MSc Research Project

Employee Attrition Prediction: University of Hull

Submitted By: Olatunde E. Ibrahim

Abstract

The ability to predict employee attrition in advance is significant and can help organizations take proactive measures to retain valuable talent(s), plan succession for institutional knowledge retention and improve workforce management. This research explores regression-based machine learning models to quantify and rank the attrition tendencies of each employee relative to other employees; this being a major limitation of classification-based researches, where focus has often being on 'leave or not' attrition prediction. Experimenting four different supervised machine learning regressor (namely Catboost, Extreme Gradient Boosting, Light Gradient Boosting Machine and Random Forest Regressor) and optimizing Tweedie loss function, a probability distribution function for modeling non-negative continuous target variables, the Extreme Gradient Boosting regressor showed leading predictive ability and ranked how well an employee tends towards attrition in the pool of all employees, having R-Square scores of 0.87 and 0.73, Root Mean Square Error (RMSE) value of 0.17 and 0.25, Mean Average Error (MAE) value of 0.10 and 0.15 for training and validation respectively. This research reveals the possibility and potentials of regression-based machine learning methods in quantifying attrition tendency of an employee relative to another and offers insights into enhancing human resources management strategies.

1.0 Introduction and Background

One of the major concerns for many organizations globally is employee turnover. Oftentimes, the departure of talented and skilled employees does not only disrupt the continuity of business operations but also incurs substantial costs for recruitment, training, and loss of institutional knowledge. According to Oxford Economics (2014), the average financial impact imposed on an organization through the loss (attrition) of an employee earning at least £25,000 annually is approximately £31,000. These values range between £20,113 and £39,887 for retailers and legal firms respectively. In a study by Remote (2023) and confirmed by HRreview (2023), average employee attrition rate (in United States of America and United Kingdom) is set to hit 41.4 percent, with US and UK currently experiencing 46.8 and 35.6 percent respectively. These discoveries among others have led researchers and organizations to become increasingly focused on devising effective strategies to predict and mitigate attrition. Through the identification of flight risks among employees, employers and human resource specialists can intervene with targeted support and personalized engagement strategies, thus enhancing employee satisfaction and organizational performance (Rahmadani et al., 2020; Robertson-Smith, 2009).

In recent years, diverse research have been conducted in an attempt to investigate and predict employee attrition through the use of statistical tools and machine learning models. While attrition can be classified as voluntary or involuntary, most researchers have examined/assumed voluntary attrition and predominantly focused on classification. Yedida et al. (2018) studied employee attrition prediction using varieties of supervised learning models on an unnamed

dataset obtained from Kaggle. The research was concluded with the adoption of K-Nearest Neighbors (KNN) classifier as the best performing model with a record accuracy of 94% at k-value of 6. This model is quite limited as a result of the number of data points used in its training. Experimentally, KNN algorithm sometimes struggle to find relevant neighbors with increasing number of dataset features (dimensions), and in turn produce poor performance in high-dimensional spaces and often computationally expensive (Hu et. al, 2020). Punnoose and Pankaj (2016) investigated employee turnover prediction using various machine learning classifiers on a Human Resource (HR) dataset (containing 33 features) obtained from a global retail organization. The research showed leading performance by Extreme Gradient Boosting (XGBoost) among other classifiers such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), KNN etc. Using area under the receiver operating characteristic curve (ROC-AUC) as performance metrics, top values of 0.88 and 0.86 were recorded for training and holdout respectively. Sikaroudi et al. (2015), utilized statistical tools such as Apriori and Pearson Chi Square test as well as machine learning techniques in the prediction of employee turnover on a set of HR data from an automotive parts manufacturing company. The research which carried out in-depth evaluation of each feature importance and contribution to the overall prediction classification concluded with Random Forest Classifier having the best performance with an average accuracy of 90.6% for the different numbers of k-fold cross validation experimented. Alao and Adeyemo (2013) studied how decision tree can identify features that influence employee turnover using sets of data on employees' demographical information and personnel records. Zhao et al. (2019) also carried out extensive research on the prediction of employee attrition using various splits of HR dataset from IBM. The research utilized ten different supervised machine learning classifiers, with focus on performance metrics other than accuracy. Notably, these previous research works have shown novel discoveries and contribution to knowledge, however, they were primarily focused on classification models to determine a potential leaver and have been met with the limitation of quantifying likelihood of such employee or groups of employees leaving the organization(s) relative to any other employee.

Zhu et al. (2017) researched how univariate and multivariate time series methods could forecast employee attrition. According to the authors, over 11-years of monthly turnover data consisting of about 8000 records of active and terminated employees were used for model building and validation. In comparison with other models/methods experimented in the research, a dynamic regression model (using lag7 CLI as predictor) with additive trend and seasonality was adopted having R-squared values of 0.77 and 0.59 for training and holdout respectively. Despite the interesting outcome of this research, it fails to adopt a machine learning approach but rather chose a statistical method through its use of Number Cruncher Statistical System (NCSS), Statistical Analysis System (SAS), and R. Also, the model can be said to overfit, owing to the wide difference between the training and holdout R-squared value, thus possessing poor generalization to new data.

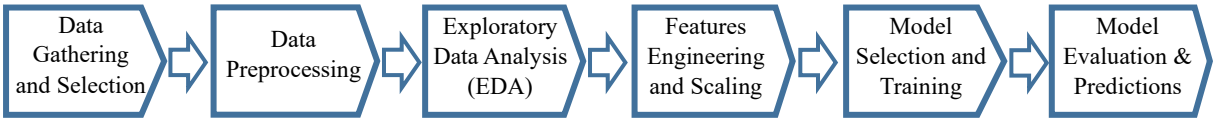
According to Zhao et al. (2019), irrespective of the interesting research outcomes discussed above, the prediction of employee turnover that results from using statistical approaches and/or machine learning methods are often problem-specific and not easy to generalize. In this study, we aim to research employee attrition prediction at the University of Hull by critically examining regression-based predictive machine learning models, such that employees' attrition tendencies will be measured as continuous values between lower and upper limit of 0 and 1

respectively. The lower limit 0 corresponds to status of active current staff and upper limit 1 corresponds to exit from the institution. Summarily, the research aims to determine the measure of how each employee attains status of 1 (Exit).

With the choice of CatBoost Regressor, Extreme Gradient Boost (XGBoost) Regressor, Light Gradient Boosting Machine (LGBM) regressor and Random Forest Regressor, this research provides a more nuanced perspective, thus enabling the University of Hull to quantity attrition prediction of each staff member and by extension groups of employees. Without any doubt, past advancements in predictive analytics have paved the way for the adoption of these regression models.

2.0 Methodology

This section presents detailed outline of the steps taken for the development of models for employee attrition prediction in University of Hull using supervised machine learning techniques. Summarily, the approach involves data collection, cleaning and harmonization, exploratory data analysis, feature engineering, data encoding and appropriate model selection for training and evaluation as shown below.



2.1 Data Preprocessing

In order to achieve maximum understanding of trends in the dataset as well as development of high-performing model(s), the data selection/collection, cleaning and harmonization processes have been carefully carried out. In agreement with Chien and Chen (2008), HR data is often noisy, inconsistent and contains missing information, hence the need for cleaning and preprocessing. This research utilized combination of two main records of employees of the University for its analysis. These records are namely the leavers' dataset (Employees who have left the institution through resignation, retirement and other exit means between 2016 and 2023) and the current staffs dataset. In line with ethical consideration, the datasets were anonymized during collection/selection process by removal of personal details such as sexual orientation, personal address etc. The data cleaning focuses on removal of duplicated and less relevant columns and records (e.g. contract types with no significant employment benefits) and filling of null values in each columns, e.g. using insight from relative columns. Data harmonization ensures that columns and values are renamed or replaced for consistency and uniformity, date columns and other data types are formatted correctly. Additionally, columns available in current staffs' dataset but unavailable in leavers' data were created using insights from relative columns. A new binary-value column named Attrition was also created to distinguish between leavers and current employees, where leavers have value of 1 and current staff members have value of 0. A unified dataset 'combined_data' was subsequently formed by the integration of the two prior datasets using concatenate function in Pandas library.

The combined data has 4743 records (consisting of 2503 leavers and 2240 current staffs) and 16 features. A total of 14 features were used for model building while the other two were used for Exploratory Data Analysis (EDA) Only. The details are shown in Table 1.

Variables	Variable Type	Unique Values	Use
Leaving Reason	Categorical	8	Exploratory Data Analysis (EDA) Only
End Year	Numerical	8	
Staff Category	Categorical	2	EDA and Model Building
Attrition	Numerical	2	
Contract Type	Categorical	3	
Division	Categorical	10	
Faculty	Categorical	45	
Contract Basis	Categorical	2	Model Building Only
Grade	Categorical	18	
Start Year	Numerical	56	
FTE	Numerical	80	
School Name	Categorical	159	
Subject Group	Categorical	253	
Length of Service	Numerical	1457	
Job Title	Categorical	1582	
Position Number	Categorical	4463	

Table 1: Overview of the Dataset showing variable types and numbers of unique values. Leaving Reason and End Year are only peculiar to leavers; hence, they were unused for model building.

2.2 Exploratory Data Analysis (EDA)

According to Mohd et al. (2022), the prediction of employee attrition by Human Resource departments can be aided through the evaluation of historical data and graphical data mining methods. In order to gain insights into the workforce trends and to establish grounds for relative validation of proposed machine learning models, Exploratory Data Analysis was conducted on the employees' attrition data.

First, the analysis focused on calculating the attrition counts and rates per year. The data revealed that attrition rates remained within relatively close range between 2018 and 2022 (except for a surge in 2019), having an average of 13% per year; a figure slightly higher than 10.6% estimated by the Universities and Colleges Employers Association (UCEA, 2018).

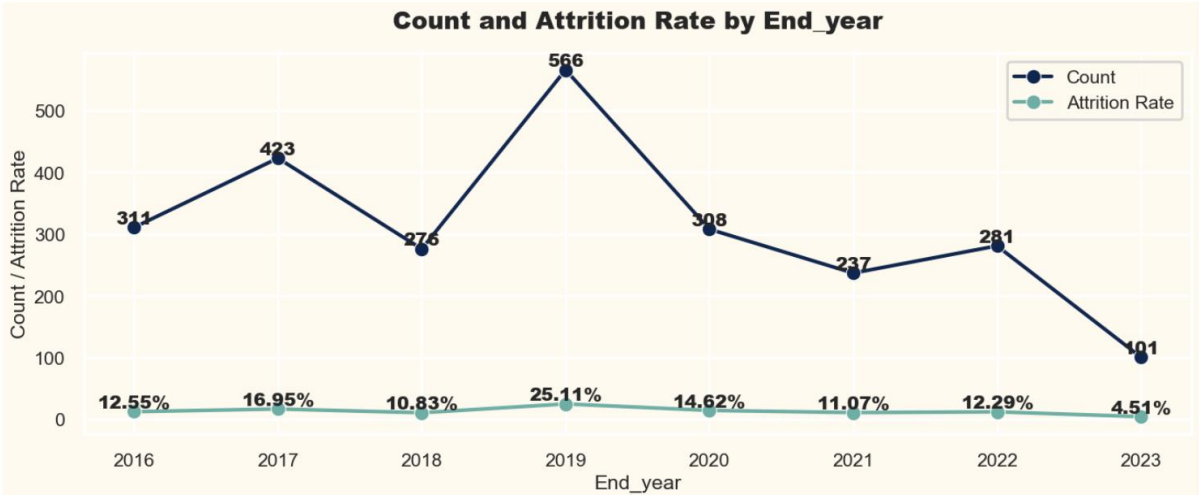


Figure 1: Plot of University of Hull Employee Attrition counts and rates per year from 2016 to early 2023.

Mathematically, percentage annual turnover (attrition) rate is expressed as;

$$\text{Annual Turnover Rate}(\%) = \frac{Number\ of\ Employees\ who\ left\ in\ a\ year}{Average\ Number\ of\ Employees\ in\ the\ year} * 100 \quad \ldots (1)$$

The analysis was further extended to evaluation of attrition rates based on contract types. The findings showed that employees on temporary contracts were 1.2 times and 1.8 times more likely to leave the institution compared to fixed-term contract type and continuing staffs respectively. In harmony with the hypothesis made by Liu et al. (2022), temporary staff members may be facing peculiar challenges or may have different motivations to leave the institution compared to their counterparts on other types of contracts.

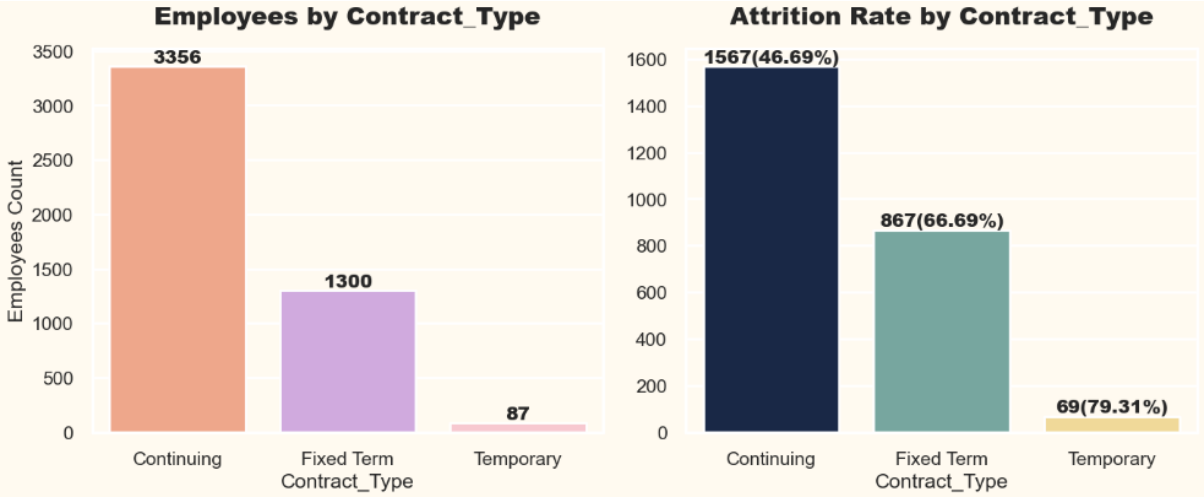


Figure 2: Plots of University of Hull Employees by Contract types and attrition rates by Contract types between 2016 and 2023.

Furthermore, the analysis delved into the attrition patterns based on staff categories, which included Professional and Academic staff. The analysis outcome revealed that there was no significant difference in attrition rates between these two categories. Both professional and academic staff had comparable attrition rates over the period examined, suggesting that factors influencing employee attrition may not be heavily dependent on staff categories.

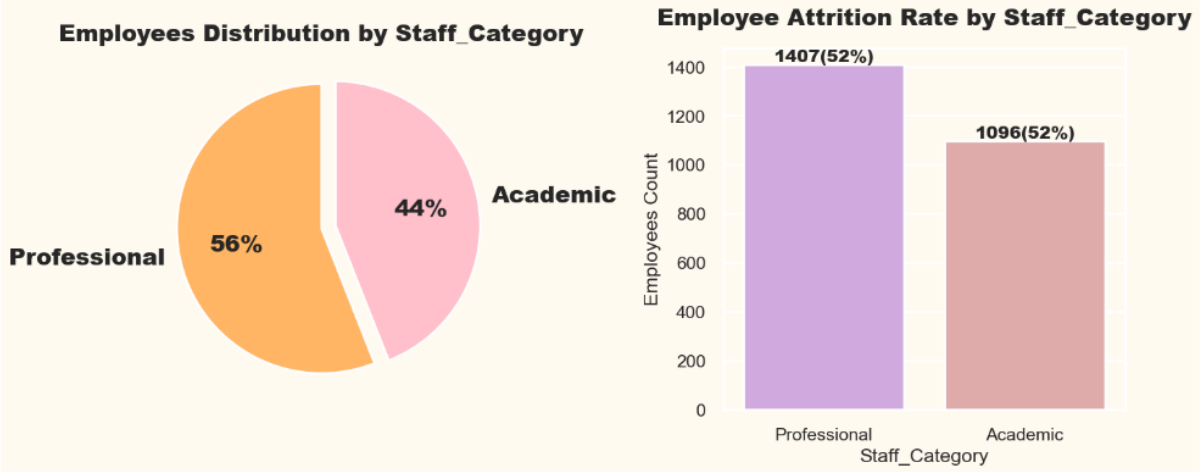


Figure 3: Plots of University of Hull Employees by Staff Categories and Attrition Rate by Staff Categories between 2016 and 2023

Finally, the reasons for employee departures were explored to understand the proportions of different leaving reasons. The data was analyzed to calculate the percentage of attrition by leaving reasons. The results demonstrated that employees were ten times more likely to resign than to retire. On average, 7 out of 10 employees who left the institution did so voluntarily through resignation, while a smaller proportion left due to retirement or other reasons. This finding highlights the importance of understanding and addressing the factors that contribute to employee resignations to retain talent and maintain a stable workforce.

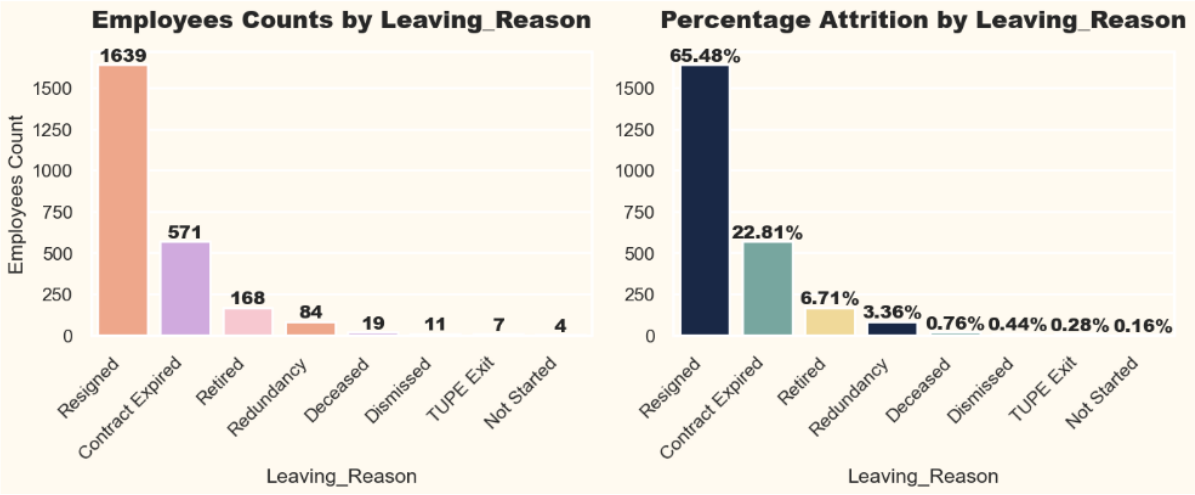


Figure 4: Plots of University of Hull Employees by Leaving Reasons and percentage Attrition by Leaving Reasons between 2016 and 2023.

2.3 Feature Engineering

2.3.1 Data Encoding

In the world of data analysis and machine learning, effective handling of categorical data is crucial to ensuring accurate model predictions (Masateru et al., 2012). This research incorporates both label encoding and target encoding techniques to different categorical columns based on their observed unique characteristics. The categorical data which exhibit nominal relationship with no observed inherent order were label encoded. While target encoding was used to encode ordinal categorical data, thus, Contract_Type feature was separately target encoded using Attrition as the target variable. Target encoding leverages target variable to assign numerical values, using the mean of the target variable for each category.

Target encoding is expressed as $\hat{x}_l = \frac{\sum_i x_i^{train}{}_{-l} y_i^{train}}{N_l}$ (Pargent et al., 2022) ...(2)

$\hat{x}_l$ represents the encoded mean value for a specific category l in the categorical feature x, x_i^{train} *and* y_i^{train} represents the value of the categorical feature x and the target variable for *the i-th* instance in the training dataset respectively. N_l corresponds to the total number of instances in the training dataset where the categorical feature x has the value l.

Encoder	Label Encoder	Target Encoder
Number of Encoded Features	9	1

Table 2: Numbers of categorical features encoded.

2.3.2 Data Augmentation

Data augmentation techniques play a pivotal role in addressing class imbalance in datasets. Although, the dataset used in this analysis shows a fair level of balance with an imbalance ratio of 1.12 for the target variable, however SMOTE (Synthetic Minority Over-sampling Technique) was used to augment the data by generating synthetic instances of the minority class ('Attrition' = 0). This helps the training process to becomes more robust, thus, enhancing the predictive ability of models used.

2.4 Model Selection

Four major regressions models namely Catboost, Extreme Gradient Boost (XGBoost), Light Gradient Boosting Machine (LGBM) and Random Forest Regressors were experimented in this study. The choice of these models was influenced by the nature and size of the dataset, problem-specific requirements, computation time etc. Also, according to Mohammed et al. (2022), adoption of different algorithms is imperative for comparison of performance metrics as a result of their individual strengths and limitations.

2.4.1 CatBoost Regressor

CatBoost is a very powerful gradient boosting framework designed specifically for handling categorical feature(s) in training and evaluation dataset, that is, it does not require explicit encoding like many other machine learning models. Able to absorb the ten categorical variables in the dataset, it achieves this by employing its innate techniques known as Ordered Target Statistic and Ordered Boosting (Prokhorenkova et al, 2017), which naturally allows processing of categorical data by considering the statistical properties of these features during the boosting process (Hancock & Khoshgoftaar, 2020). This significantly reduces the need for preprocessing and feature engineering.

The Ordered Target Statistics uses the numerical approximation below (with input values x^i , and expected output values y for i ε{1 . . . k}) in place of one-hot encoding.

$$\hat{x}_k^i \approx E(y|x^i = x_k^i) \qquad \ldots(3)$$

Using this regressor, a simple and efficient model was built with a focus on optimizing Tweedie loss function. Tweedie distribution is a family of probability distributions that is useful for modeling non-negative continuous target variables (Bonat and Kokonendji, 2017). This is particularly valuable in this prediction task where target prediction must be between 0 and 1 (Predicting how well each current employee tends from 0 towards 1). Setting Tweedie variance power parameter (p) to 1.4, the developed model is privileged to utilize an approximate distribution that is between the Poisson distribution (p = 1) and the Gamma distribution (p = 2). The model trains at a learning rate of 0.05, making the training process more stable. It utilizes L2 regularization term 0.05 for 1000 boosting iterations.

2.4.2 Extreme Gradient Boosting (XGBoost)

Similar to CatBoost, XGBoost belongs to the family of gradient boosting algorithms. It iteratively combines the predictions of multiple weak learners (typically decision trees) to create a strong ensemble model (Chen and Guestrin, 2016). Renowned for exceptional predictive accuracy and ability to handle missing or null values in dataset, XGBoost also employs L1 (Lasso) and L2 (Ridge) regularization terms, which helps prevent overfitting. The

model architecture developed for this research reveals the use of regularization, subsampling, and column sampling hyperparameters, which help to prevent overfitting and enhance the model's ability to generalize to new data. Additionally, the Tweedie loss function at variance power 1 ensures less sensitivity to outliers.

2.4.3 Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine (LGBM) is a popular machine learning algorithm under ensemble learning category. Specifically designed for gradient boosting tasks, it is known for efficiency, speed, and high predictive performance. According to Ke et al. (2017), LGBM shows accelerated training process by up to over 20 times on multiple public datasets relative to other supervised machine learning models, while achieving almost the same accuracy. Also, Hancock and Khoshgoftaar (2020) claimed that LGBM supports automatic encoding of categorical features, however, this research trained the model with encoded data.

2.4.4 Random Forest

Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to produce a robust and accurate predictive model. It achieves this through random subsampling of data and feature selection. During the training process, subsets of the original dataset are randomly selected with replacement, creating diverse training samples for each tree.

In all the models, the Attrition variable was set as the target while other variables constitute the input features. Using the split function in sklearn library, the dataset was split into training and evaluation (test) set at ratio 8:2 respectively. With the choice of K-fold cross validation at k=5 and GridSearchCV hyperparameter tuning, the best hyperparameters were used and each model (Except CatBoost) trained with ‘best_model.fit’ function. Summarily, common parameters and hyperparameters of choice for each model are represented below.

Model	Loss Function	N_ estimators	Random State	Learning Rate	Data Type(s) Used
CatBoost	Tweedie, p=1.4	1000	2	0.05	Categorical, Numerical
XGBoost	Tweedie, p=1	1000	2	0.01	Numerical (Encoded)
LGBM	Tweedie, p=1.1	500	2	0.01	Numerical (Encoded)
RF	Unsupported	1000	42	Unsupported	Numerical (Encoded)

Table 3: Report of some parameters common to the models and their values.

3.0 Results

3.1 Feature Importance

The SHAP (SHapley Additive exPlanations) explainer tool was used in this research to evaluate individual feature’s contribution to each of the model’s predictive ability. It specifically offers visual description that is easy to read and comprehend. Table 4 shows details of the result and order of features’ contribution to each the model’s predictive ability.

	CatBoost	XGBoost	LGBM	Random Forest
↓	Start Year	Start Year	Length of Service	Position Number
	Length of Service	Position Number	Start Year	Length of Service
	Faculty	Length of Service	Position Number	Start Year
	School Name	School Name	School Name	Contract Type
	Subject Group	Contract Type	Contract Type	School Name
	Contract Type	Faculty	Faculty	Faculty
	Division	Job Title	Job Title	Staff Category
	Job Title	Contract Basis	Subject Group	Subject Group
	Position Number	Subject Group	Division	Division
	Grade	Staff Category	Staff Category	Job Title
	FTE	Division	Grade	FTE
	Staff Category	FTE	FTE	Grade
	Contract Basis	Grade	Contract Basis	Contract Basis

Table 4: Descending Order of features' importance to the contribution of models' predictive ability

3.2 Performance Evaluation

As a regression task, this study evaluates performance of each model using the R-Squared Score (R2), Root Mean Square Error (RMSE) and the Mean Average Error Score.

3.2.1 The R-Squared Score

The R-Squared Score also known as the coefficient of determination measures the proportion of the variance in the target variable (Attrition) that is predictable from the input features. The XGBoost Regressor shows the best performance with record value of 0.8767 and 0.7311 for training and validation respectively. This implies that the model utilized approximately 88% of the attributes in the features during training process for its predictions. Additionally, the model's ability to generalize to new data is very high and does not overfit. LGBM performed closely to XGBoost with training value of 0.8463 and validation value 0.7225. The third and least performer are, Catboost and Random Forest Regressor respectively. The details of individual model's performance are shown in Table 5 and the graphical representation of their training and validation R-score is shown in Figure 5.

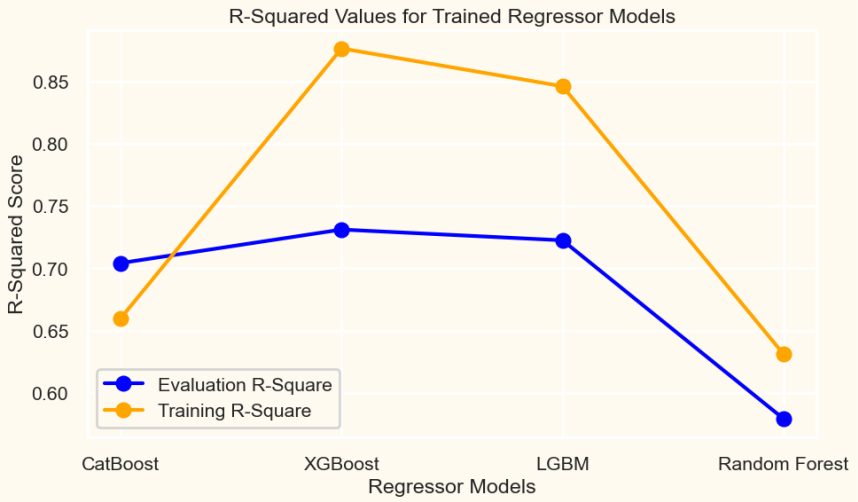
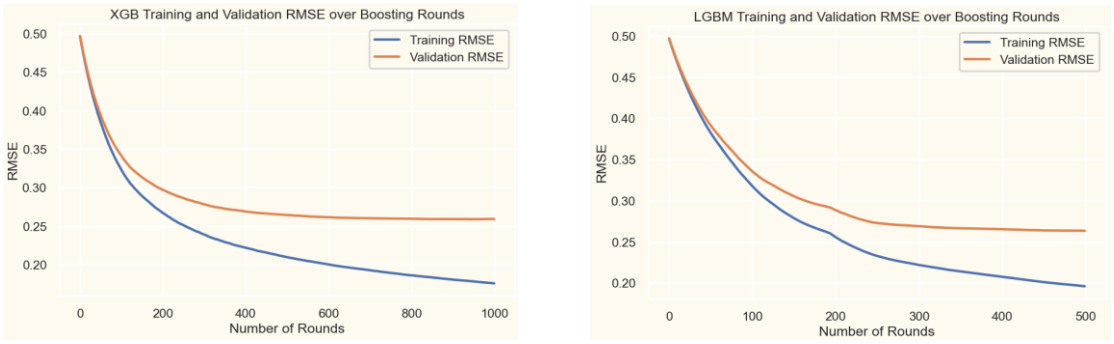


Figure 5: Plot of training and validation R-Squared values for the four experimented models.

3.2.2 Root Mean Square Error (RMSE)

The RMSE helps to quantify the average of the magnitude of errors between predicted and actual values. It has been used in this research to measure how the prediction values align with the actual target values. The XGBoost led with the best training RMSE of 0.1756, seconded by LGMB having 0.1960, following by CatBoost having 0.2961 and Random Forest trailing with value of 0.3034. The RMSE plot during training and validation for the two leading performers are shown in Figure 6.



(a) *(b)*

Figure 6: Plot of training and validation RMSE values for XGB(a) and LGBM(b) regressors.

Table 4 shows the details of performance metrics scores for the training and validation processes. The results of the top performers (XGBoost and LGBM) exceed the outcomes generated by Zhu et al. (2017); the leading regression-based employee turnover prediction research (in terms of model performance) known at the time of this study and research by Thaden et al. (2010)

Model	RMSE Score		MAE Score		R-Square Score	
	Training	Validation	Training	Validation	Training	Validation
CatBoost	0.2961	0.2735	0.2110	0.1737	0.6476	0.7008
XGBoost	**0.1756**	**0.2593**	**0.1042**	**0.1533**	**0.8767**	**0.7311**
LGBM	0.1960	0.2634	0.1172	0.1554	0.8463	0.7225
RF	0.3034	0.3243	0.2112	0.2259	0.6318	0.5794

Table 5: The table shows the training and validation performances of the experimented models. The top performance in each evaluation metrics is boldened.

With these results, it can be further established that using and tuning the Tweedie Loss function (as in Table 3) aided the performance of other models except Random Forest where Tweedie is unsupported in its architecture. Additionally, data encoding using label and target encoder, as well as data augmentation aided the performances of XGBoost and LGBM ahead of CatBoost Regressor where data was encoded using its inbuilt Ordered Target Statistic and Ordered Boosting. As a result of the leading performances exhibited by XGBoost, its prediction is adopted and utilized.

4.0 Discussion (Predictions and Interpretation)

Experimentally, each of the models predicted continuous numerical attrition values for the current staffs between 0 and 1 (the primary focus of this research). These predicted values imply how each employee tends towards attainment of the status of 1 (Exit) relative to the other

employees. The highest prediction value produced by XGBoost (The adopted model) is 0.8362; this corresponds to the employee most likely to exit the University relative to any other, while the lowest prediction value 0.0086 corresponds to the employee least likely to exit the University relative to others.

Utilizing deduction from the EDA in Figure 1, an average of 300 employees exit the University annually. Since 101 has left already in 2023, features of the top 200 predicted potential leavers which include Job Titles, School Names, Grade, Faculties, and Divisions would be considered and interpreted.

4.1 Job Title/Roles

Evaluation of the predictions based on job titles can avail valuable insights into the potential reasons behind employee turnover within the university. This analysis reveals that research-related roles (roles that contain research in their titles) account for the largest portion of the top 200 predicted leavers, being 22%. This suggests that factors such as research challenges or career opportunities may contribute to a higher attrition rate among employees engaged in research activities. Academic researchers would likely choose institution(s) with conducive and research-enabling environments over other factors (Weinstein et al., 2021). Lecturers follow closely, constituting 16% of the top predicted leavers. This could indicate potential challenges relating to teaching loads/workloads, work-life balance, career growth are affecting this sect of the University employees. Similarly, factors within administrative functions, such as workload or frequency of organizational changes, could be major influence on Administrators or Administrative officers.

Surprisingly, managerial roles account for 10.5% of the top predicted leavers. These are heads of departments and high ranked personnels. This prediction outcome might suggest challenges in leadership, work stress, or other managerial aspects that could impact the retention of managers within the university.

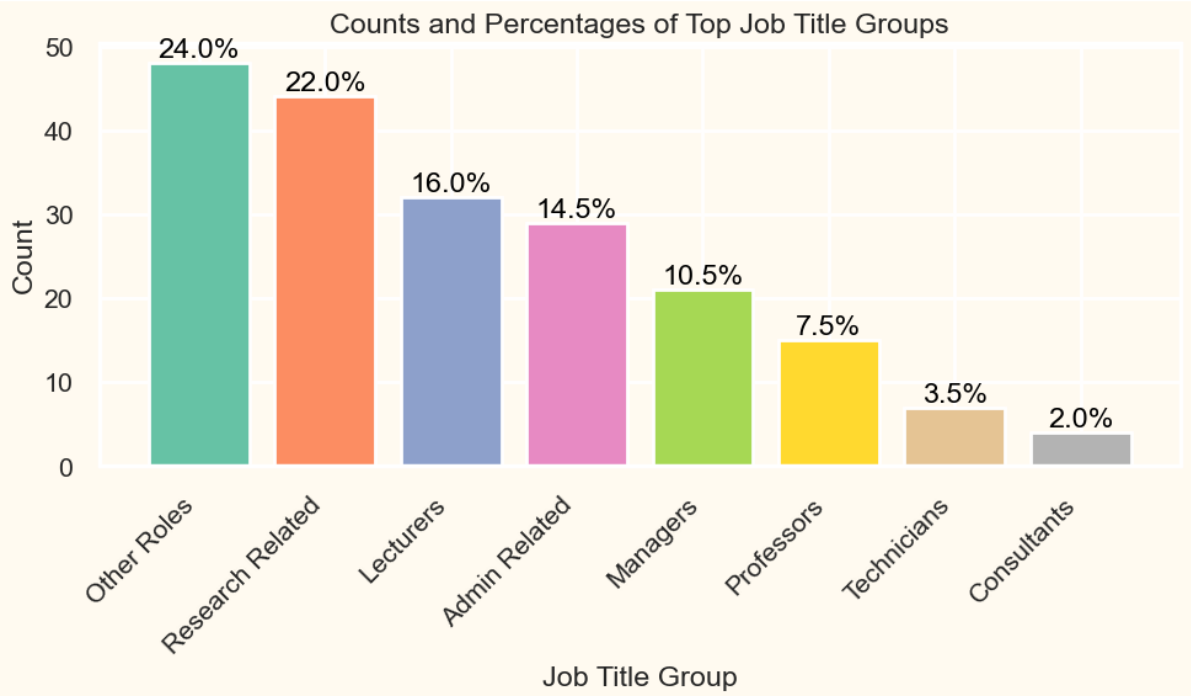


Figure 7: Job Titles of the predicted top 200 potential leavers

4.2 School Names

Here, school name corresponds to departments. Health related school names top the list with a combined value of 20%. Hull York Medical School emerges as a significant contributor to the predicted leavers, constituting 12%, while Faculty of Health and Sciences Office follows with 8% of the total. This observation could indicate unique challenges such as demanding workloads or perhaps growing regional, national and international opportunities within the medical field. According to the United Kingdom (UK) government (GOV.UK, 2023), "Employment in life sciences in the UK has seen a continuous upward trend between 2012 and 2020" with a further 4% increase in employee number in 2021. This, among other increasing local and international opportunities could lead to loss of health-related staff members to a more promising role across public and private sectors. Similarly, staff members in the Engineering and Business schools are more likely to be attracted by many professional and academic opportunities than counterparts in Humanities or Art. Creation of enabling and conducive environment can help retain these staff members.

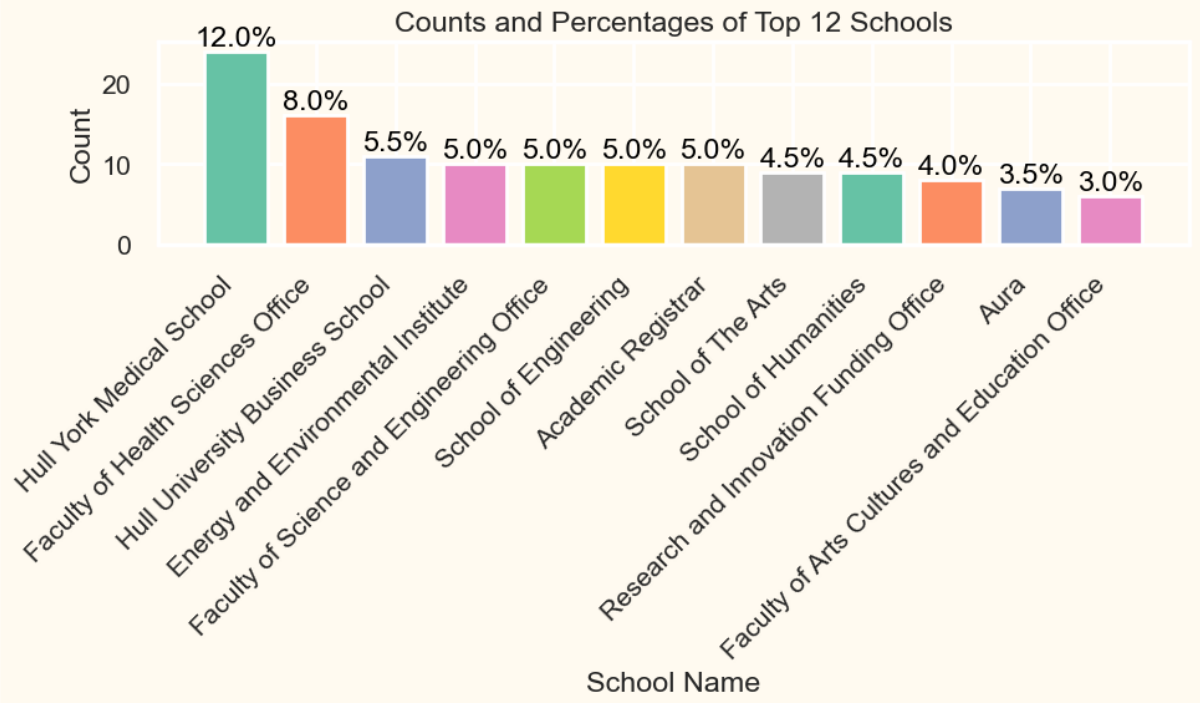


Figure 8: Job Titles of the predicted top 200 potential leavers

4.3 Grade

The analysis outcome by grade reveals higher attrition possibilities in mid senior employees. These employees potentially have better grasp of the institutional knowledge, skills and know-how and are often instrumental in the grooming of new or junior employees. Development of tailored and innovative solution would help in their retention, thus preventing erosion of institutional knowledge.

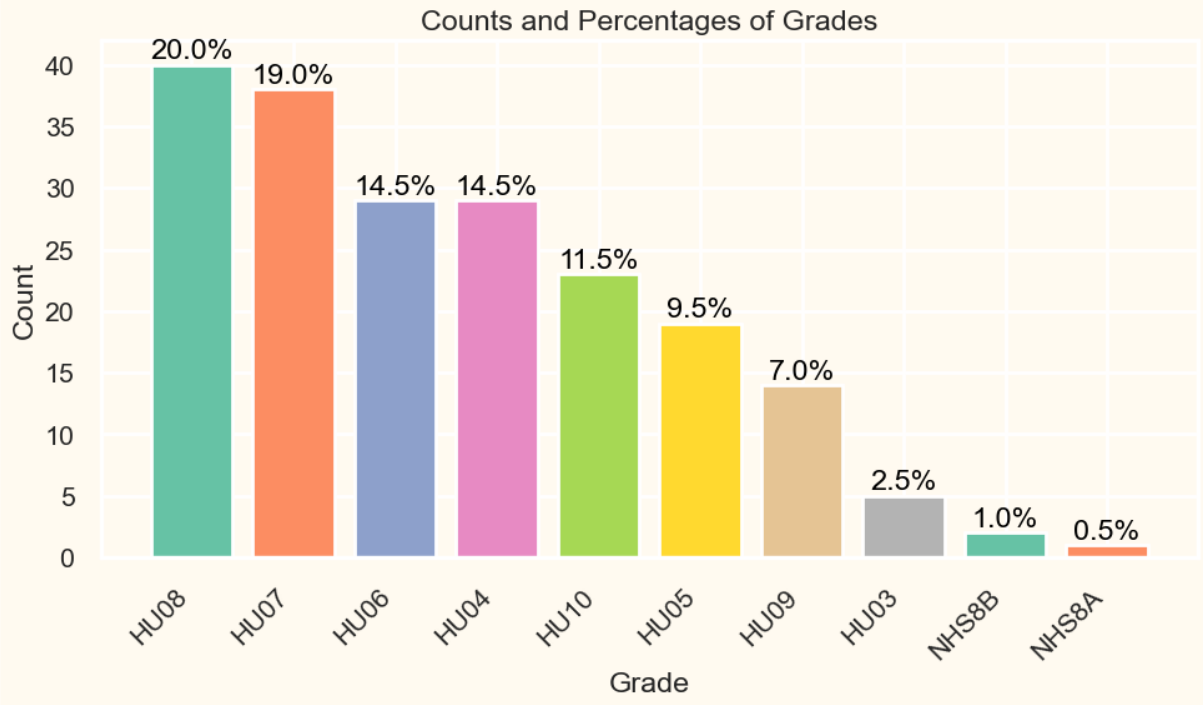


Figure 9: Percentages of grades of the predicted top 200 potential leaver

5.0 Conclusion

Using four supervised machine learning regressor models, prediction of how well an employee at the University of Hull tends towards attrition relative to other employees have been established. The performances of these predictive regressor models have also revealed the role of optimizing Tweedie loss function as well as the use of encoded numerical values in place of categorical variables in the overall model performances and prediction outcomes. The research, among other discoveries, revealed leading attrition tendencies among research-related job titles, health-related schools. These predictive outcomes are not only understandable but usable in the development of targeted solutions to employees' management, talent retention and overall enhancement of HR analytics.

5.1 Future Work and Recommendations

This research has demonstrated significant success in quantifying and ranking the attrition tendencies of an employee in the pool of total employees, however further evaluation of the attrition prediction outcomes relative to actual attrition is essential to determine if the predictions hold. Integration of the designed model to an existing HR or a new platform is recommended. Also, the research needs to be examined using other approaches such as time series, survival analysis or deep learning.

References

Alao, D & Adeyemo, A (2013) Analyzing employee attrition using decision tree algorithms, *Computing, Information Systems & Development Informatics Journal*, 4(1), 17-28.

Bonat, WH., Jørgensen, B., Kokonendji, C.C., Hinde, J. & Demétrio C.G.B. (2018) Extended Poisson–Tweedie: Properties and regression models for count data. *Statistical Modelling*, 18(1), 24-49. Available online: https://doi.org/10.1177/1471082X177157 [Accessed 13/8/2023].

Chen, T. & Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22*nd* ACM SIGKDD international conference on knowledge discovery and data mining, 785–794. Available online: https://doi.org/10.1145/2939672.2939785 [Accessed 13/8/2023]

Chien, C.F. & Chen, L.F. (2008) Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. *Expert Systems with Applications,* 34(1), 280-290.

Hancock, J. & Khoshgoftaar, T. (2020) CatBoost for big data: an interdisciplinary review. *Journal of Big Data* 7, 94. Available online: https://doi.org/10.1186/s40537-020-00369-8 [Accessed 13/8/2023].

Hu, Y., Peng, G., Wang, Z., Cui, Y., & Qin, H. (2020) Partition Selection for Large-Scale Data Management Using KNN Join Processing. *Mathematical Problems in Engineering*, 2020, 1-14. Available online: https://doi.org/10.1155/2020/7898230 [Accessed 13/8/2023].

HRreview (2023) Employee Turnover Rates have Increased by 9% Since 2019. Available online: https://hrreview.co.uk/hr-news/recruitment/employee-turnover-rates-have-increased-by-9-since-2019/150788#:~:text=Employee%20turnover%20rates%20are%20set,UK%20is%20hitting%2035.6%20percent. [Accessed 17/8/2023].

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017) Lightgbm: a highly efficient gradient boosting decision tree. Advances in neural information processing systems. *New York: Curran Associates*. 3146–54.

Liu, X., Qin, C., Liu, S. & Lu, W. (2022) Why and When Temporary Workers Engage in More Counterproductive Work Behaviors with Permanent Employees in Chinese State-Own Enterprise: A Social Identity Perspective. *International Journal of Environmental Research and Public Health*. 19(13), 1-18. Available online: https://doi.org/10.3390/ijerph19138030 [Accessed 17/8/2023]

Mohammed, E., Alsaadi, E., Khlebus, S., Alabaichi, A. (2022). Identification of human resource analytics using machine learning algorithms. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 20, 1004-1015. Available online: https://doi.org/10.12928/TELKOMNIKA.v20i5.21818 [Accessed 13/8/2023].

Mohd, A., Reddy, S. & Rama, M. (2022) Exploratory Data Analysis (GEDA): A Case Study on Employee Attrition. 7(9), 1-11. Available online: https://doi.org/10.46243/jst.2022.v7.i09.pp01-11

Oxford Economics (2014) The Cost of Brain Drain; Understanding the Financial Impact of Staff Turnover. https://www.oxfordeconomics.com/wp-content/uploads/2023/05/cost-brain-drain-report.pdf

Pargent, F., Pfisterer, F., Thomas, J. et al. (2022) Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput Stat* 37, 2671–2692. Available online: https://doi.org/10.1007/s00180-022-01207-6 [Accessed 13/8/2023].

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. & Gulin, A. (2018) Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* 31, 6638–6648.

Punnoose, R., & Pankaj, A., (2016) Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*. 5(10).

Rahmadani, V.G., Schaufeli, W.B. & Stouten, J., (2020). How engaging leaders foster employees' work engagement. *Leadership and Organization Development Journal*, 41(8), 1155-1169.

Remote (2023) How to reduce employee turnover with a strong talent retention strategy. Available online: https://remote.com/blog/employee-turnover [Accessed 17/8/2023].

Robertson-Smith, G., & Markwick, C. (2009) Employee Engagement: A Review of Current Thinking. *Brighton, UK: Institute for Employment Studies*.

Sikaroudi, E., Rouzbeh, G. & Sikaroudi, A. (2015) A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, 8(4), 106-121.

Thaden, E., Jacobs-Priebe, L., & Evans, S. (2010) Understanding Attrition and Predicting Employment Durations of Former Staff in a Public Social Service Organization. *Journal of Social Work*, 10(4), 407–435. Available Online: https://doi.org/10.1177/1468017310369606 [Accessed 17/8/2023].

Tsunoda, M., Amasaki, S. & Monden, A. (2012) Handling categorical variables in effort estimation. 99-102. Available Online: https://doi.org/10.1145/2372251.2372267 [Accessed 17/8/2023].

Universities and Colleges Employers Association (2018) Staff numbers on open-ended academic contracts have increased. Available Online: https://www.ucea.ac.uk/news-releases/14dec18/ [Accessed 17/8/2023].

United Kingdom Government, UK.GOV (2023) Bioscience and health technology sector statistics 2021, Updated 14 June 2023. Available Online: https://www.gov.uk/government/statistics/bioscience-and-health-technology-sector-statistics-2021/bioscience-and-health-technology-sector-statistics-2021 [Accessed 13/8/2023].

Weinstein, N., Chubb, J., Haddock, G., & Wilsdon, J. (2021) A conducive environment? The role of need support in the higher education workplace and its effect on academics' experiences

of research assessment in the UK. *Higher Education Quarterly*, 75(1), 146–160. Available Online: https://doi.org/10.1111/hequ.12259 [Accessed 17/8/2023].

Yedida, R., Reddy, R., Vahi, R., Jana, R., Gv, A. & Kulkarni, D. (2018) Employee Attrition Prediction.

Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B. & Zhu, X. (2019) Employee Turnover Prediction with Machine Learning: A Reliable Approach. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Systems and Applications. *IntelliSys 2018. Advances in Intelligent Systems and Computing*, 869, 737–758. Available online: https://doi.org/10.1007/978-3-030-01057-7_56 [Accessed 13/8/2023].

Zhu X., Seaver W., Sawhney R., Ji, S., Holt, B., Sanil, G. & Upreti, G. (2017) Employee turnover forecasting for human resource management based on time series analysis. *Journal of Applied Statistics*, 44(8), 1421–1440. Available online: https://doi:10.1080/02664763.2016.1214242 [Accessed 17/8/2023].