Pattern Identification of Foreign Direct Investment from the Ethiopian Investment Agency Data: Predictive Model in Support of Policy Advocacy Roles

Mahlet Tiruneh
HiLCoE, Computer Science Programme, Ethiopia
almahlet@gmail.com

Tibebe Beshah
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Predicting sustainability status of Foreign Direct Investment (FDI) cancellation has several benefits to the economic development of the country in general and to the Ethiopian Investment Agency (EIA) in particular. This research followed a pattern discovery and data mining approach. The classification modeling technique was found to be the most suitable methodology in mining patterns. Applying experimentation setups using J48, PART and Naïve Bayes in combination with 4-fold, 10-fold and 75/25% data splitting test options, nine different models have been developed.

Further from novel pattern identification, a mining model evaluation technique which is biased to the task of investment policy advocacy is discovered and applied. Thus; the nine predictive models were ultimately evaluated for Accuracy, Error Rate, Receiver Operating Characteristic (ROC), F-Measure and Kappa statistics value. Among them, a model with PART algorithm with 10-fold splitting system was found to be the "Best" in terms of the above listed evaluating measures except the ROC value. Finally, based on eight selected rules from the "Best" performing model, a prototype to a Decision Support System (DSS) was developed. The DSS can assist the policy advocacy tasks to be undertaken by the investment policy advocacy experts at the EIA. In addition to the sustainability status prediction of FDI, the research output is a significant contribution to other researchers and experts in investment domain, mainly FDI.

Keywords: Prediction; Classification; EIA; FDI; Cancellation; Evaluation Parameters; Prototype

1. Introduction

In Ethiopia, significant numbers of FDI investors started investment projects locally but cancel their investment license permit for various reasons. On the other hand, Investment Policy Advocators (IPAs) like the EIA, collect large volume of data from their FDI or international customers in the form of "required information". Here, investment policy advocacy is one task of IPAs which is lobbying or advising the government on policy measures needed to create an attractive investment climate for investors. Therefore, these agencies advocate policies between the investment environment and investment policy makers.

This research tries to address the following problems:

a) Regardless of the government's efforts to create an attractive investment climate for FDI, a significant number of investors ceased their investment project;
b) Although policy advocacy is one of the key IPA's roles, this task is not supported by business intelligence, knowledge discovery techniques, data mining technologies and DSS that could have helped such huge task.

The following list of factors motivated us to work on this data mining research:

a) Inefficiency of the existing non-automated information processing techniques in assisting the EIA's policy advocacy roles;
b) Availability of large volume of foreign investment data in EIA's database in a manner that each record has a number of descriptive attributes;
c) The hope to assist the country's struggle to fulfill the aimed grand strategies through investment.

This study follows both quantitative and qualitative methods. The quantitative method is used to collect and analyze foreign investor's data. On the other hand, the qualitative aspect is used to understand the business operation by making a close relationship with domain experts and the responsible body such as the database administrator of the organization.

Also, this research employs the Cross Industry Standard Process (CRISP) data mining process model in which it strictly follows the six process model named as Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. To implement the experimentation, the classification modeling technique using the J48, PART as well as Naïve Bayes algorithms in Waikato Environment for Knowledge Analysis (WEKA) data mining tool was applied.

2. Related Work

FDI can be defined as an investment made by a firm or an entity based in one country into a firm or entity based in another country. According to the World Bank, FDI is defined as an investment made to acquire a lasting management in an enterprise operating in a country other than that of the investor [1]. FDI is not just a capital movement. In addition to capital, a controlled subsidiary often receives direct input of managerial skills, technology and other tangible and intangible assets. Unlike portfolio investors, direct foreign investors have substantial control over the management of foreign subsidiary [2].

On the other hand, according to [3], FDI cancellation refers to an investment that has been terminated after receiving an investment license by the host country investment agency. It has an impact in terms of political instability, social crises as well as moral aspect of an investor. Investors fail to carry out projects in different parts of the country after they were awarded licenses worth hundreds of millions of dollars.

The issues of FDI cancellation and the necessity of investment policy advocacy, in order to combat withdrawal from FDI business have not been sufficiently explored up to the extent of the problem. Very few works can be quoted. A highly related work was done in [4]. We have studied the benefits of investment policy advocacy in FDI sector selection in Ethiopia. The class association rule mining technique was applied to generate six mining models. The models are capable of FDI sector selection. Finally, the study has concluded that patterns from investment datasets have a lot to provide for the better of the investment climate and policy advocacy.

Thus, regardless of the negative impact of FDI cancellation to the economic growth as well as GDP of the nation, as opposed to other well-examined fields like bankruptcy prediction, financial distress or market prediction, research on the application of data mining techniques for the purpose of predicting FDI permit cancellation has been rather minimal.

3. The Proposed Solution

3.1 Business and Data Understanding

Business understanding attempts are made to understand the core business objective of the organization under study through descriptive statistics of all possible values of the data explaining attribute in the collected investment dataset. Furthermore, discussion with key policy advocacy experts at the agency enabled us to understand the business objective, the problem and limitation they

have in classifying the types of foreign investors who would more likely cancel their investment permit. Based on this, currently EIA uses both non-automated and automated methods aiming at investment policy advocacy area.

The non-automated methods are the following.

a) Investors' front office suggestion box analysis;
b) Researches on specific investment complaints;
c) Collecting & addressing comments about the operation of investment promotion activities, as received from investors and their officers.

The automated method is simple statistical analysis of investment data using MS Excel.

The problems with the above listed methods of policy advocacy are the following.

a) The non-automated information processing methods may only help to make some "quick fixes" around the front office operations of the agency through improving its day to day operational activities than effectively assisting the task of policy advocacy.
b) The only automated information processing method, that is statistical manipulation of investment data using MS Excel spreadsheets, and reporting same, is somehow assisting the agency, in policy advocacy so far.

Thus, those methods which are currently in use by the agency have little or no power in mining interesting and novel patterns, finding non-trivial classification, and discovering hidden knowledge from the huge accumulation of investment data.

When it comes to data understanding, facts in the EIA's investment dataset are:

a) Collected in the form of MS Excel spreadsheets;
b) Contains 20 years (Jan. 1992-Dec. 2013) of investment data having 61,059 individual investment records;
c) Collected from Addis Ababa EIA (main investment office);
d) Each record represents individual investment projects;
e) Each record contains 13 data explaining attributes;
f) Each attribute contains one or more possible values;
g) Each attribute has categorical (text) data type.

3.2 Preprocessing

At this stage, data cleaning routines are applied to fill in missing values (with the mean value), smooth out noise (by removing the records), and detect outliers (by removing or substituting with mean values) in the data. The cleaned data is further processed by feature selection consulting the domain experts and the WEKA attribute selection preprocessing techniques (to reduce dimensionality) and derivation of new attributes. The result of these processes generates the dataset for training and testing the classification approach selected in this study.

Twenty years data (January, 1992-December 2013) with a total of 61,059 individual investment records were collected from the EIA. But, we have taken a reduced sample of 37,126 records even before cleaning due to sampling techniques of choosing recent five years data (Jan. 2008-Dec.2013). Once the attribute selection has been done, two types (Dimensional/Horizontal or Data Size/Vertical) of data reduction also have been applied on the dataset.

Here, the data reduction phase is dependent upon the attribute selection which makes the data to further lower to 6,050 records with selected attribute number of five. All the remaining attributes are believed to be less relevant to the research due to the fact that it is not self-explanatory/descriptive attribute.

Thus, even if the crude data had a total of 13 explanatory attributes; only five of them (Form_of_Company, Type_of_Investor, Type_of_Investment_License, Economic_Sector, Investment_Status) were found to be relevant for the purpose of the study.

Therefore, the attributes of the dataset with their data type and descriptions are as follows (see Table 1).

Table 1: Attribute of Investors Profile Data

Attribute	*Possible values*	*Data type*	*Description*
Form_of_Company	Sole/PLC/Share/ Public	Categorical	Categorizes the source of investment capital
Type_of_Investor	Domestic/Wholly Foreign/Joint with Domestic/Public/Ethiopian by Birth/Foreign but Domestic	Categorical	Categorized investment projects based on owner ship nationality
Investment _Type	Domestic/Foreign/Public	Categorical	Categorized the source of investment capital
Type_of_Investment_License	New/Expansion	Categorical	Categorized license of investment in terms of initial or modified
Economic_Sector	Primary/Secondary/Tertiary	Categorical	Categorized investment projects based on the sector they invest
Investment_Status	Pre-Implementation/ Implementation/Operation	Categorical	Categorized status of investment in terms of operation
Termination	Yes/No	Categorical	Type of Foreign Investor (terminate/Not terminate)

Similarly, for research of this type, it is necessary to see the regional state of Ethiopia as one major factor for the foreign investment license cancellation. Therefore, from attribute type Region, attribute name "Region_Type" is created. Hence, the new attribute of final dataset based on foreign direct investment permit cancellation are summarized in Table 2.

Table 2: Attribute of "Final Dataset"

No.	*Attribute Name*	*Attribute Description*	*Attribute type*	*Values*
1.	Form_of_Company	Source of Investment capital	Categorical	Plc/Share/Sole
2.	Type_of_Investor	Categorized investment projects based on ownership nationality	Categorical	Joint with Domestic, Wholly Foreign
3.	Type_of_Investment_License	Initial or modified investment	Categorical	New, Expansion
4.	Economic_Sector	Sector of Economy	Categorical	Primary, Secondary, Tertiary
5.	Investment_Status	Operation of investment	Categorical	Operation, Implementation, pre-Implementation
6.	Region_Type	Geographical location of area based on budget allocation	Categorical	City Administration, Developing Regions, Multi- Regional, Others
7.	Termination	Sustainability in the project	Categorical	Yes, No

Since the selected tool (WEKA 3.6.4) doesn't accept the data in Excel format, we converted the data in comma separated version (CSV) of text file in Attribute Relation File Format (ARFF) and finally the data was used for the experimentation.

3.3 Experimental Settings and Results

In general, an experiment is conducted by finding the steps needed to utilize J48, PART as well Naive Bayes algorithms for classifying FDI dataset, discovering rules generated from the dataset and the

meaning of them. For this reason, the whole cleaned investment dataset (6050) is fed to the WEKA tool using the 4 fold, 10 fold and 75/25% splitting techniques.

According to [5], test dataset is used to estimate accuracy of classification rules. So the data has to be different from the training dataset; otherwise over fitting will occur. For this reason, for the test dataset also, three consecutive data mining experimentations, each with three interrelated sub-experiments, are made. Each experiment is conducted on different dataset of the investment data. Finally, the experimentation result is summarized in Table 3.

Table 3: Experimentation Result Generated from the Nine Predictive Models

Experiment no.	*Experiment undertaken*		*Model Results*				
	Algorithms	*Test options*	*Accuracy*	*Error rate*	*ROC area*	*F-measure*	*Kappa statistics*
Model 1	J 48	4-fold	4621/76.38%	1429/23.62%	0.698	0.726	0.226
Model 2	J 48	10-fold	4621/76.38%	1429/23.62%	0.698	0.726	0.226
Model 3	J 48	75/25%	4621/76.38%	1429/23.62%	0.698	0.726	0.226
Model 4	PART	4-fold	4622/76.39%	1428/23.60%	0.697	0.726	0.226
Model 5	PART	10-fold	4622/76.39%	1428/23.60%	0.697	0.726	0.226
Model 6	PART	75/25%	4622/76.39%	1428/23.60%	0.697	0.726	0.226
Model 7	Naïve Bayes	4-fold	4578/75.67%	1472/24.33%	0.721	0.715	0.194
Model 8	Naïve Bayes	10-fold	4578/75.67%	1472/24.33%	0.721	0.715	0.194
Model 9	Naïve Bayes	75/25%	4578/75.67%	1472/24.33%	0.721	0.715	0.194

3.4 Model Evaluation and Selection

Further from novel pattern identification, a mining model evaluation technique which is biased to the task of investment policy advocacy is discovered and applied. Therefore, the nine predictive models were ultimately evaluated for Accuracy, Error Rate, ROC, F-Measure and Kappa statistics value.

In general, the PART generated models are found to be the "Best" in terms of their highest accuracy rate (4622/76.3967%), lowest ("Best") error rate (1428/23.6033%), lowest ("Worst") ROC area (0.697), "Best" F-measure (0.726) and Kappa statistics (0.226; similar to J48) in all three test options (see Tables 4 and 5).

Based on Table 4, conclusion has been made to select the "Best" model (see Table 5).

Table 4: Model Result Measurements

Evaluation Parameter	*Best*	*Better*	*Worst*
Accuracy	4622/76.39%	4621/76.38%	4578/75.67%
Error	1428/23.60%	1429/23.62%	1472/24.33%
ROC	0.721	0.698	0.697
F-measure	0.726	None	0.715
Kappa Statistics	0.226	None	0.194

Table 5: Model Evaluation

Evaluation Parameter	*Model 1 (J48-4 Fold)*	*Model 2 (J48-10 fold)*	*Model 3 (J48-75/25%)*	*Model 4 (PART-4 fold)*	*Model 5 (PART-10 fold)*	*Model 6 (PART-75/25%)*	*Model 7 (Naïve Bayes-4 fold)*	*Model 8 (Naïve Bayes-10 fold)*	*Model 9 (Naïve Bayes-75/25%)*
Accuracy	Better/ 76.38%	Better/ 76.38%	Better/ 76.38%	Best/ 76.39%	Best/ /76.39%	Best/ 76.39%	Worst/ 75.67%	Worst/ 75.67%	Worst/ 75.67%
Error	Better/ 23.62%	Better/ 23.62%	Better/ 23.62%	Best/ 23.60%	Best/ 23.60%	Best/ 23.60%	Worst/ 24.33%	Worst/ 24.33%	Worst/ 24.33%
ROC	Better/ 0.698	Better/ 0.698	Worst/ 0.698	Worst/ 0.697	Worst/ 0.697	Worst/ 0.697	Best/ 0.721	Best/ 0.721	Best/ 0.721
F-measure	Best/ 0.726	Best/ 0.726	Best/ 0.726	Best/ 0.726	Best/ 0.726	Best/ 0.726	Best/ 0.715	Worst/ 0.715	Worst/ 0.715
Kappa	Best/ 0.226	Best/ 0.226	Best/ 0.226	Best/ 0.226	Best/ 0.226	Best/ 0.226	Worst/ 0.194	Worst/ 0.194	Worst/ 0.194

Accordingly, even though models 4, 5 and 6 have similar "Best" results in all evaluation parameters, comparing the test options, k-fold validation usually results in better performance than using percentage split [6]. Therefore, the model with 4 and 10 fold test options (i.e., models 4 and 5) are selected to be better than the one with the percentage split (i.e., model 6). Further evaluation also has been necessary to select one of model 4 or 5.

Here, as several empirical findings indicate, a common choice for K is 10 (10 fold is believed to be default for several experiments). In [6] with larger folds the bias of true error rate estimator will be small (the estimator will be very accurate), the variance of the true error rate estimator will be large and the computational experiments will be many. Then, cross validation 10 folds, K results, from the folds can be averaged (or otherwise combined) to produce a single estimation.

Therefore, model 5 (i.e., PART algorithm with 10-fold splitting system) is selected as the final model due to its "Best" evaluation result as well as the test option it was created on. The final eight WEKA generated PART Rules are shown in Figure 1.

1. Investment_Status = Pre_Implementation AND Economic_Sector = Tertiary: No (1941.0/486.0)
2. Investment_Status = Pre_Implementation AND Economic_Sector = Primary: No (1189.0/224.0)
3. Investment_Status = Pre_Implementation AND Region_Type = Others: No (732.0/225.0)
4. Investment_Status = Operation: No (602.0/13.0)
5. Investment_Status = Implementation: No (583.0/74.0)
6. Region_Type = City_Administration AND Type_Of_Inv_License = New AND Form_of_Company = Sole: Yes (483.0/231.0)
7. Region_Type = City_Administration: No (349.0/131.0)
8. Region_Type = Multi_Regional: Yes (159.0/41.0)

Figure 1: The Final Selected Eight Classification Rule from the "Best" Model

Finally, based on the chosen model, a prototype to a DSS which has the capacity to assist policy advocacy tasks of investment promotion experts at the EIA is developed

3.4 Model Deployment

This is the final process of the CRISP data mining life cycle and called deployment stage. The step consists of developing tools and techniques of where

and how the discovered knowledge will be used. Finally, the whole knowledge endeavor becomes fruitful.

Therefore, under this deployment section, the final chosen model is designed and FDI sustainability status predicator system is developed by converting the output of the model to a prototype system using Java programming language.

The system functionalities provided by the system are shown in Figure 2.

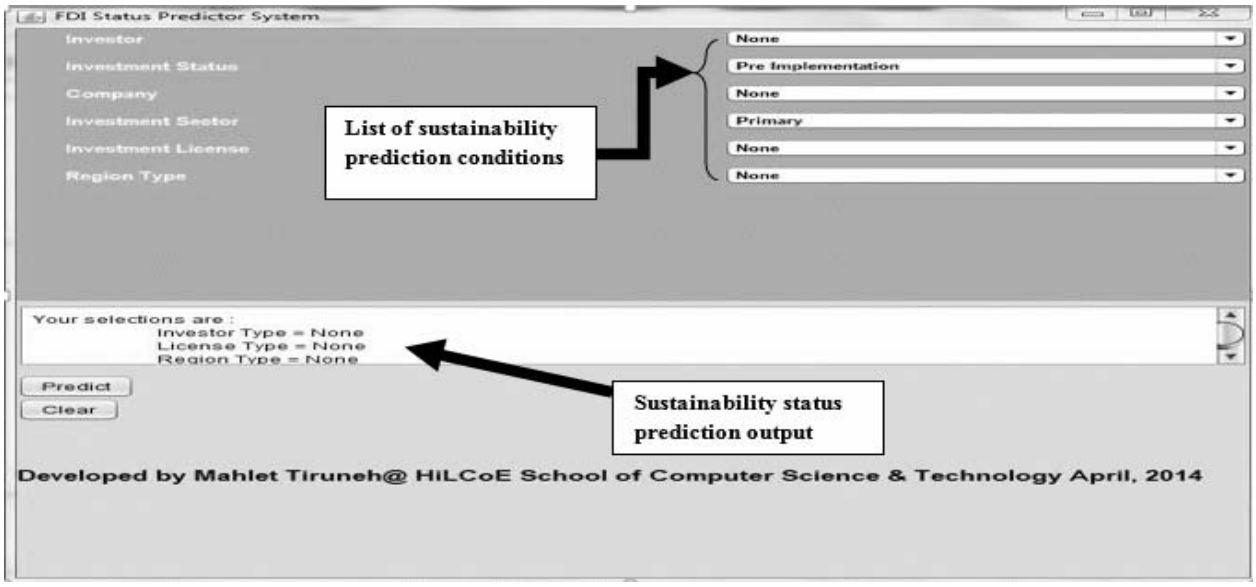

Figure 2: Snapshot Graphical User Interface of the FDI Sustainability Status Predictor Prototype System

4. Conclusion and Recommendations

4.1 Conclusion

The ultimate objective of the research is identification of novel mining patterns from investment dataset, building a predictive model, and finally to build a DSS which is capable of predicting sustainability status of FDI.

The investigation is carried out using CRISP data mining process model which strictly followed the six process model named as Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

The data used in the research has been gathered from the main database of the EIA covering the period from 1992 to 2013. Then, the data has been preprocessed and prepared in a format suitable for the data mining tasks by applying all the relevant pre-processing tasks to it. The final cleaned dataset constituted 6050 investment records and seven attributes. Among them, one is newly created and originally non-existent in the dataset.

Applying the J48, PART as well as Naïve Bayes algorithms using 4 fold, 10 fold as well as 75/25% data splitting techniques on the EIA's investment dataset, nine mining models were built and they all are generated using the WEKA data mining tool. Among the nine mining models, each of J48 generated, PART generated and Naïve Bayes generated models were used.

Once the nine predictive prediction classification models were developed, the selection of one "Best" model in turn demanded the use of applying an evaluation ground through comparison of the performance of Accuracy, Error Rate, ROC area, Kappa statistics and F-measure for each of the nine predictive models. Several empirical findings indicate these evaluation parameters are useful in evaluating and selecting a model which is most biased towards investment policy advocacy. Thus, the nine mining models are rated based on the extent to which they address the five constraints, based on the evaluation parameters PART algorithm with 10 fold model is selected.

Finally, as a solution to the decision making problem in the business domain, a prototype of a decision support system is developed based on the results from the mining models. The developed DSS can assist EIA's investment policy advocators in their policy advocacy roles in order to minimize FDI investment cancellation.

4.2 Recommendations

Data mining offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. The availability of new data mining algorithms, however, should be met with caution. First of all, these techniques are only as good as the data that has been collected. Good data is the first requirement for good data exploration. Assuming good data is available, the next step is to choose the most appropriate technique to mine the data. However, there are tradeoffs to consider when choosing the appropriate data mining technique to be used in a certain application. There are definite differences in the types of problems that are conductive to each technique. The "Best" model is often found by trial and error: trying different technologies and algorithms

Furthermore, the current computer science and data mining technologies can be integrated with the tasks of investment, in general and investment policy advocacy roles undertaken by IPAs in particular. Therefore, a number of investment business oriented technology tools, decision support systems, novel patterns, and applicable models should be created and applied to the required purposes.

Also, as an IPA, the EIA should strive for the better of its policy advocacy rolls. This in turn can be achieved only when more research and exploration to the huge accumulation of data at the agency's custody is undertaken. The more the data is explored in detail, the better the solutions to any existent problems in the investment advocacy and other investment related areas can be discovered.

As discussed earlier, the sustainability of investors in FDI projects will attract other foreign investors to come and invest in the country. Ultimately, this will enhance the country's economic development through knowledge and technology transfer to the host country. Therefore, the agency should strive to demand and support a lot of researches of this kind.

References

[1] F. I. A. Service, Investment Promotion Training: Data Collection and Analysis Course Manual, July 2002.

[2] J. C. Anyanwu, Promoting Investment in Africa, African Development Bank, UK and USA, 2006.

[3] OECD, "Policy Framework for Investment User's Toolkit Chapter 2. Investment Promotion and Facilitation," ed: A publication of the Investment Division of the OECD Directorate for Financial and Enterprise Affairs, 2011.

[4] W. J. Biruck, "Mining Patterns from Investment Data: DSS in support of EIA's Investment Policy Advocacy Roles," HiLCoE School of Computer Science, Addis Ababa, 2014.

[5] S. M. Veronica, "Towards the use of C4.5 Algorithm for Classifying Banking Dataset," Vol. 8, October 2003.

[6] P. Ozer, "Data Mining Algorithms for Classification," Artificial Intelligence, Radboud University, Nijmegen, 2008.