0

**Measure of Anomaly for field First name per Datapoint**
<table>
<thead>
<tr>
<th>Company name</th>
<th>Birth date</th>
<th>Income</th>
<th>Income class</th>
<th>Tax class</th>
<th>Children</th>
<th>Education</th>
<th>Revenue</th>
<th>Employees</th>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11/19/1875</td>
<td>540000</td>
<td>high</td>
<td>full</td>
<td>no</td>
<td>Higher education</td>
<td>90000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>08/19/2020</td>
<td>270000</td>
<td>high</td>
<td>full</td>
<td>yes</td>
<td>Incomplete higher</td>
<td>1546020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10/12/1964</td>
<td>180000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
<td>Secondary / secondary special</td>
<td>1029658.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>01/22/1957</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
<td>Higher education</td>
<td>3092300</td>
<td>31</td>
<td>1631698</td>
</tr>
<tr>
<td>5</td>
<td>01/30/1976</td>
<td>112500</td>
<td>medium</td>
<td>reduced</td>
<td>yes</td>
<td>Higher education</td>
<td>521280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>08/28/1953</td>
<td>135000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
<td>Secondary / secondary special</td>
<td>229500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>04/21/1958</td>
<td>67500</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
<td>Secondary / secondary special</td>
<td>343800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>01/23/1982</td>
<td>126000</td>
<td>medium</td>
<td>full</td>
<td>no</td>
<td>Lower secondary</td>
<td>563269.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>916470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1400258</td>
<td>1400</td>
<td>1000000</td>
</tr>
<tr>
<td>11</td>
<td>02/24/1952</td>
<td>81000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
<td>Secondary / secondary special</td>
<td>132444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>07/09/1980</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
<td>Lower secondary</td>
<td>225000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>08/05/1971</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>no</td>
<td>Secondary / secondary special</td>
<td>942300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10/27/1977</td>
<td>90000</td>
<td>low</td>
<td>reduced</td>
<td>yes</td>
<td>Higher education</td>
<td>50940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>08/03/1966</td>
<td>292500</td>
<td>high</td>
<td>full</td>
<td>no</td>
<td>Secondary / secondary special</td>
<td>450000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**KPIs for selected field**

- **True Positives**: 3
- **True Negatives**: 97
- **False Positives**: 0
- **False Negatives**: 0

**Measure of Anomaly for field Revenue per Datapoint**
### Demo Revenue

#### Select Field for Anomaly Detection
- First name
- Last name
- Company name
- Birth date
- Income
- Tax class
- Children
- Education
- Revenue

#### KPIs for selected field

**True Positives**: 3
- Remove Filter
**True Negatives**: 97
- Filter Table
**False Positives**: 0
- Filter Table
**False Negatives**: 0
- Filter Table

#### Measure of Anomaly for field Revenue per Datapoint

<table>
<thead>
<tr>
<th>e</th>
<th>First name</th>
<th>Last name</th>
<th>Company name</th>
<th>Birth date</th>
<th>Income</th>
<th>Income class</th>
<th>Tax class</th>
<th>Children</th>
<th>Education</th>
<th>Revenue</th>
<th>Employees</th>
<th>Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>FreeSeas Inc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3092300</td>
<td>31</td>
<td>1631698</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>Peregrine Pharmaceuticals Inc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1400258</td>
<td>1400</td>
<td>1000000</td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td>Palmetto Bancshares, Inc. (SC)</td>
<td>02/15/1943</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>299504</td>
<td>1</td>
<td>2638759</td>
</tr>
</tbody>
</table>
## Reusability of Pre-Processing and Model Setup

Setup per feature type

<table>
<thead>
<tr>
<th>Type of variable</th>
<th>Pre-processing</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
<td>One-hot encoding of characters</td>
<td>Variational autoencoders with LSTM cells</td>
</tr>
<tr>
<td>Categorical</td>
<td>One-hot encoding</td>
<td>Complete autoencoder with regularization</td>
</tr>
<tr>
<td>Date</td>
<td>Numerical features from digits</td>
<td>Complete autoencoder with regularization</td>
</tr>
<tr>
<td></td>
<td>Normalization</td>
<td></td>
</tr>
<tr>
<td>Numerical</td>
<td>Normalization</td>
<td>Undercomplete autoencoder with custom loss</td>
</tr>
</tbody>
</table>
Talk Outline

1. What’s Wrong with Data Quality?
2. Error Detection using ML
3. Demo & Features
4. Error Remediation using RPA
### Identifying Where DQ Exception Occurred

**Searching for Correct Data using the ID**

#### Example: CRM data at a bank

<table>
<thead>
<tr>
<th>Customer Number</th>
<th>Name</th>
<th>Street</th>
<th>Number</th>
<th>City</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>65465456</td>
<td>Jane Doe</td>
<td>Riverstreet</td>
<td>32</td>
<td>Zurich</td>
<td>8004</td>
</tr>
<tr>
<td>12564321</td>
<td>John Doe</td>
<td>Central Town</td>
<td>i</td>
<td>Zurich</td>
<td>8001</td>
</tr>
<tr>
<td>56543651</td>
<td>Peter Smith</td>
<td>Quiet Street</td>
<td>2</td>
<td>Zurich</td>
<td>8045</td>
</tr>
<tr>
<td>36364448</td>
<td>Lisa Miller</td>
<td>Busy Street</td>
<td>3</td>
<td>Zurich</td>
<td>8022</td>
</tr>
</tbody>
</table>

**AUTOENCODER**
Finding Correct Data via “Intelligent” Robots
Computer Vision or NLP Help Find and Read Correct Data

<table>
<thead>
<tr>
<th>Customer Number</th>
<th>Name</th>
<th>Street</th>
<th>Number</th>
<th>City</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>12564321</td>
<td>John Doe</td>
<td>Central Town</td>
<td>1</td>
<td>Zurich</td>
<td>8001</td>
</tr>
<tr>
<td>56543651</td>
<td>Peter Smith</td>
<td>Quiet Street</td>
<td>2</td>
<td>Zurich</td>
<td>8045</td>
</tr>
<tr>
<td>36364448</td>
<td>Lisa Miller</td>
<td>Busy Street</td>
<td>3</td>
<td>Zurich</td>
<td>8022</td>
</tr>
</tbody>
</table>
Remediation Proposal
RPA Robot Proposes DQ Exception Correction
Automation of Data Quality Remediation

Process Overview

Client Data Repository

Error Detection via ML
Extract DQ exceptions using autoencoders

RPA robot finds correct data using NLP and computer vision, and proposes remediation of DQ exceptions

Cognitive Services

SME reviews, and approves or rejects

DQ resolutions are used to improve error detection and correction

Corrected data flows back into repository

RPA robot corrects DQ exceptions

Automation of Data Quality Remediation Process Overview
Key Findings from Projects and Experience

**Extension:** ML can replicate rule-based DQ checks **and** find **new** errors

**Multivariate relationships:** Detection of interdependencies

**Unsupervised learning!** Training data quality matters

**High reusability:** Only one-time customization effort per data type

**Cost savings:** Automate interactions with existing IT infrastructure

**High scalability:** Operations can be performed in parallel

**Competency:** Subject Matter Experts can focus on value-adding tasks