A model to Determine Business Opportunities

Ricardo Moncayo
B.Sc. Electronic Engineer , M.Sc. Biomedical Engineering

June 3, 2022

The problem:

The challenge: To build a predictive model to determine which business opportunities are more likely to win, using data integration from different sources

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

An AdaBoost classifier is trained using all data without the opportunity number

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

Relevant features are selected using mRMR (minimum Redundancy Maximum Relevance), and an AdaBoost validated in the complete dataset

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

Without country index data, relevant features are selected, and an AdaBoost validated

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

Proposal

Three Approaches were tested:

- A model trained from the raw data (Raw model)
- A model enhanced by selecting relevant features
- A model with relevant features without the country index (naive decision)
- A model-based on a supervised clustering of the country data relationships

A model is trained (with the most relevant features) using the data from a country and evaluating in the remaining countries independently. The higher results correspond to countries with similar decision processes

A model trained from the raw data (Raw model)

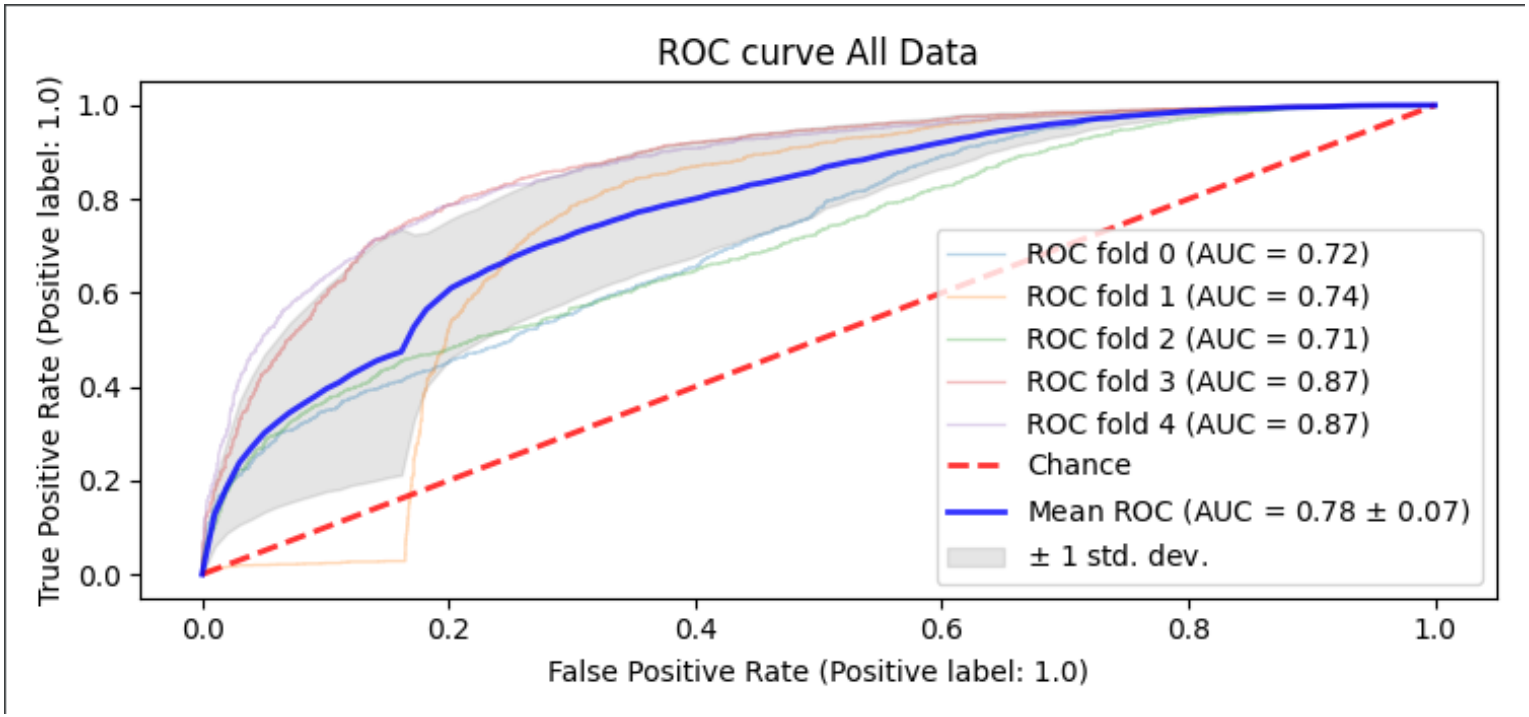
- The opportunity number is removed from the data
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation

A model trained from the raw data (Raw model)

- The opportunity number is removed from the data
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation

A model trained from the raw data (Raw model)

- The opportunity number is removed from the data
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation



A Model Enhanced by Selecting Relevant Features

- mRmR is used to select 10 most relevant features
- Stable relevant features are finally selected
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation

$$score_i(f) = \frac{relevance(f|target)}{redundancy(f|featuresselecteduntili-1)}$$

A Model Enhanced by Selecting Relevant Features

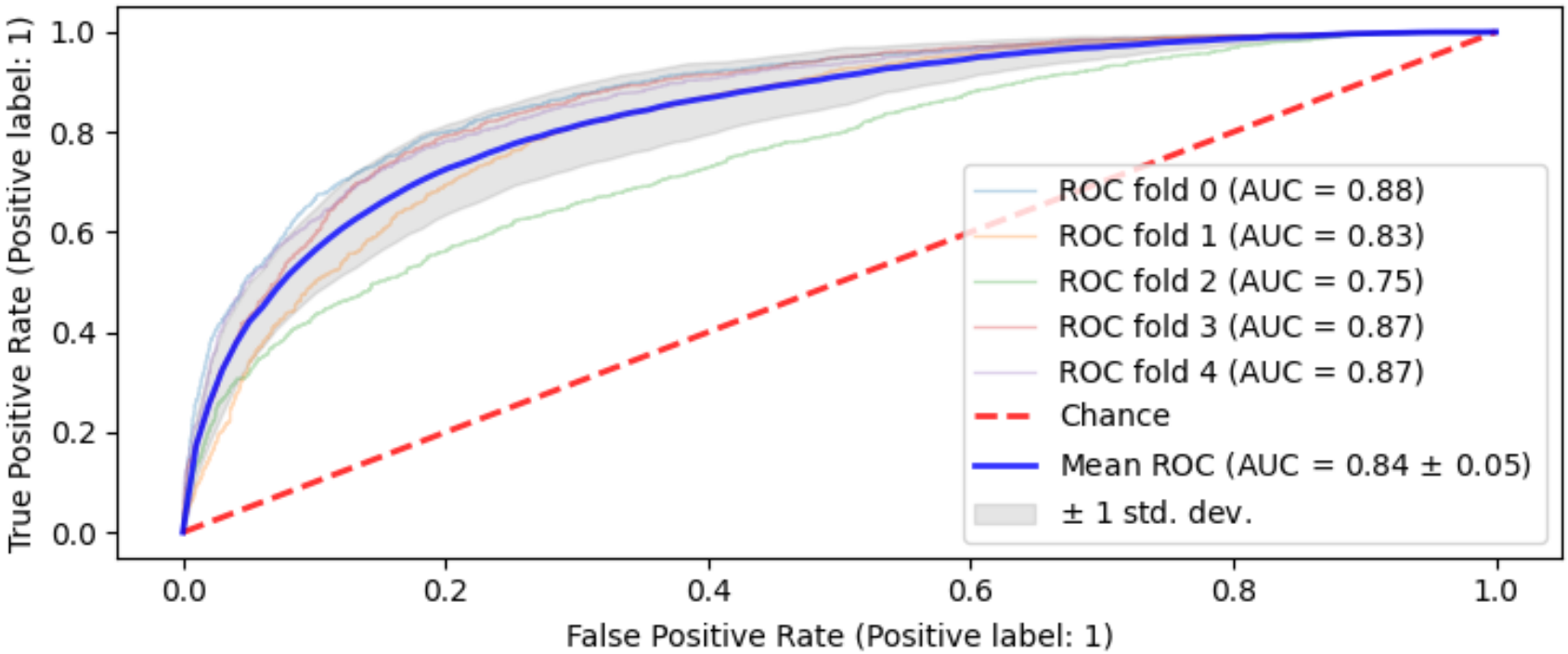
- mRmR is used to select 10 most relevant features
- Stable relevant features are finally selected
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation

A Model Enhanced by Selecting Relevant Features

- mRmR is used to select 10 most relevant features
- Stable relevant features are finally selected
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation

Best Features: CountryCode,Deal Size Category,Opportunity Amount USD,Revenue From Client Past Two Years,Sales Stage Change Count,Supplies Group,Supplies Subgroup,Total Days Identified Through Closing,Total Days Identified Through Qualified

Results using Relevant Features



A model Without country index and the most relevant features

- Country indexs are removed from the data (mRmR)
- mRmR is used to select 10 most relevant features and stable features are finally selected
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation in all data and in each country independently

A model Without country index and the most relevant features

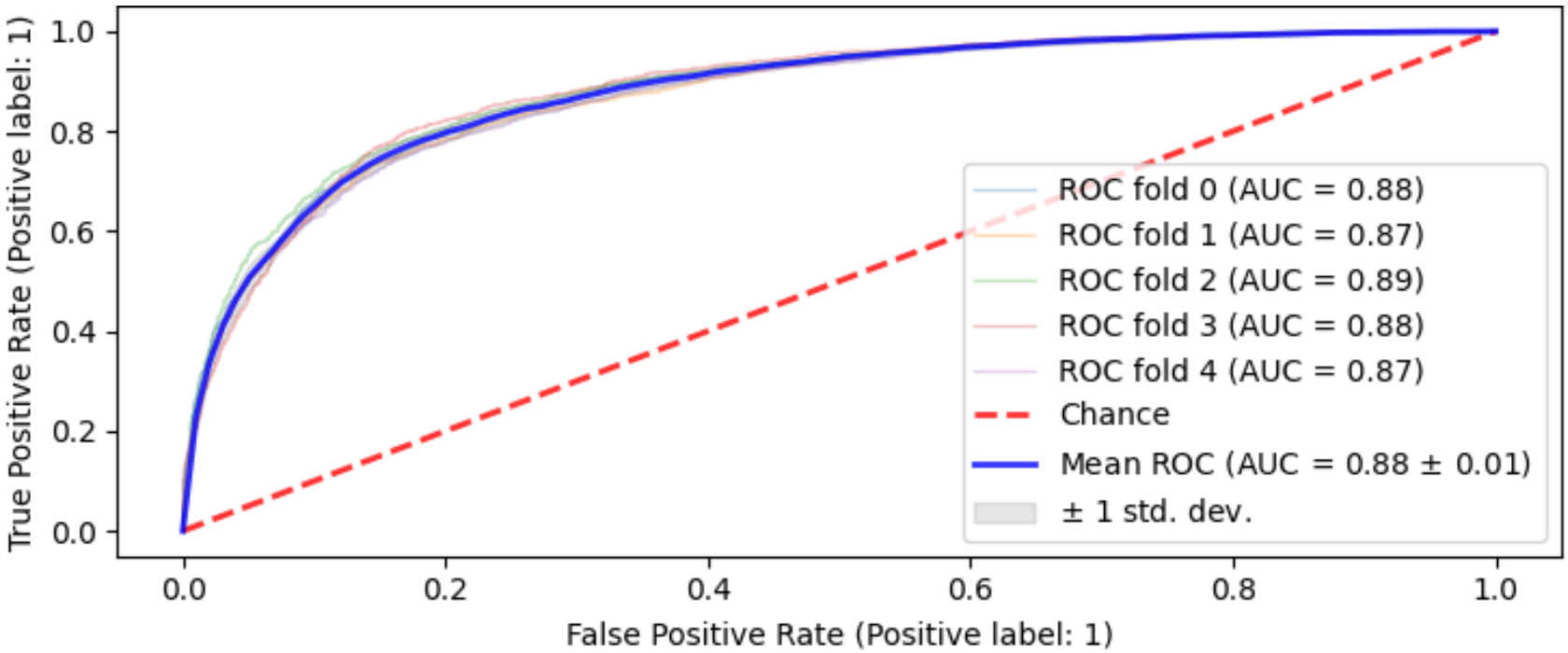
- Country indexs are removed from the data (mRmR)
- mRmR is used to select 10 most relevant features and stable features are finally selected
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation in all data and in each country independently

Best Features:Client Size By Employee Count,Deal Size Category,Opportunity Amount USD,Revenue From Client Past Two Years,Sales Stage Change Count,Supplies Group,Supplies Subgroup,Total Days Identified Through Closing,Total Days Identified Through Qualified

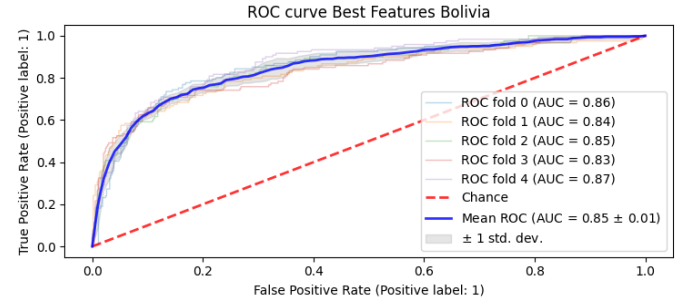
A model Without country index and the most relevant features

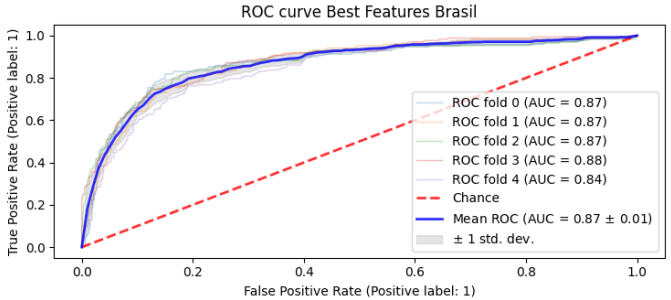
- Country indexs are removed from the data (mRmR)
- mRmR is used to select 10 most relevant features and stable features are finally selected
- An AdaBoost classifier is trained and testing using a 5-fold cross-validation in all data and in each country independently

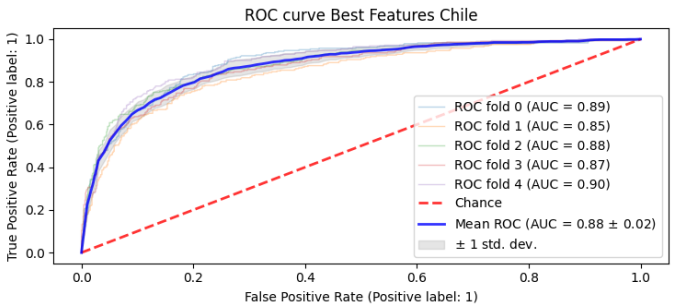
Results-Relevant Features without Country index - All Data

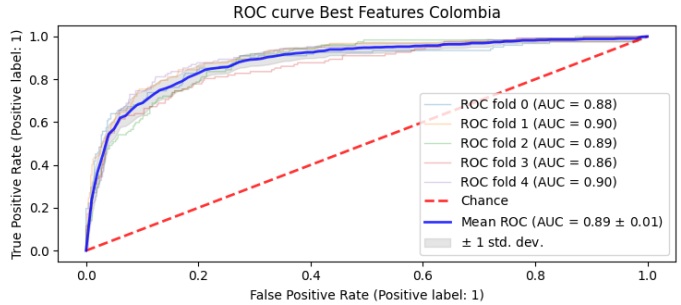


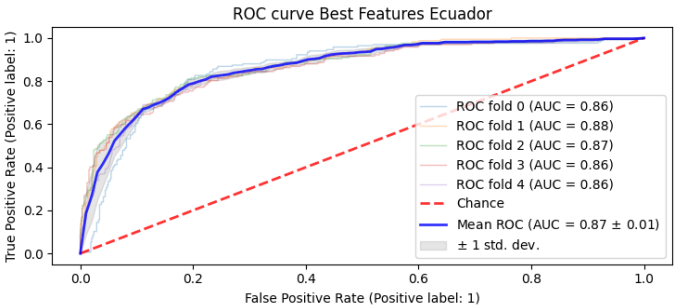
Results relevant features model per country

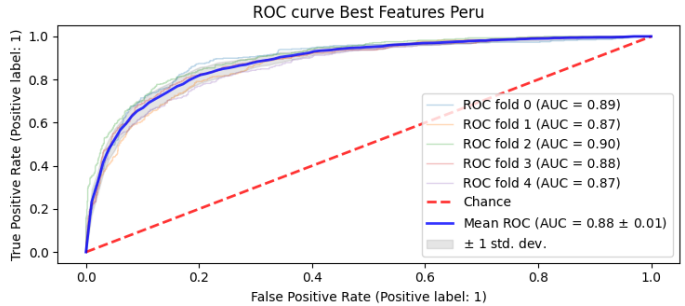


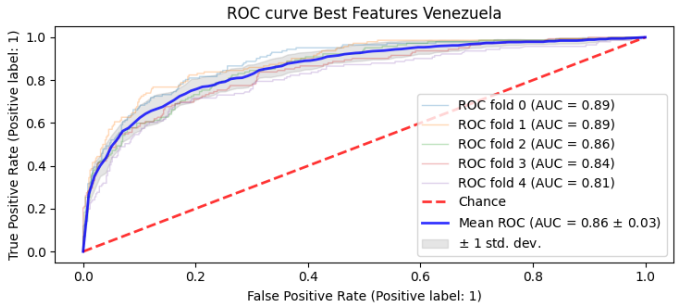












A model-based in a supervised clustering of the country data relationships

- A model is trained using the data from a country and evaluating in the remaining countries (i.e. Training from Brasil evaluating in Colombia,Training from Brasil evaluating in Peru, Training from Colombia evaluating in Brasil...)
- The AUC is computed in each evaluation
- A matrix of all AUC is obtained, and a threshold is applied using as reference the AUC values obtained previously in the feature selection process

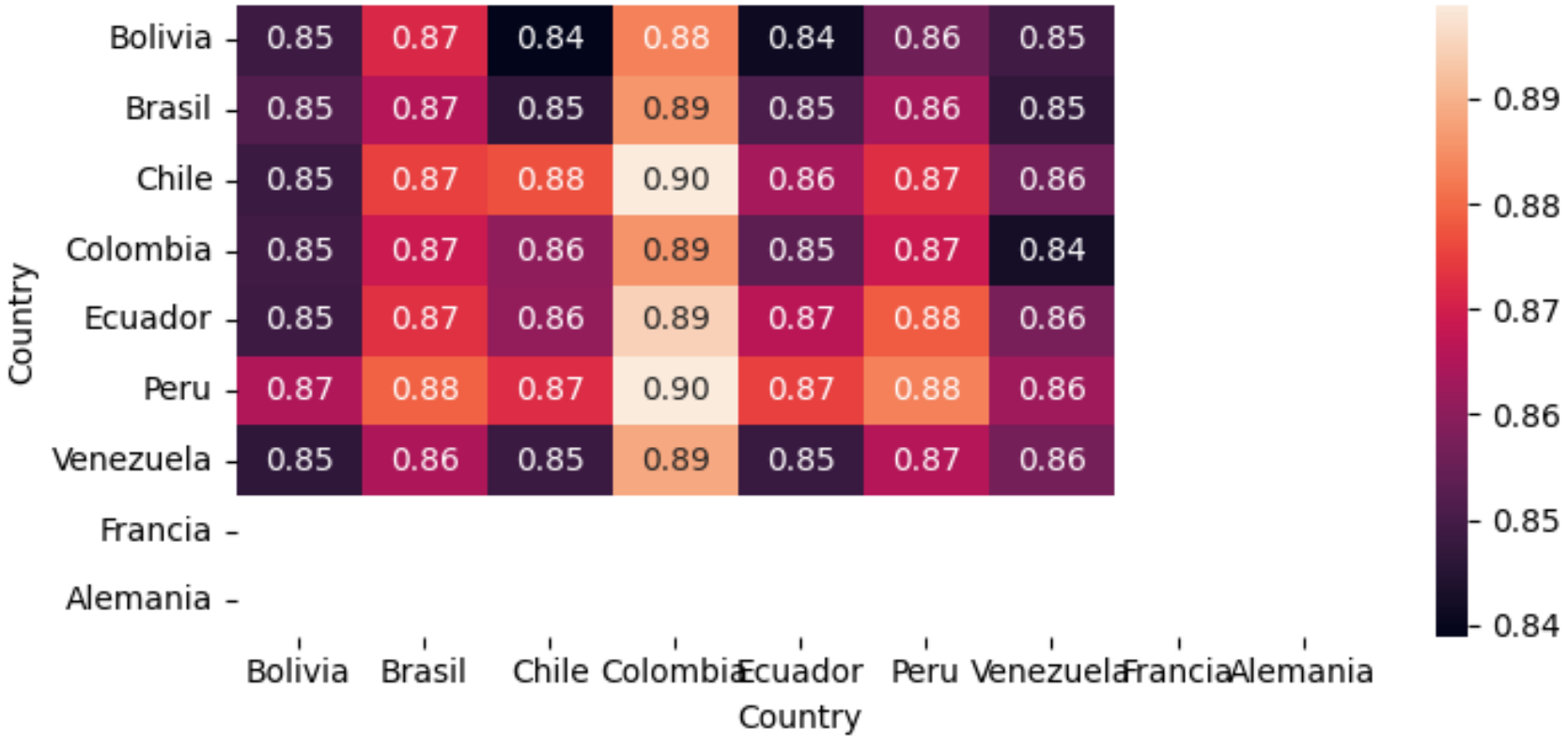
A model-based in a supervised clustering of the country data relationships

- A model is trained using the data from a country and evaluating in the remaining countries (i.e. Training from Brasil evaluating in Colombia,Training from Brasil evaluating in Peru, Training from Colombia evaluating in Brasil...)
- The AUC is computed in each evaluation
- A matrix of all AUC is obtained, and a threshold is applied using as reference the AUC values obtained previously in the feature selection process

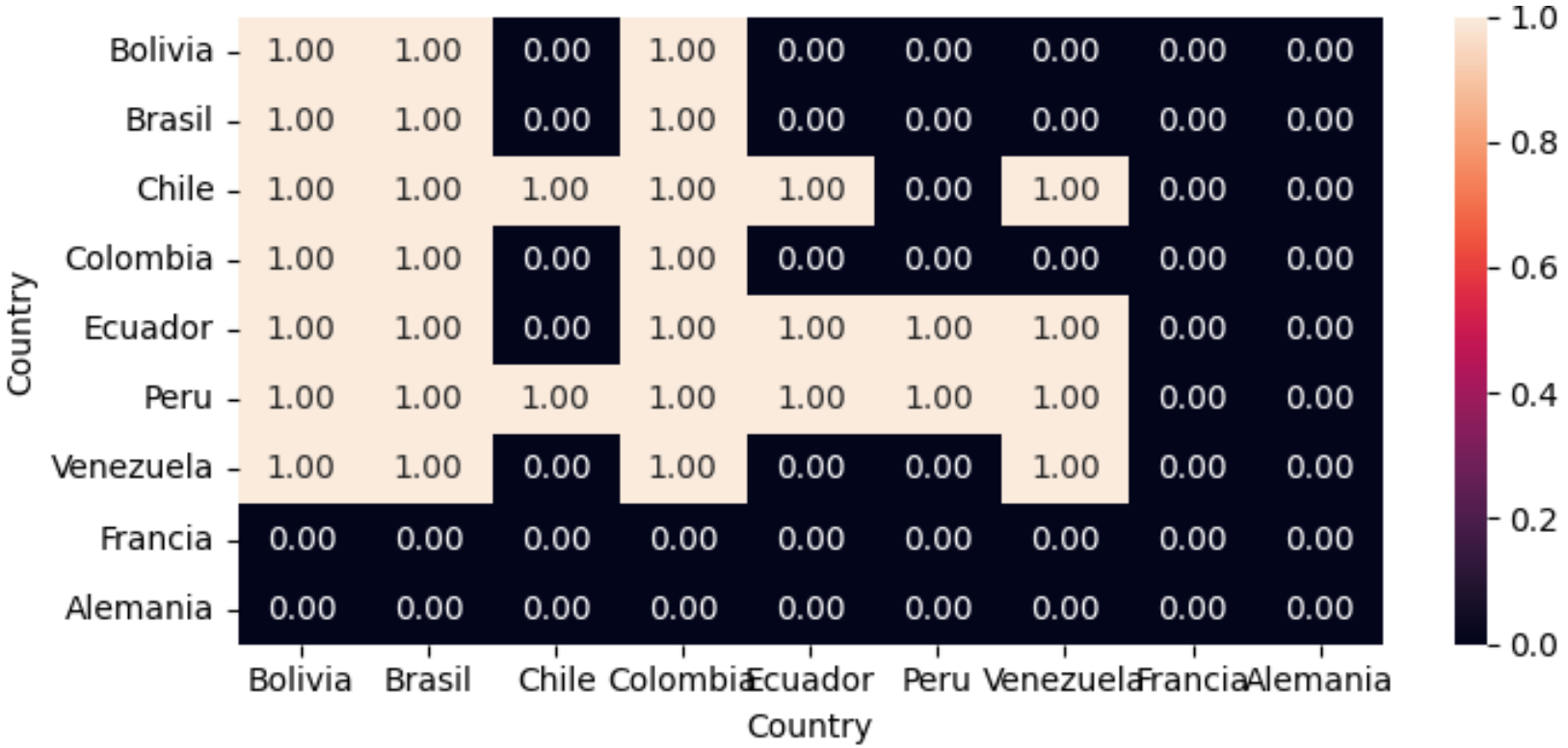
A model-based in a supervised clustering of the country data relationships

- A model is trained using the data from a country and evaluating in the remaining countries (i.e. Training from Brasil evaluating in Colombia,Training from Brasil evaluating in Peru, Training from Colombia evaluating in Brasil...)
- The AUC is computed in each evaluation
- A matrix of all AUC is obtained, and a threshold is applied using as reference the AUC values obtained previously in the feature selection process

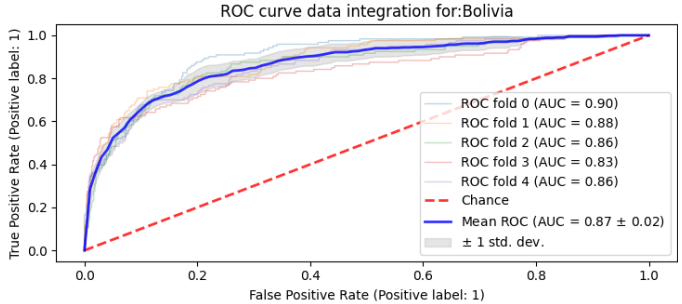
Results of the AUC matrix- Data Relations

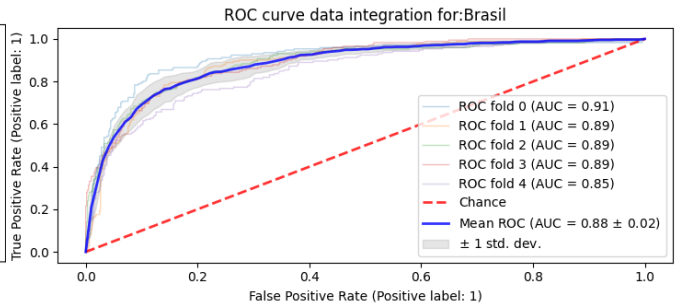


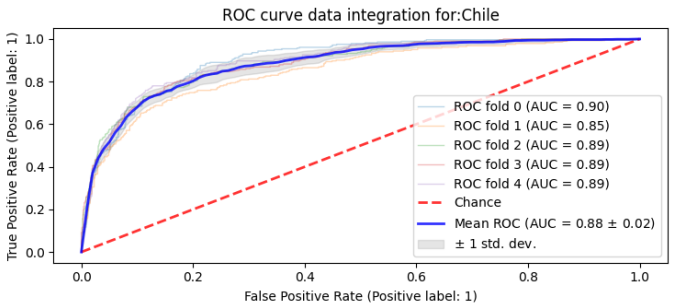
Relationships are revealed after a threshold operation

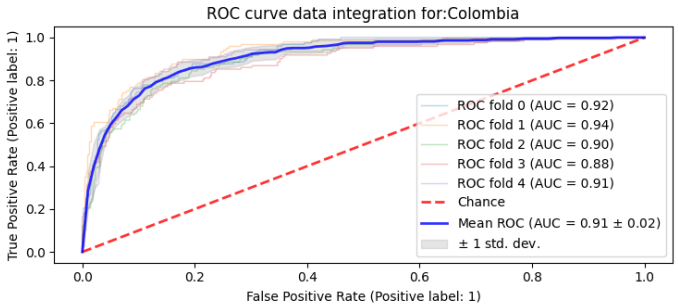


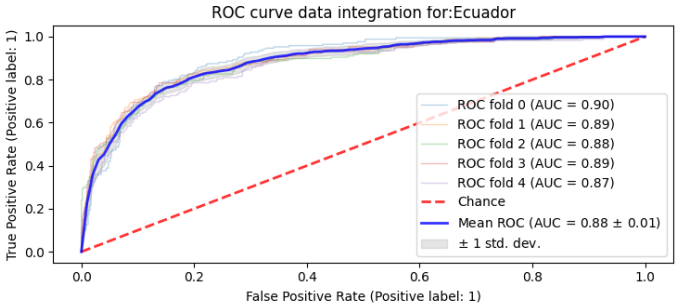
Results per country

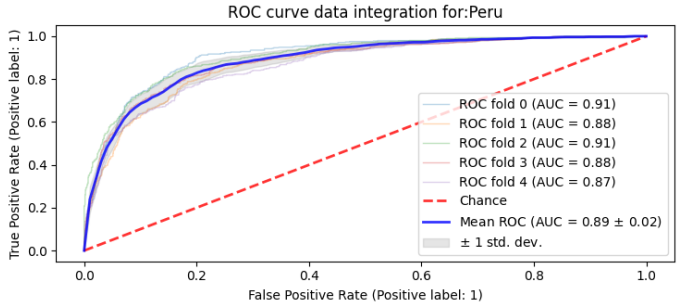


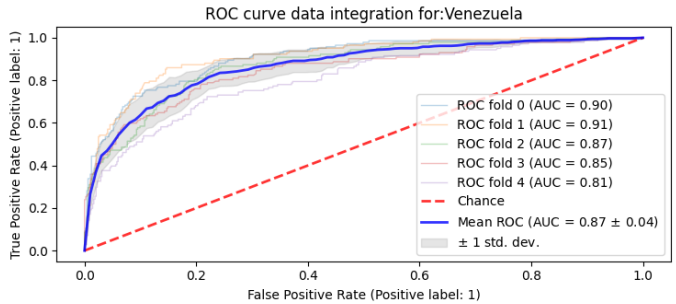




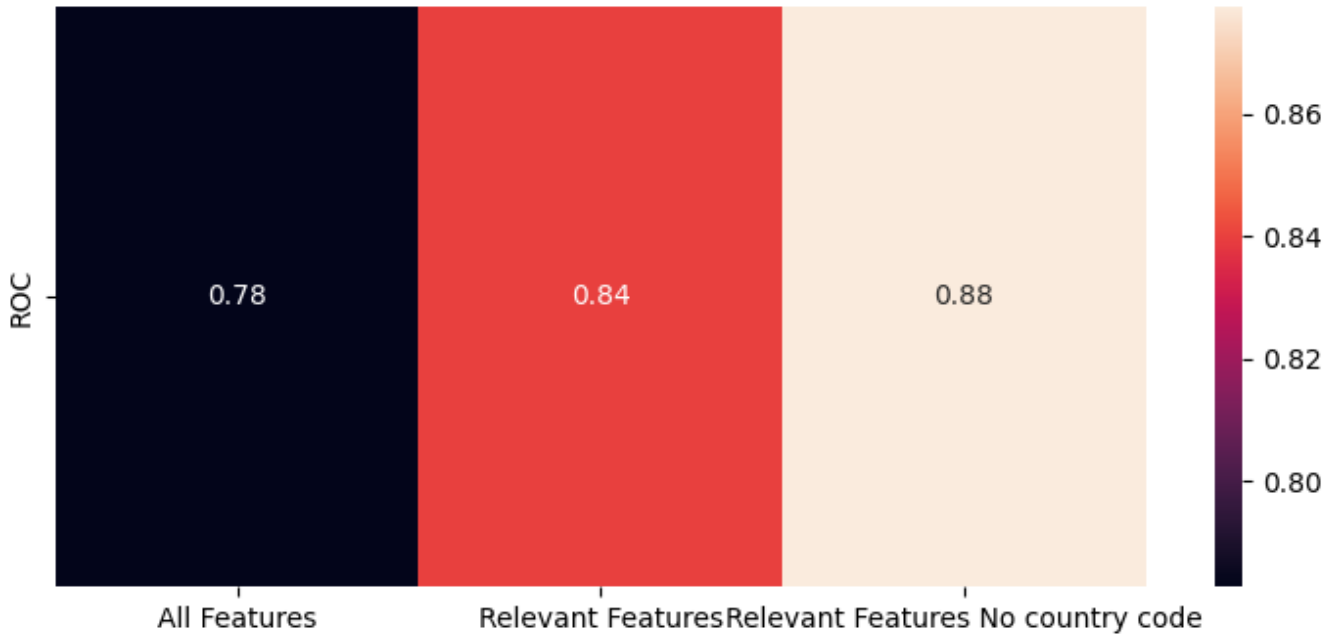




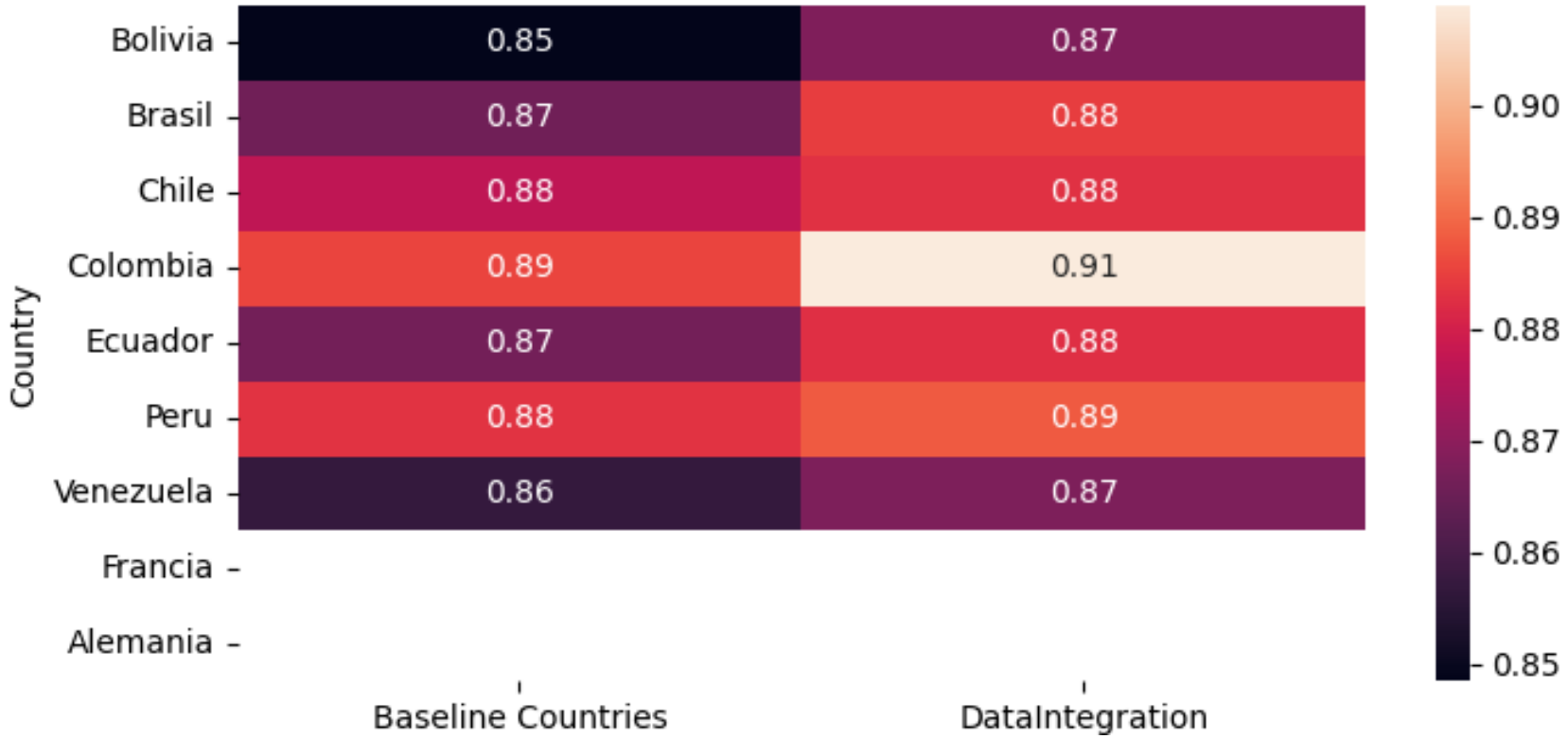




Summary - Results Relevant Features Selection



Summary - Results Data Integration per Country



Results of the Blind Data

	Relevant Features	**Data Integration**
Win	5247	5390
Lose	26576	26433

Conclusions

- Relevant Features enhance the data representation
- Data integration show an improvement in the representation
- Setting a manual threshold can be improved