COST OPTIMIZATION OF HUMAN RESOURCE MANAGEMENT THROUGH THE IMPLEMENTATION OF STRATEGIES TO REDUCE EMPLOYEE TURNOVER

[a]ZDENKO STACHO, [b]KATARÍNA STACHOVÁ
[c]ALEXANDRA BAROKOVÁ

University of Ss. Cyril and Methodius, Institute of Management, 91701, Trnava, Slovakia
email: [a]zdenko.stacho@ucm.sk, [b]katarina.stachova@ucm.sk
[c]barokova.alexandra@gmail.com

The result was created in solving the VEGA (No. 1/0038/22) Application of competitive digital games for the team cohesion development and social adaptation of Generation Z and project (KEGA 012UCM-4/2022) Human Resources Management in a Digital World A Bilingual (Slovak -English) Course Book with E-learning Modules based on Multimedia Content.

Abstract: Employee turnover has long been a concern for businesses of all sizes and industries. To address this challenge, businesses increasingly turn to data-driven approaches to predict and prevent employee turnover. The article's authors focus on the effectiveness of processing large data through a predictive machine learning model in the context of employee turnover. The real dataset used for company research contained 14999 data points and 16 elements, while each row represented one employee. The Python programming language was used for the development of the prediction model. The study's results showed the relevance of HR analytics for companies, accurately predicting employee departures using a machine learning model and achieving significant cost savings by implementing the model.

Keywords: Big Data Approach Employee Turnover, Human Resource, Replacement Cost, Cost Analysis

1 Indroduction

Human Resource (HR) analytics has become a key strategic tool for organizations, enabling them to gain valuable insights into their workforce and make data-driven decisions. One of the key areas where HR analytics can be applied is in the prediction of employee turnover. Employee turnover can have significant financial and organizational impacts, including the loss of valuable skills and knowledge, decreased productivity, and increased recruitment and training costs. Predictive analytics can help organizations to identify employees who are at risk of leaving and take proactive measures to retain them, ultimately reducing the overall cost of turnover.

The data approach in HR involves the use of data to make informed decisions about various HR functions, such as recruitment, training, performance management, compensation. and engagement (Cappelli, Meister 2018).

Big data refers to the vast and complex sets of data that are too large and dynamic to be processed using traditional data processing techniques (Maurya, Sandip 2019). With the advancements in technology, organizations can now collect and store massive amounts of data on various HR-related aspects (Boudreau Cascio 2017). By leveraging big data analytics techniques, HR professionals can identify patterns, correlations, and insights that were previously impossible to detect, thus enabling them to make more informed decisions. As such, the adoption of big data in HR has the potential to revolutionize HR practices and significantly enhance organizational performance. The analysis of big data requires specialized skills and technologies, including machine learning and data mining (Petry Jäger 2021).

1.1 Cost of employee turnover

The data approach is becoming increasingly important in employee turnover, and organizations that fail to embrace it risk falling behind their competitors (Tursunbayeva et al. 2018). By analyzing data on employee turnover, companies can detect factors that contribute to high turnover rates and develop strategies to address these issues (Boukottaya, Khemakhem 2019). Therefore, it can be concluded that there is a research opportunity to investigate the connection between HR data approaches and employee turnover.

Employee turnover is a major challenge for human resource managers and has a significant impact on organizational performance, productivity, and profitability (De Winne et al. 2019). Understanding the causes and factors contributing to employee turnover can help organizations develop effective strategies to mitigate the negative impact of employee turnover (Hom et al. 2019).

employee turnover can be voluntary or involuntary. With this perspective authors looked at the opportunities and threat that are associated with the fact of losing the employee (Gupta, Bhatia 2021). The defined opportunities were – opportunity for fresh talent, cost savings and increased diversity.

When employees leave, the organization has the opportunity to bring in new talent with fresh ideas and perspectives. This can lead to innovation and creativity within the organization. New employees may also have specialized skills that the organization previously lacked, which can help the organization stay competitive in the marketplace. If employees who have been underperforming or are highly compensated leave, the organization may be able to save money on salaries and benefits. This can help the organization allocate its resources more efficiently. Additionally, if the organization has excess staff due to changes in the business or a slowdown in operations, turnover can help reduce the number of employees without the need for layoffs. New hires can bring a greater diversity of skills, experiences, and backgrounds to the organization. This can help the organization become more competitive in its industry and better serve its customers. A more diverse workforce can also help the organization better understand the needs of different market segments and adapt to changing customer demands (Kehoe, Wright 2013).

When employees leave, it can take time for the remaining employees to adjust to new roles or for new hires to get up to speed. This can lead to decreased productivity, which can have a negative impact on the organization's performance. When experienced employees leave, the organization may lose valuable knowledge and expertise that is difficult to replace. This can lead to decreased efficiency and productivity and can also lead to increased costs associated with training new employees. Frequent employee turnover can have a negative impact on employee morale and can lead to a perception that the organization does not value its employees. This can lead to decreased engagement and productivity, as well as increased turnover in the long term (Kehoe, Wright 2013).

Tracking the costs associated with employee turnover provides valuable data that companies can use to improve their bottom line, retain top talent, and gain a competitive advantage in their industry (Cascio 2018).

Employee turnover can be costly for companies due to the direct and indirect costs associated with it. By tracking these costs, companies can identify areas where they can reduce costs and improve their bottom line. Understanding the costs associated with employee turnover can help companies make strategic decisions related to recruitment, training, and compensation (Leigh, DePaul-Haddock 2017). By identifying which areas of the organization have the highest turnover rates and associated costs, companies can develop targeted strategies to address these issues and improve overall performance. Companies that effectively manage employee turnover can gain a competitive advantage over their peers. By retaining top talent and minimizing the associated costs, companies can improve their reputation, attract more qualified candidates, and improve their overall performance (Noe et al. 2017).

Employee turnover can be costly for organizations, both in terms of direct and indirect costs. Direct costs include expenses incurred due to the departure of an employee, such as separation pay, recruitment costs, and training expenses for new hires.

Indirect costs are more difficult to quantify and can include decreased productivity, decrease morale, and lose institutional knowledge (Leigh, DePaul-Haddock 2017).

Direct costs are easier to calculate as there are precise numbers. There are several methods that companies are using to enumerate the costs of turnover. To calculate the indirect costs of employee turnover, organizations can take several approaches, including analyzing the workloads of remaining employees before and after an employee's departure to estimate the additional work they have to do, calculating the time it takes for new employees to reach the same level of productivity as the departing employee based on historical data and industry benchmarks, and gathering feedback from employees to estimate the impact of decreased morale and job satisfaction (Aldaihani, Alduais 2021). Organizations can also estimate the cost of lost institutional knowledge and relationships by analyzing the value of those assets and the time it takes to rebuild them (Mondore, Douthitt 2021). Overall, calculating the indirect costs of employee turnover requires a combination of quantitative and qualitative data analysis, and it may require input from various stakeholders within the organization.

It is used to determine the total cost associated with hiring a new employee, including recruitment costs, advertising, interviewing, and other administrative costs. This method is helpful in analyzing the effectiveness of the company's recruitment process, evaluating recruitment strategies, and assessing the impact of recruitment costs on the company's bottom line (Phillips 2016). The formula for calculating the cost-per-hire is as follows:

Cost-per-hire = (Internal costs + External costs) / Total number of hires

However, there are some limitations to the cost-per-hire method. One major limitation is that it does not take into account the quality of the new hires or the long-term impact of the recruitment process on the company's bottom line. Also, the method can be time-consuming and resource-intensive to implement, especially for companies with a large number of hires (Cascio 2018).

The turnover rate method is a common way to calculate employee turnover, which measures the number of employees who leave an organization in a given time period, divided by the average number of employees in the same time period. It is a simple and straightforward formula that provides a quick way to assess employee turnover rates (Noe et al. 2017). The formula for calculating turnover rate is:

Turnover rate = (Number of employees who left during the period / Average number of employees during the period) x 100

The replacement cost method is based on the premise that the cost of replacing an employee is equivalent to the sum of the costs associated with recruiting, selecting, and training a new employee to perform the same duties and responsibilities as the former employee. Replacement cost analysis considers the direct and indirect costs of employee turnover to the organization. The formula for calculating the replacement cost is as follows (De Cieri, Kramar 2015):

Replacement Cost = (Cost of recruiting + Cost of selection + Cost of training) * Number of employees lost

As a benefit this method provides a comprehensive estimate of the cost associated with employee turnover, can inform retention strategies, and improve recruitment effort, and helps organizations to quantify the financial impact of employee turnover (Baker, Baker 2018). However, the method may not capture all of the indirect costs associated with employee turnover, such as lost productivity, decreased morale, and decreased customer satisfaction and may not consider the unique circumstances of each employee who has left the organization.

Cost analysis is a method used to calculate the direct and indirect costs associated with employee turnover. The theory behind this method is that companies need to account for all the costs related to losing an employee to make informed decisions about how to reduce turnover and its associated costs. The formula for cost analysis is (Vidal 2017):

Total cost of turnover = (Cost of replacing an employee + Cost of lost productivity + Cost of lost knowledge and expertise) / Total number of employees

One of the benefits of the cost analysis method is that the method provides a comprehensive picture of the costs associated with employee turnover, including both direct and indirect costs. It also enables companies to make informed decisions about retention strategies and investments in employee development and training. Another benefit is that the method helps to identify the root causes of turnover and develop targeted solutions to reduce it (Sturman, Clarke 2019).

The limitations of the cost analysis method include potential difficulties to accurately calculate some indirect costs. Also, the fact that it is a retrospective method that does not provide insight into the factors that contribute to turnover (Ramlall 2017).

The costs of employee turnover can be connected to big data in HR analytics by leveraging data to understand and mitigate the drivers of turnover. With the help of big data, organizations can track and analyse a variety of factors, such as employee engagement, satisfaction, and performance, that contribute to employee turnover. By collecting and analyzing this data, companies can identify patterns and trends to predict and prevent future turnover. This can lead to significant cost savings, as the costs of replacing employees can be substantial. In addition, big data can help companies optimize their recruitment and retention strategies, allowing them to attract and retain the best talent while minimizing costs associated with turnover (Marr 2019).

1.2 Big data approach to employee turnover

The big data approach to employee turnover involves the use of data analytics to identify and predict factors contributing to employee turnover (Lienert 2018). The following are the steps involved in the big data approach to employee turnover – data collection, data analysis, identification of factors, prioritization of factor, intervention development and intervention implementation.

The first step in the big data approach to employee turnover is data collection. Organizations collect data on employee demographics, job roles, performance, compensation, benefits, job satisfaction, turnover rates, and other relevant variables. The second step is data analysis, which involves the use of statistical techniques, machine learning, and predictive modelling to identify patterns, correlations, and trends in the data. The third step is the identification of factors that contribute to employee turnover. Based on the data analysis, HR professionals can identify the factors that are most strongly associated with employee turnover. Some common factors include low job satisfaction, inadequate compensation, lack of opportunities for growth and development, poor management, and work-life balance issues (Silva et al. 2019). Once the contributing factors have been identified, the next step is to prioritize them based on their importance and impact on employee turnover. This allows HR professionals to focus their efforts on the most critical factors and develop targeted interventions to address them. The fifth step is the development of interventions to address the factors contributing to employee turnover. These interventions may include changes to compensation and benefits packages, improvements in work-life balance policies, management training, and other initiatives aimed at improving employee satisfaction and reducing turnover rates (Kouloumpis et al. 2019). The sixth step is the implementation of the interventions developed in step five. This involves rolling out the changes to the organization and monitoring their effectiveness in reducing employee turnover rates. The final step is the evaluation of the effectiveness of the interventions. HR professionals can use data

analytics to track changes in employee turnover rates and identify whether the interventions were successful in reducing turnover rates.

Overall, the big data approach to employee turnover is an effective way for organizations to identify and address factors that contribute to employee turnover. By using data analytics to gain insights into employee behavior and preferences, HR professionals can develop targeted interventions that improve employee satisfaction and reduce turnover rates. This not only helps to retain valuable employees but also saves the organization money on recruitment and training costs (Lienert 2018).

The research published so far focuses primarily on exploring the possibilities of collecting and storing vast amounts of data on various aspects related to human resources (Boudreau, Cascio, 2017), prerequisites such as specialized skills and technologies, including machine learning and data mining at work from Big Data (Petry, Jäger 2021), or analysis employee turnover and its impact on organizational performance, productivity, and profitability (De Winne et al. 2019; Noah et al. 2017, De Cieri & Kramar 2015; Vidal, 2017). However, there is no study analyzing the actual accuracy of identifying the most common reasons for employee departures through a predictive machine learning model. There are also no studies analyzing the impact of deploying a predictive model in relation to reducing fluctuations in the company's turnover rate or the impact of implementing a predictive model on the costs associated with human resources management.

The research presented in this paper expands the knowledge base mainly by demonstrating answers to the authors' sanitized research questions:

- To what extent does the predictive model developed using machine learning techniques accurately identify the most common reasons for employee departures based on the input data provided?

The authors deduced that the predictive model using machine learning techniques can accurately find the most common reasons for employee departures. This question implies that such a model can provide valuable insights into why employees leave their jobs and help companies take appropriate action to retain their employees.

- Is compensation identified as the primary reason for employee departures, based on a higher frequency of this factor compared to other reasons in the data?

The authors presumed that compensation would be the primary reason for employee departures, as indicated by a higher frequency of this factor compared to other reasons in the data. This question implies that compensation is a significant factor in employee turnover and that addressing compensation-related issues can help reduce employee turnover.

- Does the deployment of the predictive model, within a 6-month result in a reduction of fluctuations in the company's turnover rate, thereby supporting the company in adjusting internal initiatives to lower turnover?

The aim of this question is to test whether implementing a predictive model will reduce fluctuations in the company's turnover rate. The authors assumes that the initial results of the predictive analytics will help the company adjust internal initiatives to lower turnover. Based on an agreement with internal HR, the expected turnover will be 10% lower due to the usage of information from the predictive model.

- To what extent does the implementation of the predictive model result in savings from the planned HR budget, as indicated by a reduction of costs over a specified period?

The authors investigate whether implementing the predictive model will result in cost savings for the company. It implies that implementation of the predictive model will result in savings by a reduction in recruitment costs over a specified period.

2 Methodology

The authors contacted an international company that agreed to cooperate on the model development and implementation. This company operates in multiple markets and employs more than a million employees across the globe. This company desired to stay anonymous due to the sensitivity of the shared data. After initial contact the authors agreed with HR representatives what type of data should be tracked and collected 12 months after this initial meeting. After this period the dataset has been firstly reviewed by the internal HR department and cleaned based on GDPR regulations. The cleaning meant substitution of an employee name by random employee number and deleting all sensitive information. The authors worked with data connected to the number besides employee name. The real company dataset contained 14999 data points (rows) and 16 elements (columns). Each row represented one employee. The turnover tracked in this dataset was 26%. The dataset tracked Age, Job level (seniority), Years at company, Total working years, Overtime, Promotion within the last three years and results from internal employee survey. This survey tracked job satisfaction, relationship satisfaction and work life balance on the scale from one to ten, one being the lowest and 10 being the highest value. The survey was conducted in May 2021 by internal tool. The data were tracked from three countries – Slovakia (14%), Czech Republic (28%) and Germany (58%).

Description of the selected software: Python is a high-level, interpreted programming language that was created in the late 1980s by Guido van Rossum. It is widely used for web development, data analysis, artificial intelligence, scientific computing, and more. Python is known for its clear syntax and readability, making it easy for developers to write and maintain code (Hakeem, Haris 2020). The authors decided to work with Python programming language, because of several reasons. Firstly, it is an open-source language, which means that anyone can use it, modify it, and distribute it without any restrictions. It is also a cross-platform language, which means that it can run on various operating systems such as Windows, Linux, and macOS. Secondly, the system has numerous libraries and frameworks that make it easy to work with data (Ramalho 2015). Some of the most popular libraries for data analysis and predictive analytics in Python include NumPy, Pandas, Matplotlib, and Scikit-learn. Thirdly, in predictive analytics, Python can be used for a range of tasks, including data pre-processing, feature engineering, model selection and training, and model evaluation. In conclusion, Python is a versatile programming language that has become an essential tool for predictive analytics. As the demand for data-driven insights continues to grow, Python is likely to remain a popular choice for predictive analytics for years to come (Ramalho 2018).

For the prediction model development Team Data Science Process (TDPS) framework was followed. It is generally used for data science projects that was developed by Microsoft (Taylor et al. 2019).

predictive model was built based on machine learning algorithms. The authors used four modelling techniques - logistic regression, decision tree classifier, k-nearest neighbours, and random forest:

- Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable.
- K-nearest neighbours is a non-parametric method used for classification and regression.
- Decision tree classifier divides the data into smaller subsets based on the values of the input features, creating a tree-like structure.
- Random forest is an ensemble learning method that constructs a multitude of decision trees.

The data were processed using each of the four modelling techniques in the same manner. The dataset was split into testing

and training data, and the training set was used to fit the model by adjusting its parameters to minimize the error between predicted and actual data. The aim of fitting the training set to the model was to find the set of parameters that best capture the relationship between the input features and the output variable, enabling the model to make accurate predictions on new, unseen data. The result of each modelling technique was a confusion matrix, which is a table that compares the predicted values of the model with the actual values of the testing data. The confusion matrix provides information on the true positive, false positive, true negative, and false negative predictions made by the model, which is used to evaluate its performance (Zhang et al. 2021).

To compare the performance of these models, several metrics were used, including accuracy, precision, recall, the area under the curve (AUC) curve, and the F1 score. Accuracy measures the percentage of correctly classified instances, while precision measures the percentage of true positives among all predicted positives. Recall measures the percentage of true positives among all actual positives. The AUC (Area under the curve) is a metric that measures the overall performance of the model across all possible thresholds, and it represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. The F1 score is the harmonic mean of precision and recall. All these characteristics were displayed in a table for the comparison based on which one of the four models (logistic regression, decision tree classifier, k-nearest neighbours, and random forest) was selected for further analysis.

Both the qualitative and quantitative methods were used. Qualitative methods were used for gathering the employee attributes and demographics for predicting turnover as well as finding the retention factors for retaining a valuable employee, whereas the quantitative methods were used for weighting the factors influencing the turnover and analyzing the accuracy of the prediction model built for predicting turnover.

Exploratory data analysis was used to examine the dataset gathered from the real company. It involved analyzing the data to identify patterns, outliers, and other features that could be used to inform the development of the predictive model.

Factor analysis identified the fundamental factors that drive observed patterns in a dataset. By identifying these factors, it was possible to predict future outcomes based on the relationships between these essential factors and the outcome variable.

Regression analysis predicted the value of a dependent variable based on one or more independent variables, such as linear regression, logistic regression, and multiple regression.

Correlation analysis: This method was used to identify relationships between the variables in the dataset. Correlation analysis measures the degree to which two variables are related, with values ranging from -1 to +1. A positive correlation indicates that the variables are positively related, while a negative correlation indicates that the variables are negatively related.

Machine learning algorithms: This method involves using algorithms to train a model to predict outcomes based on input data. Machine learning algorithms that were used are logistic regression, decision tree classifier, K-nearest neighbours, random forest.

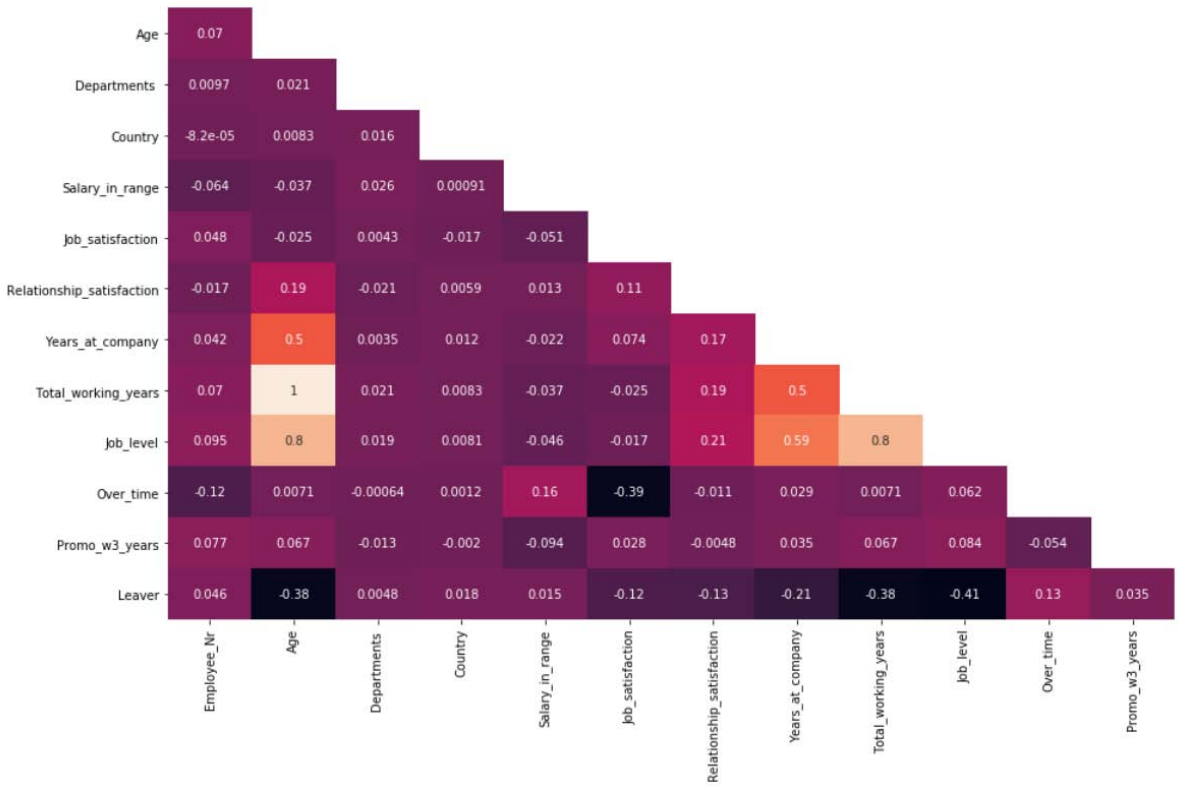
Modelling involved building and training a machine learning model using a dataset of employee data, and then using the trained model to predict which employees were likely to leave the company in the future.

3 Results

3.1 Correlation matrix

To gain a better understanding of the relationships between variables in the dataset, a correlation matrix was generated (Graph 1). On the Graph 1 two columns in correlation matrix were missing “Type of leaver” and “Reasons for leaver”, both columns contained text and therefor were not part of this comparison, moreover the correlation with column leaver is obvious. The correlation matrix also was interpreted visually using colours, with brighter colours indicating stronger correlation, black color illustrated negative correlation. A positive correlation was indicated by a plus value, meaning that as one variable increases, so did the other, while a negative correlation was indicated by a minus value, meaning that as one variable increased, the other decreased.

Graph 1: The correlation matrix of the real company dataset.



The matrix shows that there was a positive correlation between age and total working years (1), as well as between job level and total working years (0.8). This indicated that as employees got older or progressed to higher job levels, they tended to accumulate more working experience. Another positive correlation was between over-time and leaver (0.13), which suggested that employees who work overtime were more likely to leave the company.

Furthermore, the matrix reveals a moderate negative correlation between age and leaver (-0.38), indicating that younger employees were more likely to leave the company compared to their older counterparts. This trend was also reflected in the correlation between total working years and leaver (-0.38), suggesting that employees who worked for shorter periods of time were more likely to leave the company compared to those who had been with the company for longer periods.

The matrix also reveals a positive correlation between promo_w3_years and job level (0.084), which suggests that employees who received promotions within three years tended to progress to higher job levels. This may indicate that the company had a strong internal promotion system in place, which could motivate employees to stay with the company for longer periods of time.

Overall, this correlation matrix provided valuable insights into the relationships between various employee-related variables. By understanding these relationships, organizations can make informed decisions to retain employees and improve job satisfaction, which can lead to a more productive and successful workforce. This correlation matrix shows the relationships between various employee-related variables. It was important to understand how these variables are related to one another in order to identify patterns and make informed decisions that can benefit both the employees and the organization. Having completed the exploratory data analysis (EDA), the authors started to develop machine learning models to predict employee turnover

3.2 Develop machine learning models to predict employee turnover

Four different algorithms were explored- logistic regression, decision tree, k-nearest neighbours (KNN), and random forest - to identify the best approach for predicting whether an employee is likely to leave the company or not. Each of these models has its strengths and weaknesses, and by comparing their performance, the best algorithm for the dataset was selected. The authors used programming language Python for this analysis. For the better transparency, the authors summarized the results of the models in one table (Table 1).

Table 1: Summary of the models' performance

	Accuracy (on test set)	Precision	Recall	F1 score	AUC curve
Logistic regression	0.99	0.97	1.0	0.99	0.99
Decision Tree Classifier	0.99	1.0	1.0	1.0	0.99
KNN	0.79	0.66	0.48	0.58	0.80
Random Forest	0.99	0.99	0.99	0.99	0.99

Logistic Regression and Decision Tree Classifier have excellent performance in all metrics, with an accuracy of 0.99, precision of 0.97-1.0, recall of 1.0, F1 score of 0.99-1.0, and AUC curve of 0.99. This suggests that these two models are very good at predicting employee turnover and can be relied upon for accurate predictions. KNN has the lowest accuracy and F1 score of 0.79 and 0.58 respectively, indicating that this model may not be the best choice for predicting employee turnover. However, it has a relatively high AUC curve of 0.80, suggesting that it is still somewhat effective in distinguishing between positive and negative cases. Random Forest has excellent performance in all metrics, similar to Logistic Regression and Decision Tree Classifier. Its precision, recall, F1 score, and AUC curve are all 0.99, which suggests that it is also a good choice for predicting employee turnover. When everything is considered, choosing a single machine learning model for further analytics is a practical and effective approach that helped to streamline the efforts and made more confident decisions. By using logistic regression for further analytics, the authors was able to predict the probability of leaving, reasons for leaving, and associated costs with greater accuracy and confidence.

When building a predictive model for employee turnover, the machine learning algorithm looked for patterns in the data that could help predict whether an employee was likely to leave. The algorithm considered all 15 features (columns) that might impacted employee retention. To assign weights to each of these factors based on their importance, the algorithm used a technique called feature selection and method mutual information. This method measured the mutual dependence between each feature and the outcome variable, and assigned weights based on their level of correlation. The weights were assigned based on the correlation between each feature and the target variable (i.e., the variable " Leaver" we want to predict). Features that had a strong correlation with the target variable were assigned a higher weight, while features that had a weak correlation with the target variable were assigned a lower weight. Table X12 shows the list of features and assigned weights (Table 2).

Table 2: Assigned weights to the features of the training dataset.

Feature	Weight
Employee_Nr	0.00
Age	-0.07
Departments	-0.09
Country	0.02
Salary_in_range	0.00
Job_satisfaction	-0.00
Relationship_satisfaction	-0.01
Work_life_balance	-0.02
Years_at_company	0.07
Total_working_years	-0.26
Job_level	-0.33
Over_time	0.11
Promo_w3_years	0.04
Type_of_leaver	1.10
Reasons_for_leaving	2.89

As it is visible in Table 2, there are 2 strong correlations. Firstly, the "Job_level" feature has a weight of -0.33, indicating that employees with higher job levels are less likely to leave. Lastly, the "Total_working_years" feature has a weight of -0.26, indicating that employees with more total working years are less likely to leave. Once the model had assigned weights to each feature, it used these weights to calculate the final probability of an employee leaving. The values of predicted reasons for leaving are listed in Table 3. These reasons were assigned to the 11,153 employees who were still with the company at the time of evaluation.

Table 3: Counted values for predicted reasons of leaving.

Reason	Count
Overtime	5019
Compensation	3345
Internal transfer	1673
Work environment	558
Different department	446
Relocation	112
Career advancement	0
Personal reasons	0

Table 3 shows the counted values for predicted reasons of leaving. The majority of employees, 45%, were predicted to leave the company due to overtime, with a total of 5019 employees identified as potential leavers for this reason. Compensation was the second most common reason, with 30% of predicted leavers (3345 employees) citing this as their reason

for leaving. Work environment was the next most frequent reason, with 558 predicted leavers (5% of the total), followed by different department with 446 predicted leavers (4% of the total). The remaining predicted reasons, including career advancement and personal reasons, did not have any identified potential leavers. Relocation was the reason for leaving for 1% of predicted leavers (112 employees).

3.3 The costs calculation of the predicted turnover

As a last step of the predictive analytics, the authors prepared a calculation of how the costs of turnover could be determined. Calculating the cost of recruiting and onboarding new employees is an important factor to consider when evaluating the costs of employee turnover. To accurately calculate the cost of recruiting and onboarding new employees, companies should consider the following factors – job posting, recruiter fees, background checks and pre-employment testing, orientation, and training.

Total Job Postings Cost = Cost per Job Posting x Number of Job Postings

Recruiter Fees = First-year Salary x Recruiter Fee Percentage

Total Background Check and Pre-employment Testing Cost = Cost per Check/Test x Number of Checks/Tests

Total Orientation and Training Cost = Cost per Employee x Number of Employees

Total costs of onboarding = (yearly salary/12 months) x amount of training months

For example, if it takes an average of two months for a new employee to reach the same level of productivity as the employees who left for the entry level roles in Slovakia, and the average salary for that role is $15,000 per year, then the estimated cost of lost productivity would be (15 000€ / 12 months) x 2 months = 2500€. This calculation assumes that the new employee will be paid the same salary as the employee who left and that the level of productivity of the new employee will be the same as the employee who left. However, if the new employee requires additional training or support, or if the job is particularly complex, the estimated cost of lost productivity could be higher.

To determine the salary ranges, the authors analyzed salary data from various web pages. For Slovakia the authors got information from two sources – platy.sk and Ministry of Labour, social affairs, and family of Slovak republic. For Germany and Czech Republic the authors used these web pages Payscale.com, Glassdoor.com and Salaryexpert.com. Then created salary ranges for each job level and country combination based on this analysis (Table 4).

Table 4: Salary ranges based on the country and level.

Level	Country	Salary range
3	Germany	€25,000 - €45,000
	Czech Republic	€10,500 - €21,000
	Slovakia	€10,000 - €20,000
4	Germany	€35,000 - €60,000
	Czech Republic	€16,500 - €29,000
	Slovakia	€20,000 - €30,000
5	Germany	€50,000 - €90,000
	Czech Republic	€25,000 - €50,000
	Slovakia	€30,000 - €50,000
6	Germany	€75,000 - €120,000
	Czech Republic	€41,500 - €75,000
	Slovakia	€50,000 - €80,000
7	Germany	€100,000 - €200,000+
	Czech Republic	€62,000 - €125,000+
	Slovakia	€80,000 - €150,000+

The authors assigned these salary ranges to each employee based on their job level and country (Table 4). Once the costs were calculated for each employee, they were added to the database under a new column named Total cost. As a result of the authors' prediction analytics three new values were assigned to each employee (Table 5) - probability of leaving, reasons for leaving, and total costs of turnover.

Table 5: Final table of the predictive analytics.

Employee Nr	Probability_of leaving	Reasons_for leaving	Total costs
3500	9.27%	Compensation	10206
3501	0.00%	Internal transfer	15206
3052	6.04%	Compensation	3956
3053	15.30%	Overtime	16456
3054	64.20%	Overtime	11456
2830	51.30%	Internal transfer	17106
2696	24.20%	Compensation	8956
2562	72.80%	Overtime	3956

Only by preventing predicted turnover with the probability of leaving higher than 50% displayed in Table 5, company could have saved 32518 Euros.

3.4 Evaluation of research questions

Research question 3: To what extent does the predictive model developed using machine learning techniques accurately identify the most common reasons for employee departures based on the input data provided?

Based on the results of the calibration curve and Brier score, the accuracy of the predictive model developed using machine learning techniques was evaluated. The accuracy was found to be higher than 60%. However, the model was not entirely accurate, as indicated by the Brier score of 0.36. The Brier score indicated that there was still room for improvement in the predictive model's accuracy. A lower Brier score would indicate better accuracy, with a score of 0 indicating perfect accuracy. However, it's important to note that the level of accuracy required for the predictive model to be considered useful might vary depending on the specific use case and context. Additionally, while the model might not be perfectly accurate, it still provided valuable insights and information that informed decision-making and helped improve HR practices.

Research question 4: Is compensation identified as the primary reason for employee departures, based on a higher frequency of this factor compared to other reasons in the data?

Initially, the authors presumed that compensation would be identified as the primary reason for employee departures based on the frequency of this factor compared to other reasons in the data. This question implied that compensation was a significant factor in employee turnover and that addressing compensation-related issues helped reduce employee turnover. Based on the logistic regression prediction model, the authors initially identified overtime as the strongest reason for employee departures, followed by compensation as the third reason after internal transfer. However, after verifying the model with new data, it resulted in compensation being the strongest reason for employee departures. This was attributed to the company's implementation of three internal initiatives to reduce the number of overtimes.

Research question 5: Does the deployment of the predictive model, within a 6-month result in a significant reduction of fluctuations in the company's turnover rate, thereby supporting the company in adjusting internal initiatives to lower turnover?

The authors aimed to achieve a reduction in fluctuations in the turnover rate of the company after 6 months of deploying the predictive model. After the comparison of the turnover rate prior to model deployment and the rate after 6 months, it was found that there was a 4% reduction. The authors realized that the

reduction was not 10%, as initially planned, as it was overly ambitious, given the short timeframe of 6 months.

Research question 6: To what extent does the implementation of the predictive model result in savings from the planned HR budget, as indicated by a reduction of costs over a specified period?

After implementing the predictive model, the company calculated the costs of all turnover that probabilities were higher than 80% and compared them to the HR budget. The results showed that the savings achieved were 16%. Therefore, it could be concluded that the implementation of the predictive model did result in significant savings from the planned HR budget.

In conclusion, the research successfully tested the research questions and achieved the stated objectives

Authors was able to confirm the relevance of HR analytics for companies, identify the prevalence of manual data collection and analysis methods among HR representatives, accurately predicted employee departures using a machine learning model, and achieved significant cost savings through the implementation of the model. However, given the broad scope of the topic, there remains ample opportunity for further research and improvement in the field of HR analytics.

4 Discusion

Predictive analytics is a technique to identify the likelihood of future outcomes based on historical data (James et al. 2013). The goal of predictive analytics is to go beyond understanding what has happened in the past to provide a best assessment of what will happen in the future (Gandomi, Haider 2015). The outcome of predictive analytics is typically a predictive model or set of predictive models. These models are designed to make predictions or forecasts about future events or behaviours based on historical data (Kanchana, Chinnadurai 2020).

In the context of employee turnover, predictive analytics can be used to identify the factors that contribute to employee turnover and to predict the likelihood of an employee leaving the organization. This can be valuable information for HR managers who can take proactive measures to retain their valuable employees and reduce the negative impact of employee turnover. Predictive analytics can be used to identify employees who are at risk of leaving an organization. This can help companies take proactive measures to retain their best employees and reduce the costs associated with turnover (Fallucchi et al.2020). Predictive analytics can also help companies identify the factors that contribute to employee turnover and develop strategies to address them.

A study by Gupta and Sagar (2022) found that predictive analytics can accurately predict employee turnover with an accuracy of 82.2%, which can help organizations take proactive measures to retain employees. Another study by Braganza and Bharati (2021) found that predictive analytics can help identify the factors that contribute to employee turnover, such as job satisfaction, work-life balance, and career growth opportunities. In the context of employee turnover, the outcome of predictive analytics could be a model that identifies the factors that contribute to employee turnover and predicts the likelihood of an employee leaving the organization. This model can then be used by HR managers to take proactive measures to retain valuable employees and reduce the negative impact of turnover (Hogan et al. 2020).

Two relevant studies that have focused on similar topics are "Predicting employee turnover in the banking sector (Dávid et al. 2019): A comparative study of data mining techniques" and "Employee turnover prediction using machine learning: A case study on Indian IT sector" (Mohapatra et al. 2020). In the first study, the authors applied several data mining techniques, including decision trees, neural networks, and logistic regression, to predict employee turnover in the banking sector. They found that decision trees and neural networks were the most effective techniques for predicting employee turnover. In the second study, the authors used machine learning techniques to predict employee turnover in the Indian IT sector. They collected data on several variables, including employee age, gender, job role, performance ratings, and length of service. The authors found that the machine learning models were able to accurately predict employee turnover and identified several key factors that contributed to employee turnover, including low job satisfaction and poor performance ratings. The conclusions of these studies are in line with the results of our research.

4.1 Implementing actions based on model outputs

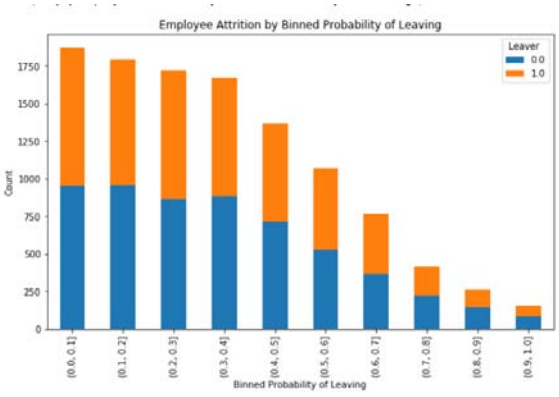
The results of the predictive analytics led to new information being assigned to each employee, including their probability of leaving, reasons for leaving, and turnover costs. HR representatives worked with these data and agreed with management to address the issue of overtime. Overtime was identified as a strong predictor of employee turnover, with 45% of all employees predicted to leave because of it. After presenting these results, the company took three initiatives to manage employee workload and reduce the likelihood of turnover due to overwork and burnout. First, the company became more transparent and proactive in managing workload, with the aim of retaining top talent, improving productivity, and maintaining a positive and engaged workforce. The predictive model highlighted the need for the company to be more transparent with employees about their expected workload and potential overtime. Secondly, the company worked on providing more detailed information about job requirements and workload during the recruitment process. This involved hiring additional staff to distribute the workload more evenly. Lastly, to manage overtime more effectively, the company monitored overtime hours and provided employees with more flexibility in managing their work schedules. This involved offering flexible working hours, allowing employees to work from home, or providing time off in lieu of overtime hours worked.

Six months later, the same employees were tracked to determine if they were still with the company or if they had left. If an employee had left, the authors verified the predicted probability of leaving against the actual outcome using the results obtained from the predictive model and the reason for leaving.

The new dataset contained 11,093 values, of which 3,349 employees left the company.

For a better interpretation of the predicted and true values of whether employees stayed or left the company, below is a graphical representation (Graph 2).

Graph 2: Ratio of predicted and true values.



The x-axis shows the ranges of predicted probabilities, which have been grouped in bins of size 0.1. For example, the first bin "0.0-0.1" represents employees whose predicted probability of leaving was between 0.0 and 0.1. The y-axis shows the proportion of employees who actually left or were predicted to leave within each bin.

In the histogram with grouped probabilities, the orange bars represent the actual distribution of the leavers, while the blue bars represent the predicted probability distribution of the leavers. From the graph, it is visible that the model tends to overestimate the probability of employees leaving, as the proportion of predicted leavers is generally higher than the proportion of actual leavers within each bin. Furthermore, the model is more accurate in predicting employees who are less likely to leave (i.e., lower predicted probabilities), as the proportions of predicted and actual leavers become more similar in the lower probability bins.

Overall, the graph helps to assess the model's performance in predicting employee turnover and could be used to identify areas for improvement in the model or in the data preprocessing. Table 6 below shows the updated list of reasons for leaving the company. The company followed the authors' advice by taking three initiatives to promote work-life balance and keep employees. As a result, overtime was no longer the main cause of employees leaving the company.

Table 6: Updated list of reasons of leaving.

Compensation	6896
Overtime	3510
Internal transfer	2140
Work environment	1274
Career advancement	210
Different department	24
Personal reasons	20
Relocation	19
Involuntary	7

Interestingly, the authors' predictions regarding compensation as the strongest justification for leaving were confirmed by the data. This suggests that offering competitive compensation packages and opportunities for growth and development can be effective in retaining employees. Additionally, the updated list of reasons for leaving can provide valuable insights for HR departments and management in identifying areas for improvement in the company. By addressing the root causes of turnover, companies can reduce the costs associated with employee turnover and increase employee satisfaction and retention. The model was verified using new data and by comparing the predicted and actual values. Overall, the results were valuable to the company that participated in this research. Additionally, a new list of reasons for leaving the company was generated based on data collected after six months. By comparing the situation before and after the model was applied, the authors found differences in the reasons for leaving the company. This indicated that the model had an impact on the company and helped to identify issues that needed to be addressed. Another positive outcome of the model implementation was a decrease in the turnover rate from 26% to 22%. This suggests that the model was successful in identifying factors contributing to employee turnover and in helping the company take appropriate measures to retain employees. The company used the predictive model to calculate the costs of turnover and compared them to the HR budget. The HR representative selected all employees with predicted turnover higher than 80% and added their assigned costs together. The final number was then compared to the yearly HR budget and the results showed savings of 16%.

Overall, the model verification process demonstrated the value of using predictive analytics in HR to address employee turnover. The results provided useful insights that could be used to make data-driven decisions and improve the company's HR practices.

4 Conclusion

High levels of employee turnover can have significant financial and operational implications, including increased costs associated with recruiting and training new employees, decreased productivity, and decreased employee morale. By using data mining and predictive analytics techniques to identify the factors that contribute to employee turnover, businesses can proactively take steps to address these issues and retain their valuable employees.

Predictive analytics is a complex topic that has various possibilities. The authors worked on a list of potential research topics that could be elaborated further. Firstly, additional features could be explored. The current research used 16 features to predict employee turnover, but there may be other relevant factors that were not included in the analysis. By adding more features, the model may be able to make more accurate predictions and provide deeper insights into the reasons for turnover. For example, the culture and values of a company, employee engagement and managerial style could be an important factor in predicting turnover. These could be measured through surveys or by examining factors such as the level of collaboration and communication within the organization. Further research could delve into the data and conduct both qualitative and quantitative investigations specifically targeting the reasons behind employee departures. This comprehensive approach would provide deeper insights into the underlying factors contributing to turnover and facilitate the development of more targeted retention strategies. In addition, the model can be used to identify the most valuable employees to the organization based on their level of contribution and the costs associated with replacing. The model's performance could be evaluated across different subgroups. The current research does not explore how the model's performance varies across different subgroups of employees, such as those with different departments or levels of experience. By examining how the model performs for different subgroups, researchers can identify areas where the model may need to be refined or adjusted. Further research could investigate the impact of different algorithms. The present examination used four different machine learning algorithms to create a confusion matrix. It may be worth exploring other algorithms, such as support vector machines (SVM) or neural networks (NN), to compare their performance and determine which algorithm is most effective for this particular dataset. Assessment of the impact of external factors could be another valuable point of view. The authors did not explore how external factors, such as economic conditions or industry trends, affect employee turnover. Another perspective on predictive analytics could be a cost-benefit analysis. The authors calculated the costs of turnover for each employee, but they did not explore the potential cost savings associated with interventions to reduce turnover.

Overall, the predictive model provides the company with valuable insights into the factors that contribute to employee turnover, allowing them to take proactive steps to retain their top talent. By using data-driven approaches to HR, companies can improve their retention rates, reduce costs associated with turnover, and maintain a competitive edge in the marketplace.

Literature:

1. Aldaihani, A., & Alduais, H.: Estimating the Costs of Employee Turnover in an Organization: A Case of Kuwait's Banking Sector. *Journal of Economics and Business*, 2021. 4(1), 61-77.
2. Baker, R., & Baker, G.: Recruitment economics: How much does recruitment really cost? *Human Resource Management International Digest*, 2018. 26(7), 15-17.
3. Boudreau, J. W., & Cascio, W. F.: Human resource analytics: Why we need it more than ever. *Journal of Organizational Effectiveness: People and Performance*, 2017. 4(1), 2-14.
4. Boukottaya, A., & Khemakhem, M.: HR Analytics and Big Data: The Rise of Talent Intelligence. In *Handbook of Research on Human Resources Strategies for the New Millennial Workforce* 2019. pp. 157-174. IGI Global.
5. Braganza, A., & Bharati, P.: Employee turnover prediction using predictive analytics: An empirical study. *International Journal of Information Management*, 2021. 56, 102200.

6. Cappelli, P., & Meister, J. C.: *Big data in human resources and talent management: Emerging practices.* Routledge. 2018.
7. Cascio, W. F.: *Managing human resources: Productivity, quality of work life, profits.* McGraw-Hill Education. 2018.
8. Dávid, L., Lengyel, I., & Mészáros, G.: Predicting employee turnover in the banking sector: A comparative study of data mining techniques. *Acta Polytechnica Hungarica*, 2019. 16(2), 99-119
9. De Cieri, H., & Kramar, R.: *Human resource management in Australia: Strategy, people, performance*. McGraw-Hill Education. 2015.
10. De Winne S., Marescaux E., Sels L., Van Beveren I., Vanormelingen S.: The impact of employee turn-over and turnover volatility on labor productivity: a flexible non-linear approach, *International Journal of Human Resource Management*, 2019. 30(21), pp.3049-3079
11. Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E.: Predicting Employee Turnover Using Machine Learning Techniques. *Computers*, 2020. 9, 86
12. Gandomi, A., & Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 2015. 35(2), 137-144.
13. Gupta, A., & Sagar, M.: Predictive analytics for employee turnover: A case study in the IT industry. *Journal of Business Research*, 2022. 141, 87-95.
14. Gupta, P., & Bhatia, R.: Employee turnover: A review of opportunities and threats. *Management Science Letters*, 2021. 11(6), 1453-1462.
15. Hakeem, A., & Haris, N. A.: Python in cyber security: A review. *International Journal of Advanced Computer Science and Applications*, 2020. 11(7), 68-74.
16. Hogan, J., Chamorro-Premuzic, T., & Kaiser, R. B.: The People Analytics Code of Conduct. *Harvard Business Review Digital Articles*, 2020. 2-7.
17. Hom P.W., Allen D.G., Griffeth R.W.: *Employee retention and turnover: Why employees stay or leave, Employee Retention and Turnover: Why Employees Stay or Leave*, 2019. pp.1-328
18. James, G., Witten, D., Hastie, T., & Tibshirani, R.: *An introduction to statistical learning: with applications* in R. Springer. 2013.
19. Kanchana, R., & Chinnadurai, S.: Data preprocessing techniques for predictive analytics: A review. *Journal of Ambient Intelligence and Humanized Computing*, 2020. 11(10), 4033-4048.
20. Kehoe, R. R., & Wright, P. M.: The impact of high-performance human resource practices on employees' attitudes and behaviours. *Journal of Management*, 2013. 39(2), 366-391
21. Kouloumpis, E., Georgiadis, P., Lekakos, G., & Matsatsinis, N.: Predicting employee turnover using machine learning algorithms and big data analytics. *Journal of Business Research*, 2019. 98, 387-398.
22. Leigh, J. P., & DePaul-Haddock, A. M.: The direct and indirect costs of employee turnover: A review and synthesis. *Human Resource Management Review*, 2017. 27(2), 288-303.
23. Lienert, J.: Big data for predicting employee turnover. *International Journal of Human Resource Management*, 2018. 29(18), 2608-2626.
24. Marr, B.: *Data-driven HR: How to use analytics and metrics to drive performance.* Kogan Page Publishers. 2019.
25. Sandip, M.: *Impact of HR analytics in Industry 4.0.* 2019.
26. Mohapatra, S., Sahoo, S., & Sahoo, S. K.: Employee turnover prediction using machine learning: A case study on Indian IT sector. *Journal of Business Research*, 2020. 119, 562-575.
27. Mondore, S., & Douthitt, S.: The Cost of Turnover: Financial Impact Analysis. *Journal of Business and Psychology*, 2021. 36(2), 259-271.
28. Noe, R. A., Hollenbeck, J. R., Gerhart, B., & Wright, P. M.: *Human resource management: Gaining a competitive advantage*. McGraw-Hill Education. 2017.
29. Petry T.., Jäger W.: *Digital HR, Haufe Fachbuch*, 2021. ISBN 9783648147511
30. Phillips, J. J.: *Investing in human capital: A capital budgeting approach to employee development.* Routledge. 2016.
31. Ramalho, L. *Fluent Python: Clear, concise, and effective programming.* O'Reilly Media, Inc. 2015.
32. Ramalho, R. S. Python 3 object-oriented programming, 3rd edition. *International Journal of Information Management*, 2018. 40, 311-312.
33. Ramlall, S.: Cost-benefit analysis of employee retention programs. *International Journal of Human Resource Management*, 2017. 28(5), 753-775.
34. Silva, A., Sarmento, M., & Gomes, A.: Big data analytics in human resources management: A review of the literature. *International Journal of Information Management*, 2019. 48, 48-57.
35. Sturman, M. C., & Clarke, M. C.: The bottom-line benefits of employee turnover. *MIT Sloan Management Review*, 2019. 60(2), 44-51.
36. Taylor, B., Green, B., & Dorsey, T.: *Applied Data Science: An Introduction to the Team Data Science Process*. Apress. 2019.
37. Tursunbayeva, A., Liao, Y., & Qureshi, I.: Big Data in Human Resource Management: A Review and Future Directions. *International Journal of Human Resource Management*, 2018. 29(5), 875-912.
38. Vidal, J.: Cost of employee turnover: A review of the literature. *Human Resource Management Review*, 2017. 27(4), 575-591.
39. Zhang, L., Wang, Z., & Liu, Y.: A comparative study of machine learning models for credit scoring: Evidence from China. *PloS one*, 2021. 16(4), e0249997

Primary Paper Section: A

Secondary Paper Section: AE, JC, JD