Ingeniería Investigación y Tecnología
volumen XXIV (número 4), octubre-diciembre 2023 1-12
ISSN 2594-0732 FI-UNAM artículo arbitrado
Información del artículo: Recibido: 7 de mayo de 2020, reevaluado: 20 de septiembre de 2022, aceptado: 6 de septiembre de 2023

https://doi.org/10.22201/fi.25940732e.2023.24.4.031

Ingeniería Investigación y Tecnología

Employee profile and labor turnover in outsourcing companies: A data mining approach

Perfil de empleados y rotación laboral en empresas de outsourcing: Un enfoque de minería de datos

Márquez-Hermosillo Abigail
Instituto Tecnológico de Sonora
Dirección de Ingeniería y Tecnología
Departamento de Computación y Diseño
E-mail: mha.9600@gmail.com
https://orcid.org/0000-0002-9629-6496

Salazar-Lugo Guillermo
Instituto Tecnológico de Sonora
Dirección de Ingeniería y Tecnología
Departamento de Computación y Diseño
E-mail: guillermo.salazar@itson.edu.mx
https://orcid.org/0000-0001-7375-2658

Rodríguez Luis Felipe
Instituto Tecnológico de Sonora
Dirección de Ingeniería y Tecnología
Departamento de Computación y Diseño
E-mail: luis.rodriguez@itson.edu.mx
https://orcid.org/0000-0001-8114-0299

Borrego Gilberto
Instituto Tecnológico de Sonora
Dirección de Ingeniería y Tecnología
Departamento de Computación y Diseño
E-mail: gilberto.borrego@itson.edu.mx
https://orcid.org/0000-0001-7315-5693

Abstract

Data mining techniques can be applied to search for hidden information in large volumes of data. In human resources management, data mining is useful for identifying the reasons behind employee turnover and behavior. This knowledge makes it possible to identify employee profiles and helps improve personnel selection processes, which are appropriate means to reduce company turnover rates. In this article, we analyze the situation of a human resources outsourcing company and apply data mining techniques to classify labor turnover in low-skilled employees. We follow the methodology CRISP-DM to build and evaluate different classification models and discover a list of relevant characteristics of employee profiles prone to turnover. Furthermore, we compare the results of applied techniques to assess performance and suitability to identify factors associated with turnover and generate undesirable employee profiles. The results show that Age, Salary, Location, and Work Experience in Time and Area are key factors that help classify turnover and, therefore, can be used to suggest personnel selection policies to the company. The results obtained in this article may serve as a reference framework for companies that hire low-skilled employees, particularly those that provide human resources outsourcing services, so they can collect and analyze employee data and identify profiles prone to turnover. The significance of this work is that results: i) are presented in the context of a real human resources outsourcing company and ii) are obtained from the analysis of low-skilled employee data available in such a company, which are aspects scarcely explored in related research. A limitation of this research was the partial absence of specific socio-demographic data in the available data set and of variables related to organizational climate and culture.
Keywords: Labor turnover, employee profile, data mining, data analysis, classification techniques.

Resumen

Las técnicas de minería de datos se pueden aplicar para buscar información oculta en grandes volúmenes de datos. En la gestión de recursos humanos, la minería de datos es un enfoque útil para identificar las razones detrás de la rotación y el comportamiento de los empleados. Este conocimiento permite identificar perfiles de empleados y ayuda a mejorar los procesos de selección de personal, que son medios apropiados para reducir la tasa de rotación en las empresas. En este artículo analizamos la situación de una empresa de subcontratación de recursos humanos y aplicamos técnicas de minería de datos para clasificar la rotación laboral en empleados poco calificados. Seguimos la metodología CRISP-DM para crear y evaluar diferentes modelos de clasificación y descubrir una lista de características relevantes de los perfiles de los empleados propensos a la rotación. Además, comparamos los resultados de las técnicas aplicadas para evaluar el desempeño y la idoneidad para identificar factores asociados con la rotación y generar perfiles de empleados no deseados. Los resultados muestran que la edad, el salario, la ubicación y la experiencia laboral en tiempo y área son factores clave que ayudan a clasificar la rotación y, por lo tanto, pueden usarse para sugerir políticas de selección de personal a la empresa. Los resultados obtenidos en este artículo pueden servir como un marco de referencia para las empresas que contratan empleados poco calificados y particularmente para aquellos que brindan servicios de subcontratación de recursos humanos para que puedan recopilar y analizar datos de los empleados e identificar perfiles propensos a la rotación. La importancia de este trabajo es que los resultados: 1) Se presentan en el contexto de la situación de una empresa real de subcontratación de recursos humanos y 2) Se obtienen del análisis de los datos de empleados poco calificados disponibles en dicha empresa, que son aspectos poco explorados en investigaciones relacionadas. Una limitación para esta investigación fue la ausencia parcial de datos sociodemográficos específicos, así como de variables relacionadas con el clima y la cultura organizacional.
Descriptores: Rotación laboral, perfiles de empleado, minería de datos, analítica de datos, técnicas de clasificación.

https://doi.org/10.22201/fi.25940732e.2023.24.4.031

Introduction

Human resources (HR) have become a strategic asset for companies to achieve their business objectives (Chiavenato, 2011). In most companies, one of Human Resources Management's main objectives (HRM) objectives is to attract and retain a talented workforce by designing and monitoring strategies to reduce the turnover rate. Labor turnover is the exchange between employees leaving and coming into the company in a period, usually monitored as the turnover rate (Chiavenato, 2011). Labor turnover has negative effects such as loss of productivity, poor performance in the market, and cost increases in recruitment and training (Aranibar *et al*., 2018) due to the constant hiring of new employees.

The turnover rate behavior varies across sectors and industries. In particular, outsourcing companies are highly affected by the turnover problem (Ranganathan & Kuruvilla, 2008). The outsourcing sector includes those companies focused on providing external activities to other companies. Specifically, Human Resources Outsourcing (HRO) is dedicated to outsourcing HRM activities for an external company. The HRM often outsources activities such as personnel selection and recruitment, training, compensation, development, employee orientation, and strategies (Ji, 2016; Barrett, 2020).

Employee turnover and the decision to leave a company are diverse. These can be associated with internal and external factors such as organizational culture, labor conditions (e.g., position, payment, and career plan), and characteristics of the employee profile (e.g., age, gender, education, and previous experience) (Arokiasamy, 2013). HR managers, therefore, make efforts to control employee withdrawal and thus minimize the adverse effects for the company. Some strategies include the detection of reasons why employees are leaving the company to create better policies and improve internal labor conditions and organizational climate. Other strategies involve the detection of employee profiles prone to withdrawal to improve selection processes by avoiding these profiles and thus reduce the risk of turnover of new employees (Allen *et al*., 2010; Kim & Ployhart, 2018).

The integration of technology in companies has proven to help improve efficiency. Business intelligence tools, for example, facilitate the representation and analysis of large volumes of data. This tool helps identify relevant situations in the company (e.g., in a production line) and anticipate potential undesirable events (Martínez, 2011). In particular, HR managers use these technologies supported by machine learning techniques in daily activities such as talent selection and development (Attri, 2018) to reduce labor turnover rates by anticipating personnel performance. Business Intelligence tools involve data management, data analytics, business performance tracking, and information delivery (Dias & Sousa, 2015). However, data analytics for HRM has not been widely explored and is usually used in a descriptive and predictive scope (Pape, 2016).

Implementing data mining tools such as data analytics on HRM can bring a competitive advantage by combining technology, business processes, and people skills to make better decisions based on the information and predict the behavior of future employees and current employees (Dias & Sousa, 2015). In this context, incorporating data analytics in HRO companies becomes a need since the main asset of companies is human resources.

In this article, we build classification models for labor turnover in low-skilled employees of an outsourcing company. Various data mining techniques were tested over a data set containing employee data from an HRO company in Sonora, México. The objective was to compare the results of applied techniques and build a classification model that helps identify employee profiles prone to turnover. The structure of this article is as follows. In the next section, we present related work. Afterward, we describe the method utilized and the results obtained. Finally, we provide a discussion and some concluding remarks.

Related work

In this section, we discuss some fundamental aspects associated with labor turnover and data mining techniques applied in HRM.

Human talent management, selection process, and labor turnover

HRM contributes to achieving the company's strategic objectives by carrying out crucial activities such as fostering favorable relationships between employees, controlling labor turnover, and controlling the internal and external factors that may affect the employee's development (Dessler, 2017). In particular, in low-skilled jobs, several internal and external factors may influence the employee's decision to quit or stay in the job: organizational climate, commitment, satisfaction, flexibility, and company policies (Aranibar *et al*., 2018). Some other key factors are mentioned, such as staff benefits, working environment, motivation, leadership, career plan, age, place of residence, number of dependents, salary, training, job expectations, the influence of co-workers, and job fit and expectation in several research works (Dubey *et al*., 2016; Tam & Khoa, 2018; Al Mamun & Hasan,

2017; Kangas *et al*., 2018). Furthermore, Ranganathan & Kuruvilla (2008) analyzed three spectra that influence attrition that is related to: i) work environment, ii) employee demographic profile, and iii) psychological factors related to the employee profile.

Regarding Mexico, the manufacturing industry reports that labor turnover is associated with the economy's dynamism, which is correlated directly with the Gross Domestic Product (GDP) and inversely with unemployment (Moreno *et al*., 2015). During economic expansion, the manufacturing industry generates structural contradictions by demanding job vacancies that are not filled, which weakens the productive process, calling into question the efficiency and effectiveness of the manufacturing industry (Herrera *et al*., 2019). On the other hand, the Mexican hotel industry reports three leading causes of labor turnover, particularly in Guanajuato state (Caldera *et al*., 2019): i) compensation systems (salary and non-salary related) are not competitive, ii) personnel recruitment and selection systems are inadequate, and iii) employees are unmotivated due to lack of clarity in their tasks and functions.

Mondy & Noe (2005) suggest that HRM depends on human resources research, which tries to find the answers to the human resources problems that occur and affect the company, including labor turnover. Part of this human resources research focuses on controlling labor withdrawal and turnover rate by trying to understand its roots. Mondy & Noe also attempt to understand the use of technology to improve the HR department's process, for instance, the implementation of data analytics to measure the efficiency of HRM practices and processes and even predict events (e.g., turnover rate and employee behavior and performance).

One of the most important activities of HRM is the recruitment and selection process, which helps to create a pool of talent and potential candidates who might help to reach the business goals (Uzair *et al*., 2017). A good selection of personnel in a company brings benefits related to the organizational performance of the employees, which leads to achieving organizational objectives. A good selection of personnel in a company brings benefits related to the organizational performance of the employees, which leads to achieving organizational objectives. The recruitment and selection processes are fundamental to selecting the right person for the right job (Uzair *et al*., 2017; Almeda, 2017) because these processes shape effectiveness and performance.

Data mining applied in human resource management

The application of data mining techniques in HRM has been reported in related research using data mining as a descriptive trend and data mining as a predictive trend.

The descriptive approach analyzes current HRM data to describe insights and generate knowledge about a particular problem. With this knowledge, action strategies are proposed to solve problems or reduce impacts on the HRM and the company. For example, Gao (2017) implements data mining methods to address employee turnover problems using the algorithm CART to analyze the most important factors contributing to high employee turnover. The results included a priority table that shows those people who should be retained.

On the other hand, the predictive approach has had a more active role in current research. For example, (Sikaroudi *et al*., 2015) applied data mining techniques to predict the voluntary turnover of employees in a manufacturing plant in Arak, Iran. This study proposed a decision support system for the human talent department through data mining and evaluating employee characteristics such as age, technical skills, and work experience. In particular, this study attempted to infer the performance of candidates for positions in the company to reduce losses by avoiding hiring unstable and poorly trained candidates.

The research made by Attri (2018) proposed a model to predict the probability that an employee leaves a position based on previous information from the human talent department. The primary purpose of this model was to find and apply strategies that prevent these behaviors. This proposal is based on the fact that the performance of the employees is predictable, and this knowledge would help to avoid the attrition of valuable employees for a company. Attri (2018) used criteria given by experts in the area, such as age, sex, department to which the worker belongs, job satisfaction, and education, among many others. It proved that the model could predict employee behavior by 80 %.

Also, Ribes *et al*. (2017) implemented common data mining techniques on a data set of 1000 employees and created a turnover prediction model. The resulting model showed that employees low performance, compensation schemes, and business units particularities were factors that predict turnover. The results led to the design of retention policies divided into reassignment programs, rotation of positions, and binding of new employees. Similarly, the studies mentioned above seek to solve the problem of labor turnover through data mining techniques. However, other studies have attempted to apply these techniques to predict emplo-

yee behavior and performance. For example, the study by Kirimi & Moturi (2016) proposes applying data mining classification techniques to predict the performance of existing employees in a government enterprise in Nairobi. A job performance classification model was created based on sociodemographic information and previous performance analyzes. This model was used to improve the process of job performance evaluation. As a result, it was obtained that experience was the factor with the most significant positive impact on employee performance, followed by factors such as age, grades, sex, marital status, and performance evaluation score.

As seen in this section, there is still a series of challenges to HR analytic: research efforts still focus on descriptive and predictive analytics (Pape, 2016), but there is an opportunity to implement prescriptive analytics. It is noteworthy that most research is focused on high-skilled employees, disregarding data about low-skilled workforce and outsourcing companies, a sector that is becoming a trend in HRM. In this work, we attempt to provide an analysis of such data by following a data mining approach.

Method

The methodology followed to achieve the objective of this article was the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, which consists of six phases (see Figure 1). This methodology was selected because it is a reference in the deployment of data mining projects and provides a framework to iterate from business understanding to evaluation phases (Shafique & Qaiser, 2014).

Business understanding

This phase aims to understand the business problem and determine an objective for the data mining project. We focused on analyzing the problems of the HRM in an Outsourcing company.

We conducted a situational analysis in Malumex, a Mexican company in Sonora State. We used the technique of Narrative Research (Hernández *et al.*, 1991), to create a story presented as a case study about the company's growth through the years, the problems presented in HR processes, and the perception of the importance of employee profiles within selection processes and turnover rate monitoring. This technique was selected because it facilitates the use of information provided by people in the company and allows them to have a more secure environment and flexible time in the sessions. Also, we used some confidential company documents, including recruitment, selection, and sales reports and statistics. These documents helped complete the story provided by employees and further understand the company. Furthermore, this helped obtain the context necessary to define a data mining objective.

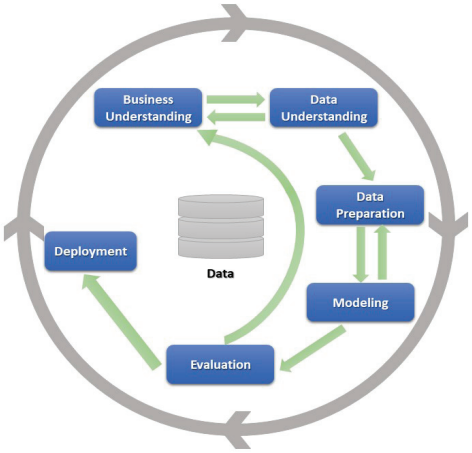


Figure 1. CRISP-DM Methodology (Shafique & Qaiser, 2014)

Meetings and documentary revision led to the creation of the case study. The company's Chief Human Resources Officer (CHRO) and Chief Executive Officer (CEO) were present in the five sessions, which lasted between 40 and 60 minutes. The topics discussed were the company's history, HRM Office's processes, problems related to the employees and turnover rate, strategies implemented to reduce turnover rate, the selection process, the systems and documents used to manage employee data, and consequences of turnover to the company. The meetings were recorded and transcribed to integrate the case study.

Data understanding

The main objective of this phase is to understand the obtained data, its sources, and its characteristics. We collected data from different sources related to employee profiles and recruitment processes of the company, which are associated with turnover according to reported literature (Dubey *et al.*, 2016; Ranganathan & Kuruvilla, 2008; Sikaroudi *et al.*, 2015; Attri, 2018; Kirimi & Moturi, 2016): i) SuperNOMINA software that included information of employees over the past 20 years, ii) job application forms that included socio-demographic information from the period 2016-2019, and iii) the digital records of job interviews from the period 2017-2019. The three data sources were integrated into a single data set by capturing the information manually in an

MS Excel spreadsheet, which resulted in a data set with 20 attributes.

We also made an exploratory analysis of this dataset to understand the types of attributes better. The descriptive analysis included the calculation of measures of central tendency and frequencies using the statistical software SPSS.

Data preparation

This phase aims to transform the obtained data into a suitable format to execute data mining algorithms and create models. One of the biggest challenges in this phase was to create a single data set from the three sources. To achieve this, we extracted a subset from the initial data set that matched the period from the other two data sources. The data set was revised to eliminate information duplicated. Furthermore, we executed five data preparation tasks to transform the variables into a suitable format for the data mining algorithms:

1. Definition of new variables: the data contained in the different data sets allowed us to define new variables such as workers' job experience and age, which have been referred to as factors that may influence the decision to quit a job (Dubey *et al*., 2016).
2. Definition of the dependent variable: we had to identify the variable that represented employees' working time in Malumex to classify labor turnover in the company's context using data mining techniques.
3. Exclusion of non-relevant variables: several variables were considered non-relevant as they did not contribute to defining employee profiles prone to turnover (e.g., employee ID).
4. Replacement of missing values: we performed this common preparation task by replacing values using the resulting median value, for example.
5. Transformation of binary and categorical attributes: transforming the type of variables into numeric values was necessary to apply data mining techniques.

Modeling

The purpose of this phase is to create a model that solves the objective of the project. Once the data set was prepared, we executed two iterations using Rapidminer software: manual and automatic, featuring a selection of variables for the training and testing subsets. The results and performance of the following data mining techniques were compared: Naive Bayes (NB) (Chhogyal & Nayak, 2016), Generalized Linear Model (GLM) (Dobson & Barnett, 2008), Logistic Regression (LR), Random Forest (RF) (Couronné *et al*., 2018), Deep Learning (DL) (Guo *et al*., 2016), Decision Tree (DT) (Jahan *et al*., 2018), and Support Vector Machine (SVM) (Van *et al*., 2016), as these techniques are the most used for classification problems in related literature. The configuration of the algorithms was set as the default parameters given by the data mining system.

A partition relation of 0.6 and 0.4 was established to create training and testing subsets, respectively. Also, an automatic sampling type with an automatic value of Local Random Seed was established. The training and validation subsets in the second iteration were created following the same rules. An Automatic Feature Selection and Generation were established to create an optimal Feature Set regarding the complexity and error. The feature set was applied to the complete training data to build the final model.

Evaluation

This phase aims to establish the criteria to evaluate the resulting model. The most common performance indicators for data mining models are accuracy and classification error. Other indicators are Area Under the Curve (AUC), sensitivity, f measure, and confusion matrix. For this project, accuracy, classification error, and AUC were observed to determine the performance of the models. An accuracy closest to 1 means that the model has a good classification performance, while a classification error closest to 0 means the possibility of error is the lowest in the model. According to Narkhede (2018), the AUC indicator measures the model's capability to distinguish between classification classes. We established that the criteria to identify the performance of a model was the AUC value due to the information it gives about the model.

We divided the data into seven batches to randomly assign a different batch to subsets of the same size. We also tried seven different iterations of the validation process through a performance evaluation for binary variables to calculate the AUC, Accuracy, and Classification error. The suitability of the model was decided in terms of interpretability. The results and information obtained from the models matched the criteria given by related works about what variables can help classify labor turnover.

Results

This section presents the results of implementing the described phases of the method.

https://doi.org/10.22201/fi.25940732e.2023.24.4.031

Case study

Malumex is an outsourcing company located in Ciudad Obregón, Mexico that dedicates about 80 % of its operations to hiring manufacturing operators for food processing companies. The activities of the most workforce hired are manual operations such as packaging and boxing, i.e., operative jobs. The profile of these employees involves a basic educational level and requires no experience or specialized skills. The labor conditions include low salaries, extended work schedules with work shifts of up to 12 hours, and a lack of ergonomics in the facilities. The contract scheme defines an initial work period of 90 days so that the company guarantees that the investment associated with the recruitment process is recovered. The contract is renewed every three or six months, depending on factors such as employee performance.

Figure 2 shows the labor turnover rate of employees that worked with the principal client of Malumex between January 2017 and March 2019. As seen, the turnover percentage tends to increase as high as 50 % despite the implemented strategies to reduce it. The fluctuation of turnover over months makes it difficult to determine patterns. These relatively high turnover rates have had adverse effects on the company, including the high costs of attracting talent, unattainable objectives of the HRM department, and the work overload associated with the lack of the required workforce. This situation has also affected the financial situation of Malumex, as the lack of human talent makes it difficult to close negotiations with new clients. Labor turnover also affects Malumex's relationship with current clients due to the impossibility of maintaining the minimum requirement for personnel operations. The side effects of this situation were the high cost of liquidation of the employees, the penalty imposed by the client, and a reduction of income.

The strategies that Malumex has implemented to reduce the labor turnover rate include a bonus for hiring and rewards for the assistance and permanence of employees in the company. Another strategy implemented in the summer of 2017 was to analyze the profile of the successful employee to improve the hiring process. For instance, the minimum hiring age was increased from 18 to 23 years based on the opinion that people under 23 years were unsuitable for this job. Nevertheless, these strategies did not have the desired success. In this context, understanding the profile and behaviors of employees prone to leave the job becomes evident. In the case of Malumex, before the initial contract is completed.

Once we analyzed the situation in Malumex, it was found that the turnover rates present increments that negatively impact the company. Malumex has already identified that many employees leave their job before the 90-day work period. Therefore, analyzing and classifying the employee profiles was crucial for this project. In this sense, the objective of the data mining project was to create a classifier model for employees who resign before 90 days of hiring, as they are not profitable for Malumex.

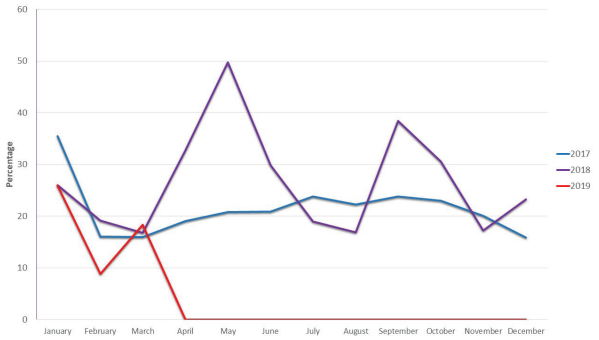


Figure 2. Monthly turnover rate at Malumex in 2017-2019

Data understanding

The initial data set extracted from superNOMINA software contained information on employees from 1998 to March 2019. This software manages information about job characteristics and socio-demographic data. Job applications and digital records of job interviews were also utilized to complement the data set. In general, all these data sources contained missing values.

The final data set included 20 variables and 3735 records (see Table 1). As part of the preparation tasks, new variables and their corresponding values were calculated and included in the dataset: Age and Seniority. These two variables represented the socio-demographic information of employees. A descriptive analysis of SPSS helped to understand the data set better. Table 2 shows measures of the central tendency for these attributes. The average Seniority is 188.50 days, with an s.d. of 422.57, which implies a vast dispersion of data. The median of the distribution is equal to 73 days. We decided to represent the Seniority variable with the median value.

Figure 3 shows the frequency of some categorical and binary variables: Gender, Education, Job experience in Time and Sector, and Job Resignation Reason. We can observe that the male population is higher (58 %) than the female population (42 %). Also, the most frequent education category is Junior High School (65.41 %), while Unjustified absence (33.22 %) is the most common resignation reason. Meanwhile, 48.08 % of the em-

ployees have previous experience in factories, and their experience in time is often less than three months (38.62 % of the population).

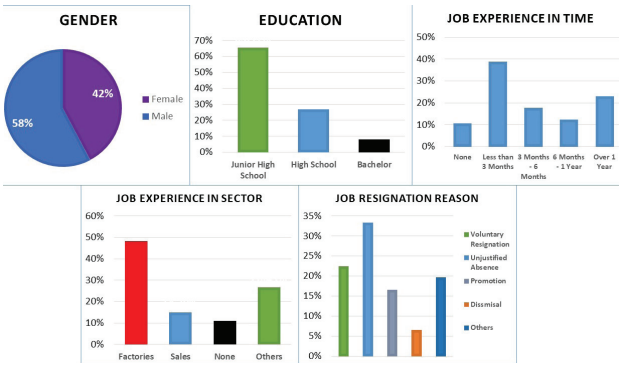
Table 1. Variables included in the final merged data set

Attribute	Description	Type	Missing values	Data source	Decision
Id Employee	Employee's Id	Integer	0	SuperNOMINA	Excluded
Work Center	Site of work	Categorical	0	SuperNOMINA	Included
Salary	Category of salary	Categorical	0	SuperNOMINA	Included
Team Work	Team where the employee worked	Categorical	0	SuperNOMINA	Included
Job title	Name of the role played	Categorical	0	SuperNOMINA	Included
Hiring Date	Date the employee started working	Date	0	SuperNOMINA	Excluded
Resignation Date	Date the employee end the job relationship	Date	93	SuperNOMINA	Excluded
Work shift	Nature of work schedule	Binary	0	SuperNOMINA	Included
Date of birth	Date the employee was born	Scale	0	SuperNOMINA	Excluded
Gender	Gender of the employee	Binary	0	SuperNOMINA	Included
Type of contract	Nature of work contract	Categorical	0	SuperNOMINA	Included
Working day	Time period during the employee is at work	Binary	0	SuperNOMINA	Included
Job separation reason	Reason of the end of job relationship	Categorical	674	SuperNOMINA	Included
Location	Place where the employee lives	Binary	608	Job Applications	Included
Progeny	Condition if the employee had progeny or not	Binary	586	Job Applications	Included
Marital status	Condition of marital status	Categorical	586	Job Applications	Included
Education	Educational level	Categorical	584	Record of interviews and Job Applications	Included
Previous Experience in time	Duration of the last job	Categorical	630	Record of interviews and Job Applications	Included
Previous Experience in sector	Area of the last job	Categorical	586	Record of interviews and Job Applications	Included
Rehired	Condition if the employee worked before with the company	Binary	0	SuperNOMINA	Included

Table 2. Measures of central tendency for scale type attribute

Attribute	Min	Max	Average	Median	S.D.
Age	14	65	29.34	26.52	9.56
Seniority	0	7759	188.50	73	422.57

Figure 3. Frequencies of nominal attributes in the data set utilized



Data preparation

We first defined two new attributes: Age and Seniority. The values of the second attribute were calculated utilizing information from the available data about periods working at Malumex. The latter attribute was established as the dependent variable since it represents the working time of employees and therefore results helpful to define whether an employee left the job before the initial contract of 90 days (i.e., related to undesirable employee profile).

Variables such as the ID of employees and dates such as date of birth, resignation, and hiring were excluded since these did not contribute to defining employee profiles prone to turnover or were already considered in the values of some other variable (e.g., Age). We also transformed binary and categorical attributes into numeric attributes so that the data mining techniques could be applied. Regarding missing values, the values of Seniority were filled with the median value equal to 73. For the attributes Experience in Time, Experience in the sector, Job resignation reason, Marital status, Progeny, Education, and Location, the missing values were not substituted not to skew the data. Finally, we transformed the dependent variable Seniority into a binary attribute to distinguish the employees who worked less than 90 days from those who worked a more extended period. The categories for this attribute were set as true and false, and the target for the data mining problem was labeled as false.

Modeling

In the first iteration, the techniques listed below were utilized to build classification models using 17 total attributes in the data set. The attribute Job separation reason was excluded since this variable would not be available in a new hiring scenario.

The parameters set for each model are as follows:

- Random Forest: Number of Tree = 20, Maximal Depth = 4.
- Support Vector Machine: Gamma = 1.0E-4, C = 10.0, Total number of support vectors = 1575, Bias (offset) = –0.944.
- Deep Learning: Model Metrics = Type Binominal, model id: rm-h2o-frame-model559065, MSE = 0.23713148, R2^ = 0.0033993945, logloss = 0.6670185.
- Decision Tree: Maximal Depth = 3.
- Generalized Linear Model: Model Metrics Type = Binomial GLM, Description = N/A, model id = rm-h2o-model-model-917445, frame id = rm-h2o-frame-model-460884, MSE = 0.23554558, R2^ = 0.010064466, logloss = 0.6638515.
- Logistic Regression: Model Metrics Type = Binomial GLM, Description: N/A, model id = rm-h2o-model-lr-58768, frame id = rm-h2o-frame-lr-471698, MSE = 0.23652673, R2^ = 0.005940958, logloss = 0.6659203.
- Naive Bayes: Simple Distribution = Distribution model for label attribute Seniority, Class false (0.610) 2 distributions, Class true (0.390) 2 distributions.

A second iteration was carried out using the three models with the highest AUC value (see Table 4) to evaluate whether the attribute with a high correlation with the dependent variable "Job separation reason" had a significant impact, where an automatic feature selection and generation was established for finding the best feature sets. The parameters for the second iteration are given in Table 3.

Evaluation

In the first iteration (Table 4), the results showed that the models tested had an accuracy within the range of .613 to .621, where SVM was the model that presented the highest accuracy. Nevertheless, the RF presented an AUC value equal to .625 while the rest of the models had an AUC around .500 - .550; an AUC near .500 means the model could not distinguish between classes. With those evaluations and interpretations, the model with the best performance on the first iteration was the RF.

We utilized the three models with the highest AUC value for the second iteration. We tested an automatic feature selection of the training and evaluation subsets to evaluate the impact on the performance of the AUC value. The results (see Table 5) showed an accuracy value higher than the first iteration for the three models, but only the SVM had a performance above .500 on the AUC value. The accuracy and classification error values did not present significant differences (as the variation between iterations was less than 0.100, and it did not get close to 1).

Although these two models (RF with no automatic feature selection and SVM with automatic feature selection) are the ones with an AUC above .500, it can be considered either RF or SVM as the best models. The weights of the variables selected for both models are presented in Table 6.

The SVM algorithm performed better when an automatic feature selection and generation were set, with an accuracy of 70.1 % and AUC of 0.645. This model is not wholly appropriate for profiling new employees, but it can be beneficial to track and predict the performance of retired personnel. The reason is that most attributes refer to job characteristics rather than socio-demographic information. The RF algorithm presented an accuracy of 62 % and AUC of 0.625. This model could help classify the future performance of new prospects and candidates as it includes attributes that refer to the socio-demographic profile of workers.

Table 3. Parameters of the classification models built in the second iteration

RF	SVM	DL
		Model metrics type: Binomial
	Gamma = 1.0E-4	Model id: rm-h2o-model-model-35472
Number of trees = 100	C = 10.0	Frame id: rm-h2o-frame-model-309724
Maximal depth = 4	Total number of support vectors: 1574	MSE: 0.23721279
	Bias (offset): –0.838	R^2: 0.0030576016
		logloss: 0.6673893

Table 4. Evaluation of classification models built in the first iteration

Model	Accuracy	Classification error	AUC
Naive bayes	.613	.387	.511
Generalized linear model	.611	.388	.525
Logistic regression	.614	.385	.507
Deep learning	.614	.385	.549
Decision tree	.613	.387	.500
Random forest	.620	.379	.625
Support vector machine	.621	.378	.550

Table 5. Evaluation of classification models with automatic selection and generation

Model	Accuracy	Classification error	AUC
Deep learning	.622	.378	.585
Random forest	.639	.361	.617
Support vector machine	.701	.299	.645

Table 6. Attribute weights of the classification models with highest accuracy and AUC values

Random forest		Support vector machine	
Attribute	Weight	Attribute	Weight
Age	.192	Job resignation reason = Promotion	.382
Team work	.134	Work experience in time = 3 to 6 months	.062
Work shift	.130	Work center = Biscuits	.052
Marital status	.080	Teamwork = Security	.040
Gender	.069	Work Experience in sector = Factory	.035
Work experience in time	.051	Experience in time = Over 1 year	.034
Work center	.050	Experience in sector = Office	.030
Working day	.041	Work shift = 12 hours	.027
Type of contract	.037	Marital status = Married	.017
Location	.028	Job resignation reason = Dismissal	.008
Salary	.022		
Rehired	.015		

Discussion

The results of the RF and SVM models show insights about the employee profiles in the company Malumex and their relation with turnover and withdrawal. The attributes of Salary, Location, Age, Teamwork, and Work Experience resulted in key factors involved with the employee profile, which are also mentioned in related literature as relevant attributes for the prediction of labor turnover (Dubey *et al*., 2016; Ranganathan & Kuruvilla, 2008; Sikaroudi *et al*., 2015; Attri, 2018; Kirimi & Moturi, 2016). Nevertheless, most of these studies focus on sectors such as IT and sales, not related to factors involved in turnover in low-skill positions and outsourcing companies.

The results of the models were limited by the partial absence of socio-demographic information, as well as variables like organizational climate and the culture of the company (Aranibar *et al*., 2018; Sikaroudi *et al*., 2015; Yousaf & Bhulai, 2016). We acknowledge that the results of the models could improve if this information were included. Furthermore, the variable Seniority was associated with 90 days (due to the initial contract of Malumex), and results could be different if different periods are considered.

In this case study, attributes such as Type of contract, Rehired, Job resignation reason = dismissal and promotion were found as employee profile attributes significant for predicting labor turnover and key for a classification model. The type of contract has also been reported as a labor turnover factor in the Latinamerica zone. Particularly, when the contract is temporary (Beccaria & Maurizio, 2020), which is frequently associated with low skilled jobs.

Recruitment policies could be suggested for the company of this study, taking into account these results and particularly the characteristics mentioned in Table 6 (relevant to classify whether an employee might leave the job before 90 days or not):

- RF showed that Work Experience in Time is relevant to classifying turnover. In contrast, SVM showed that Work Experience in Time equal to 3-to-6-months and over one year and Work Experience in sectors like Factory and Office are both relevant. Based on this, we can say that the company can avoid candidates who present those characteristics.
- RF presents the attribute Rehired as a relevant attribute, and SVM showed the attributes Job resignation reason = promotion and dismissal as relevant attributes as well. Therefore, a selection requirement would be to avoid candidates who have previously worked with the company, especially when the candidate leaves the company due to promotions and dismissal cases.
- SVM presented the attributes Work Center = Biscuits and Teamwork = Security as other characteristics that help classify turnover. These two variables represent areas of the company, which could mean that an employee who works in these areas has a high probability of leaving the company before 90 days. These insights are not enough to make a statement, but they can advise about opportunities for the company.
- Both SVM and RF presented Work shift as a key attribute, so an opportunity area for the company could be evaluating and improving this characteristic. In this line, RF presented Salary and Type of Contract as key attributes.

Even though we presented a case of an outsourcing company with low-skilled employee profiles, these results can be taken as a framework for similar companies to start their data mining project and identify the factors involved with labor turnover. In fact, in the Mexican hotel industry, the recruitment system has been reported as a key factor in decreasing labor turnover (Caldera *et al*., 2019). Thus, this new area could be the first candidate for applying the method presented in this work.

Some of the factors found in this case study could be presented in other companies, e.g., age and work experience (Attri, 2018; Sikaroudi *et al*., 2015). However, there must be many other factors that are not presented in this study due to the different information loaded into the models and the unique characteristics of the job profiles, employees, and companies.

For other companies, this study could be a guide to evaluate the variables related to Labor Turnover of Low Skills Employee Profiles. However, it does not mean the results can be normalized for every company. A particular case study and model should be constructed for every different company.

Conclusions

In this study, we utilized data mining techniques to build classification models for labor turnover based on a real data set of an outsourcing company located in Sonora, México. The objective was to classify labor turnover in low-skilled employees and evaluate different classification models to discover a list of relevant characteristics of employee profiles prone to turnover. The results showed that the support vector machine and random forest models presented the highest accuracy and AUC values. Furthermore, the attributes of Age, Team Work, Work Shift, Marital status, Experience in time, and Experience in the sector are key factors that contribute to classifying whether an employee might leave the job before 90 days. The comprehension of this type of characteristic of employee profiles can lead the company to establish withdrawal control and retention strategies directed towards a specific and known profile. For example, in future hiring processes, specific characteristics of the employee profile could be examined in detail or be helpful in the final selection decision phase.

The classification models and the factors associated with the low-skilled employee profiles discovered in this research can lead to a framework for Human Resources Outsourcing organizations and Human Resource managers when developing data mining projects. Furthermore, this research can help identify variables that should be collected in the selection processes of this type of organization.

As future work, we plan to:

a) Integrate different company attributes in the models (e.g., organizational culture, organizational climate).
b) Integrate additional employee attributes in the models (e.g., results of psychometric tests, job evaluations, and technical and professional skills.
c) Build an approach to apply and improve the model systematically in different companies. Furthermore, a business intelligence solution will be implemented to help decision-makers classify future employees' performance.

Acknowledgments

This work was supported by CONACYT. This work was supported by PROFAPI 2022_0047.

References

Al-Mamun, C. A., & Hasan, M. N. (2017). Factors affecting employee turnover and sound retentionstrategies in business organization: A conceptual view. *Problems and Perspectives in Management*, 63-71. http://dx.doi.org/10.21511/ppm.15(1).2017.06

Allen, D., Bryant, P., & Vardaman, J. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. A*cademy of management Perspectives,* 24, 48-64.

Almeda, C. (2017). La rotación de personal: Todo lo que debes saber sobre ella. Retrieved on may 25, 2019 on http://blog.talentclue.com/rotacion-de-personal

Aranibar, M. F., Melendres, V. D., Ramírez, M. C., & García, B. R. (2018). Los factores de la rotación de personal en las maquiladoras de exportación de Ensenada, B. C. *Revista Global de Negocios,* 6(2), 25-40.

Arokiasamy, A. R. (2013). A qualitative study on causes and effects of employee turnover in the private sector in Malaysia. *Middle-East Journal of Scientific Research*, 16(11), 1532-1541.

Attri, T. (2018). Why an employee leaves: Predicting using. (Ph.D. dissertation). National College of Ireland, School of Computing, Dublin. Retrieved on https://norma.ncirl.ie/id/eprint/3434

Barrett, B. (2020). Should human resources consider outsourcing human resource development. *Journal of Social Science and Humanities*, 3(1), 23-30. http://doi://10.26666/rmp.jssh.2020.3.4

Beccaria, L., & Maurizio, R. (2020). Rotación laboral en América Latina: Intensidad y diferencias entre países. *Revista Internacional del Trabajo,* 139, 171-204. https://doi.org/10.1111/ilrs.12160

Caldera, D., Zárate, L., & Arredondo, M. (2019). Rotación de personal en la industria hotelera en el estado de Guanajuato, México. *Revista Ibero Americana de Estrategia,* 18(4), 615-629. https://www.redalyc.org/articulo.oa?id=331267304006

Chhogyal, K., & Nayak, A. (2016). An empirical study of a simple Naive Bayes classifier based on ranking functions. Australasian Joint Conference on Artificial Intelligence, 324-331.

Chiavenato, I. (2011). *Administración de recursos humanos: El capital humano de las organizaciones*. Mcgraw-Hill Interamericana.

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, 19, 270. https://doi.org/10.1186/s12859-018-2264-5

Dessler, G. (2017). *Human Resource Management*. Pearson Education.

Dias, I., & Sousa, M. (2015). *Business intelligence applied to human resources management.* New Contributions in Information Systems and Technologies. Springer.

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models.* CRC press.

Dubey, R., Gunasekaran, A., Altay, N., Childe, S., & Papadopoulos, T. (2016). Understanding employee turnover in humanitarian organizations. *Industrial and Commercial Training*, 48(4), 208-214.

Gao, Y. (2017). *Using decision tree to analyze the turnover of employees*. (Ph.D. dissertation). Uppsala University.

Guo, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187, 27-48. https://doi.org/10.1016/j.neucom.2015.09.116

Hernández, R., Fernández, C., & Baptista, P. (1991). MCGRAW-HILL.

Herrera, P. A., Sánchez, M. L., Escobar, D. M., & Esparza, O. A. (2019). The maquiladora industry labor market in Mexico, an oligopsony: Effects of the new international division of labor. *Economía: teoría y práctica*, (51), 45-72. https://doi.org/10.24 275/etypuam/ne/512019/herrera

Jahan, R., Khan, N., Suman, P., & Singh, D. K. (2018). Classification & prediction techniques in data mining: A review. *International Journal of Advanced Research in Computer Science*, 9, 43.

Ji, S. (2016). Human resource outsourcing and risk management for SMEs. 6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016). Retrieved on https://doi.org/10.2991/mcei-16.2016.23

Kangas, M., Kaptein, M., Huhtala, M., Lämsä, A. M., Pihlajasaari, P., & Feldt, T. (2018). Why do managers leave their organization? Investigating the role of ethical organizational culture in managerial turnover. *Journal of Business Ethics*, 153, 707-723. https://www.jstor.org/stable/45022842

Kim, Y., & Ployhart, R. E. (2018). The strategic value of selection practices: antecedents and consequences of firm-level selection practice usage. *Academy of Management Journal*, 61, 46-66. https://psycnet.apa.org/doi/10.5465/amj.2015.0811

Kirimi, J. M., & Moturi, C. A. (2016). Application of data mining classification in employee performance prediction. *International Journal of Computer Applications*, 146, 28-35. https://doi.org/10.5120/ijca2016910883

Martínez-Luna, G. L. (2011). Minería de datos: Cómo hallar una aguja en un pajar. *Ingenierías*, 14, 55-63.

Mondy, R. W., & Noe, R. M. (2005). *Human resource managament*. Pearson Education.

Moreno, L. R., López, G. G., & Marín, M. (2015). Comportamiento de la tasa de rotación laboral en la industria maquiladora en Mexicali, Baja California, 2009-2013. *Revista Global de Negocios*, 3(4), 11-26. https://ssrn.com/abstract=2658732

Narkhede, S. (2018). Understanding AUC-ROC Curve. Towards Data Science. Retrieved on https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Pape, T. (2016). Prioritising data items for business analytics: Framework and application to human resources. *European Journal of Operational Research*, 252, 687-698. https://doi://10. 1016/j.ejor.2016.01.052

Ranganathan, A., & Kuruvilla, S. (2008). *Employee turnover in the business process outsourcing industry in India*. On Management practices in high-tech environments. IGI Global.

Ribes, E., Touahri, K., & Perthame, B. (2017). *Employee turnover prediction and retention policies design: A case study*. arXiv e-prints. https://doi.org/10.48550/arXiv.1707.01377

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12, 217-222.

Sikaroudi, E., Mohammad, A., Ghousi, R., & Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, 8, 106-121.

Tam, V. W., & Khoa, N. L. (2018). Power spectral and bispectral study of factors affecting employee turnover. Organization, *Technology & Management in Construction: An International Journal*, 10, 1727-1734. https://doi.org/10.2478/otmcj-2018-0006

Uzair, S., Majeed, A., & Shakeel, S. (2017). Recruitment, selection policies and procedure. *International Journal of Multidisciplinary and Current Research*, 5, 525-529.

Van-Belle, V., Van-Calster, B., Van-Huffel, S., Suykens, J., & Lisboa, P. (2016). Explaining support vector machines: a color based nomogram. *PLOS ONE*, 11(10). https://doi.org/10.1371/journal.pone.0164568

Yousaf, H. M., & Bhulai, S. (2016). Analysing which factors are of influence in predicting the employee turnover. Tech. Rep., Vrije Universiteit Amsterdam.

Cómo citar:

Márquez-Hermosillo, A., Rodríguez L. F., Salazar-Lugo, G. M. A., & Borrego-Soto, G. (2023). Employee profile and Labor turnover on outsourcing companies: A data mining approach. *Ingeniería Investigación y Tecnología*, 24 (04), 1-12. https://doi.org/10.22201/fi.25940732e.2023.24.4.031