International Journal of Engineering & Technology, 7 (4.32) (2018) 90-92

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predictive Analytics as A Service on Moroccan Tax Evasion

Houda Jihal[1*], Mohamed Amine Talhaoui[1], Abderrahmane Daif[1], Mohamed Azzouazi[1]

[1] LTIM Laboratory, Faculty of science Ben Msik, Hassan II University Casablanca, Morocco
**Corresponding author E-mail: houda.jihal@gmail.com*

Abstract

Tax evasion is a global problem in governments. It affects society by damaging public accounts and compromising government performance. The government must take a multidisciplinary approach to face this phenomenon. Analytics on big data enables government organizations to improve existing processes and operations and engage in entirely new types of analyses that weren't possible before. Predictive analytics combines the capabilities of machine learning, statistical analysis and data mining to forecast the future and allows tax authority to prevent tax fraud, reduce the cost of managing taxes and optimize public spending.
The purpose of this paper is to predict income from direct Moroccan taxes based on the linear regression model as a first step in the fight of the tax evasion.

Keywords: *Predictive Analytics, linear regression, Statistical Computing, Tax evasion.*

1. Introduction

1.1 Moroccan tax administration

The general tax administration (DGI) is part of the Ministry of Economy and Finance as the central authority of tax policy and administration in Morocco. Ensure tax revenue plays a critical part in the development of the country. The Moroccan system is generally declarative; the mission of control therefore occupies a very important dimension to guarantee a quality service to the citizen.
DGI is also an institution that suffers from a significant lack of resources in relation to its mission and the actual fiscal potential of its environment. The number of human resources indicated in the end of the activity report is very low compared to other countries around the Mediterranean. It is in Morocco of 1.4 for 10,000 inhabitants. It is 9 for 10,000 inhabitants in France and 4 for 10,000 inhabitants in Tunisia.
Predicting income and evaluating the cost of managing tax can be an interesting track to explore to better optimize public spending. Nevertheless, a multidisciplinary approach is essential to better understand the resistance and to act in the right direction.
Managing taxation need full visibility of all activity related to taxpayers. Since 2010, the DGI has started the process of automating these different services. In 2016 the DGI set up a risk analysis system for the automatic selection of declarations with high scores revaluing a significant fiscal risk. The taxpayer files his declaration in the integrated taxation system SIT. The SIR system is responsible for grouping all the information of the taxpayer from the partners of the DGI. Then it is up to the SAR analytical system to process the information and establish a score that will allow officers to decide on the type of control or the compliance of the taxpayer. But the system had a false positive problem.
The paper is organized as follows: Section 1.2 presents related works analysis in tax fraud. Section II is the results and discussion part. Section III describes the different tools and experiments carried out to obtain the results. In section VI, we summarize the main conclusions and future research.

1.2 Related Work

There is an evidence that every government count on tax to improve the social level of citizens and to cover public spending. Tax is a tool of social policy and equity between citizens. M.Amori [1] Test the impact of the tax system on economic growth in Morocco. He analyzes the interaction relationship between changes in tax revenues and GDP, test empirically the impact of taxation by an econometric model of economic growth. In his model, two variables were adopted to represent the effect of taxation: Tax-ToGDP Ratio and tax reform effect.
E. Earley [2] explain how data analytics applies to financial statement audits and why it could represent a change in how audits are conducted and provide a context for researchers in terms of problems to be addressed related to data analytics. G.L. Gray and R.S. Debreceny [3] explore the application of data mining techniques to fraud detection in the audit of financial statements and propose a taxonomy to support and guide future research. V. Chandola and al [4] provide a structured and comprehensive overview of the research on anomaly detection.
Wu and al [5] Used association rules data mining to develop a screening framework to detect possible non-compliant value-added tax (VAT) reports that may be subject to further auditing. F. Tian and T. Lan [6] generate a Taxpayer Interest Interacted Network (TPIIN) with employing a graph-based method to characterize property that describes two suspicious relationship trails with a same antecedent node behind an Interest Affiliated (IAT) and propose a colours network-based model (CNBM) for characterizing economic behaviours, social relationships and the IATs between taxpayers. D. Bogdanov and al [7] build a tax fraud detection system prototype that uses secure multi-party computation (SMC) to remove the companies concerns over confidentiality. S. Basta and al [8] proposed an auditing methodology for detecting Italian fraudulent VAT credit claims based on a rule-based system, which is capable of trading among conflicting issues, such as maximizing audit benefits, minimizing false positive audit predictions, or deterring probable upcoming frauds. S.Kishore

Babu [9] proposed a solution to the application fraud detection in income tax data based on a model that uses Gaussian process with varying hyper parameters. S. Yaram [10] focuses on the implementation of both document clustering algorithm and a set of classification algorithms (Decision Tree, Random Forest and Naïve Bayes), along with fraud detection. Rahimikia and all [11]used a hybrid intelligent system that combines multilayer perceptron (MLP) neural network, support vector machine (SVM), and logistic regression (LR) classification models with harmony search (HS) optimization algorithm to detect corporate tax evasion for the Iranian National Tax Administration .

2. Results and Discussion

Figure 1 shows the process followed to choose the appropriate model, after data collection, we chose to divide the Dtax column on 100, then we prepared the training and testing data for model running, finally we used two performance measures (NRMSE, COD) to choose the best model in our case.

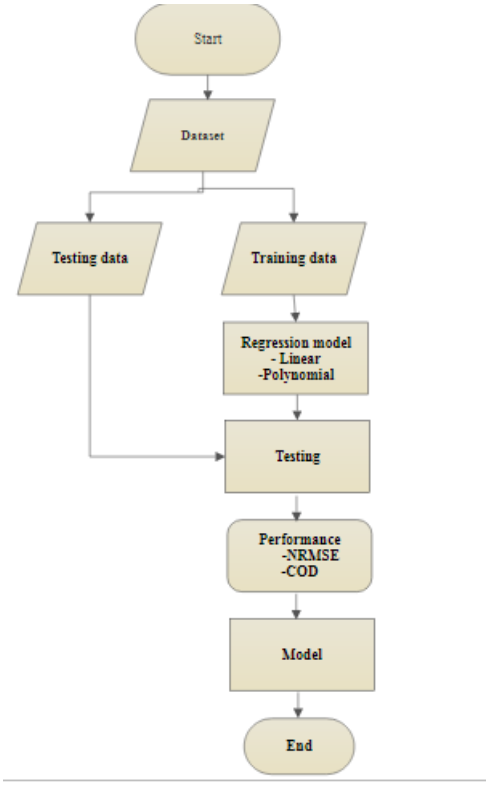

Fig.1: Flow chart of the system

Waiting for collaboration with the tax authorities, we chose in the first place to create a tax revenue database from the activity report available at the DGI portal.
The dataset contain CSV file, this is a brief overview of the dataset attributes:

Table1: Tax Income

Features	Description
year	Year [1996-2016]
Dtax	Direct tax = Corporation tax+ income tax
Vat	Value added tax
Dti	Tax on stump duty

The training data sets are taken into Xtrain and Ytrain vectors. Xtrain vector contains the year column values of training data set and Ytrain vector contains the direct income data of training data set. The testing data set also have been into Xtest and Ytest vectors. A sequence of points is generated, at which linear regression model Fig2, and Polynomial regression model Fig3 functions will be applied.

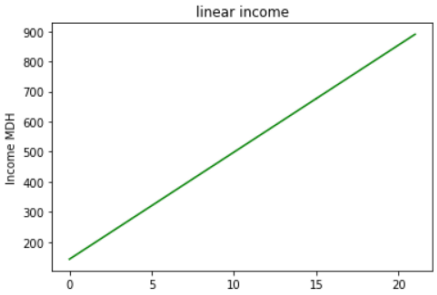

Fig.2: Linear graph plot

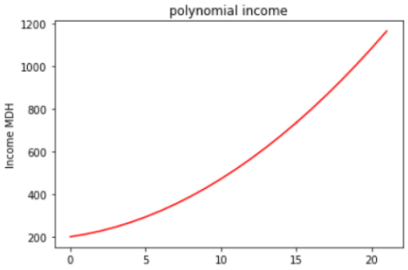

Fig.3: Polynomial graph plot

After running the models, we utilize the NRMSE and COD test to test the performance. We obtain this table:

Table2: NRMSE and COD values with linear and polynomial model

Model	NRMSE	COD		
Linear	25.81	0.82		
Polynomial	25.81	-4.07		

Table2 compares the performance of the two models.
By observing this table, the model product the same NRMSE value but different COD.
In this study the linear model yielded good results.

3. Experimental

3.1 predictive analytics

Data analytics, or DA, is a process for analyzing sets of data to guide business decisions and test scientific theories using techniques from statistics, machine learning, artificial intelligence and data mining. Data analytics can be categorized into predictive, descriptive and prescriptive models. As explained in [12], predictive models can find relationship between outcome and dependent variables. Many techniques have been developed for predictive modeling such as SVM, Bayesian methods, neural networks, regression models, k-NN, uplift models and decision trees. But ensemble models proved to be achieving good accuracy when compared to others, reason being, they train several similar models and combines results so that a best model can be derived to predict new data.

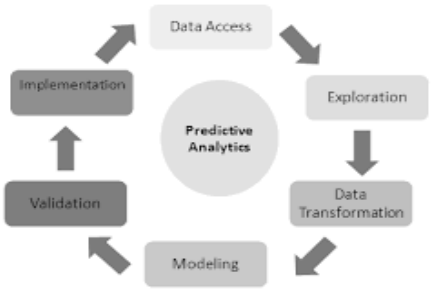

Fig.4: Predictive analytics process

3.3 linear Regression

In this study LRT is being used to estimate the tax incomes for next years. LRT is appropriate to evaluate the strength of a relationship between two variables [12].
In general, regression is the problem of estimating a conditional expected value. While as "linear" refers to the assumption of a linear relationship between y and x. Thus, in statistics, linear regression is a method of estimating that linear relationship between the input data and the output data. The common formula for a linear relationship used in this model is (1).

$$Y = \alpha + \beta x \tag{1}$$

Where Y is the response variable, X is the predictor variable, α is the bias coefficient and β is the coefficient for the predictor column.
A learning technique is used to find a good set of coefficient values. Once found, different X values can be plugged to predict the Y values.

3.3 Polynomial regression

Polynomial regression is a form of linear regression in which the relationship between the independent variable *x* and the dependent variable *y* is modelled as an *nth* order polynomial.

3.4 performance measures

Normalized Root Mean Square Error (NRMSE can be calculated using Eq(2).

$$NRMSE = \sqrt{\frac{1}{N}\frac{\Sigma_i (d_i - y_i)^2}{\Sigma_i (d_i)^2}} \tag{2}$$

Coefficient of determination (COD) can be calculated using Eq(3).

$$COD = R^2 = 1 - \frac{\Sigma_i (x_i - y_i)}{\Sigma_i \left(x_i - \bar{x}_i \right)^2} \tag{3}.$$

4. Conclusion

This paper presented a study about building a predictive model for income tax. Linear and polynomial model were successfully used and the performance was measured by the Normalized Root Mean Square Error (NRMSE) and the Coefficient of determination (COD) to compare and select the best analysis, in our case the linear model was the best.
This paper is a first step in the tax evasion prevention. In future we will work on the optimization of the detection fraud system, ensuring confidentiality of the data and false positive elimination.

References

[1] Amori Mohammed, El Mokhtar Zbair (2016) 'Système Fiscal et Croissance Économique, Etude Empirique : Cas Du Maroc', Sciences Economiques, Université Mohammed V, and Sciences Juridiques 18 (2): 438–44.

[2] Earley, Christine E (2015) 'Data Analytics in Auditing: Opportunities and Challenges '*Business Horizons* 58 (5). "Kelley School of Business, Indiana University": 493–500.

[3] Gray, Glen L., and Roger S. Debreceny (2014) 'A Taxonomy to Guide Research on the Application of Data Mining to Fraud Detection in Financial Statement Audits', *International Journal of Accounting Information Systems* 15 (4). Elsevier Inc.: 357–80.

[4] Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009) 'Anomaly Detection', *ACM Computing Surveys* 41 (3): 1–58.

[5] Wu, Roung Shiunn, C. S. Ou, Hui Ying Lin, She I. Chang, and David C. Yen(2012) 'Using Data Mining Technique to Enhance Tax Evasion Detection Performance', *Expert Systems with Applications* 39 (10): 8769–77.

[6] Tian, Feng, Tian Lan, Kuo Ming Chao, Nick Godwin, Qinghua Zheng, Nazaraf Shah, and Fan Zhang(2016) 'Mining Suspicious Tax Evasion Groups in Big Data', *IEEE Transactions on Knowledge and Data Engineering* 28 (10): 2651–64.

[7] Dan Bogdanov, and J Marko (2017)'Financial Cryptography and Data Security',9603: 227–34.

[8] Basta, Stefano, Fabio Fassetti, Massimo Guarascio, Giuseppe Manco, Fosca Giannotti, Dino Pedreschi, Laura Spinsanti, Gianfilippo Papi, and Stefano Pisani (2009) 'High Quality True-Positive Prediction for Fiscal Fraud Detection', *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, 7–12.

[9] Babu, S Kishore, and S Vasavi (2017) 'Predictive Analytics as a Service on Tax Evasion Using Gaussian Regression Process'.

[10] Yaram, Suresh. (2016) 'Machine Learning Algorithms for Document Clustering and Fraud Detection', *2016 IEEE International Conference on Data Science and Engineering (ICDSE)*.

[11] Rahimikia, Eghbal, Shapour Mohammadi, Teymur Rahmani, and Mehdi Ghazanfari (2017) 'Detecting Corporate Tax Evasion Using a Hybrid Intelligent System: A Case Study of Iran', *International Journal of Accounting Information Systems* 25. Elsevier Inc.: 1–17.

[12] Bakar, Zuriana Abu, Rosmayati Mohemad, Akbar Ahmad, and Mustafa Mat Deris(2006) 'A Comparative Study for Outlier Detection Techniques in Data Mining'