REDECA
REVISTA ELETRÔNICA DO DEPARTAMENTO DE CIÊNCIAS CONTÁBEIS & DEPARTAMENTO DE ATUÁRIA E MÉTODOS QUANTITATIVOS DA FEA

EMPLOYEE TURNOVER INTENTION - MAPPING PROFILES UNDER A DECISION TREE PERSPECTIVE

INTENÇÃO DE ROTATIVIDADE DE FUNCIONÁRIOS - MAPEAMENTO DE PERFIS SOB UMA ÁRVORE DE DECISÃO

Vinícius Gomes Soares [1]

José de Jesús Pérez Alcázar [2]

Fernando Fagundes Ferreira[3]

ABSTRACT

This work aims to map some profiles having more propensity to quit prematurely a company. The analysis is important because affects the productivity of employees and it represents a high cost for companies around the world. The research applies a decision tree model in a study database of public domain with 1470 records, where it is possible to group profiles under 38 different variables to understand what can influence more the turnover. The result is a model with 81% of accuracy which has identified employees working overtime and new hires in the sales executive position with a higher risk of quitting prematurely the company. In some modeling approaches it is necessary focusing more on interpretability over performance. As the goal of this research is to map and understand key factors of turnover, the decision tree model is ideal. However, the model has a recall of 27%, which means that can predict about 1/3 of turnover cases. This paper contributes with a true modeling application towards People Analytics, sharing openly the model performance and discussing the features related to turnover. Companies can adapt this study in their databases in order to trace employees in turnover risk groups.

Keywords: People Analytics, HR Analytics, Turnover, Decision Trees.

RESUMO

Este trabalho tem como objetivo mapear alguns perfis com maior propensão a sair prematuramente de uma empresa. A análise é importante porque afeta a produtividade dos funcionários e representa um alto custo para empresas em todo o mundo. A pesquisa aplica um modelo de árvore de decisão em um banco de dados de estudo de domínio público com

[1] Msc. Student. Center for Interdisciplinary Research in Complex Systems, University of São Paulo, São Paulo – Brazil.

[2] Professor. Center for Interdisciplinary Research in Complex Systems, University of São Paulo, São Paulo – Brazil

[3] Professor. Center for Interdisciplinary Research in Complex Systems, University of São Paulo, São Paulo – Brazil; School of Philosophy, Sciences and Letters, University of São Paulo, Ribeirão Preto-Brazil. Corresponding author e-mail address: ferfff@usp.br (Fernando F Ferreira).

1470 registros, onde é possível agrupar perfis sob 38 variáveis diferentes para entender o que pode influenciar mais na rotatividade. O resultado é um modelo com 81% de acerto que identificou funcionários que trabalham horas extras e novos contratados na função de executivo de vendas com maior risco de desligamento prematuro da empresa. Em algumas abordagens de modelagem é necessário focar mais na interpretabilidade do que no desempenho. Como o objetivo desta pesquisa é mapear e entender os principais fatores de rotatividade, o modelo de árvore de decisão é o ideal. No entanto, o modelo tem um retorno de 27%, o que significa que pode prever cerca de 1/3 dos casos de rotatividade. Este artigo contribui com uma verdadeira aplicação de modelagem para Análise de Pessoas, compartilhando abertamente o desempenho do modelo e discutindo as características relacionadas à rotatividade. As empresas podem adaptar este estudo em seus bancos de dados para rastrear funcionários em grupos de risco de rotatividade.

Palavras-chave: Análise de Pessoas, Análise de RH, Rotatividade, Àrvores de Decisão.

1. Introduction

Employee turnover can be defined as the proportion of employees leaving an organization voluntarily or involuntarily. The voluntary leaving is attributed to whom is seeking better career opportunities in other companies. The involuntary leaving usually refers to low performance issues, layoffs or any restructuring in the company. The employee turnover may be positive for staff renew in order to get more high performing people in the team, but when these high performing decide to leave, there is a problem. Regardless the reasons, employee turnover has been proven to be costly and disruptive to any company (Lim, et al., 2017).

The employee turnover is one of the major tasks of HR managers and executives in general. The undesirable turnover results not only in high financial costs, but also in loosing intellectual capital and cumulative knowledge acquired for the employees. Therefore, manage and control the turnover are top priorities to keep the business growing sustainably (Lim, et al., 2017).

Since 2000s the amount spent with turnover was discussed. The Food Market Institute estimates that grocers in the U.S 5.8 billion in turnover costs annually (Holtom, et al. 2005). The expenses are associated to directed costs whenever an employee leaves and additional costs for training and new hires. The prediction cost is 6 to 9 month's salary on average each time a company replaces an employee, resulting for the U.S employer an amount of US$ 600 billion in 2018 (Chiat and Panatik, 2019).

Another associated cost was calculated by the United States Department of Labor, estimating that one-third of a new hire's salary is used to fill the vacant position. To replace a manager, supervisor, or a technical position the cost can range from 50% to 300% of position's annual salary (Belete, 2018).

The voluntary turnover can be caused by autocratic leaders in the work environment, demographic issues, like age, experience or marital status and other reasons linked to corporate questions like cultural fit, work-life balance, salary and career progression (Alam, et al., 2018).

Understanding deeply the main reasons for the employee turnover can incentive the companies take preventive actions to protect their staff and avoid unnecessary financial costs. Employing analytical techniques and data analysis is an efficient way to reach this goal and help to control the employee turnover.

The objective of this work is using the interpretative feature of the decision trees to understand the main reasons that may lead an employee leave its job. In the following section there are some examples of works using machine learning approaches trying to trace the turnover pattern. In Section 3 the method is thoroughly described with the database, variables, correlations, and the algorithm. The results are shown in section 4, the discussion in section 5 and the conclusion in section 6..

2. Related Works

The work of Alam (Alam, et al., 2018) uses a public turnover database to compare six different techniques: decision trees, random forest, support vector machine, multi-layer perceptron, which is a simpler neural network, naive bayes and k-nearst neighbor (knn). There are some important evaluation results to measure the model's efficiency and precision such as ROC curve and variables importance.

Aswale (Aswale and Mukul, 2020) explores a database from a hypothesis that employee turnover is directly linked to the job satisfaction and builds correlations between variables in a regression model.

DOI: 10.23925/2446-9513.2022v9id58575

Islam (Islam, et al., 2018) has a similar work using a public database but focusing only on the Random Forest technique. Sisodia (Sisodia, et al., 2017) also use the same public database with employee turnover information to compare 5 techniques, but shows more evaluation metrics such as accuracy, sensitivity and specificity.

The work of Zhao, (Zhao, et al., 2018) is one of the most complete analytical approaches regarding employee turnover. The publication brings an analysis of 2 different databases and compares techniques such as random forest, gradient boosting, extreme gradient boosting, random forest, decision trees, logistic regression, support vector machine and neural networks and k-nn. The goal is having distinctive KPIs of each model to study the performance and behavior of the techniques under different datasets.

There are 2 works dealing with bigger databases, above 10.000 records, got from internal companies' data collections. Cahyani and Budiharto (Cahyani and Budiharto, 2017) use a database with around 50.000 records, but don't go deeply in predictive and analytical approaches. Rombaut and Guerry (Rombaut and Guerry, 2018) explore a 13.000 records database from a Belgian company applying regression techniques and analyzing the variables qualitatively.

The works described above are more focused in comparing different techniques under the turnover perspective. This work intends to analyze the reasons behind turnover. To do that the decision tree algorithm is suitable because it is possible to interpret the results and define clusters where the turnover is significantly high.

3. Method

The method is divided into three subsections. Two of them are related do data, the first one has some descriptive statistics about the variables, besides the source of the database. The second one shows the manipulations needed to employ the decision tree method. The last subsection aims to share details about the decision tree algorithm.

a. Data

To employ the analysis is mandatory having a database with independent variables and a target or response variable which is the employee turnover. Kaggle[4] provides a public database for studies called IBM HR Analytics Attrition Performance.

The dataset has 1470 records of employees, where 237 (16.12%) left the company. There are 38 independent variables related to performance, work experience and demographic issues of each employee. The list of variables was separated into numerical and non-numerical ones. Table 1 shows the numerical variables and the mean, minimum and maximum values, most of the variables are self-explanatory, maybe the more difficult to interpret are 'EmployeeCount' and 'EmployeeNumber', which are just Ids for the employee, 'HourlyRate' and

[4] A community of databases and data science studies, competitions and hints in algorithms applications: https://www.kaggle.com

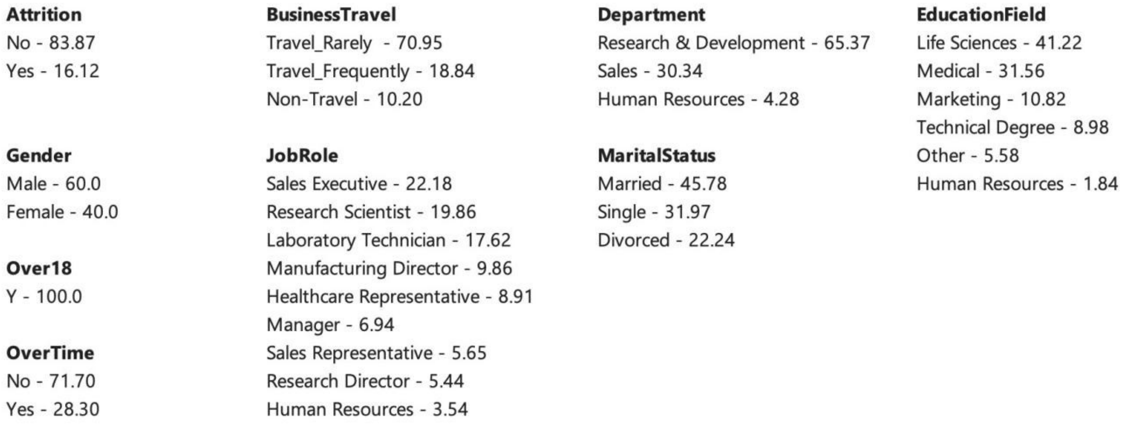
‘MonthlyRate’ which are related to the company’s cost with the employee and ‘PercentSalaryHike’ which is the percentage of the salary increased from one year to another.

Figure 1 shows the frequency of each category in the non-numerical variables. Another important visualization is the Pearson correlation map between variables showed in Figure 2. The redder the square is the lower is the correlation. It is possible to see that the predominant color is red, which is good for the modeling process because it means that there are more variability in the input variables. When the correlation between them is higher, it means that those variables can extract the same tendency.

Variable	**Mean**	**Max**	**Min**
Age	37	60	18
DistanceFromHome	9	29	1
Education	3	5	1
EmployeeCount	1	1	1
EmployeeNumber	1025	2068	1
EnvironmentSatisfaction	3	4	1
HourlyRate	66	100	30
JobInvolvement	3	4	1
JobLevel	2	5	1
JobSatisfaction	3	4	1
MonthlyIncome	6503	19999	1009
MonthlyRate	14313	26999	2094
NumCompaniesWorked	3	9	0
PercentSalaryHike	15	25	11
PerformanceRating	3	4	3
RelationshipSatisfaction	3	4	1
StandardHours	80	80	80
StockOptionLevel	1	3	0

TotalWorkingYears	11	40	0
TrainingTimesLastYear	3	6	0
WorkLifeBalance	3	4	1
YearsAtCompany	7	40	0
YearsInCurrentRole	4	18	0
YearsSinceLastPromotion	2	15	0
YearsWithCurrManager	4	17	0

Table I: Numerical variables of the study database



Attrition
No - 83.87
Yes - 16.12

Gender
Male - 60.0
Female - 40.0

Over18
Y - 100.0

OverTime
No - 71.70
Yes - 28.30

BusinessTravel
Travel_Rarely - 70.95
Travel_Frequently - 18.84
Non-Travel - 10.20

JobRole
Sales Executive - 22.18
Research Scientist - 19.86
Laboratory Technician - 17.62
Manufacturing Director - 9.86
Healthcare Representative - 8.91
Manager - 6.94
Sales Representative - 5.65
Research Director - 5.44
Human Resources - 3.54

Department
Research & Development - 65.37
Sales - 30.34
Human Resources - 4.28

MaritalStatus
Married - 45.78
Single - 31.97
Divorced - 22.24

EducationField
Life Sciences - 41.22
Medical - 31.56
Marketing - 10.82
Technical Degree - 8.98
Other - 5.58
Human Resources - 1.84

Figure 1: Non-numerical variables and frequency of each category

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis &Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

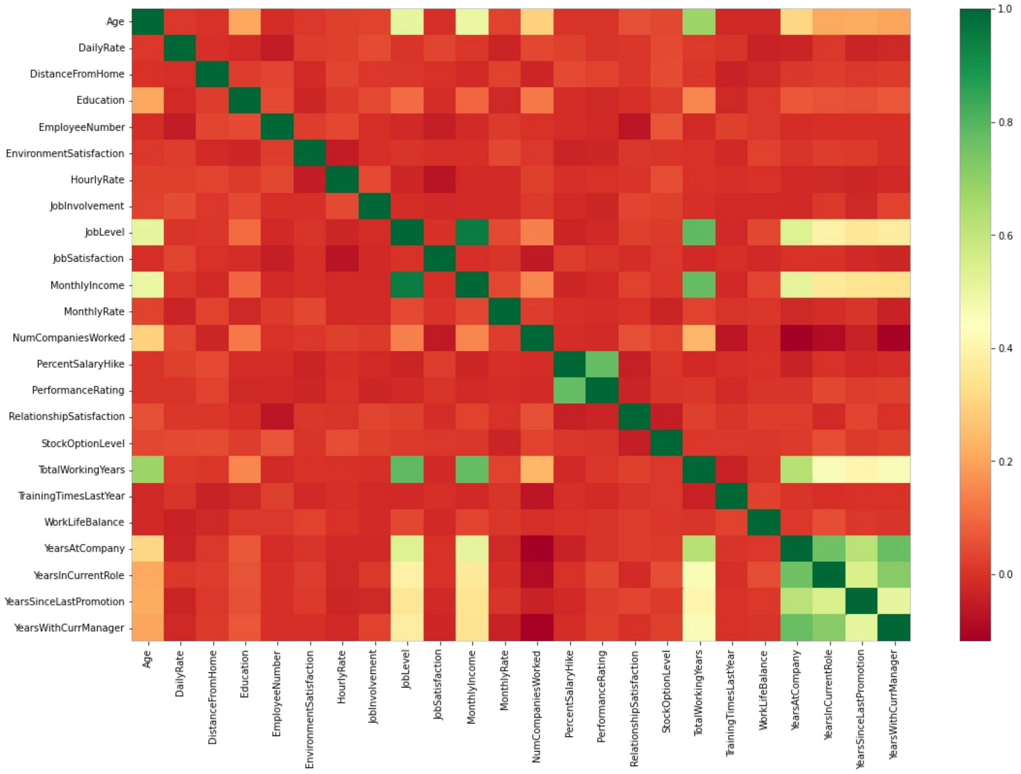


Figure 2: Correlation map between variables

b. Data Pre-processing

Before running the decision tree algorithm, it has been done a missing data check. For this database there were no variables with missing data, so it was not necessary applying any treatment on data.

The next step was splitting the dataset into dependent and independent variables. It was created a vector with the variable 'Attrition' to be the dependent variable and a vector without 'Attrition', 'EmployeeCounts', 'StandardHours', 'Over18' and 'EmployeeNumber' to be the independent variables. Those 3 variables besides 'Attrition' were removed because they don't explain the turnover event based on their features.

The third step was applying one-hot encoding on categorical variables in order to get a suitable input for the decision tree model. This technique creates an extended matrix, with the dimension of the matrix being the number of states or categories, and each dimension represents a category (Yu, *et al.,* 2022) One-hot encoding has been applied in the variables 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus' and 'OvertTime'. Figure 3 shows an example of one-hot encoding in the variable Business Travel.

DOI: 10.23925/2446-9513.2022v9id58575

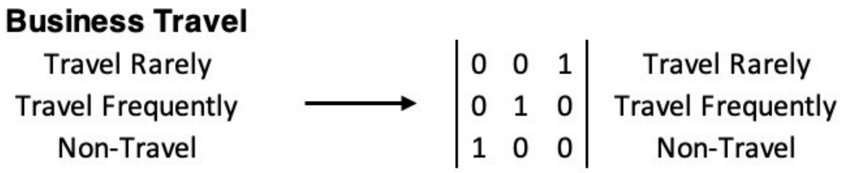


Figure 3: Example of One-Hot Encoding

Another step was splitting the database into training and test. 80% of the dataset has been used to train the model and 20% for test.

c. Decision Trees

Decision trees are based in divide-and-conquer approach and can be used to discover features and extract patterns in datasets. These characteristics, coupled with their intuitive interpretation, are some of the reasons decision trees have been used extensively for both exploratory data analysis and predictive modeling applications since 80's (Myles, *et al.,* 2004).

The main concepts of a decision tree are the nodes, branches, splitting, stopping criteria and prune. All of them are described below (Song and Ying, 2015).

The nodes can be divided into three different groups, root, which divide all records into two or more mutually exclusive subsets, internal nodes, representing one of the possible choices available at that point in tree structure having child nodes as a result of the above splitted parent nodes, and leaf nodes, representing the final combination of decisions or events (Song and Ying, 2015).

Branches create the hierarchy in which a decision tree model is formed and represent chance outcomes or occurrences that emanate from root nodes and internal nodes. Splitting means have input variables in parent nodes splitted into purer child nodes, this purity degree includes some criteria like information gain or Gini index (Song and Ying, 2015).

Stopping criteria must be included in order to get the best prediction or classification as possible and avoid problems like overfitting. Some of the stopping criteria are minimum number of records in a leaf, minimum number of records in a node prior to splitting, and the depth of any leaf from the root node. Lastly, pruning is used when the stopping criteria does not work well, and it is necessary removing nodes (Song and Ying, 2015).

Table 2 shows a comparison of decision tree algorithms according to Decision Trees Scikit-learn documentation (2022). The implementation of this work uses the 'DecisionTreeClassifier' function from the machine learning Python package scikit-learn with a maximum depth of 4 and the entropy criteria. This function uses an optimized version of CART algorithm for decision trees.

Methods	CART	C4.5	CHAID	QUEST
Measure used to select input variable	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square for categorical variables; ANOVA for continuous/ordinal variables
Pruning	Prepruning using a single pass algorithm	Prepruning using a single pass algorithm	Prepruning using a Chi-square test for independence	Post-pruning
Dependent variable	Categorical and continuous	Categorical and continuous	Categorical	Categorical
Input variables	Categorical and continuous	Categorical and continuous	Categorical and continuous	Categorical and continuous
Split at each node	Binary; split on linear combinations	Multiple	Multiple	Binary; split on linear combinations

Table II: Comparison of decision tree algorithms.

4. Results

This work used the confusion matrix method to evaluate the model itself and the tree structure to analyze the most important variables influencing on turn over. Confusion matrix is a two-way frequency table with two binary variables Actual (positive or negative) and Predicted (positive or negative), the four elements are called true negative (TN), true positive (TP), false negative, (FN), false positive (FP) (Zeng, 2020). Table 3 shows the confusion matrix of this decision tree model with the four elements used to evaluate the model metrics.

	Predicted Negative	**Predicted Positive**
Actual Negative	221 (TN)	15 (FP)
Actual Positive	42 (FN)	16 (TP)

Table III: Confusion Matrix.

As defined in the method section the database for test was defined as 20% of the total, which means 294 records. Table 4 shows 5 metrics to evaluate the model based on the results of the confusion matrix.

Method	**Formula**	**Value**
Sensitivity (Recall)	TP/(TP+FN)	27%
Specificity	TN/(FP+TN)	93%
Accuracy	(TP+TN)/N	81%
Precision	TP/(TP+FP)	52%
F-Score	2x(PxS)/(P+S)	0,35

Table IV: Evaluation metrics of the decision tree. N (total elements), P (precision), S (sensitivity).

The accuracy of the model is 81% and recall is 27%, which means that is possible to predict one quarter of employees that are going to quit the company. The accuracy is high because the model has a better performance in predicting who is not going to quit the company. The tree structure is represented in Figure 4 the bluer a leaf is the more turnover proportion it has. Both Figure 4 and 5 aim to show the structure and the pathway leading to more turnover. The leaves details are showed in Figure 6.

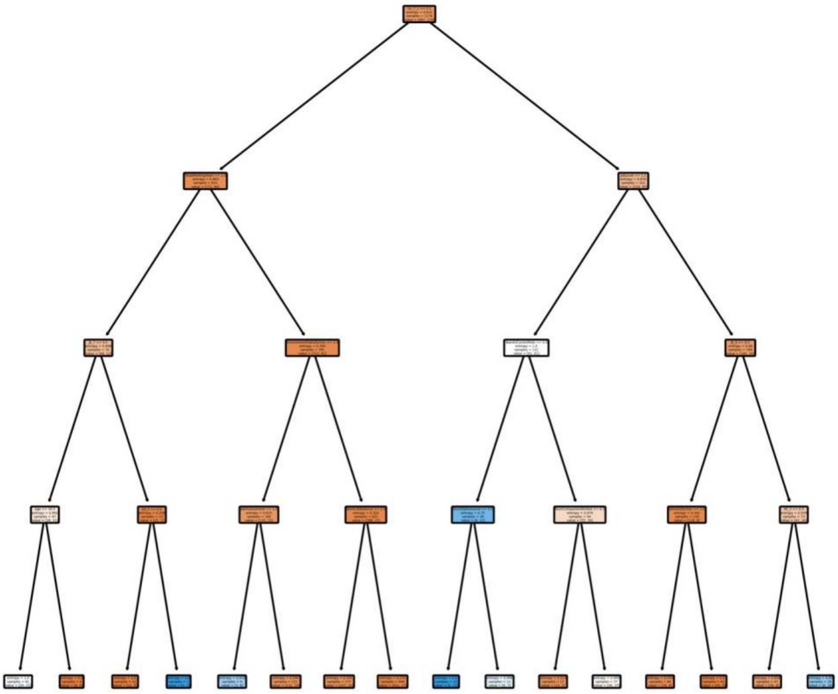


Figure 4: Tree Structure.

5. Discussion

The interpretability of decision tree models allows a discussion over the splits and can help to understand the turnover reasons for this database. Figure 5 shows the path that will be discussed. The goal is analyzing the nodes and splits until the bluer leaves. There are two leaves that will not be analyzed because the sample inside is very low.

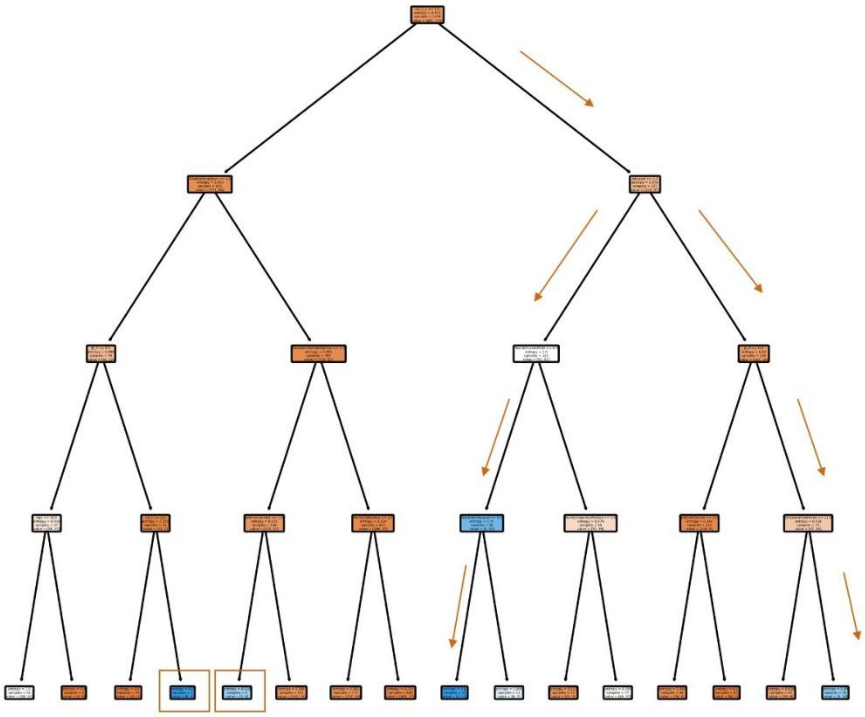


Figure 5: Tree Structure.

The interpretability of decision tree models allows a discussion over the splits and can help to understand the turnover reasons for this database. Figure 4 shows the path that will be discussed. The goal is analyzing the nodes and splits until the bluer leaves. There are two leaves that will not be analyzed because the sample inside is very low.

Each leaf in the selected paths is shown in the Figure 6. The first split is the variable ‘Overtime’, which makes sense. In addition to the fatigue factors of employees, overtime results in various effects on the job performance and significantly affects the turnover intention (Junaidi, *et al.,* 2020).

The second split was the job level, which means that there is a segregation between the lower job levels and the others. The third split was the job role, there is a segregation between sales executives and the others. The fourth split was the marital status and segregated singles from other status. This path described all the arrows in the right on Figure 5.

After the job level split there is another path going to the left. The next split on this path is years in current role segregating who has less than a year of the others and the next split segregates who has the lower stock option level of the other group.

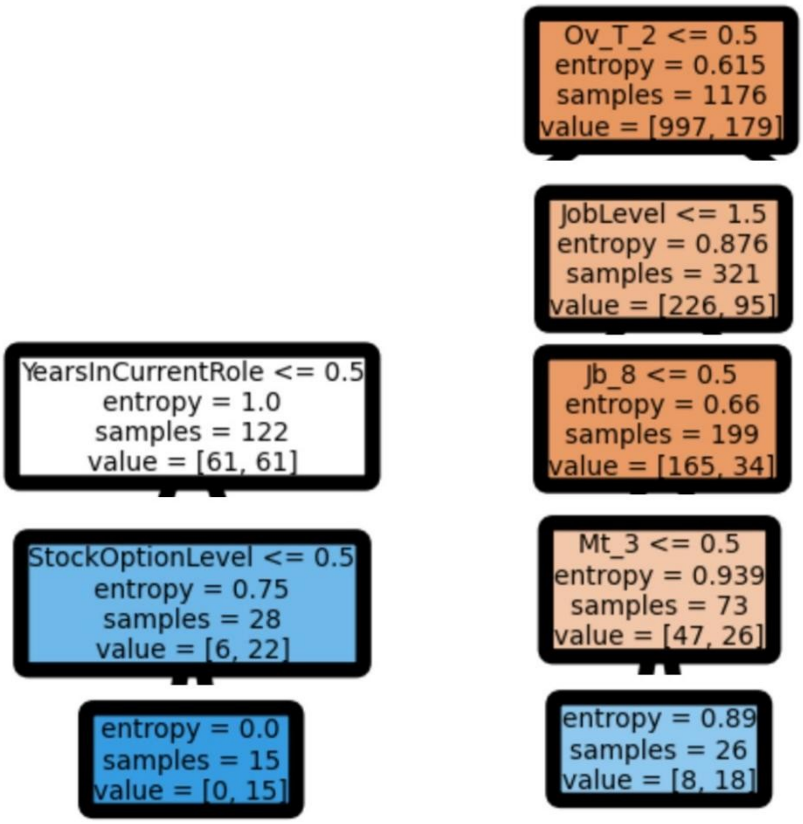


Figure 6: Selected leaves of the decision tree model.

6. Conclusion

This decision tree model shows that who is a single sales executive, working overtime in a starting position or who is working overtime in a low-level position and has not completed a year in the company and does not receive any stock option level has more chance to quit the company.

These mapped profiles could have more attention of the HR in order to prevent and avoid the turnover. Each company can have a different result, but the analysis of the turnover influencing factor of this database instigates a reflection about the mapped profiles.

Usually new employees who don't work in key positions are not close to company's strategies and are more susceptible to leave or accept an offer from the market. People overworking tend to have less work-life balance and be unhappy, being an important decision point in the process of leaving a company. Sales executive is a position with very strict goals and high pressure most of times, also contributing to the turnover process.

DOI: 10.23925/2446-9513.2022v9id58575

References

A. Belete. (2018), "Turnover intention influencing factors of employees: an empirical work review", Journal of Entrepreneurship & Organization Management.

A. D. Cahyani, W. Budiharto. (2017), "Modeling intelligent human resources systems (irhs) using big data and support vector machine (svm)", in: proceedings of the 9th International Conference on Machine Learning and Computing, pp. 137–140.

A. Junaidi, E. Sasono, W. Wanuri, D. Emiyati. (2020), "The effect of overtime, job stress, and workload on turnover intention", Management Science Letters 10 pp. 3873–3878.

A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, S. D. Brown. (2004), "An introduction to decision tree modeling", Journal of Chemometrics: A Journal of the Chemometrics Society 18 pp. 275–285.

A. Lim, J. Loo, P. Lee. (2017), "The impact of leadership on turnover intention: The mediating role of organizational commitment and job satisfaction", Journal of Applied Structural Equation Modeling 1 pp. 27–41.

Chiat, L. C., & Panatik, S. A. (2019). "Perceptions of employee turnover intention by Herzberg's motivation-hygiene theory: A systematic literature review", Journal of Research in Psychology, 1(2) pp. 10-15.

D. S. Sisodia, S. Vishwakarma, A. Pujahari. (2017), "Evaluation of machine learning models for employee churn prediction", in: International Conference on Inventive Computing and Informatics (ICICI), IEEE, pp. 1016–1020.

E. Rombaut, M.-A. Guerry. (2018), "Predicting voluntary turnover through human resources database analysis", Management Research Review.

G. Zeng. (2020), "On the confusion matrix in credit scoring and its analytical properties", Communications in Statistics-Theory and Methods 49 pp. 2080–2093.

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Inderrieden, E. J. (2005). "Shocks as causes of turnover: what they are and how organizations can manage them." Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 44(3), pp. 337-352.

L. Yu, R. Zhou, R. Chen, K. K. Lai, (2022) "Missing data preprocessing in credit classification: One-hot encoding or imputation?", Emerging Markets Finance and Trade 58 pp. 472–482.

M. K. Islam, M. M. Alam, M. B. Islam, K. Mohiuddin, A. K. Das, M. S. Kaonain. (2018), "An adaptive feature dimensionality reduction technique based on random forest on employee turnover prediction model", in: International Conference on Advances in

Computing and Data Sciences, Springer, pp. 269–278.

M. M. Alam, K. Mohiuddin, M. K. Islam, M. Hassan, M. A.-U. Hoque, S. M. Allayear. (2018), "A machine learning approach to analyze and reduce features to a significant number for employee's turn over prediction model", in: Science and Information Conference, Springer, pp. 142–159.

N. Aswale, K. Mukul. (2020), "Role of data analytics in human resource management for prediction of attrition using job satisfaction", in: Data Management, Analytics and Innovation, Springer, pp. 57–67.

Scikit-learn in Python: Decision trees (2022)., https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart, 1.10

Y.-Y. Song, L. Ying. (2015), "Decision tree methods: applications for classification and prediction", Shanghai archives of psychiatry 27 p.130.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018), "Employee turnover prediction with machine learning: A reliable approach", in Proceedings of SAI intelligent systems conference pp. 737-758. Springer, Cham.

Recebido: 29/06/2022
Aceito: 07/07/2022