Paradigm International Journal of Mathematics and Computer in Engineering 3(2) (2025) 325–348

International Journal of Mathematics and Computer in Engineering

https://reference-global.com/journal/IJMCE

Original Study

Staff churn and lifetime prediction using machine learning

Hasan Hüseyin Yurdagül[1], Hatice Özdemir[1], Adem Seller[1], Fatma Ceren Ulus[2], Mehmet Fatih Akay[3†]

[1]Universal Software Incorporated Company, İstanbul, 34746, Türkiye

[2]EFA Innovation Consultancy and Technology, Adana 01330, Türkiye

[3]Department of Computer Engineering, Faculty of Engineering, Çukurova University, Adana 01330, Türkiye

Communicated by Haci Mehmet Baskonus; Received: 21.03.2024; Accepted: 16.07.2024; Online: 14.12.2025

Abstract

The aim of this study is to develop machine learning based models for staff churn and staff lifetime prediction. Three different approaches are used in the development of these models. In the first approach, prediction models are developed without feature selection. In the second approach, prediction models are developed using the Minimum Redundancy Maximum Relevance (mRMR) feature selection algorithm. In the third approach, prediction models are developed using the Principal Component Analysis (PCA) feature selection algorithm. Two different datasets are used in the development of the models. For predicting staff churn in both datasets, Logistic Regression (LR), Categorical Boosting (CatBoost), and Extreme Learning Machine (ELM) are used. To predict staff lifetime, Support Vector Machine (SVM), Gradient Boosting Machine (LightGBM), and K-Nearest Neighbors (KNN) are used. In order to evaluate the performance of the prediction models, Accuracy and F-Score are used for classification-based models, while Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are used for regression-based models. The results obtained in this study show that the feature selection algorithms have no significant effect on the performance of the models.

Keywords: Staff churn, staff lifetime, machine learning.
AMS 2020 codes: 68Txx; 68T07; 68T01.

1 Introduction

Information Technologies (IT) have become a sector characterized by rapid and continuous technological developments and turnover rates in the IT sector have reached high levels. The most important asset of companies in the IT sector are experienced and talented staffs with technological knowledge. Therefore, companies strive to recruit and retain the best possible team members to ensure they are productive and successful in achieving their business goals. However, according to a study conducted by Linkedin in 2020, the average turnover rate in companies in the IT sector in the US is 13.2% [1]. This rate is quite high compared to other sectors and represents a huge cost in terms of time and resources for companies operating in this sector. A staff working in the IT sector leaves his or her job due to many factors such as technological developments, competitive labour market, lack of work-life balance, lack of promotion opportunities and salary. The company cost of departing staff ranges from almost 30% to 200% of the gross annual salary. In addition, when a staff member leaves the company, he takes his own knowledge and the company's know-how with him. Therefore, companies have

†Corresponding author.
Email address: mfakay@cu.edu.tr

Paradigm
ISSN 2956-7068 doi:10.2478/ijmce-2025-0024

to spend time and resources on training and recruiting new staff. This reduces the company's efficiency and increases its costs. High staff attrition rate has a negative impact on a company's competitiveness. The churn of qualified and experienced staff can lead to a decrease in the quality of service that the company provides to its customers, which in turn can lead to a decrease in customer satisfaction. In addition, the churn of staff can lead to companies having to increase the benefits and incentives for their staff. High turnover rates force companies to offer higher salaries, bonuses and other incentives to retain staff. This situation increases companies' costs and negatively affects their profit margins. In consequence, staff turnover can have a negative impact on a company's productivity, performance and competitiveness. Therefore, companies often need to take various measures such as recruitment strategies, staff retention programs and staff development opportunities to reduce or minimize staff churn. In this strategic planning process, data analytics has become increasingly important. Data analytics applications in the IT sector help companies to make their workforce management strategies more efficient and competitive. For this reason, IT organisations need to invest in software such as attrition scoring and life expectancy prediction, and incorporate this software into their strategic planning processes. Staff churn scoring allows organisations to predict which staff is most likely to leave in a given period. These predictions help organisations identify the staff at high risk of leaving and take preventative action. In this way, companies can protect their existing staff and increase their productivity, rather than looking for new staff to replace those who have left. On the other hand staff lifetime allows companies to estimate how long a particular staff will stay in the workplace. These predictions help companies develop appropriate strategies to increase staff engagement.

In the above mentioned context, the aim of this study is to provide more accurate and efficient decisions in the planning and strategic decision-making processes of organisations by predicting staff turnover in the IT sector and estimating staff life expectancy. In this way, companies can anticipate staff churns and minimise the negative effects of staff churns by taking the necessary precautions. In addition, by predicting staff life expectancy, organisations can predict how long their staff will stay with the company and determine appropriate strategies.

This study is organized as follows: Section 2 covers relevant literature. The datasets used are described in Section 3. The methodology is presented in Section 4. Section 5 presents results and discussion. Section 6 concludes the paper by giving main novelties of this paper.

2 Related work

Recently, many studies on staff turnover prediction have been introduced to literature. In [2], a study of the techniques used in predicting attrition rates has been presented by examining 36 studies between 2009 and 2019. The study states that techniques such as Decision Tree (DT), Artificial Neural Networks (ANN), LR and SVM have been used and that factors such as feature selection/data balance also have an effect. In [3], an attempt has been made to demonstrate the performance of the Chi-Squared Automatic Interaction Detector (CHAID) in estimating staff turnover in the retail industry using real-time data. The performance evaluation of CHAID has been based on Accuracy, Confident Matrix, CHAID-DT and Area Under Curve (AUC). Authors in [4] proposed an approach model for estimating the probability of staff turnover in an organisation. Prediction models have been built by using LR, Random Forest (RF), Extreme Gradient Boosting (XGBoost). The best performing models have been combined using the Stacking Algorithm. The results showed that the proposed model had a higher Accuracy rate. In [5], machine learning algorithms have predicted the probability of staff turnover using a dataset from the Human Resources (HR) department of a company. The article also examined the effect of different characteristics, such as age, gender, working hours, performance, on turnover using a real company dataset. As a result, it has been found that the use of machine learning algorithms achieve high accuracy rates, which can help companies to reduce their turnover rate. In [6], the use of artificial intelligence applications in predicting staff absenteeism and promotion has been explored using a case study of International Business Machines (IBM) staff data. The research results showed that artificial intelligence-based methods provide higher Accuracy rates compared to traditional methods, and thus can be used more effectively in HR management of companies. In [7], CatBoost, LightGBM and XGBoost methods have been compared using k-fold cross validation to determine the probability of staff leaving their organisations. The results showed that the model has been high accuracy value and performance. Authors in [8] presented an XGBoost-powered attrition prediction framework using Genetic Algorithm (GA) based parameter optimisation to optimise the Synthetic Minority Over-Sampling Technique

model for unbalanced data. In [9], an approach has been proposed to predict staff turnover intentions during the recruitment process. The proposed approach has been based on K-Nearest Neighbour (KNN) and RF machine learning algorithms. A staff turnover dataset created by IBM has been used. The KNN-based model performs better than the RF-based model. In [10], LR, KNN, Naive Bayes, DT, SVM, XGBoost and RF have been used to estimate the churn of sales staff. The results showed that RF outperforms in the prediction of salesperson loss. In [11], an approach has been proposed in which DT, KNN and SVM have been used for the prediction of staff turnover. As a result of training and testing on the IBM dataset, the Accuracy of the machine learning algorithms has been calculated. Authors in [12] examined the impact of objective factors on staff turnover, identifies the factors that influence a staff's decision to leave, and predicts whether a particular staff will leave using machine learning algorithms. In [13], RF, XGBoost and ANN, have been used to predict staff turnover. In this study, IBM HR Analytics staff Attrition and performance dataset on Kaggle have been used. The results showed that XGBoost had the best Accuracy value and RF had the best Precision value. ANN has been found to have the best Sensitivity and F1-Score. In [14], a preliminary analysis of the application of machine learning methods for predicting staff turnover has been carried out. At the same time, different classification models have been compared to find the most interpretable one. Among the methods, LR shows the best performance with 88% Accuracy and 85% AUC - Receiver Operator Characteristic (AUC-ROC) values. Authors in [15] experimented with traditional methods such as LR, DT, RF and SVM to estimate staff turnover. The results showed that the SVM model had promising results in predicting staff turnover. In [16], a model for estimating high-risk staff turnover has been presented using HR data from a pharmaceutical company in Iran. A Gradient Boosting (GB) based model with high Accuracy (89%) has been developed using Data Mining algorithms and interviews with HR managers. The results also include the impact of the COVID-19 pandemic and remote working scenarios on staff turnover. In [17], staff turnover prediction has been performed using LR and SVM. The LR model achieved 67% Accuracy, 65% Precision and 73% AUC, while the SVM achieved 93% Accuracy, 98% Precision and 96% AUC. Authors in [18] aimed to analyse staff turnover based on factors such as education, work experience, gender and department. In addition, GB predicted the staff turnover rate using the feature selection and Randomised Grid Search approach by two-way elimination. The cross-validation method has been used to ensure the validity of the results. In [19], ANN, SVM, DT, RF and hybrid ANN-SVM based methods have been used to estimate staff turnover. The performance of the models has been evaluated using various evaluation metrics. It has been found that the SVM and the ANN have a superior performance that has been compared to the other methods.

While this study shares similarities with other studies in the literature in terms of subject matter, it also contains some significant differences. Generally, the literature focuses on staff churn prediction, whereas this study includes both staff churn and staff lifetime predictions. Additionally, previous studies often use the IBM dataset for their experimental work. However, this study is conducted using real and up-to-date data obtained from two different business centers, one from the public sector and one from the private sector. The most notable distinction of this study from other studies in the literature is the use of two different feature selection algorithms. These algorithms identify the most important and influential features in the datasets, which are then used to build the prediction models.

3 Datasets

In this study, two different datasets are used. The first data set used within the scope of the study consists of data belonging to Universal Software Company's (USC) own personnel. The second data set consists of data obtained from a public institution that is a customer of Universal Software. While short-term employees are taken into account in the first data set, longer-term employees are taken into account in the second data set.

This study leveraged two datasets provided by USC. The attributes of the first dataset are listed in Table 1. The first dataset contains 22 staff attributes and 286 pieces of data that include 108 churn and 178 active staff.

Table 1 Attributes in the first dataset.

Attribute	Description
ID	Staff ID
Gender	Gender of the staff
Marital status	Marital status of the staff
Birth Date	Birth date of the staff
Education level	Education level of the staff
Graduated School	School the staff graduated from
Graduated Department	Department the staff graduated from
Department	Department of staff in the company
Duty	Duty of the staff in the company
Total Working Years	Total years of experience of the staff
Work Experience 1 - Working Time	Years of work experience in the first company
Work Experience 2 - Working Time	Years of work experience in the second company
Work Experience 3 - Working Time	Years of work experience in the third company
Universal Entry Date	Entry date of the staff to Universal
Current Salary	Current salary of the staff
Weekly Working Hours	Hours per week the staff works
Job Satisfaction	Staff satisfaction score with his job
Internal Relationship Satisfaction	Internal relationship satisfaction score of the staff
Business Travel Frequency	The number of business trips of the staff
Address - District	District where the staff lives

The attributes of the second dataset are listed in Table 2. The second dataset contains 13 staff attributes and 2512 pieces of data that include 1381 churn and 1131 active staff.

Table 2 Attributes in the second dataset.

Attribute	Description
Staff Type	Type of the staff
Staff Subtype	Subtype of the staff
Gender	Gender of the staff
Marital Status	Marital status of the staff
Agreement	Agreement between company and staff
Education level	Education level of the staff
Department 1	The department to which the staff is affiliated
Department 2	The department where the staff works
Insured Unit	Insurance policy of the staff
Duty	The duty assigned to the staff
Address - Province	Province where the staff lives
Address - District	District where the staff lives
Age	Age of the staff

4 Methodology

In this study, analytical developments have been provided using the Python programming language. Numpy and Pandas libraries have been used for data operations. Matplotlib library has been used for data visualization. Tensorflow, PyTorch and Scikit-learn libraries have been used for modeling.

4.1 Minimum redundancy maximum relevance

mRMR ensures that the selected traits are not only minimally associated with the input traits but also strongly associated with the output variable. The mRMR approach aims to select the best features while diluting the weak ones [20]. The approach works iteratively, determining the most significant feature for the output variable and the least associated with any of the features in the dataset during each iteration. After a feature is chosen, it is removed from the original dataset. This process is repeated until K iterations have been completed.

4.2 Principal component analysis

PCA is a statistical method used to reduce the dimensions of a dataset while preserving most of its variance. It transforms the original variables into new, uncorrelated principal components. In our study, PCA involved calculating the eigenvalues and eigenvectors of the data covariance matrix to identify the components that capture the maximum variance. We select the principal components based on the cumulative explained variance ratio, retaining enough components to capture a significant amount of the total variance. This dimensionality reduction simplifies the data and enhances computational efficiency, making PCA an essential step in our data preprocessing approach.

4.3 Logistic regression

LR is a statistical analysis technique used to predict binary outcomes, such as yes or no, based on past observations from a dataset. It is based on the logistic function, which converts probability into a value between 0 and 1. The model is fitted with an optimization procedure that maximizes the likelihood of the observed data given the predicted probabilities. To be able to pick best parameters for the algorithm, we employ the grid search cross-validation to determine the best hyperparameters. The primary focus was on tuning the regularization strength (C) and the type of penalty (l1 or l2). We test several values of C from 0.01 to 100 to understand the trade-off between bias and variance, aiming to prevent overfitting while maintaining the model's ability to generalize.

4.4 Support vector machines

SVM is a supervised learning method that examines data for classification and regression by identifying a hyperplane in an N-dimensional space that unambiguously classifies data points. Several hyperplanes may be used to split the two categories of data points. The algorithm aims to locate the plane with the greatest distance between data points from both classes. To be able to pick best parameters for the algorithm, we employ the grid search approach with a specific emphasis on the kernel type (linear, polynomial, RBF), the penalty parameter C, and the kernel coefficient gamma. Given the high dimensionality and the nature of our data, special attention was given to the RBF kernel to explore non-linear boundaries for predicting staff lifetime.

4.5 Light gradient boosting machine

LightGBM is a gradient-boosting framework that uses tree-based learning techniques. It has several advantages over other boosting algorithms, including high processing speed, low Random Access Memory usage, and parallel and Graphics Processing Unit (GPU) learning support. LightGBM is a histogram-based method that discretizes continuous-value variables to minimize processing costs. The training time of decision trees is determined by the number of divisions required for calculation. LightGBM reduces resource utilization and shortens training time by separating the tree per leaf and growing it by selecting the leaf with the greatest delta loss. However, this approach may complicate the model and lead to overfitting on short samples. LightGBM was optimized by adjusting the number of leaves, the learning rate, and the number of boosting iterations. We utilized LightGBM's gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) to efficiently handle large data scales and high dimensionality, focusing on minimizing overfitting and improving predictive performance.

4.6 Categorical boosting

Catboost is an open-source machine learning technique based on GB. It stands out from other methods due to its rapid learning speed, ability to handle numeric, categorical, and text data, GPU support, and visualization

options. Additionally, it can handle empty data and code categorical data without requiring a separate coding step during data preparation. CatBoost's parameters were optimized using the model's built-in tools for handling categorical features effectively. We focused on the depth of the trees, learning rate, and the number of trees. Additionally, CatBoost's feature importance scores were utilized to iteratively refine the features included in the final model.

4.7 Extreme learning machine (ELM)

ELM is a training algorithm for a single hidden layer Feed Forward Network (FFNN) that converges much faster than traditional methods and provides promising performance [21]. Unlike standard FFNN learning techniques, such as the Back Propagation Algorithm (BPA), ELM does not employ a gradient-based methodology. This procedure sets all parameters once, eliminating the need for iterative training. For the ELM, our main tuning parameters were the number of hidden neurons and the type of activation function. We experimented with different neuron counts to find a balance between processing time and model accuracy, using a validation set to gauge performance.

4.8 K-Nearest neighbors regressor (KNN)

K-Nearest Neighbors Regressor (KNN) is a straightforward yet effective non-parametric algorithm used for predictive modeling based on the similarity of features. It predicts the output variable by averaging the values of its K nearest neighbors, making it well-suited for regression tasks involving continuous data. To optimize the KNN model, the main parameter adjusted was the number of neighbors (K). The optimal K value is critical, as a low K can lead to a model sensitive to noise, while a high K may result in over-smoothing and potential under fitting. We utilized a validation curve to explore K values from 1 to 20, enabling us to select the most appropriate K for our data.

5 Results and discussion

In this study, three different approaches have been used to develop the prediction models. The first approach involves developing prediction models without feature selection on the datasets. The second approach involves developing prediction models using the mRMR feature selection algorithm. The third approach involves developing prediction models using one of the dimensionality reduction algorithms called Principal Component Analysis algorithm.

5.1 The prediction models developed using the first dataset and their corresponding results

In the first dataset, prediction models are developed to score staff churn using 5-fold cross-validation with LR. For the prediction of staff lifetime, prediction models are developed using 5-fold cross validation with SVM. The features selected by the mRMR feature selection algorithm are Job Satisfaction, Internal Relationship Satisfaction, Universal Check-in Day, Graduated Department, Work Experience 1 - Working Time, Work Experience 3 - Working Time, Universal Entry Month, Birthday, Total Working Years and Graduated School.

Parameter values of LR used in the developed staff churn prediction model is given in Table 3.

Table 3 Parameter values staff churn prediction model developed using LR, CatBoost and ELM.

Algorithm	Parameters	Values
LR	Penalty	12
LR	C	10^5
CatBoost	N_estimators	100
CatBoost	Max_depth	6
CatBoost	Iterations	500
CatBoost	Learning_rate	0.03
ELM	Alpha	10^{-5}
ELM	N neurons	64
ELM	Activation Function	tanh

The complexity matrix of the staff churn prediction model developed using the first approach with LR, Catboost and ELM are given in Figures 1, 2 and 3.

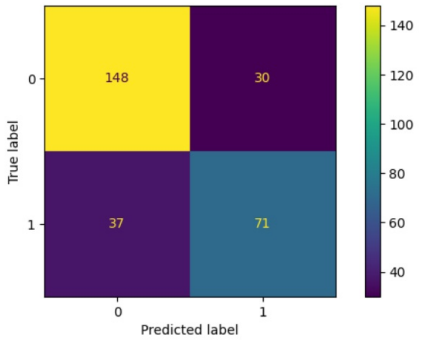

Fig. 1 Complexity matrix obtained with LR.

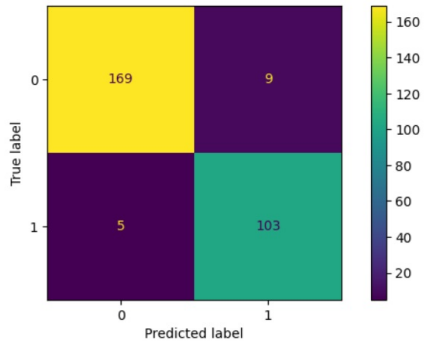

Fig. 2 Complexity matrix obtained with Catboost.

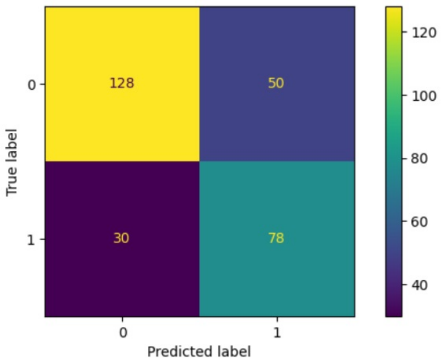

Fig. 3 Complexity matrix obtained with ELM.

The complexity matrix of the staff churn prediction model developed using the second approach with LR, Catboost and ELM are given in Figures 4, 5 and 6.

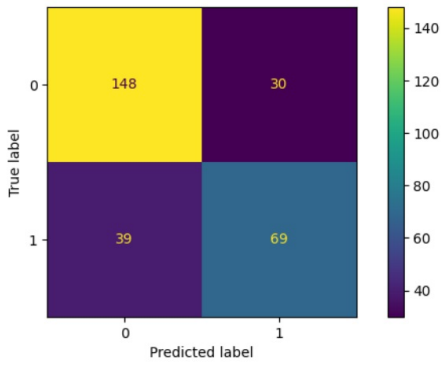

Fig. 4 Complexity matrix obtained with LR.

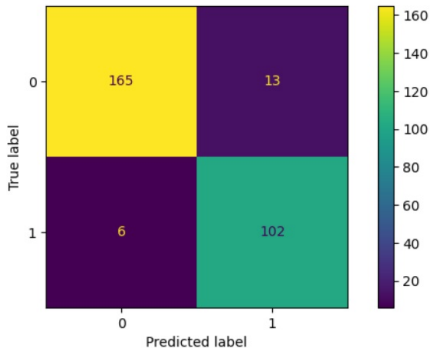

Fig. 5 Complexity matrix obtained with Catboost.

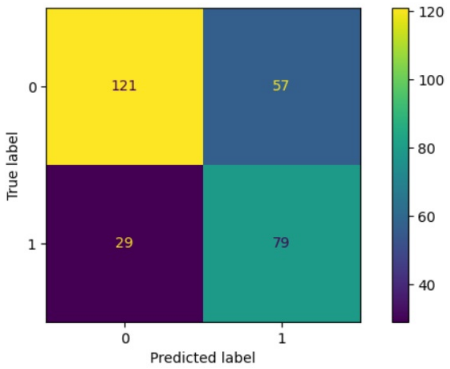

Fig. 6 Complexity matrix obtained with ELM.

The complexity matrix of the staff churn prediction model developed using the third approach with LR, Catboost and ELM are given in Figures 7, 8 and 9.

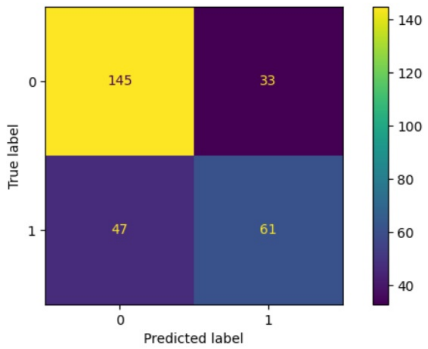

Fig. 7 Complexity matrix obtained with LR.

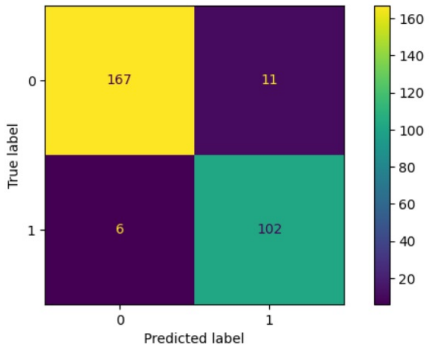

Fig. 8 Complexity matrix obtained with Catboost.

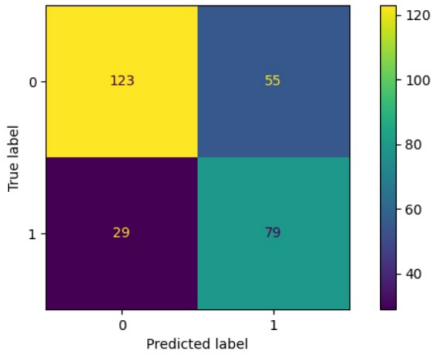

Fig. 9 Complexity matrix obtained with ELM.

The comparison of Accuracy and F1-Score values of the results obtained with the staff churn prediction models developed using LR, Catboost and ELM are given in Table 4.

Table 4 Accuracy and F1-score values of staff churn prediction models developed using LR, Catboost and ELM.

Approaches	Algorithms	F1-Score	Accuracy
First Approach	LR	0.68	0.77
First Approach	CatBoost	0.94	0.95
First Approach	ELM	0.66	0.72
Second Approach	LR	0.67	0.76
Second Approach	CatBoost	0.91	0.93
Second Approach	ELM	0.65	0.70
Third Approach	LR	0.60	0.72
Third Approach	CatBoost	0.92	0.94
Third Approach	ELM	0.65	0.71

Based on the results:

-The First Approach is most effective, particularly for CatBoost, which suggests that the methodology or data handling in this approach synergizes well with CatBoosts algorithm.

-In the Second Approach, all algorithms show a decrease in performance with CatBoost dropping by about 0.03 in F1-Score and 0.02 in Accuracy compared to the First Approach. However, it still maintains strong performance relative to LR and ELM.

-The Third Approach sees further deterioration in the performance of LR, with a drop to an F1-Score of 0.60 and an Accuracy of 0.72. Meanwhile, CatBoost and ELM maintain relatively stable performances, indicating resilience or better suitability of these algorithms for the conditions or treatments specific to this approach.

Parameter values of SVM used in the developed staff lifetime prediction model is given in Table 5.

Table 5 Parameter values of staff lifetime prediction model developed using SVM, LightGBM and KNN.

Methods	Parameters	Values
SVM	C	10^{10}
SVM	Kernel	RBF
SVM	Gamma	Scaled
LightGBM	N_estimators	50
LightGBM	Num_leaves	30
LightGBM	Max_depth	None
LightGBM	Learning_rate	0.1
KNN	N_neighbours	5

The comparison graph of predicted and actual values for staff lifetime prediction models developed using the first approach with SVM, LightGBM and KNN are given in Figures 10, 11 and 12.

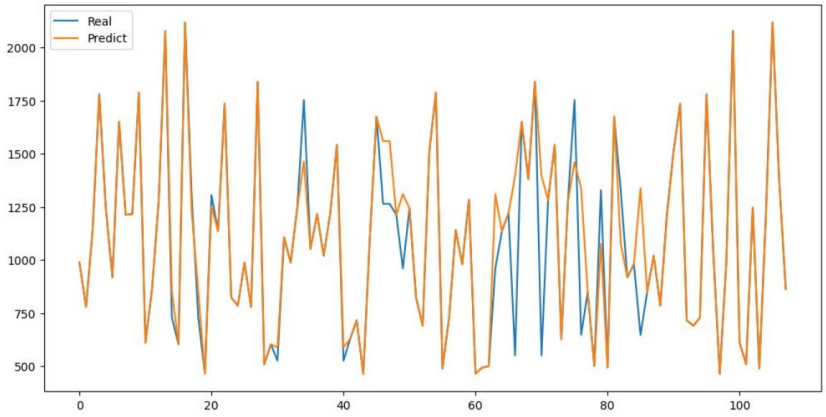

Fig. 10 Predicted values and actual values generated by SVM.

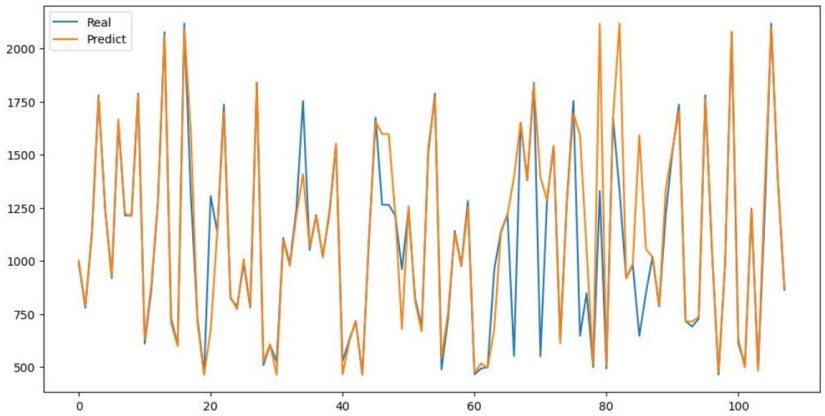

Fig. 11 Predicted values and actual values generated by LightGBM.

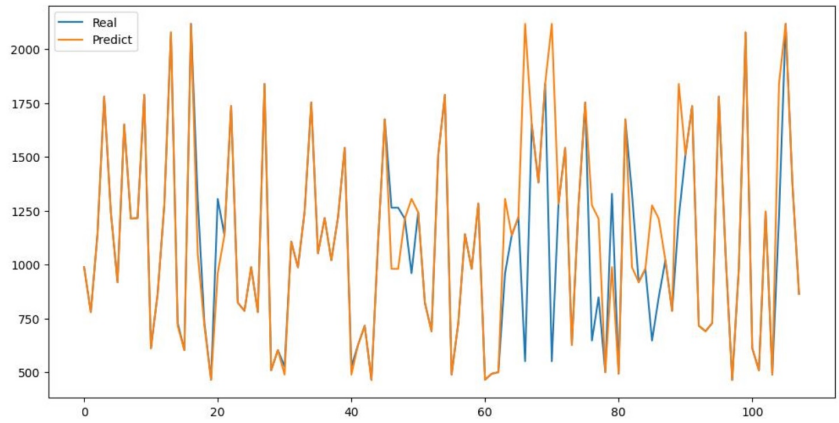

Fig. 12 Predicted values and actual values generated by KNN.

The comparison graph of predicted and actual values for staff lifetime prediction models developed using the second approach with SVM, LightGBM and KNN are given in Figures 13, 14 and 15.

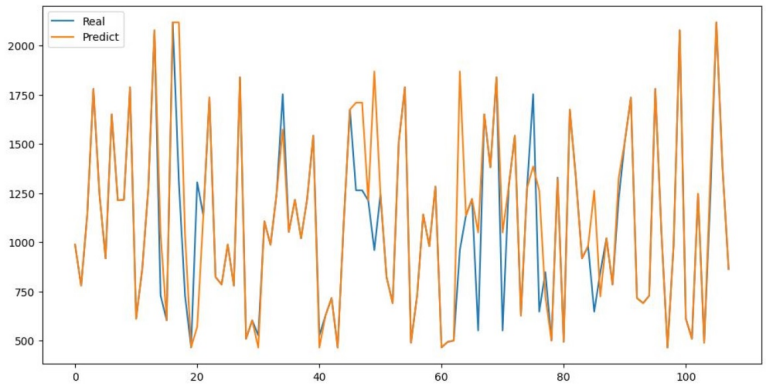

Fig. 13 Predicted values and actual values generated by SVM.

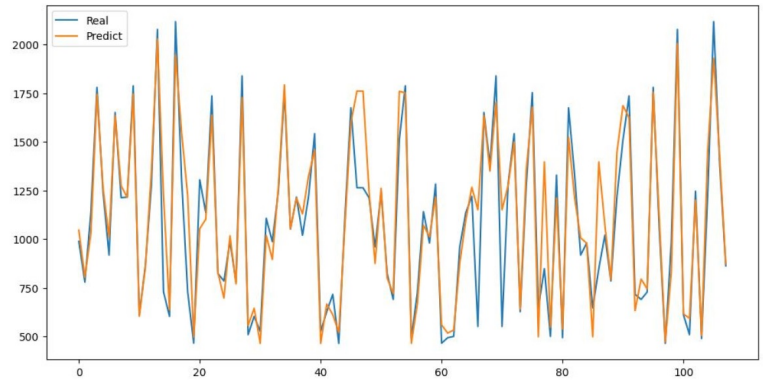

Fig. 14 Predicted values and actual values generated by LightGBM.

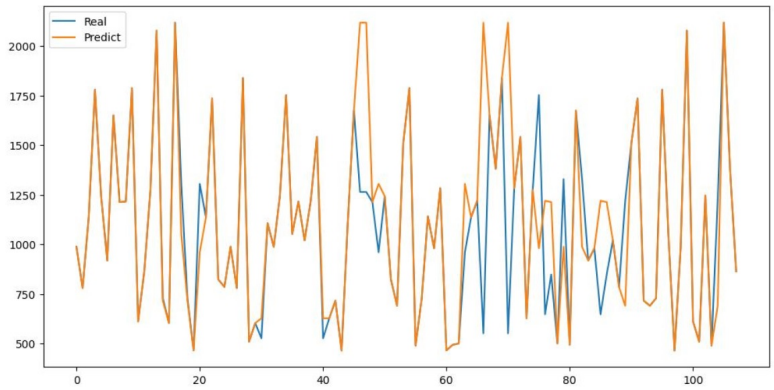

Fig. 15 Predicted values and actual values generated by KNN.

The comparison graph of predicted and actual values for staff lifetime prediction models developed using the third approach with SVM, LightGBM and KNN are given in Figures 16, 17 and 18.

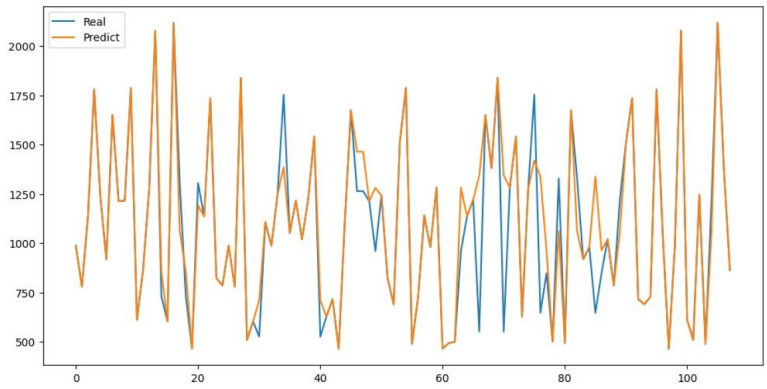

Fig. 16 Predicted values and actual values generated by SVM.

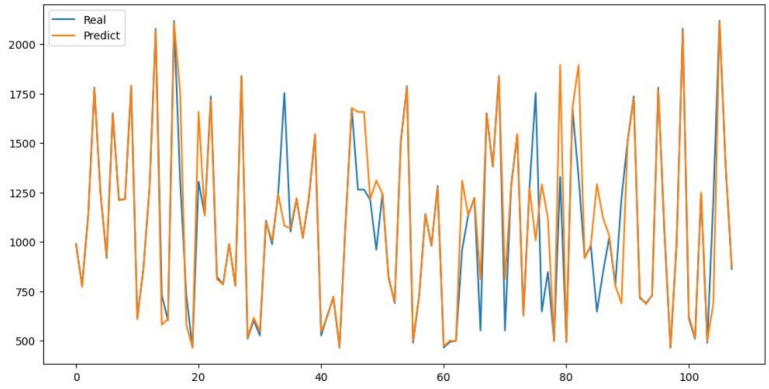

Fig. 17 Predicted values and actual values generated by LightGBM.

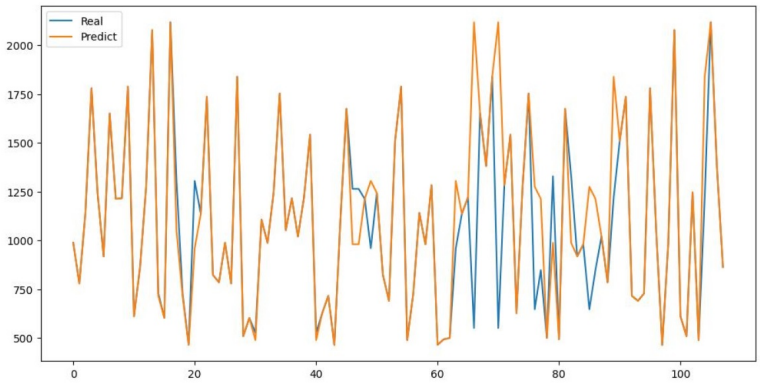

Fig. 18 Predicted values and actual values generated by KNN.

The comparison of MAE and MAPE values of the results obtained with the staff lifetime prediction models developed using SVM are given in Table 6.

Table 6 Accuracy and F1-score values of staff churn prediction models developed using LR, Catboost and ELM.

Approaches	Algorithms	MAE	MAPE (%)
First Approach	SVM	55.44	7.24
First Approach	LightGBM	87.97	10.44
First Approach	KNN	83.33	10.95
Second Approach	SVM	76.54	8.69
Second Approach	LightGBM	102.86	11.37
Second Approach	KNN	99.53	12.12
Third Approach	SVM	62.70	8.01
Third Approach	LightGBM	83.98	8.41
Third Approach	KNN	83.33	10.95

Based on the results:

- The First Approach is highly effective for SVM, which achieves the lowest MAE (55.44) and MAPE (7.24%), suggesting a strong alignment between the algorithm and the methodology or data handling in this scenario.

- In the Second Approach, there is a noticeable performance decline for all algorithms, with LightGBM experiencing the most significant increase in error metrics, indicating potential misalignment between the algorithm capabilities and the approach's complexity.

- The Third Approach sees improvements in error metrics for LightGBM and SVM, demonstrating better optimization or alignment with this approach's requirements. SVM consistently maintains lower error rates, underscoring its effectiveness across different settings.

5.2 The prediction models developed using the second dataset and their corresponding results

In the second dataset, prediction models are developed for staff churn scoring using LR, and CatBoost. For the prediction of staff lifetime, prediction models are developed using SVM, LightGBM and KNN. When developing the models on the second dataset, results have been obtained using 10-fold cross validation. The features selected by the mRMR feature selection algorithm are Staff Type, Workplace, Mission, Agreement,

Address District, Staff Subtype and Age. Parameter values of the staff churn prediction models developed using LR, Catboost and ELM are given in Table 7.

Table 7 Parameter values of staff churn prediction models.

Algorithms	Parameters	Values
LR	Penalty	12
LR	C	10^5
CatBoost	N_estimators	100
CatBoost	Max_depth	6
CatBoost	Iterations	1000
CatBoost	Learning_rate	0.1
ELM	Alpha	10^{-3}
ELM	N neurons	64.32
ELM	Activation function	tanh

The complexity matrices of the staff churn prediction model developed using the first approach with LR, CatBoost and ELM are given in Figures 19, 20 and 21.

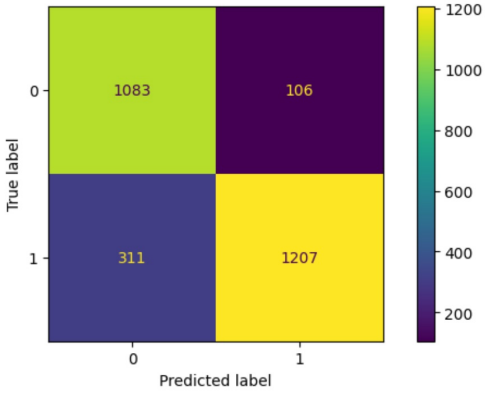

Fig. 19 Complexity matrix obtained with LR.

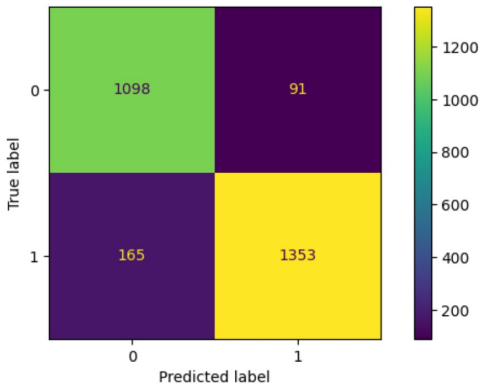

Fig. 20 Complexity matrix obtained with Catboost.

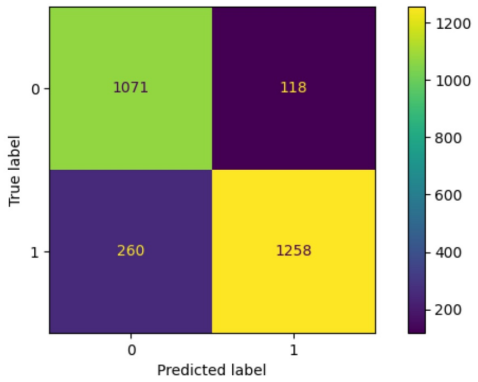

Fig. 21 Complexity matrix obtained with ELM.

The complexity matrices of the staff churn prediction model developed using the second approach with LR, CatBoost and ELM are given in Figures 22, 23 and 24.

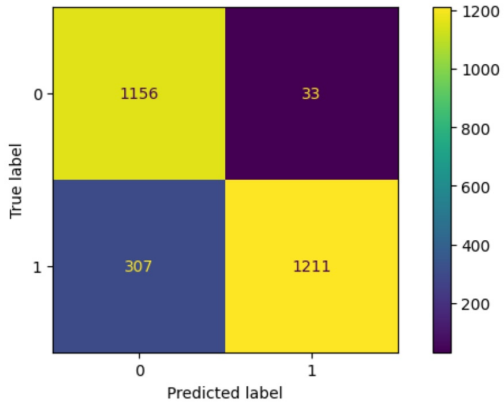

Fig. 22 Complexity matrix obtained with LR.

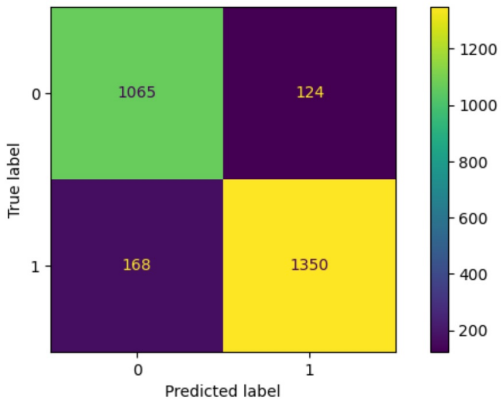

Fig. 23 Complexity matrix obtained with Catboost.

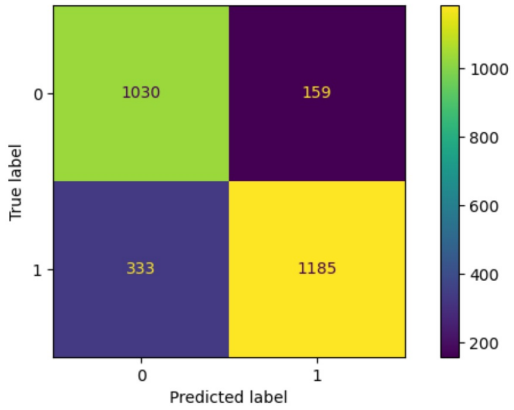

Fig. 24 Complexity matrix obtained with ELM.

The complexity matrices of the staff churn prediction model developed using the third approach with LR, CatBoost and ELM are given in Figures 25, 26 and 27.

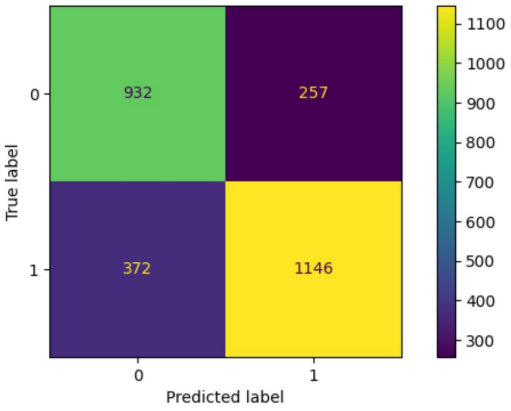

Fig. 25 Complexity matrix obtained with LR.

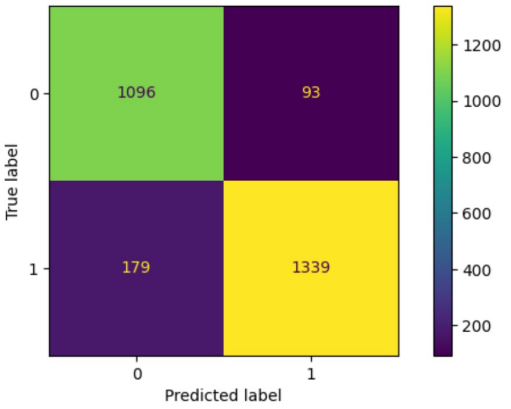

Fig. 26 Complexity matrix obtained with Catboost.

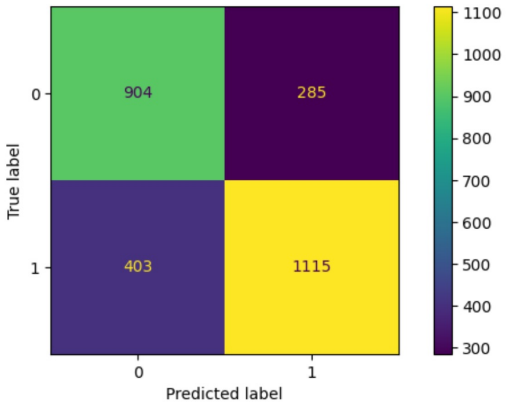

Fig. 27 Complexity matrix obtained with ELM.

The comparison of Accuracy and F1-Score values of the results obtained with the staff churn prediction models developed using LR, CatBoost and ELM are given in Table 8.

Table 8 Accuracy and F1-score values of staff churn prediction models developed using LR, Catboost and ELM.

Approaches	Algorithms	F1-Score	Accuracy
First Approach	LR	0.85	0.85
First Approach	CatBoost	0.91	0.91
First Approach	ELM	0.87	0.86
Second Approach	LR	0.88	0.87
Second Approach	CatBoost	0.90	0.89
Second Approach	ELM	0.83	0.82
Third Approach	LR	0.78	0.77
Third Approach	CatBoost	0.91	0.90
Third Approach	ELM	0.76	0.75

Based on the results

- The First Approach yields strong results, especially for CatBoost, which achieves an F1-Score and Accuracy of 0.91, suggesting optimal compatibility between the algorithm and the methodology used in this approach.

- The Second Approach shows a slight decline in performance for all algorithms, with CatBoost and LR experiencing small drops in their F1-Scores and Accuracy compared to the First Approach, indicating that the adjustments or complexities introduced in this approach may slightly hinder the effectiveness of these algorithms.

- The Third Approach leads to a stabilization in performance for CatBoost, maintaining high scores similar to the First Approach, while LR and ELM see further declines. This suggests that CatBoost is more robust or better suited to the conditions or treatments specific to this approach, whereas ELM struggles, showing a noticeable decrease in performance. Parameter values of the staff lifetime prediction developed using SVM, and LightGBM are given in Table 9.

Table 9 Parameter values of staff lifetime prediction models.

Methods	Parameters	Values
SVM	C	10^5
SVM	Kernel	RBF
SVM	Gamma	Scaled
LightGBM	N_estimators	100
LightGBM	Num_leaves	50
LightGBM	Max_depth	None
LightGBM	Learning_rate	0.1
KNN	N_neighbours	5

The comparison graph of predicted and actual values for staff lifetime prediction models developed using the first approach with SVM, LightGBM and KNN are given in Figures 28, 29 and 30.

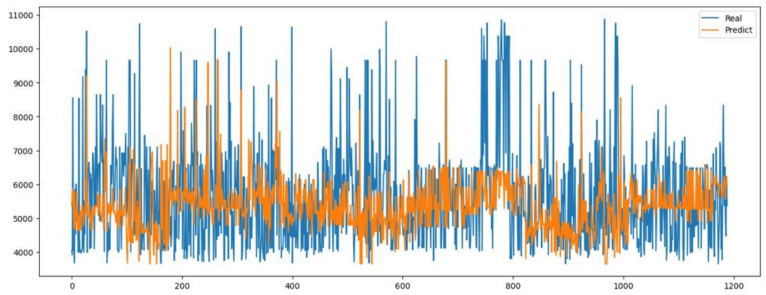

Fig. 28 Predicted values and actual values generated by SVM.

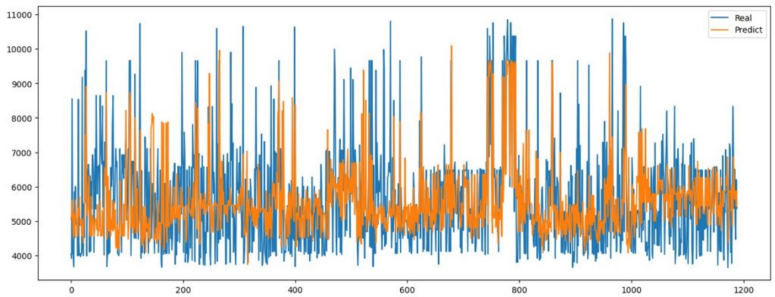

Fig. 29 Predicted values and actual values generated by LightGBM.

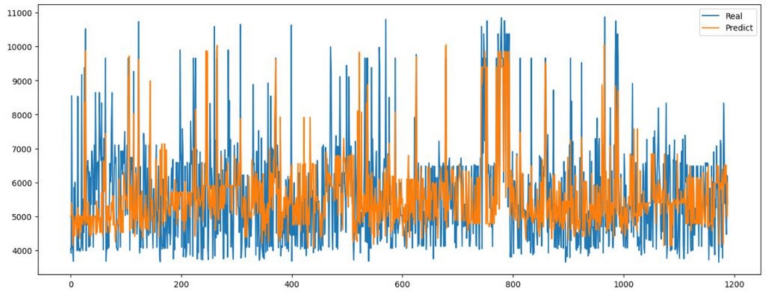

Fig. 30 Predicted values and actual values generated by KNN.

The comparison graph of predicted and actual values for staff lifetime prediction models developed using the second approach with SVM, LightGBM and KNN are given in Figures 31, 32 and 33.

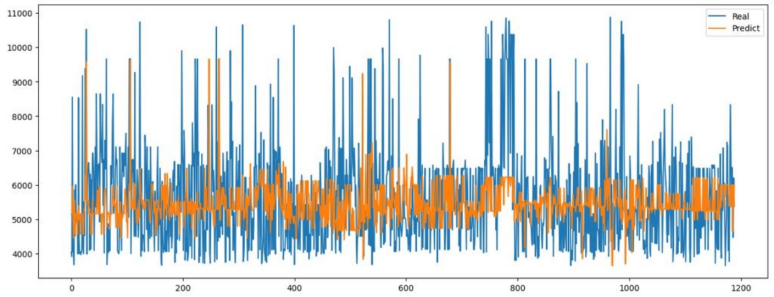

Fig. 31 Predicted values and actual values generated by SVM.

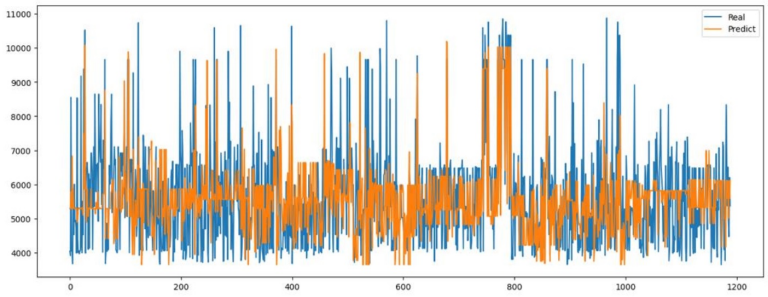

Fig. 32 Predicted values and actual values generated by LightGBM.

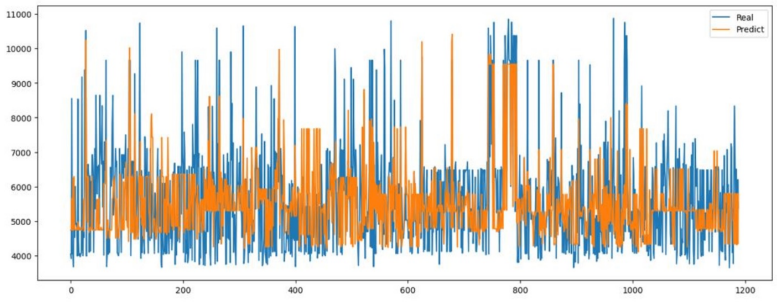

Fig. 33 Predicted values and actual values generated by KNN.

The comparison graph of predicted and actual values for staff lifetime prediction models developed using the third approach with SVM, LightGBM and KNN are given in Figures 34, 35 and 36.

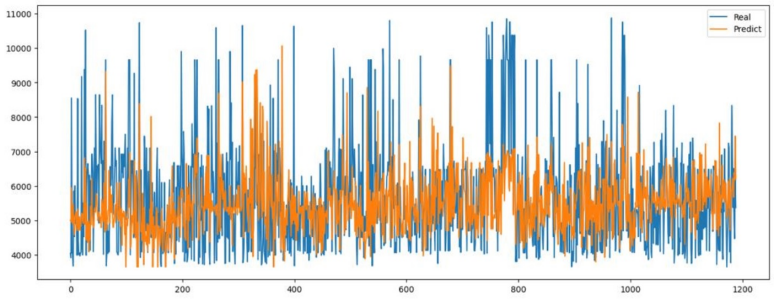

Fig. 34 Predicted values and actual values generated by SVM.

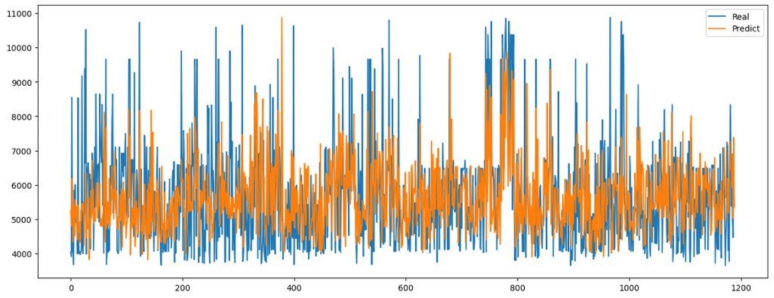

Fig. 35 Predicted values and actual values generated by LightGBM.

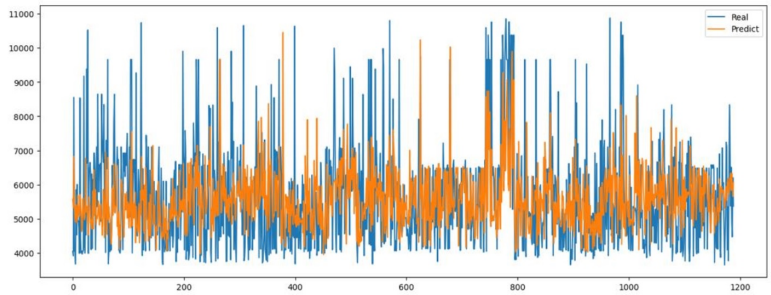

Fig. 36 Predicted values and actual values generated by KNN.

The comparison of MAE and MAPE values of the results obtained with the staff lifetime prediction models developed using SVM, LightGBM and KNN are given in Table 10.

Table 10 Accuracy and F1-score values of staff churn prediction models developed using LR, Catboost and ELM.

Approaches	Algorithms	MAE	MAPE (%)
First Approach	SVM	1158.25	19.94
First Approach	LightGBM	1011.81	18.37
First Approach	KNN	999.96	18.35
Second Approach	SVM	1124.01	19.68
Second Approach	LightGBM	1038.58	19.03
Second Approach	KNN	1044.15	18.89
Third Approach	SVM	1087.65	19.07
Third Approach	LightGBM	1045.42	18.85
Third Approach	KNN	1068.99	19.19

Based on the results

- The First Approach shows the best performance for KNN with the lowest MAE (999.96) and a low MAPE (18.35%), indicating that KNN is particularly effective in handling the specific data and methods used in this approach.

- In the Second Approach, all algorithms exhibit a slight increase in both MAE and MAPE, with SVM showing the highest MAE (1124.01) and MAPE (19.68%), suggesting that the complexities introduced in this approach may affect the precision of the predictions.

- The Third Approach sees a further increase in MAE for all algorithms except SVM, which shows a decrease to 1087.65, maintaining a relatively stable MAPE (19.07%). This indicates some recovery in performance for SVM, while the other models continue to struggle with the approach's conditions.

- When the staff churn prediction models developed with the first dataset and the second dataset are compared,

* Two implemented approaches perform close to each other.

* The feature selection algorithm exhibited negligible impact on model performance.

- When the staff lifetime prediction models developed with the first dataset and the second dataset are compared,

* The implementation of the mRMR feature selection algorithm led to performance deterioration in LightGBM

models.

* The SVM-based model developed through the second and third approach demonstrated a 1.06% decrease in MAPE value compared to the SVM-based model developed through the first approach.

6 Conclusion

With the growth of the IT sector, the employment rate in the sector is increasing at the right rate. However, for a variety of reasons, the tenure of staffs in the IT sector is shorter than in other sectors. Therefore, companies are engaged in various activities to retain their talented staffs. This is because it takes time for the staff to be replaced by a departing staff to learn skills similar to the technical or business expertise of the departing staff. This study proposes models for predicting turnover and life expectancy using machine learning techniques. The results show that staff churn and life expectancy can be predicted with reasonable error rates using machine learning algorithms. In general, it has been shown that staff turnover and staff life can be predicted using machine learning methods, and the results are promising for future studies. However, this study may present various challenges regarding the integration with HR systems and data privacy. Firstly, the development and implementation process of the model requires ensuring data privacy and the protection of personnel information. Therefore, data usage must comply with legal standards. Additionally, the accuracy and timeliness of the data play a crucial role in determining the model's performance. It is essential that incoming data is accurate and up-to-date during model development. Another issue is ensuring continuous data flow and compatibility when applying machine learning models to a system. Managing the information flow between the system and the model, as well as securing and managing the necessary resources for integrating various software applications, is of utmost importance.

7 Declarations

7.1 Conflict of interest:

Authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7.2 Author's contributions:

H.H.Y.-Conceptualization, Data Curation. H.Ö.-Software, Data Curation, Writing-Original Draft. A.S.-Software, Data Curation. F.C.U.-Conceptualization, Methodology, Writing. M.F.A.-Supervisior, Methodology, Writing-Original Draft. All authors read and approved the final submitted version of this manuscript.

7.3 Funding:

Not applicable.

7.4 Acknowledgement:

Not applicable.

7.5 Data availability statement:

The article includes all the data that support the findings of this study.

7.6 Using of AI tools:

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

References

[1] Linkedin, Industries with the Highest (and Lowest) turnover rates, https://www.linkedin.com/business/talent/blog/talent-strategy/industries-with-the-highest-turnover-rates, Accesed: July 25, 2023.

[2] Akasheh M.A., Malik E.F., Hujran O., Zaki N., A decade of research on data mining techniques for predicting employee turnover: a systematic literature review, Expert Systems with Applications, 121794, 2023.

[3] Chaudhary M., Gaur L., Chakrabarti A., Who's next: evaluating employee churn in retail using machine learning algorithm CHAID, Journal of Survey in Fisheries Sciences, 10(1), 339–351, 2023.

[4] Chung D., Yun J., Lee J., Jeon Y., Predictive model of employee attrition based on stacking ensemble learning, Expert Systems with Applications, 215, 119364, 2023.

[5] Shaik S., Kumar P.S., Reddy S.V., Reddy K.S.S., Bhutada S., Machine learning based employee attrition predicting, Asian Journal of Research in Computer Science, 15(3), 34–39, 2023.

[6] Sharma S., Bhat R., Implementation of artificial intelligence in diagnosing staffs attrition and elevation: a case study on the IBM employee dataset, The Adoption and Effect of Artificial Intelligence on Human Resources Management, Part A, 197–213, 2023.

[7] Kakulapati V., Subhani S., Predictive analytics of employee attrition using K-Fold methodologies, International Journal of Mathematical Sciences and Computing, 1, 23–36, 2023.

[8] Konar K., Das S., Das S., Employee attrition prediction for imbalanced data using genetic algorithm-based parameter optimization of XGB classifier, 2023 International Conference on Computer Electrical and Communication Engineering, IEEE, 20-21 January 2023, Kolkata, India.

[9] Atef M., Elzanfaly D.S., Ouf S., Early prediction of employee turnover using machine learning algorithms, International Journal of Electrical and Computer Engineering Systems, 13(2), 135–144, 2022.

[10] Cömert G.D., Özcan T., Kaya T., Towards Industry 5.0, (Chapter 3: Salesperson churn prediction with machine learning approaches in the retail industry), International Symposium for Production Research, 6-8 October 2022, Antalya, Türkiye.

[11] Ganthi L.S., Nallapaneni Y., Perumalsamy D., Mahalingam K., Proceedings of International Conference on Data Science and Applications, (Chapter 44: Employee attrition prediction using machine learning algorithms), International Conference on Data Science and Applications, 10-11 April 2021, Kolkata, India.

[12] Garigipati R.K., Raghu K., Saikumar K., Handbook of Research on Technologies and Systems for E-Collaboration During Global Crises, (Chapter 9: Detection and identification of employee attrition using a machine learning algorithm), IGI Global, USA, 2022.

[13] Gim S., Im E.T., Big Data, Cloud Computing, and Data Science Engineering, (Chapter 5: A study on predicting employee attrition using machine learning), 7th IEEE/ACIS International Conference on Big Data, Cloud Computing, Data Science and Engineering (BCD 2021) 4-6 August 2022, Danang, Vietnam.

[14] Guerranti F., Dimitri G.M., A comparison of machine learning approaches for predicting employee attrition, Applied Sciences, 13(1), 267, 2023.

[15] Maria-Carmen L., Education Research and Business Technologies, (Chapter 27: Classical machine-learning classifiers to predict employee turnover), 20th International Conference on Informatics in Economy, 14-15 May 2021, Bucharest, Romania.

[16] Mozaffari F., Rahimi M., Yazdani H., Sohrabi B., Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data, Benchmarking: An International Journal, 30(10), 4140–4173, 2023.

[17] Maharjan R., Employee Churn Prediction using Logistic Regression and Support Vector Machine, https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=2043&context=etd_projects, Accessed: July 17, 2024.

[18] Mate Y., Potdar A., Priya R.L., Artificial Intelligence and Technologies, (Chapter 10: Ensemble methods with didirectional feature elimination for prediction and analysis of employee attrition rate during COVID-19 pandemic), 3rd International Conference on Recent Trends In Advanced Computing Artificial Intelligence and Technology, 17-18 December 2020, Chennai, India.

[19] Poornappriya T.S., Gopinath R., Employee attrition in human resource using machine learning techniques, Webology, 18(6), 2844–2856, 2021.

[20] Ding C., Peng H., Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology, 03(02), 185–205, 2005.

[21] Wang J., Lu S., Wang S.H., Zhang Y.D., A review on extreme learning machine, Multimedia Tools and Applications, 81(29), 41611–41660, 2022.