Detecting Atypical Behaviors of Taxpayers with Risk of Non-Payment in Tax Administration, A Data Mining Framework

Ordóñez, José[1]; Hallo, María[1,*]

[1]*Escuela Politécnica Nacional, Facultad de Ingeniería de Sistemas, Quito, Ecuador*

Abstract: One of the primary processes in tax administration is debt collection management. The objective of this process, among others, is to recover economic resources that have been declared by taxpayers. Due to limitations in tax administration such as staffing, tools, time, and others, tax administrations seek to recover debts in the early stages of control where collection costs are lower than in subsequent stages. To optimize the debt collection management process and contribute to decision-making, this study proposes a deep learning-based framework to detect atypical behaviors of taxpayers with a high probability of non-payment. Normal and atypical behavior groups were also analyzed to identify interesting events using association rules.

Keywords: Data mining, debt management analysis, machine learning, patterns of taxpayer behaviors

Detección de Comportamientos Atípicos de Contribuyentes con Riesgo de no Pago en una Administración Tributaria, Un Marco de Trabajo de Minería de Datos

Resumen: Uno de los principales procesos en la administración tributarias es la gestión de cobranza. El objetivo de este proceso, entre otros, es la recuperación de los recursos económicos que han sido declarados por los contribuyentes. Debido a las limitaciones de las administraciones tributarias, tales como: personal, herramientas, tiempo, etc., las administraciones tributarias buscan la recuperación de las deudas en las etapas tempranas de control, donde el costo de recaudación es menor que en las etapas posteriores. Para optimizar el proceso de gestión de cobranza y contribuir a la toma de decisiones, este trabajo propone un marco de trabajo basado en aprendizaje profundo para detectar comportamientos atípicos de contribuyentes con alta probabilidad de no pago. Grupos de comportamiento normal y atípico fueron también analizados para encontrar eventos de interés usando reglas de asociación.

Palabras clave: Minería de datos, análisis de gestión de deuda, aprendizaje automático, patrones de comportamiento de contribuyentes tributarios

1. INTRODUCTION

In the process of discovering knowledge, data mining serves as the foundation for extracting valuable insights from data. This is accomplished through a variety of methods, techniques, and algorithms falling under the categories of classification, regression, clustering (Fayyad & Piatetsky-Shapiro, 1996), and summarization (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Tax administrations have also leveraged data mining to analyze historical information and extract knowledge (Ordóñez & Hallo, 2019; Rad & Shahbahrami, 2016; Gonzáles & Velásquez, 2013).

Evasion is one of the risks identified in the debt management process of tax administration; therefore, the process seeks to recover the debts calculated based on the declarations or based on legal evidence. Debt recovery is a critical aspect of tax administration that involves the identification and mitigation of risks such as evasion. Recovering debts can generate costs due to the need for bank collection systems, databases, and computer programs. The cost of managing owed values is generally lower in the early stages of collection than in later stages (Alink, 2000). Tax administrations are also striving for greater efficiency in their debt management processes to improve office performance (Huang, Yu, Hwang, Wei, & Chen, 2017). Utilizing information technology such as data mining can contribute to this goal by providing evidence-based data for business analysis and decision-making (Seddon, Constantinidis, & Tamm, 2016). In this context, tax authorities have attempted to analyze the risk of non-payment, but thus far have been unable to predict high-risk debtor taxpayers with uncollectible debt (Wu, Ou, Lin, Chang, & Yen, 2012). Additionally, a literature review does not identify any relevant

*maria.hallo@epn.edu.ec
Recibido:26/10/2023
Aceptado: 28/05/2023
Publicado en línea: 01/08/2023
10.33333/rp.vol52n1.04

studies that can predict tax debtors with a high risk of default in the coming years at an appropriate time (Ordóñez & Hallo, 2019).

The aim of this study is to carry out a data mining framework to detect atypical behaviors of tax debtors with a high risk of non-payment in tax administrations. This study will assist information professionals and knowledge engineers in identifying atypical debtor taxpayers, enabling them to build models that predict which debtor taxpayers are likely to default on their payments. The data mining process and developed models may also be applicable in departments related to debt management, or those that assess non-payment risk in tax administrations.

The proposed data mining framework is based on an adapted version of the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology (Chapman, et al., 2000).
The primary outcome of this study was the identification of normal and anomalous behavior (outliers) groups among tax debtors. A clustering algorithm was employed for this task. The presence of a cause-and-effect relationship in the identified outlier events rendered them events of interest, which is a desirable characteristic for determining whether they require further investigation for decision-making model development. Conversely, when outliers lack this cause-and-effect relationship, they may be removed from the analysis database, and further analysis may be disregarded (Senator, Goldberg & Memory, 2013).

To demonstrate the proof of concept, this study utilized data from the payment management department of the Internal Revenue Service of Ecuador.
This paper is structured as follows: Section 2 outlines background information on tax administration and outlier detection; Section 3 presents the proposed data mining framework; Section 4 describes the results of the framework's implementation; and, finally, Section 5 summarizes the conclusions and future work.

2. BACKGROUND

This section provides information on tax administration, outlier detection, and related work using machine learning techniques.

2.1 Tax Administration

Tax administrations have several primary functions, such as processing statements, determining taxes, controlling the application of tax and non-fiscal legislation, conducting inspections, managing debt, and providing services and communication. The risk of taxpayers evading their taxes is present in these functions, and the administration has limited resources to control the process (Ordóñez & Hallo, 2019; Alink, 2000). The tax administration has two objectives:

- To provide services that meet the needs of taxpayers.
- To investigate and control only high-risk taxpayers.

One identified risk is in the debt management process, which supports the collection of primary function and aims to recover values based on taxpayer declarations or legal evidence. Additionally, the debt management process records all financial transactions with taxpayers, providing a source of historical information to extract knowledge using analytic techniques. Finally, the process takes action to manage the recovery of values when taxpayers refuse to pay voluntarily (Alink, 2000).

2.2 *Outlier Detection*

Outliers are observations that significantly deviate from the norm. Methods are used to model normal data, but deviations from the norm are considered as outliers (Aggarwal, 2017). Data that includes outlier observations have large gaps between outliers and inliers (Hawkins, 1980).

Outlier analysis allows for the interpretation of results (Aggarwal, 2017), and a relationship between the condition and effect pair can be identified in outlier data. Furthermore, when the condition-effect is explained, outlier data can be transformed into events of interest (Senator, Goldberg, & Memory, 2013).

The dataset has the following characteristics:

- Data set is not labeled.
- Some attributes in the dataset are correlated. For example, business taxpayers have higher incomes and generate higher taxes. These types of attributes are called covariates.
- The dataset also includes attributes such as payment time and the amount of the fine that are not correlated.
- The statistical distribution of the dataset is unknown.

Given the features of the dataset in the debt management process, a method that reduces the number of variables and evaluates reconstruction error was chosen. The method is named a replicator or autoencoder because in the first step the data set is encoded using a function φ and the second part of the data set is decoded to the data by the function ψ (Hawkins, He, Williams, & Baxter, 2002). Function parameters are estimated using a neural network, and their architecture needs to be determined.

2.3 Outlier Analysis

The aim of outlier analysis is to convert atypical events into events of interest. This conversion helps to identify cause-effect relationships in anomalous data and to detect behavior patterns that are relevant to the application domain (Senator, Goldberg, & Memory, 2013). This process can be divided into three steps:

a) Identify outliers using techniques that suit the domain problem (Mandhare & Idate, 2017; Aggarwal, 2017; Souiden, Omri, & Brahmi, 2022; Domingues, Filippone, & Michiar, 2018).
b) Detect events of interest in outliers using methods that explain the findings (Senator, Goldberg, & Memory, 2013; Mokoena, Celik, & Marivate , 2022; Herskind Sejr & Schneider-Kamp, 2021).
c) Filter or remove outliers without patterns that are relevant to the problem domain. Additionally, there are algorithms sensitive to noise and outliers (Senator, Goldberg, & Memory, 2013; Yang,

Rahardja, & Fränti, 2021; Chen, Wang, Hu, & Zheng, 2020; Ramos, Watanabe, Traina, & Traina, 2018)

In this paper, the normal and unusual behaviors (outliers) of indebted taxpayers are analyzed and the results are interpreted to discover events of interest for the tax administration domain using a machine learning approach. The patterns of outlier data are explained using natural language labels to determine which outlier data correspond to an event of interest. In another related study, outliers with no event of interest for the problem domain are filtered out (Ordóñez, Hallo, & Luján-Mora, 2020).

3. DATA MINING FRAMEWORK

This section presents a data mining framework proposed to detect outliers, adapted from the CRISP-DM methodology (Chapman et al., 2000). The framework employs five phases for conducting a data mining project, which are iteratively performed. These phases include business domain understanding, data understanding, data preparation, modeling, and evaluation. The models are developed using the Python programming language.

3.1 Business Domain Understanding

The first phase of the data mining process focuses on the business perspective, during which the following steps are taken:

- Identification of the types of transactions involved in the debt management process in the tax administration.
- Definition of the payment time range and analysis of the statements of the debt management process.
- Identification of the data sources of debt management process information.

3.2 Data Understanding

The second phase involves the analysis of raw data to become familiar with it and to resolve quality problems in the raw data. The following steps are taken:

- Description of the type of attributes available in the data sources.
- Identification of the attributes that characterize the debt management.

3.3 Data preparation

This section involves all the activities necessary to prepare the final dataset. To achieve this objective, the following steps are proposed:

- Define interest attributes.
- Eliminate the transactions with quality problems.

3.4 Modeling

In this phase, a deep learning approach is used to create the clustering model. Thus, the following steps are carried out to identify data with normal and unusual behavior using an unsupervised approach.

- Definition of the architecture of the neural network to detect the outliers.
- Selection of the technique to create the clusters of outliers.
- Division of data to create, evaluate and test the models.
- Tuning of hyper-parameters using training and validation data.
- Evaluation of the performance of the learning process considering the error of reconstruction.

3.5 Evaluation

In this phase, the resulting groups are analyzed using the following steps:

- Division of outliers into groups with similar characteristics by determining the cluster number after applying the model step.
- Evaluation of the quality of the outlier clusters.
- Identification of the interest events according to concerns of tax administration.

4. RESULTS

In this section, the results of the data mining framework are presented.

4.1 Business Understanding

The debt management database in the tax administration of the study case consists of three types of transactions:

- Self-determined statements: Taxpayers determine the values of tax to pay.
- Determined statements: Tax administration determines the values owed by a taxpayer using administrative records.
- Suspended statements: Statements that are in an appellate court and await a court ruling.

To characterize the debts in the tax administration according to domain expert's criteria, attributes from three additional databases were added to the raw dataset. These databases include:

- Income: Database with assets, liabilities and income of taxpayers.
- Difficulty to collect: Database with taxpayers that are difficult to collect tax.
- Remission: Database with debts that have been paid in stages where the tax administration has forgiven fines and interest. This is an unwanted behavior, but it allows to obtain money from old debts.

Taxpayers were grouped in two types: a) large economic groups that produce the highest collection amounts and, b) small taxpayers such as natural persons, small businesses, and so on. The tax administration defines which taxpayers belong to each group every year.

4.2 Data Understanding

Raw data was collected from four databases. The main variables are presented in Table 1.

Table 1. Attributes that characterize the problem

Database	Attribute
Debt management	Tax code, description of the tax class, taxpayer status, description of the tax state, tax group, type of tax, fiscal month, province, region, tax year, tax fine, and payment indicator.
Income Collection difficulty	Income Collection difficulty

4.3 Data Preparation

In this step correlated data were reduced. Redundant attributes were eliminated and the values outside the possible range and null values were reviewed to eliminate invalid transactions. The number of days before payment was calculated for each transaction. The type of taxes was aggregated in 4 categories (direct, indirect, fines and others). The number of transactions of debt management analyzed for the debts of the self-determined statements was 2.1E+06 of which 1.9E+06 debts were paid and 1.2E+05 were not paid.

4.4 Modeling

The dataset was analyzed to obtain data with normal and unusual behavior. An atypical data is considered outlier; however, they could have events of interest. The events of interest were found using associated rules over the groups detected.

The outliers were found using a method that reduces the number of variables and evaluates the reconstruction error considering the features of dataset in debt management process. The method is named a replicator or auto-encoder because in the first step the dataset is encoded using a function φ and the second step of the dataset is decoded to the data by the function ψ (Hawkins, He, Williams, & Baxter, 2002). Function parameters are estimated using neural network and their architecture is shown in Figure 1.

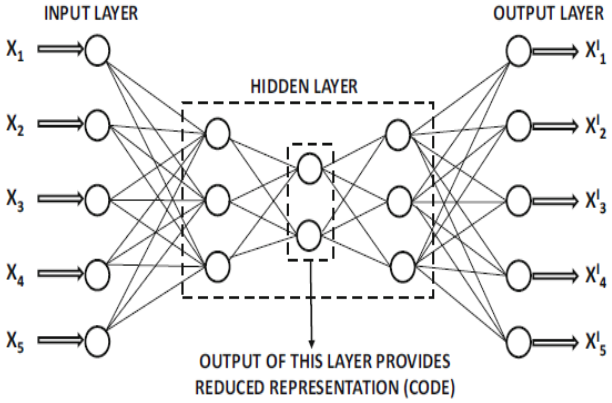

Figure 1. Auto-encoder architecture

The objective of training process is to minimize the aggregate error of reconstruction (Aggarwal, 2017). The error of reconstruction is evaluated using Equation (1).

$$error\ of\ reconstruction = \sum_{i=i}^{m}(x_i - x_i')^2 \qquad (1)$$

where

- m: input dimensions
- x_i: input.
- x_i': reconstructed input for ith dimension.

The library pyod.models.auto_encoder of python 3.6 was used to obtain data with normal and unusual behavior. The parameters used were:

- hidden_neurons=[8, 2, 8]
- random_state=10
- epochs=15
- batch_size=128
- contamination=0.1
- validation_size=0.3
- hidden_activation: relu function
- l2_regularizer= 0.1

The analysis produced the following results:

- 2.1E+05 records with outlier behavior.
- 1.8E+06 records with normal behavior.

The type of data found in the debt management process is shown in the Figure 2.

4.5 Evaluation

The data acquired during the modeling phase, comprising both normal and atypical behavior, underwent scrutiny for events of interest for tax administration. These events were classified as transactions with outstanding features, which are elaborated upon in subsequent sections.

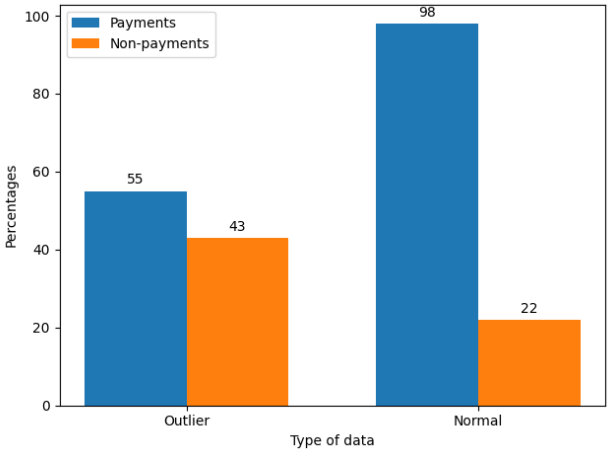

Figure 2. Type of data in the identified groups

a) Groups of outliers with similar characteristics

The condition-effect relationship identified in the outlier data was employed to interpret the results (Aggarwal, 2017). This allowed us to transform outlier data into events of interest (Senator, Goldberg, & Memory, 2013).

Number of clusters in outlier data

Clustering is the process of grouping data with high similarity to define finite sets of data categories (Fayyad & Piatetsky-Shapiro, 1996), thus enabling analysis of groups with similar patterns or characteristics.

The k-means technique is a commonly used distance-based clustering method. The technique is one of the most used methods to optimize K-means algorithm (Umargono, Suseno, & Gunawan, 2020; Shi et al., 2021). The elbow method determines the number of clusters that provides the most information by plotting the sum of the squared distances between each point and the centroid in a cluster. The point at which the metric drops suddenly guides the selection of the best-fitting model (Han, Kamber, & Pei, 2012).

The skkearn.cluster.KMeans python library was employed to implement the k-means technique and elbow method. The following parameters were configured:

- Number of cluster k: 1 to 10.
- Cost for every k group: Sum of squared distances. Figure 3**Error! Reference source not found.** shows the sum of squared distances for k=1 to 10 using k-means technique.
- The rest of the parameters of skkearn.cluster.KMeans were used in their default value.

The Figure 3**Error! Reference source not found.** also shows two elbow point when k = 2 and k = 3 added more information as a number of cluster than the other values of k, therefore the quality of the cluster have to be evaluated.

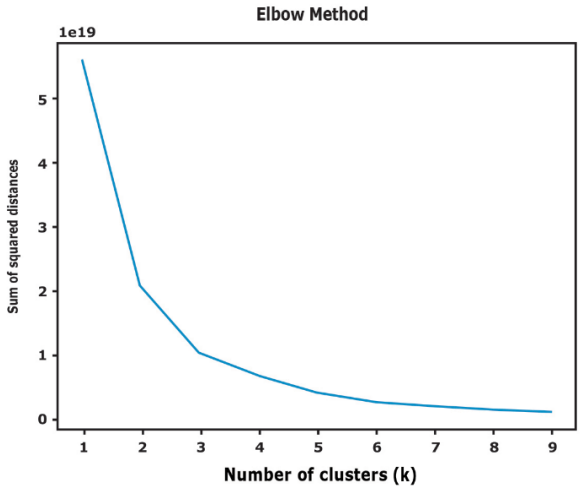

Figure 3. Number of clusters in outlier data

b) Quality of the outlier clusters

Cluster quality assesses the efficacy of a cluster in terms of the cluster separation and cohesion of each group. Cluster quality evaluation can be accomplished using supervised or unsupervised methods. The Silhouette coefficient, an intra-cluster metric, is an unsupervised method that evaluates the quality of a cluster, with its value ranging from -1 to 1, where 1 represents the best value (Han, Kamber, & Pei, 2012). The Silhouette coefficient have been used to assess the quality of k-means clusters because this coefficient has the advantage that does not need a training set to evaluate the cluster (Thinsungnoen, Kaoungku, Durongdumronchai, Kerdprasop, & Kerdprasop, 2015; Shutaywi & Kachouie, 2021). To compute the Silhouette coefficient, the Silhouette_score from the Python tool sklearn was used. The values obtained are presented in Table 2.

Table 2. Quality of outlier clusters

Number of clusters (k)	Silhouette coefficient
2	0.962
3	0.911
4	0.881
5	0.851

According to Figure 3**Error! Reference source not found.** and Table 2, when the number of cluster k=2, the quality cluster is the best. Therefore, the outlier data was divided in:

- Group 0: 2.09E+05 records.
- Group 1: 1.2E+03 records.

c) Identify the interest events

After identifying the groups, patterns in each group were analyzed to determine the events of interest from the outliers. The process of extracting knowledge from the database utilizes

data mining tasks. One of the data mining rules is association rules, which enables analysts to interact easily with data and mining results (Hipp, Güntzer, & Nakhaeizadeh, 2002).

Format data to apply association rules

Before applying association rules to discover the most important relationships between variables, each characteristic of the dataset was mapped into binned categories or cut points. The optimal cut point technique based on the entropy metric was used. This technique divides continuous values into multiple intervals (Fayyad & Irani, 1993) and is regarded as one of the most important methods for discretization (Grzymala-Busse & Mroczek, 2016).

Moreover, the binned categories were transformed into labels using natural language, which enables analysts in tax administration to read the results more efficiently. For instance, if variable "debt" had three cut points, the labels were described as follows:

- Label 1: Debts paid in less than x days.
- Label 2: Debts paid between x to y days.
- Label 3: Debts paid between z days and over.

Table 3 shows the cut points in outlier data showing the number of categories to analyze in each variable.

Table 3. Cut points of attributes that characterize the problem in outlier data

Attribute	Units	Cut points
Time needed to pay the debt	days	4
Value of tax payable	dollars	7
Value of fine to be paid	dollars	2
Taxpayer Income	dollars	1

Table 4 shows some examples of cut points applied in variables of dataset. For instance, the cut points for the time needed to pay the debt were 7, 13, 270 and 805.

Table 4. Cut point for time needed to pay the debts

Variable	Cut point value	Natural language
Time needed to pay the debt	7	These debts are paid in in less than 7 days
Time needed to pay the debt	13	These debts are paid between 7 to 13 days
Time needed to pay the debt	270	These debts are paid between 13 to 270 days
Time needed to pay the debt	805	These debts are paid at 805 days and over
Value of tax payable	4E+03	Tax value is less than $ 4E+03 dollars
Taxpayer Income	8.6E+07	The taxpayers have an income less than $ 8.6E+07

Association rules technique for discovering patterns

Next, associate rules were applied to find patterns that characterize outlier dataset for two groups of data such as the payment frequency. Table 5 shows the outstanding patterns identified for the first group which was named group 0.

Table 5. Event of interest identified in group 0

Association rule support	Item sets
98%	These transactions are not difficult to collect, and the taxpayers have an income less than $ 8.6E+07
83.1%	These debts have not been appealed, they have not paid in remission time, and the taxpayers have an income less than $ 8.6E+07
81.1%	Tax value is less than $ 4E+03 dollars, these debts have not been appealed, and the taxpayers have an income less than $ 8.6E+07
80.8%	These debts have not been appealed, they are not large economic group, these transactions are not difficult to collect, and they have not paid in remission time

According to the domain expert, group 0 of outliers contains events of interest that were identified using data mining. Payment analysis reveals that the majority of payments are made before day 270, after which the number of payments made by taxpayers decreases significantly. However, most unpaid debts exceed the 270-day default threshold. Figure 4 shows the frequency of payments using cut points. On the other hand, group 1 of the outlier dataset does not contain any events of interest, as indicated by the domain expert, and thus will not be presented.

d) Data with normal behavior analysis

After identifying and isolating the outlier data, the data with normal behavior was obtained. This data was divided into similar groups using the k-means technique to identify patterns. To determine the number of clusters, the elbow method and intra-cluster metric were used. Association rules were then applied to understand the behavior of these groups.

Number of clusters in data with normal behavior

As per the outlier analysis, there were 1.8E+06 records with normal behavior. Similar to the outlier data, the k-means technique and elbow method were used to select the optimal number of clusters for the normal data. Figure 5 displays the outcome of the elbow method, which utilized the skkearn.cluster.KMeans Python library. Two groups were chosen since the larger intra-cluster value was obtained with k=2. The silhouette score for the intra-cluster value was 0.854.

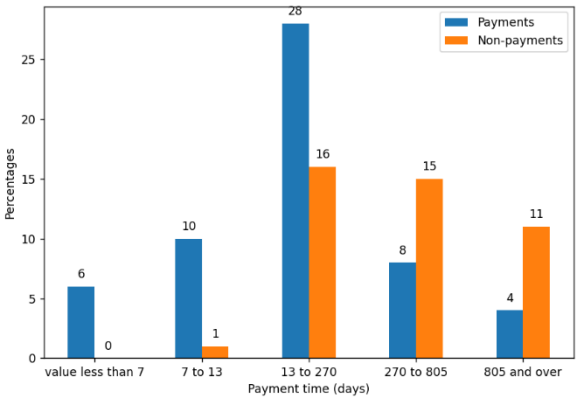

Figure 4. Frequency of payment for outlier group

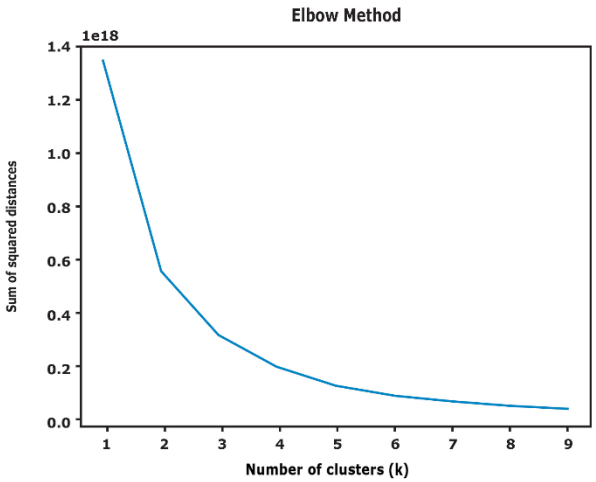

Figure 5. Number of clusters in normal behavior

The data with normal behavior was divided in:

- Group 0: 9.5E+04 records.
- Group 1: 1.7E+06 records.

The records not paid in the first were 1%, and two percent of the records not paid in group second group were identified. The frequency for payment is shown in Figure 6. Most debts are paid up to 85 days.

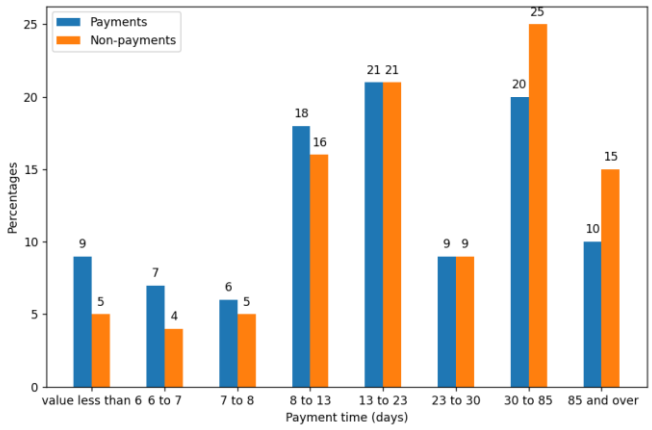

Figure 6. Frequency of payment for normal behavior

The results were communicated to the authorities of tax administration section who could use it to develop strategies to get the payment. Additionally, these results have been used to select data pre-processed as input of models to predict debts with high risk of non-payment in tax administration in short periods of time (Ordóñez, Hallo, & Luján-Mora, 2020).

5. CONCLUSION

With the data mining framework and with the data of the tax administration of Ecuador, a model to find atypical comportment groups of taxpayer's debtors with high risk of non-payment was developed using machine learning techniques. Given the conditions of the dataset, a technique was applied to find the unusual and normal behaviors of debts. The result of this analysis determined two groups with unusual behavior. Two groups with normal behavior were also identified.

With the time estimated by the models on the outlier group, tax administrations can determine the debts that belong to uncollectible debt. For example, for group 0 with unusual behavior (Figure 4), until day 270 debts are mostly canceled. From that threshold the probability of no payment begins to rise. The patterns of this data were also found using associated rules to find event of interest (Table 5). With the knowledge generated, tax administrations can make decisions regarding communication for the collection of obligations, guidelines to grant payment facilities, select records that need to be audited, among others. Additionally, tax administrations could use the results to select data as input for other models in future works.

REFERENCES

Aggarwal, C. (2017). *Outlier Analysis.* Cham: Springer Nature. https://doi.org/10.1007/978-3-319-47578-3

Alink, V. (2000). *Handbook for Tax Administrations Organizational structure and management of Tax Administration.* The Netherlands: CIAT. https://www.ciat.org/Biblioteca/DocumentosTecnicos/Ingles/2000_handbook_for_ta_netherlands_ciat.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0.* USA: CRISP-DM consortium. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Chen, C., Wang, Y., Hu, W., & Zheng, Z. (2020). Robust multi-view k-means clustering with outlier removal. *Knowledge-Based Systems. 210*(2020), 1-12. https://doi.org/10.1016/j.knosys.2020.106518

Domingues, R., Filippone, M., & Michiar, P. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition. 74*(2028), 406-421, https://doi.org/10.1016/j.patcog.2017.09.037

Fayyad, P., & Piatetsky-Shapiro, G. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, P., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 82-88. https://dl.acm.org/doi/10.5555/3001460.3001477

Fayyad, U., & Irani, K. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1022-1027. https://hdl.handle.net/2014/35171

Gonzáles, P., & Velásquez, J. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Syst. Appl., 40*(5), 1427-1436. https://doi.org/10.1016/j.eswa.2012.08.051

Grzymala-Busse, J. W., & Mroczek, Teresa. (2016). A Comparison of Four Approaches to Discretization Based

on Entropy. *Entropy*, *8*(3), 69 https://doi.org/10.3390/e18030069

Han, J., Kamber, M., & Pei, J. (2012). 10 - Cluster Analysis: Basic Concepts and Methods. *The Morgan Kaufmann Series in Data Management Systems*, 2012, 443-495. https://doi.org/10.1016/B978-0-12-381479-1.00010-1

Hawkins, D. (1980). *Identification of Outliers*. Springer Netherlands. https://doi.org/1010.1007/978-94-015-3994-4

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier Detection Using Replicator Neural Networks. *Springer Berlin Heidelberg*, 170-180. https://doi.org/10.1007/3-540-46145-0_17

Herskind Sejr, J., & Schneider-Kamp, A. (2021). Explainable outlier detection: What, for Whom and Why? *Machine Learning with Applications*, *6*(2021), 100172. https://doi.org/10.1016/j.mlwa.2021.100172

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2002). Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. In: Perner, P. (eds) *Advances in Data Mining. Lecture Notes in Computer Science*, (vol. 2394). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-46131-0_2

Huang, S., Yu, M., Hwang, M., Wei, Y., & Chen, M. (2017). Efficiency of Tax Collection and Tax Management in Taiwan's Local Tax Offices. *Pacific Economic Review*, *22*(4), 620–648. https://doi.org/10.1111/1468-0106.12235

Mandhare, H., & Idate, S. (2017). A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques. *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 931-935. https://doi.org/10.1109/ICCONS.2017.8250601

Mokoena, T., Celik, T., & Marivate , V. (2022). Why is this an anomaly? Explaining anomalies using sequential explanations. *Pattern Recognition*, *121*(2022), 108227 https://doi.org/10.1016/j.patcog.2021.108227

Ordóñez, J., & Hallo, M. (2019). Data Mining Techniques Applied in Tax Administrations: A Literature Review. *2019 Sixth International Conference on eDemocracy eGovernment (ICEDEG)*, 224-229. https://doi.org/10.1109/ICEDEG.2019.8734342

Ordóñez, J., Hallo, M., & Luján-Mora, S. (2020). Detection of Taxpayers with High Probability of Non-payment: An Implementation of a Data Mining Framework. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. https://doi.org/10.23919/CISTI49556.2020.9140837

Rad, M., & Shahbahrami, A. (2016). Detecting high risk taxpayers using data mining techniques. *2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, 1-5. https://doi.org/10.1109/ICSPIS.2016.7869895

Ramos, J., Watanabe, C., Traina, C., & Traina, A. (2018). How to speed up outliers removal in image matching. *Pattern Recognition Letters*, *114*(2018), 31-40. https://doi.org/10.1016/j.patrec.2017.08.010

Seddon, P., Constantinidis, D., & Tamm , T. (2016). How does business analytics contribute to business value? *Information Systems Journal*, *27*(3), 237-269. https://doi.org/10.1111/isj.12101

Senator, T., Goldberg, H., & Memory, A. (2013). Distinguishing the Unexplainable from the Merely Unusual: Adding Explanations to Outliers to Discover and Detect Significant Complex Rare Events. *KDD 2013 Workshop on Outlier Detection and Description*, 40-45. https://doi.org/10.1145/2500853.2500861

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(31). https://doi.org/10.1186/s13638-021-01910-w

Souiden, I., Omri, M., & Brahmi, Z. (2022). A survey of outlier detection in high dimensional data streams. *Computer Science Review*, *44*(2022), *100463*. https://doi.org/10.1016/j.cosrev.2022.100463

Thinsungnoen, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., & Kerdprasopb, N. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. *Proceedings of the 3rd International Conference on Industrial Application Engineering 2015*. https://doi.org/10.12792/iciae2015.012

Umargono, E., Suseno, J., & Gunawan, V. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*, 121-129. https://doi.org/10.2991/assehr.k.201010.019

Wu, R., Ou, C., Lin, H., Chang, S., & Yen, D. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Syst. Appl*, *39*(10), 8769–8777. https://doi.org/10.1016/j.eswa.2012.01.204

Yang, J., Rahardja, S., & Fränti, P. (2021). Mean-shift outlier detection and filtering. *Pattern Recognition*, *115*(2021), 107874. https://doi.org/10.1016/j.patcog.2021.107874

BIOGRAPHIES

José Ordóñez received his master's degree in computer science at Escuela Politécnica Nacional, in 2020. He has been working for Internal Service Revenue in Ecuador since 2010 and has also been researching about knowledge process in tax administrations in the Business Intelligence Laboratory of Escuela Politécnica Nacional University since 2017. His main research interests include data mining, debt management analysis, machine learning, taxpayer behavior patterns, risk analysis, audit recommendation, and survival analysis.

Maria Hallo is a professor at the Faculty of Systems Engineering of the National Polytechnic School, Quito-Ecuador. MSc in Computer Science from Notre Dame de la Paix University. PhD in Computing Application from the University of Alicante. His areas of interest are Business Intelligence, Databases, Semantic Web, Data Mining, Information Systems.