A Statistical and Machine Learning Model to Detect Money Laundering: an Application

Dr. Miguel Agustín Villalobos
Director – Advanced Analytics Institute

Dr. Eliud Silva
Professor

Actuarial Sciences Department
Anahuac University
June 15th, 2017
Agenda

› Executive summary

› Background and key references

› Financial Institutions Needs

› The model

› Conclusions
The objective is to apply statistical learning methods to address the AML needs of medium size banks to comply to recent regulation.

- Recent regulation by Central Banks is requiring financial institutions to implement a data & statistical models based approach to detect money laundering.
- Medium and small sized banks seek to comply with regulation without or with minimal changes to their AML systems which are mostly deterministic rule-based.
- We present a three module solution that responds to the medium size banks in the region to have an easy to implement solution to:
  - Complement their current AML systems
  - Minimize changes to their current AML and transactional systems
  - Provide a “best classification models combination” to detect suspect transactions.
- In the particular case analyzed here, it is concluded that the developed model, provides a 99.6% correct classification rate on test data, leading to a reduction in the number of alerted cases from the current close to 30% of transactions to less than 1%.
Money laundering poses significant challenges to financial institutions

› For example, economic impact in Mexico is estimated to be between $10 BUSD\(^1\) and 25 BUSD\(^2\)

› High volumes and complexity

› Dynamic and fast evolving

---

Significant research has been done in the area of counter financial crimes based on data mining and statistical learning methods

› There has been a significant amount of research in the use of statistical learning and data mining to fight financial crimes. At the end of this presentation we include the most relevant recent references used to develop the model presented here.

› The key objective is to develop a model simple enough for ease of implementation and sophisticated enough to protect the financial institution and comply with regulation.

› Two main references to this applied work are:
  - Kharote & Kshirsagar (2014)
    › Clustering
    › Profile generation
    › Customer behavior extraction
  - Claudio & Joao (2016):
    › Clustering based client profiling
    › Classification rule generation
Due to the extended use of electronic transactions in different currencies, Central Banks in LA are starting to require implementation of more advanced risk models

- Current models, based on deterministic business rules need to be replaced by statistical learning models

- Risk Model Criteria
  - Inherent risk factors
  - Transactional Profile

- Data based client risk classification under well structured criteria

- Periodic validation

- Theoretical and empirical basis documentation
Most financial institutions already have an AML system in place, but need to modify it to comply with current regulations.

- For speedy implementation, a solution is needed to complement current AML business rule based systems.
- Due to security issues faced by banks IT environment, a model with minimal impact in current systems and processes must be implemented.
- First requirement imposed by new regulations is for banks to challenge their current customer risk rating process with a risk rating process based on data.
The model is structured in three modules defined to achieve simplicity and minimal financial institution systems.

- Based on the premises simplicity and minimal bank systems intrusion our model is structured in three modules and assumes no online interphases with current systems.
- The updates to the institution’s AML system are assumed to be made by their personnel, according to their standards and procedures.
- The process will be illustrated using transformed data from a Latin American financial institution.
- Two types of events are considered:
  - Alerts: transactions flagged for investigation
  - Suspicious: Reported to regulator after investigation
The Customer Risk Classification seeks to assess customer risk based on historical data to estimate alert propensity and define customer risk levels.

- Based on historical transactional information containing an indicator of whether a transaction was flagged for investigation or not, Alert Propensity Scores are developed.
- Several algorithms were used being C5.0 the one that provided the best correct classification (98.41% with the testing set).

- Propensity Scores are then used to define different levels of risk (boundaries and number of risk levels depend on the organization’s risk appetite (in this application 5 levels are used)).

Customer allocation to risk levels is done utilizing customer “static” risk factors as predictors of risk level class and updated risk classification rules are developed

A C5.0 algorithm was elected to predict Risk Level based on the different customer “static” potential risk factors, to determine relative importance of each and create the set of rules for cluster allocation (68% Correct Classification)
Developing the classification model for suspicious transactions initiates by defining key transaction profile risk factors, balancing the data and splitting into training and testing data sets.

- After descriptive analysis of customer transaction data over a year of history, the following key factors were computed to characterize transactional profile both at customer and customer/month of the year levels, both for transaction amounts and number of transactions:
  - Mean, Median
  - Standard deviation, Q1, Q3, IQR, Minimum and Maximum
  - Q1 - 1.5IQR, Q1 - 3IQR, Q3 + 1.5IQR, Q3 + 3IQR,

- Data is split into a training and testing sets (50%/50%)
- Since suspicious transactions represent less than .4% of the training data, it is needed to balance the data

![Diagram of data flow](image-url)
After adjusting for data unbalance by two different methods, several classification models were used and compared in terms of correct classification, AUC and GINI.

Five different classifiers were used with both the Boosted Balanced Data and the Reduced Balanced data, their performance was compared and the best three were selected utilizing the Boosting Balancing method to obtain a final model best combination.

<table>
<thead>
<tr>
<th>Model</th>
<th>BOOST</th>
<th>REDUCE x 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Correct Classification</td>
<td>AUC</td>
</tr>
<tr>
<td>CHAID</td>
<td>87.20%</td>
<td>0.981</td>
</tr>
<tr>
<td>C&amp;RT</td>
<td>90.03%</td>
<td>0.956</td>
</tr>
<tr>
<td>C5.0</td>
<td><strong>99.72%</strong></td>
<td><strong>0.991</strong></td>
</tr>
<tr>
<td>Neural Network</td>
<td>88.70%</td>
<td>0.992</td>
</tr>
<tr>
<td>SVM</td>
<td>89.80%</td>
<td>0.666</td>
</tr>
</tbody>
</table>
The best three models were then combined under different criteria to arrive at a “best model.”

The best three models were combined under the following criteria, to arrive at a “best model.”

<table>
<thead>
<tr>
<th>Model (CHAID + C&amp;RT + C5.0)</th>
<th>% Correct Classification</th>
<th>AUC</th>
<th>GINI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest Confidence</td>
<td>99.74%</td>
<td>0.996</td>
<td>0.990</td>
</tr>
<tr>
<td>Confidence Weighted Voting</td>
<td>95.84%</td>
<td>0.998</td>
<td>0.995</td>
</tr>
<tr>
<td>Average Propensity</td>
<td>97.80%</td>
<td>0.986</td>
<td>0.974</td>
</tr>
<tr>
<td>Raw Propensity</td>
<td>96.11%</td>
<td>0.987</td>
<td>0.977</td>
</tr>
<tr>
<td>Adjusted Propensity</td>
<td>94.26%</td>
<td>0.987</td>
<td>0.976</td>
</tr>
<tr>
<td>Voting</td>
<td>96.15%</td>
<td>0.953</td>
<td>0.973</td>
</tr>
</tbody>
</table>

It can be seen that the “Highest Confidence” criteria yields the best results.
Given the results we may decide to use the C5.0 algorithm alone since the rules generated can be easily incorporated into the Institution’s AML system.

› We can see that the differences between the blended model and de C5.0 model are minimal.

<table>
<thead>
<tr>
<th></th>
<th>% Correct Classification</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5.0</td>
<td>99.72%</td>
<td>0.991</td>
</tr>
<tr>
<td>Blended Models</td>
<td>99.74%</td>
<td>0.996</td>
</tr>
</tbody>
</table>

› Given the difference, due to ease of implementation in the financial institution, it may be decided to use the C5.0 model alone.

› Currently, the number of transactions that are alerted amount to close to 30% of total number of transactions, whereas with the approach presented here, using the blended model, flagged transactions represent less than 1%.

<table>
<thead>
<tr>
<th>True Value</th>
<th>Predicted 0</th>
<th>Predicted 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.39%</td>
<td>0.25%</td>
</tr>
<tr>
<td>1</td>
<td>0.01%</td>
<td>0.35%</td>
</tr>
</tbody>
</table>
Bibliography

Thank you!