Detecting Money Laundering Actions Using Data Mining and Expert Systems

Ekrem Duman
Dogus University
Industrial Engineering Department
Istanbul, TURKEY

Joint work with Ayse Buyukkaya
Outline

✓ Our experience in Banking
  ✓ Introduction
  ✓ Success Factors in predictive modeling

✓ Motivation for AML
  ✓ Why AML is important?
  ✓ Existing solutions
  ✓ Motivation for a better solution

✓ Suggested AML Framework
  ✓ Exploratory data analysis
  ✓ Inferential data analysis
  ✓ Expert system
Introduction

Basic Objectives in Using DM:

1. **Descriptive**
   - Clustering / Segmentation
   - Basket (association) analysis
   - Sequence (pattern) analysis

2. **Predictive**
   - Classification
   - Time series analysis - regression
Introduction

Some of the DM projects we made:

✓ Clustering / Segmentation
✓ Product specific cross and up sell models
✓ Sequence analysis
✓ Customer churn
✓ CC, Loan and OD behavior scorecards
✓ Suspicious transactions
Success Factors

Some technical factors that can affect the success of a project:

✓ The adequacy of data
✓ Selecting input variables
✓ The way of using input variables
✓ Forming the training set
✓ Determining model application period
✓ Algorithm selection
✓ Determining model assessment criteria
✓ Feeding the campaign results back
Success Factors

Everything is the same but two factors are handled differently:

<table>
<thead>
<tr>
<th></th>
<th>SA %</th>
<th>OA %</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROD A</td>
<td>1.6</td>
<td>9.2</td>
<td>5.7</td>
</tr>
<tr>
<td>PROD B</td>
<td>4.3</td>
<td>8.8</td>
<td>2.1</td>
</tr>
<tr>
<td>PROD C</td>
<td>4.1</td>
<td>21.3</td>
<td>5.2</td>
</tr>
<tr>
<td>PROD D</td>
<td>0.3</td>
<td>1.7</td>
<td>5.1</td>
</tr>
<tr>
<td>PROD E</td>
<td>2.3</td>
<td>3.6</td>
<td>1.5</td>
</tr>
<tr>
<td>PROD F</td>
<td>0.7</td>
<td>1.0</td>
<td>1.4</td>
</tr>
<tr>
<td>PROD G</td>
<td>3.0</td>
<td>3.1</td>
<td>1.0</td>
</tr>
<tr>
<td>Average</td>
<td>2.3</td>
<td>6.9</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Avg Target | 10.578 | 10.039

*The percentage of actual buyers in the model target list within the following month

Need for developing specific models for each product!
Success Factors

Some technical factors that can affect the success of a project:

- The adequacy of data
- Selecting input variables
- The way of using input variables
- Forming the training set
- Determining model application period
- Algorithm selection
- Determining model assessment criteria
- Feeding the campaign results back
Success Factors

The adequacy of data

✓ Do we have the necessary and sufficient variables in data mart?

✓ Are the data complete?
  ✓ How should we handle the missing data?

✓ What is the data quality?
Selecting input variables

✓ Few variables → interpretable, fast, less accurate (?) models
✓ Moderate number of variables → ?
✓ Many variables → non-interpretable, slow, highly accurate (?) models

How can we determine the right variables?

Our Experience:

NN  ⇒  few variables
DT  ⇒  many variables
Success Factors

The way of using input variables

✓ In original form?
✓ Transformed?
  ✓ Log
  ✓ Categorization
  ✓ Other

<table>
<thead>
<tr>
<th>Range($)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>0</td>
</tr>
<tr>
<td>1-100</td>
<td>1</td>
</tr>
<tr>
<td>100-1000</td>
<td>2</td>
</tr>
<tr>
<td>1000-10000</td>
<td>3</td>
</tr>
<tr>
<td>10000+</td>
<td>4</td>
</tr>
</tbody>
</table>

Our Experience:
- Clustering ⟷ Categorization
- Cross Sell ⟷ Original
Success Factors

Forming the training set

✓ Who should take place in the training set?
  ✓ Whole customer base?
  ✓ Some clusters only?
    ✓ How to cluster customers?

✓ What should be the ratio of positive to negative samples in the training set?

Our Experience:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Sell</td>
<td>1 - 1</td>
</tr>
<tr>
<td>Scorecard</td>
<td>1 - k (k&gt;1)</td>
</tr>
</tbody>
</table>
Success Factors

Determining model application period

✓ What is the campaign period?

✓ What is to be done for campaign timing to meet the needs of business?

✓ In which period the customer behavior will be analyzed?

Our Experience:

<table>
<thead>
<tr>
<th>Asset</th>
<th>1 month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liability</td>
<td>3 months</td>
</tr>
</tbody>
</table>
Success Factors

Algorithm selection

✓ Neural networks: slow and difficult to understand
✓ Regression: faster, understandable
✓ Decision trees: fast, understandable
  ✓ The ones working better with symbolic variables
  ✓ The ones working better with continuous variables
Success Factors

Determining model assessment criteria

✓ Accuracy (confusion) matrix
✓ Hit rate
✓ Capture rate
✓ LIFT
## Success Factors

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Conf.</th>
<th>Hit %</th>
<th>Capture %</th>
<th>LIFT</th>
<th>Accuracy (0, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART 7</td>
<td>70</td>
<td>16.2</td>
<td>41.8</td>
<td>8.3</td>
<td>54</td>
</tr>
<tr>
<td>C50</td>
<td>70</td>
<td>15.8</td>
<td>48.5</td>
<td>8.0</td>
<td>74</td>
</tr>
<tr>
<td>ECHAID7</td>
<td>70</td>
<td>19.1</td>
<td>37.9</td>
<td>9.7</td>
<td>67</td>
</tr>
<tr>
<td>QUEST</td>
<td>73</td>
<td>16.6</td>
<td>29.7</td>
<td>8.5</td>
<td>83</td>
</tr>
<tr>
<td>C50</td>
<td>80</td>
<td>13.3</td>
<td>5.8</td>
<td>6.8</td>
<td>74</td>
</tr>
<tr>
<td>ECHAID7</td>
<td>80</td>
<td>21.9</td>
<td>15.9</td>
<td>11.2</td>
<td>67</td>
</tr>
<tr>
<td>Avg (70%)</td>
<td>71</td>
<td>16.9</td>
<td>39.5</td>
<td>8.6</td>
<td>70</td>
</tr>
<tr>
<td>Avg (80%)</td>
<td>80</td>
<td>17.6</td>
<td>10.8</td>
<td>9.0</td>
<td>71</td>
</tr>
</tbody>
</table>
Success Factors (CROSS SELL)

Feeding the campaign results back

- Determining additional variables required
- Preparing and adding the variables to data mart
- Cleaning the data and increasing its quality
- Model development
- Assessment
- Implementation (Campaigns)
- New data + campaign results
### Success Factors

**After some improvement studies:**

<table>
<thead>
<tr>
<th></th>
<th>Old</th>
<th>New</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80%</td>
<td>80%</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>SA %</td>
<td>OA %</td>
<td>Ratio</td>
<td>OA %</td>
</tr>
<tr>
<td>PROD A</td>
<td>1.6</td>
<td>9.2</td>
<td>5.7</td>
<td>9.3</td>
</tr>
<tr>
<td>PROD A-12</td>
<td>1.0</td>
<td></td>
<td></td>
<td>6.8</td>
</tr>
<tr>
<td>PROD B</td>
<td>4.3</td>
<td>8.8</td>
<td>2.1</td>
<td>25.7</td>
</tr>
<tr>
<td>PROD B-6</td>
<td>1.1</td>
<td></td>
<td></td>
<td>10.0</td>
</tr>
<tr>
<td>PROD C</td>
<td>4.1</td>
<td>21.3</td>
<td>5.2</td>
<td>28.5</td>
</tr>
<tr>
<td>PROD D</td>
<td>0.3</td>
<td>1.7</td>
<td>5.1</td>
<td>0.8</td>
</tr>
<tr>
<td>PROD E</td>
<td>2.3</td>
<td>3.6</td>
<td>1.5</td>
<td>2.6</td>
</tr>
<tr>
<td>PROD F</td>
<td>0.7</td>
<td>1.0</td>
<td>1.4</td>
<td>2.2</td>
</tr>
<tr>
<td>PROD G</td>
<td>3.0</td>
<td>3.1</td>
<td>1.0</td>
<td>8.2</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>2.0</strong></td>
<td><strong>6.9</strong></td>
<td><strong>3.2</strong></td>
<td><strong>10.5</strong></td>
</tr>
</tbody>
</table>

**Old New**

**Average**
Some Campaigns We Made

First Term Deposit (TD) Model and Campaign:

✓ The customers who did not have a TD are targeted
✓ Sales/reached = 26% (lift effect = 45)
✓ Avg account size = 2.5 times the general avg

Second Term Deposit Model and Campaign:

✓ The customers who did not have a TD are targeted
✓ Sales/reached = 40% (lift effect = 98)
✓ Avg account size = 2.5 times the general avg
Some Campaigns We Made

**Bill Payment Order Campaign:**
- The customers who did not have a payment order before are targeted
- Sales/reached = 6% (lift effect = 30)
- One more bill payment order, differentiated by the sales to lower segments

**Overdraft Campaign:**
- Sales/reached = 38% (lift effect = 55)
- The actual users in the first three months are 12% more than the general average
Outline

✓ Our experience in Banking
  ✓ Introduction
  ✓ Success Factors in predictive modeling

✓ Motivation for AML
  ✓ Why AML is important?
  ✓ Existing solutions
  ✓ Motivation for a better solution

✓ Suggested AML Framework
  ✓ Exploratory data analysis
  ✓ Inferential data analysis
  ✓ Expert system
Motivation for AML

- Terrorism is the main threat to everybody.
- Before 9/11 it was related to underdeveloped countries.
- After 9/11 it was understood that it threatens everybody.
- One of the ways of struggling terrorism is to block their financial activities.
- Terrorism finance is a type of black money and AML (anti money laundering) techniques can be used to combat it.

(e.g. Kurdish terrorist group PKK and its revenues from illegal drugs)
Motivation for AML

- There are many commercial packages available for AML
- They perform standard checks
  - Is the account holder on OFAC list?
- They are mostly rule based.
  - Look at the (past) transactions
  - Identify irregularities by some predefined rules

"Which customers made EFT more than 50 times last month?"
Motivation for AML

Basic deficiency of AML commercial packages:

✓ Too few “AND” rules
  ✓ list is too big to inspect

✓ Too many “AND” rules
  ✓ an actually fraudulent transaction/person could be missed
Motivation for AML

Money laundering actions can be inspected at two levels:

✓ Individual account/person level
  ✓ look at the (past) transactions of an account/person and identify irregularities

✓ Network Level
  ✓ look at all accounts/customers and identify suspicious loops
Motivation for AML

Money laundering actions can be inspected at two levels:

- Individual account/person level
  - look at the (past) transactions of an account/person and identify irregularities
- Network Level
  - look at all accounts/customers and identify suspicious loops
Suggested Framework

A three phase AML solution framework:

✓ exploratory data analysis
  ✓ descriptive data mining (DM) to determine unusualities

✓ inferential data analysis
  ✓ predictive DM to determine cases that need to be inspected

✓ expert system
  ✓ coding the inspection process
Suggested Framework

I - Exploratory data analysis phase:

✓ used to narrow down the search list

✓ customers/accounts are clustered wrt some behavioral variables
  ✓ we prefer to consolidate all accounts of a customer and work on
    customer list

✓ Customers who are close to cluster center are taken as “normal”
  and not inspected anymore
Suggested Framework

I - Exploratory data analysis phase:

✓ How can we decide that a customer showed unusual behavior?

✓ Is a customer who made 50 EFTs unusual?

  ✓ It could be...

    ✓ If he owns a small business it may be normal

    ✓ If he always does this it may be normal

✓ To judge we need to know;

  ✓ the average value in the demographic/behavioral segment of the customer

  ✓ the routine (historic) behavior of the customer
Suggested Framework

I - Exploratory data analysis phase:

✓ The clustering to be made should take both aspects into account

✓ First aspect is handled automatically

✓ For the second aspect, we calculated the deviation values and used them in clustering

\[
\text{Deviation} = \frac{\text{value in last month} - \text{avg of the last six months}}{\text{std. dev. in the last six months}}
\]
Suggested Framework

I - Exploratory data analysis phase:

✓ Variables should be selected in accordance with the purpose:

<table>
<thead>
<tr>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD_CA_NUM_CRD_TRX</td>
</tr>
<tr>
<td>BD_CA_NUM_CRD_TRX_DEV</td>
</tr>
<tr>
<td>BD_CA_NUM_ACCT_TL_OP</td>
</tr>
<tr>
<td>BD_CA_NUM_ACCT_TL_OP_DEV</td>
</tr>
<tr>
<td>BD_DBC_NUM_TRX</td>
</tr>
<tr>
<td>BD_DBC_NUM_TRX_DEV</td>
</tr>
<tr>
<td>BD_INT_LOGON_NUM</td>
</tr>
<tr>
<td>BD_INT_NUM_TRX</td>
</tr>
<tr>
<td>EFT_IN_TRX_NUM</td>
</tr>
<tr>
<td>EFT_OUT_TRX_NUM</td>
</tr>
</tbody>
</table>
Suggested Framework

I - Exploratory data analysis phase:

✓ For a clustering with $k$ variables and $n$ clusters;

✓ Cluster centers are determined
  ✓ This is a point in $k$ dimensional space which take averages of all customers in that cluster

✓ All customers in the cluster have some deviation from the center

Variable Deviation$_k = \text{dev of the customer’s variable } k \text{ value from cluster center value}$

Customer Deviation = total of the variable deviations
Suggested Framework

I - Exploratory data analysis phase:

Customer anomaly index = (customer deviation) / (average customer deviation in the cluster)

✓ Customers are sorted with a non-increasing value of anomaly index values;

✓ The lower part of this list is the normal or usual part

✓ The upper part can be inspected more
## Suggested Framework

### I - Exploratory data analysis phase:

<table>
<thead>
<tr>
<th>An. Ind.</th>
<th>CI</th>
<th>Primary Var.</th>
<th>VCM</th>
<th>Secondary Var.</th>
<th>VCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.000</td>
<td>3</td>
<td>BD_INT_LOGON_NUM</td>
<td>0.618</td>
<td>BD_CA_NUM_ACCT_TL</td>
<td>0.216</td>
</tr>
<tr>
<td>5.000</td>
<td>1</td>
<td>BD_CA_NUM_DEB_TRX_DEV</td>
<td>0.962</td>
<td>BD_CC_NUM_ACCT</td>
<td>0.026</td>
</tr>
<tr>
<td>4.924</td>
<td>3</td>
<td>BD_CHEQUE_NUM_TRX_DEV</td>
<td>0.341</td>
<td>BD_BRANCH_NUM_TRX_DEV</td>
<td>0.29</td>
</tr>
<tr>
<td>4.589</td>
<td>1</td>
<td>EFT_IN_TRX_NUM_DEV</td>
<td>0.979</td>
<td>EFT_OUT_TRX_NUM</td>
<td>0.008</td>
</tr>
<tr>
<td>4.285</td>
<td>3</td>
<td>BD_CHEQUE_NUM_TRX</td>
<td>0.293</td>
<td>BD_CA_NUM_CRD_TRX_DEV</td>
<td>0.259</td>
</tr>
<tr>
<td>4.123</td>
<td>3</td>
<td>BD_ACTIVE_CARD_NUM</td>
<td>0.936</td>
<td>BD_CC_NUM_TRX</td>
<td>0.017</td>
</tr>
<tr>
<td>4.000</td>
<td>1</td>
<td>BD_CA_NUM_ACCT_CL_DEV</td>
<td>0.925</td>
<td>BD_CC_NUM_ACCT</td>
<td>0.052</td>
</tr>
<tr>
<td>3.923</td>
<td>3</td>
<td>BD_CA_NUM_ACCT_FX</td>
<td>0.273</td>
<td>BD_CA_NUM_CRD_TRX</td>
<td>0.169</td>
</tr>
<tr>
<td>3.758</td>
<td>1</td>
<td>BD_CA_NUM_ACCT_CL_DEV</td>
<td>0.93</td>
<td>BD_CC_NUM_ACCT</td>
<td>0.058</td>
</tr>
<tr>
<td>3.501</td>
<td>2</td>
<td>BD_CA_NUM_ACCT_CL_DEV</td>
<td>0.625</td>
<td>BD_CA_NUM_CRD_TRX</td>
<td>0.1</td>
</tr>
<tr>
<td>3.284</td>
<td>7</td>
<td>BD_CA_MAX_BAL_RATE</td>
<td>0.551</td>
<td>BD_CHEQUE_NUM_TRX</td>
<td>0.105</td>
</tr>
<tr>
<td>3.001</td>
<td>1</td>
<td>BD_CA_NUM_ACCT_CL_DEV</td>
<td>0.962</td>
<td>BD_INT_LOGON_NUM</td>
<td>0.026</td>
</tr>
<tr>
<td>2.992</td>
<td>4</td>
<td>BD_CURRENCY_PURCHASE_NUM_TRX</td>
<td>0.212</td>
<td>BD_CA_NUM_ACCT</td>
<td>0.201</td>
</tr>
<tr>
<td>2.841</td>
<td>3</td>
<td>BD_INT_LOGON_NUM_DEV</td>
<td>0.772</td>
<td>BD_DBC_NUM_ACCT</td>
<td>0.075</td>
</tr>
<tr>
<td>2.552</td>
<td>3</td>
<td>BD_BILL_NUMBER</td>
<td>0.763</td>
<td>BD_CA_NUM_DEB_TRX</td>
<td>0.074</td>
</tr>
<tr>
<td>2.528</td>
<td>4</td>
<td>BD_CA_NUM_ACCT</td>
<td>0.198</td>
<td>BD_CA_NUM_ACCT_TL</td>
<td>0.182</td>
</tr>
<tr>
<td>2.428</td>
<td>3</td>
<td>BD_BRANCH_NUM_TRX</td>
<td>0.932</td>
<td>BD_ACTIVE_CARD_NUM</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Variable contribution measure (VCM) = variable deviation / customer deviation
II - Inferential data analysis phase:

It would be very nice to have a training set to learn who are laundering money.

The known cases are too few which is far beyond being sufficient.

Let the inspectors look at the anomaly indexes and decide which ones to inspect;
Suggested Framework

II - Inferential data analysis phase:

Anomaly List: The list produced by anomaly indexes

Inspect List: The ones that the inspectors see the need to inspect
III - Expert system phase:

Inspection process is too time consuming.

An expert system could be developed to help the inspection process.

The inspectors can focus on the few cases that the expert system finds suspicious.
Summary and Conclusions

Combating ML is very important.

Available solutions have some drawbacks.

A three phase solution framework is suggested.

- exploratory data analysis
- inferential data analysis
- expert system
Future Work

To mature and implement our solution framework.

To find solutions for the network level problems.

We are open to collaboration...
Any Questions?

Thank you for listening to us.