Bond Clustering

Problem Statement

A “Risk Stripe” is a group of traders who are collectively responsible for trading a set of Corporate Bonds. Only traders which belong to a given Risk Stripe could have the permissions to trade a bond mapped to that Risk Stripe. Each bond is uniquely mapped to a Risk Stripe, and hence, there is a many to one map of bonds to Risk Stripes which we call as “Risk Stripe Map”.

In practice the Risk Stripe Map is correct for bonds that are traded actively in the market. However, overtime, as bonds mature or traders move between organizations, the mapping tends to become inaccurate. This is particularly true for bonds that are traded in less volume.

In this problem, you are given a training data set containing bonds (identified by “ISIN” Column), Bond Properties (attributes about the bond) and the corresponding Risk Stripe. You have to identify the set of bond properties that can be used for clustering similar bonds, and predict the correct Risk Stripe for each bond in the testing set.

Evaluation Strategy:
The models will be evaluated in two stages:

1. Percentage of bonds mapped correctly (on Testing Data for Final Data Set Only)
2. Subjective evaluation of the model

Top teams will be selected based on the number of bonds mapped correctly. These teams will be independently evaluated on the model/strategy they have used.

Input Format

You will be provided with **two sets of data** in the competition.

1. **Initial Data Set:** Given to you at the beginning of the competition. You can use this to build models/strategies for factor identification and clustering.
2. **Final Data Set:** This will be distributed **6** hours before the competition ends. Your model will be evaluated based on its performance on this dataset only.

Each Data Set will contain **2** csv delimited files:

- **Training Data:** Bond Data including correct Risk Stripe
- **Testing Data:** Bond Data excluding Risk Stripe

Note: Factors for clustering bonds may be different for Initial and Final Data Set.

Output Format

Please upload the following in a single zip file:

1. A CSV file named **“output.csv”**, containing the predictions of your model on the **Testing Data of the Final Data Set** given to you 6 hours before the competition closes.
2. All source code files (quoting any references you may have used)
3. A documentation of your solution. This can be in **PDF, PPT, PPTX, DOC or DOCX** formats. Do include the following aspects in your model doc:

- Assumptions
- Any mathematical simplifications/approximations
- Modelling choices and comments on appropriateness of answers
- Any plots/graphs which explain the process of your model selection

Note:

1. Please ensure consistency between output and the methodology you explain in model document.
2. If there are any additional ideas that you wish to include in the documentation which could not be implemented, please clearly mark them so.

Sample Input

Training Data

ISIN	SP_Rating	Moody_Rating	Currency	Industry_Sector	Industry_SubGroup	Issuer_Name	Ticker	Country_Of_Domicile	Risk_Stripe
ISIN0	sp rating0	moody rating0	Currency0	Industry Sector0	Industry SubGroup0	Issuer Name0	Ticker0	Country0	Stripe 0
ISIN2	sp rating2	moody rating1	Currency0	Industry Sector1	Industry SubGroup2	Issuer Name2	Ticker2	Country0	Stripe 0
ISIN3	sp rating2	moody rating1	Currency0	Industry Sector1	Industry SubGroup2	Issuer Name2	Ticker2	Country0	Stripe 0
ISIN5	sp rating3	moody rating2	Currency0	Industry Sector3	Industry SubGroup4	Issuer Name4	Ticker4	Country1	Stripe 0
ISIN13	sp rating1	moody rating0	Currency0	Industry Sector0	Industry SubGroup5	Issuer Name6	Ticker6	Country2	Stripe 1
ISIN14	sp rating1		Currency0	Industry Sector0	Industry SubGroup5	Issuer Name6	Ticker6	Country2	Stripe 1
ISIN18	sp rating4	moody rating3	Currency0	Industry Sector0	Industry SubGroup0	Issuer Name8	Ticker7	Country3	Stripe 0
ISIN19	sp rating1	moody rating4	Currency0	Industry Sector4	Industry SubGroup6	Issuer Name9	Ticker8	Country4	Stripe 0
ISIN20	sp rating5	moody rating5	Currency0	Industry Sector5	Industry SubGroup7	Issuer Name10	Ticker9	Country5	Stripe 2
ISIN21	sp rating5	moody rating6	Currency0	Industry Sector4	Industry SubGroup8	Issuer Name11	Ticker10	Country4	Stripe 0
ISIN22	sp rating6	moody rating6	Currency0	Industry Sector4	Industry SubGroup8	Issuer Name12	Ticker11	Country4	Stripe 1
ISIN23	sp rating5	moody rating6	Currency0	Industry Sector2	Industry SubGroup9	Issuer Name13	Ticker12	Country6	Stripe 0

Download Initial Training Data

Testing Data

ISIN	SP_Rating	Moody_Rating	Currency	Industry_Sector	Industry_SubGroup	Issuer_Name	Ticker	Country_Of_Domicile
ISIN26	sp rating7	moody rating8	Currency0	Industry Sector6	Industry SubGroup10	Issuer Name15	Ticker14	Country7
ISIN28	sp rating8	moody rating8	Currency0	Industry Sector0	Industry SubGroup11	Issuer Name16	Ticker15	Country7
ISIN43	sp rating6		Currency0	Industry Sector0	Industry SubGroup1	Issuer Name26	Ticker23	Country9
ISIN44	sp rating6		Currency0	Industry Sector0	Industry SubGroup1	Issuer Name26	Ticker23	Country9
ISIN61	sp rating6	moody rating6	Currency0	Industry Sector5	Industry SubGroup20	Issuer Name42	Ticker39	Country4
ISIN62	sp rating6	moody rating6	Currency0	Industry Sector5	Industry SubGroup20	Issuer Name42	Ticker39	Country4
ISIN63	sp rating6	moody rating6	Currency0	Industry Sector5	Industry SubGroup20	Issuer Name42	Ticker39	Country4
ISIN64	sp rating6	moody rating6	Currency0	Industry Sector5	Industry SubGroup20	Issuer Name43	Ticker39	Country4
ISIN67	sp rating2	moody rating9	Currency0	Industry Sector0	Industry SubGroup13	Issuer Name45	Ticker21	Country5
ISIN68	sp rating2	moody rating9	Currency0	Industry Sector0	Industry SubGroup13	Issuer Name45	Ticker21	Country4
ISIN69	sp rating10	moody rating11	Currency0	Industry Sector1	Industry SubGroup22	Issuer Name46	Ticker41	Country4
ISIN70	sp rating6	moody rating6	Currency0	Industry Sector0	Industry SubGroup13	Issuer Name47	Ticker42	Country5
ISIN71	sp rating6	moody rating6	Currency0	Industry Sector0	Industry SubGroup13	Issuer Name47	Ticker42	Country5
ISIN72	sp rating6	moody rating6	Currency0	Industry Sector0	Industry SubGroup13	Issuer Name47	Ticker42	Country5

Download Initial Test Data

Sample Output

Output.csv

ISIN	Risk_Stripe
ISIN26	Stripe 0
ISIN28	Stripe 0
ISIN43	Stripe 11
ISIN44	Stripe 11
ISIN61	Stripe 2
ISIN62	Stripe 2
ISIN63	Stripe 2
ISIN64	Stripe 2
ISIN67	Stripe 3
ISIN68	Stripe 3
ISIN69	Stripe 0
ISIN70	Stripe 3
ISIN71	Stripe 3
ISIN72	Stripe 3
ISIN73	Stripe 3

Explanation

Output.csv should include only two columns (ISIN, Risk_Stripe) in that order. The order for bonds should be same as the order in Test Data.