HRescue: A Modern ML approach for Employee Attrition Prediction

Rudresh Veerkhare[1†], Parshwa Shah[1†], Jiten Sidhpura[1†], and Sudhir Dhage[1]

[1] Department of Computer Engineering,
Sardar Patel Institute of Technology, Mumbai, India.
{rudresh.veerkhare, parshwa.shah, jiten.sidhpura,
sudhir_dhage}@spit.ac.in
[†]These authors contributed equally to this work.

Abstract. The biggest strength for any organization is its employees. Companies invest a lot of money and time in employees to retain them. A lot of opportunities are available for employees in this competitive technological world. Employees are leaving the organization for various reasons leading to a high attrition rate. It is not only the employees but also their talent, skills, and values leave the organization. Machine learning models can assist companies in attrition prediction. In this study, we have used the IBM HR Analytics dataset. Being an imbalanced dataset, we used ADASYN to balance it. Later we used state-of-the-art gradient boosting algorithms such as XGBoost and CatBoost on original and balanced datasets. We also implemented a new privacy-based approach with the help of Federated Learning that will help different companies to train a single robust model. Finally, with the help of SHAP values, we analyzed our model predictions. For model evaluation, we used F1-score and Accuracy as our core metrics. We obtained an F1-score of 0.69 and 0.94 on the imbalanced and balanced dataset, respectively. This solution can help companies to foresee employees who can resign so that they can take appropriate measures.

Keywords: Employee Attrition · Federated Learning · Shap Analysis · Gradient Boosting · ADASYN

1 Introduction

Employees are the key asset of any organization. Each company grows by cumulative efforts by all its employees. Matching the right talent with the correct job position is crucial for the HR department. Each company invests significant time and money in employee screening, interviews, hiring, onboarding, training, benefits, and retention policies. However, in this competitive world, the demand for highly skilled employees has skyrocketed, thus giving employees lucrative opportunities. Retaining proficient employees is becoming difficult day by day which has given rise to a phenomenon called attrition, employees leaving the company through resignation, retirement, or some other personal reasons.

According to [1], 31% of new hires have left the job within six months. Attrition is detrimental to the entire organization. There are various reasons for employee attrition such as not being excited about the work assigned, not having enough growth opportunities, inadequate compensation, not enough salary hikes, superficial and toxic relationships with their managers and colleagues. Not getting adequate leaves and quality family time reduces their motivation to work. So, each organization should take care of employees' professional growth, work-life balance, and physical and mental health to reduce attrition.

We are now a part of a data-driven economy, where data speaks more than anything. Various factors influence the attrition rate and companies can get a lot of insights about employee attrition from their employee data. Accordingly, companies can introduce various benefits, policies, healthcare programs, workshops, volunteering activities to enhance employee morale and thus, retain their employees. Retaining Employees eventually creates a win-win situation for both the company and its employees.

The objective of this research study is to utilize machine learning models to predict whether an employee working in an organization will continue to work or leave the organization. We have also focused on explaining the predictions of our classification models to determine how input data features affect attrition.

We have followed three approaches in our proposed methodology with the input dataset. The methodology flow diagram is mentioned in figure 1. In the first approach, we have an original imbalanced dataset. In the second approach, we have applied ADASYN oversampling technique to balance the dataset. After data preprocessing, classifier models were used to classify attrition class. Finally, we evaluate the model and perform model prediction analysis using SHAP values. In the third approach, we used federated learning approach where we train the model on multiple clients and perform aggregation to get a robust global model.

The structure of the paper is as follows: Section 2 describes the literature survey, where we discussed existing approaches. In Section 3, we have proposed our methodology. Section 4 contains information about Model Prediction Analysis. Section 5 demonstrates the experimental setup. In Section 6, we have presented our results. Finally, in section 7, we conclude our paper.

2 Literature Survey

Saeed Najafi-Zangeneh et al. [2] proposed a machine learning framework for Employee Attrition prediction. They have used the popular IBM HR dataset for evaluating their proposed methodology. Since the number of features in the dataset are very high, the authors have proposed max-out algorithm as their feature selection method. After that, they trained a Logistic Regression model on these newly selected features. The authors used F1-score for evaluating their model because the IBM dataset is imbalanced and achieved an accuracy score of 0.81 and an F1-Score of 0.56. Finally, they trained their model on multiple bootstrapped datasets, followed by computing the average and standard deviation of model parameters to validate their stability.

Shikha N. Khera and Divya [3] trained a Support Vector Machine (SVM) model to predict the Employee Attrition rate specifically for the Indian IT industry. They gathered the HR data from three Indian IT firms and developed a dataset containing 1650 employees. Their dataset was imbalanced, with around 83.87% of the data belonging to the class in which employee is working for the organization (active). The remaining data belongs to the class where the employee has left the organization (inactive). Originally there were 22 features in the dataset, but after applying the feature selection technique, there were 19 features considered. The authors then explained why they selected SVM over other models such as logistic regression, K-Nearest Neighbours, Random Forests, and Decision Trees. Their final model gave an accuracy score of 85% in predicting attrition.

Priyanka Sadana and Divya Munnuru [4] conducted surveys with current employees and alumni of an IT company and then created a dataset with 1649 records. For exploratory data analysis, they determined relationships between the target variable and the predictor variables in the dataset. Their dataset was imbalanced and they solve that problem with the help of Synthetic Minority Oversampling Technique (SMOTE) was used to balance the data. Because of encoding categorical data, the dimensionality of the dataset increased. Authors thus used Principal Component Analysis (PCA) for feature reduction. Later they trained multiple models such as Logistic Regression and Random Forest. Authors focused more on the Recall metric for evaluating their models. Random Forest after hyperparameter tuning gave a Recall score of 93%.

Sarah S. Alduayj and Kashif Rajpoot [5] have proposed various machine learning techniques to predict employee attrition. The publicly available "IBM Watson Analytics dataset" was used. It contained 32 features and 1470 employee data with 1233 "no" and 237 "yes" attrition categories. They removed two features and converted non-numerical data to numeric. They tried three different techniques. After trying each approach, they applied 5-fold cross-validation and feature selection. First, they applied machine learning models directly to the imbalanced dataset. The maximum accuracy and F1-score were 87.1% and 0.503 respectively given by Quadratic SVM. Secondly, they used Adaptive Synthetic Sampling(ADASYN) approach to balance the dataset by oversampling the minority "yes" class. The best F1 score was 0.967, given by KNN(K=1). The best F1 Scores were 0.829, 0.806, and 0.909 for 2,3 and 12 features respectively by Random Forest. Thirdly, they manually undersampled the data with 237 observations of each class. The best F1 score was 0.74 by Quadratic SVM.In the end, they achieved an F1 score of 0.909 with Random Forest after selecting 12 out of 29 features.

Praphula et al. [6] tested different machine learning algorithms to predict employee attrition. The study was done to find different factors influencing the attrition rate and their possible solutions. The publicly available "HR management dataset" was used. It contained 10 features and 14000 records. All records were divided into 10 business departments. As a part of exploratory data analysis, authors performed variable identification, univariate and Bi-variate analysis, and correlation. The rows with missing values were removed. The dataset was split into training and testing sets. They tried three machine learning algorithms like Support Vector Machine with Radial Basis Function Kernel, Decision Tree, and Random Forest. These algorithms were applied to 4 departments with more than 1000 employees with evaluation parameters like Precision, Recall, and F1 Score. Random Forest Classifier performed the best with precision 0.99, 0.97, 0.98, and 0.99 for Sales, Technical, Support, and IT departments respectively.

Francesca Fallucchi et al. [7] proposed a Machine Learning approach for employee attrition using objective features. The authors have used the IBM HR dataset for their study. In data preprocessing, redundant features are removed, categorical features are label encoded, and numerical features are scaled using a standard scaler. They conducted a feature correlation analysis to determine feature importance as well as an extensive descriptive analysis, taking Attrition as a target feature. Multiple predictive models are used based on methods such as the Bayesian method, SVM, decision trees, and Logistic Regression. The training set contains 70% of the data, while a test and validation set contains the remaining 30%. Also, the authors used Holdout and Cross-validation for the evaluation of models. Amongst all models, Gaussian Naive Bayes gave the best F1 score of 0.446 and the lowest false positive rate of 4.5% along with the highest true positive rate of 72%.

Usha et al. [8] proposed various machine learning techniques to predict employee attrition. They used publicly available "IBM HR Analytics dataset". It contained 35 features and 1470 employee data with 1233 "no" and 237 "yes" attrition categories. They used the Weka tool to apply machine learning. The irrelevant and redundant columns are discarded. Weka tool is used to convert dataset CSV file to ARFF (Attribute-Relation File Format) file for its processing. They applied Decision Tree and Naive Bayes classifier giving an accuracy of 82.449% and 78.8435%, respectively on 10-fold cross-validation. Then, they applied K-Means and Expectation Maximization clustering algorithms giving a correct classification rate of 57.2789% and 55.102%, respectively.

Rachna Jain and Anand Nayyar [9] have proposed to use gradient boosting-based methods such as XGBoost and GLMBoost. The authors have performed a thorough exploratory analysis and feature engineering. IBM dataset was used for modeling. The correlation of attributes was analyzed to filter given features. Also, a class imbalance problem is not addressed. Finally XGBoost model gave an accuracy score of 89%

Marc Guirao [10] has used the IBM HR Analytics dataset to predict employee attrition with the help of machine learning. The author has extensively used KNIME open-source data analytics software. Dataset was split into 80% for training and 20% for testing. Being an imbalanced dataset, the minority class was upsampled using Synthetic Minority Oversampling Technique (SMOTE) algorithm duirng training. The author then trained four machine learning models such as Naive Bayes, Random Forest, Logistic regression, and Gradient Boosting. Finally, the best score was given by Random Forest on the imbalanced test set with an accuracy score of 0.89 and an F1-Score of 0.59.

To summarize the literature survey, the IBM HR dataset is severely imbalanced. To tackle it, some of the papers have oversampled the minority class. Most of the existing approaches are predicting employee attrition using the centralized approach with the help of classical machine learning algorithms such as Logistic Regression, Decision Trees, and Random forest.

3 HRescue Methodology

3.1 Dataset Description

IBM's data scientists have created this publicly available dataset named IBM HR Analytics [11]. The dataset contains 1470 records and 35 features where the deciding feature about employee attrition is the 'Attrition' column containing "Yes" and "No" labels. Feature datatypes are shown in table 1. The dataset has an issue of class imbalance because 1233 records are from the "No" category, and 237 are from the "Yes" category. There were no missing values in the entire dataset. All employees in the dataset are aged above 18. The dataset has details about employees' Education, marital status, previous work history, and more. EmployeeCount, Over18, and StandardHours are some features from the dataset that were discarded as they contained only one unique value. EmployeeNumber feature had unique values for all the records so we ignored it as it did not provide any meaningful information.

3.2 Dataset Balancing

Dataset is severely imbalanced, having only 237 positives and 1233 negative samples for the target class Attrition as shown in figure 2. Class Imbalance makes it difficult for a machine learning model to generalize better on given data. So to tackle this problem, we are using Adaptive Synthetic (ADASYN) sampling [12] to upsample the minority class. ADASYN creates synthetic points by using the density distribution of the minority class. Categorical features are encoded using OneHotEncoding for upsampling and then inversely transformed after upsampling, to retain the original form of the dataset. After upsampling, our balanced dataset has 1160 Attrition and 1233 Non-Attrition samples.

3.3 Predictive Models

Gradient Boosting [13] is the technique to combine the power of many small learning models (learners). Generally, decision trees are considered learners for this technique. All the trees are connected sequentially with each subsequent decision tree learning from the errors made by the previous decision tree. While making the final prediction, every decision tree contributes. Learning rate ($alpha$) and $n\_estimators$ (number of trees) are important hyperparameters in the gradient boosting algorithm. Learning rate defines the rate at which the model learns, a model trained with a low learning rate makes it better but requires more computations. Using a high number of trees can make the model overfit.

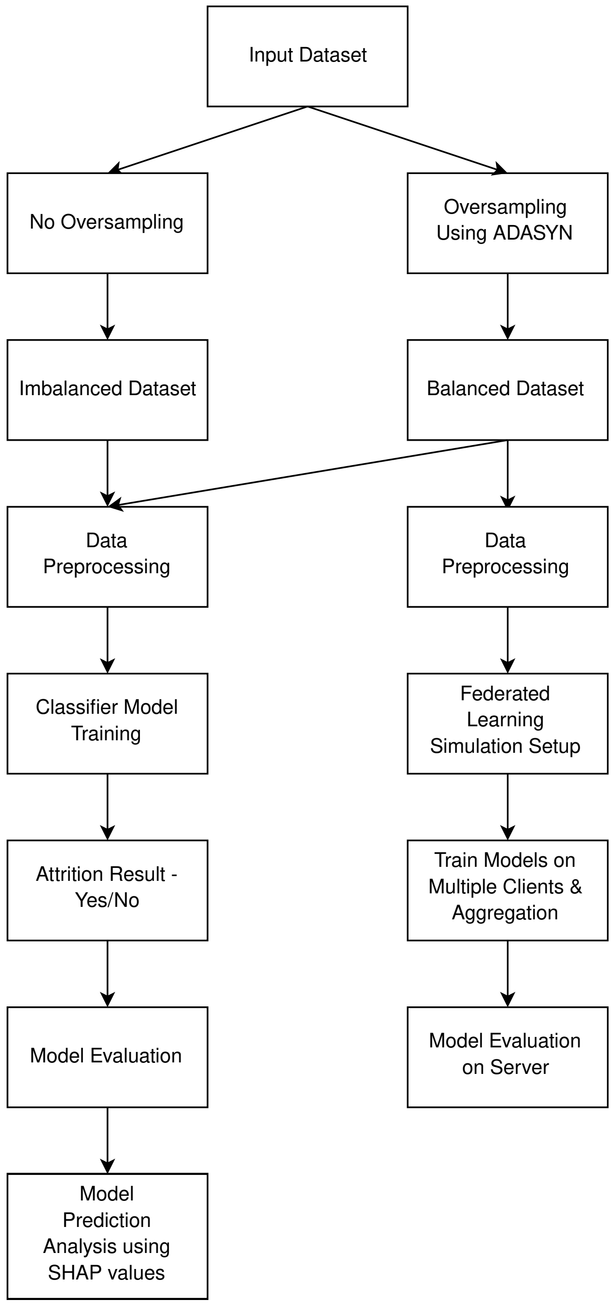

Fig. 1. HRescue Methodology Flow Diagram

Table 1. IBM HR Dataset Description

Feature	#**Type**	**Feature**	#**Type**
Age	N	MonthlyIncome	N
BusinessTravel	C	MonthlyRate	N
Daily Rate	N	YearsWithCurrentManager	N
Department	C	YearsSinceLastPromotion	C
DistanceFromHome	N	YearsInCurrentRole	C
Education	C	YearsAtCompany	N
EducationField	C	WorkLifeBalance	C
EmployeeCount	N	TrainingTimesLastYear	N
EmployeeNumber	N	TotalWorkingYears	N
EnvironmentSatisfaction	C	StockOptionLevel	C
Gender	C	StandardHours	N
HourlyRate	N	RelationshipSatisfaction	C
JobInvolvement	C	PerformanceRating	N
JobLevel	C	PercentSalaryHike	N
MaritalStatus	C	OverTime	C
JobRole	C	Over18	C
JobSatisfaction	C	NumCompaniesWorked	N
Attrition*	**C**		

* Target Variable

(N = Numerical, C = Categorical)

XGBoost : It is an implementation of gradient boosting created by Tianqi Chen [14]. It is a Stochastic Gradient Boosting with both L1 and L2 regularization. It automatically handles missing values and supports parallel tree construction. The data science community widely uses XGBoost because of its High Execution speed and Model performance.

CatBoost : It is a gradient-boosting tree-based model with out of box support for categorical features. Because of this, there is no need to perform any preprocessing on categorical features. The algorithm converts the categorical features into numerical values with the help of target statistics. It is designed by Yandex [15] and has applications in the fields of self-driving cars, weather prediction, and more. The CatBoost algorithm has performed better than other state-of-the-art gradient boosting algorithms such as XGBoost, LightGBM on Epsilon, Amazon, and other standard machine learning datasets.

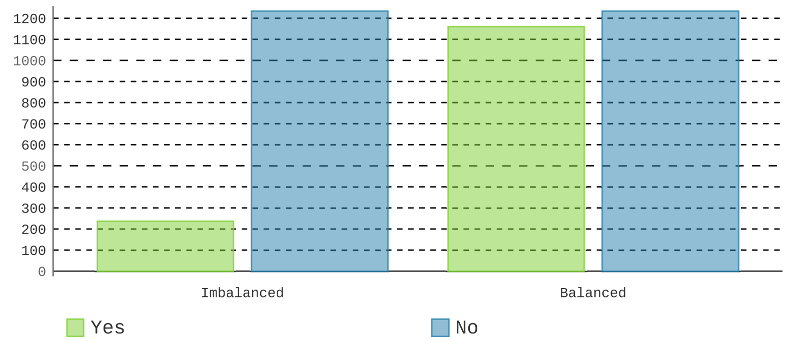

Fig. 2. Balanced Vs Imbalanced Dataset Sample Counts

3.4 Model Training

We have conducted three experiments in our methodology, one on the imbalanced dataset and the other two on the balanced dataset, sample count comparison for both datasets in given in figure 2. In all the experiments, 70% of the dataset is used for training, remainder 30% is for validation.

Imbalanced Dataset : Due to the class imbalance problem, it was difficult for the models to classify the minority class samples. To tackle this issue, we added extra weight in the Logloss (Binary Crossentropy) loss function when it makes an error while predicting the minority class sample. With categorical features in the dataset and XGBoost supporting only numerical data. These features were converted to numerical format using the LabelEncoder Class from the sklearn's preprocessing module while training with XGBoost. In the case of CatBoost, no preprocessing is performed on such categorical data. To avoid overfitting, we also used a high value of L2 Regularization in our models.

Balanced Dataset : A balanced dataset does not contain a class Imbalanced problem, so there are no extra weights used in Logloss loss function. The same preprocessing as imbalanced dataset is used.

3.5 Federated Learning Experiment

Federated Machine Learning [16] is a Distributed ML approach addressing data privacy issues. In this approach client data never leaves the client device, rather an ML model is trained on the edge device i.e on the client device. The model parameters of the trained model are then securely transferred over the network. Model parameters from multiple models of various clients are aggregated using algorithms like FedAverage [16] to create a robust global model.

Having a global system to predict employee attrition is challenging because employee data is sensitive to corporations. This can be addressed by using the Federated Machine Learning approach. Using this method, corporations can train local models on their data. Then, aggregate all these models to make a global model. This would enable utilizing a large amount of employee data from different corporations.

Dataset : For this experiment, we are using the same size samples of the IBM dataset. Before sampling, the dataset is balanced using ADASYN, and Principle Component Analysis (PCA) is performed on the data. In a real-world implementation, each client would have their own data.

Simulation Setup : For the simulation, we are using five clients each having an equal amount of data. After each round of training, model parameters are sent over the network to the server and, then the server aggregates them using FedAverage. This new model is then evaluated on test data and distributed to clients for the next iteration. We are using Logistic Regression and a simple Neural Network for this experiment.

Logistic Regression : This is a simple linear model used to classify linearly separable data. In our experiment, sklearn implementation is used and internal parameters such as coefficient and intercepts are transferred over the network for aggregation. The global model achieved an accuracy of 87% on the validation data.

Neural Network : The network consists of two hidden layers with 16 and 8 neurons, respectively, that use Relu as an activation function. An output layer consists of a single neuron with a Sigmoid as an activation function. This network has 641 trainable parameters, which are transferred over the network. Models are trained for 50 iterations. The final global model achieved an accuracy of 89% on the validation set.

3.6 Evaluation Metrics

We have used the Logloss function for our classification problem in our research study. As shown in equation (1), the Logloss value is calculated as the negative summation of the actual label (y_i) times log of the probability of class 1 (p_i) and (1 - actual label (y_i)) times log of the probability of class 0 ($1 - pi$), divided by the total number of samples (N).

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \quad (1)$$

Because of the imbalanced nature of the dataset, Accuracy does not provide information about how accurate the model is for predicting the minority class samples. Hence we have used the F1 Score to evaluate our models. This metric as shown in equation (2) is harmonic mean of Precision and Recall.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2}$$

4 Model Prediction Analysis

In any business, the reason for prediction by a model is equally important as the model predicting accurately. The Human Resources (HR) department of the company can take appropriate policies or decisions if they know why a particular employee has left. For this, we have computed SHAP (Shapley Additive Explanations) [18] values of the CatBoost model trained on the original imbalanced dataset. SHAP values are used to explain why a model has predicted for the given sample of features in a certain way. From figure 3, the features displayed on the left of the summary plot are in the decreasing order of feature importance. OverTime is the most important feature for prediction while PerformanceRating is the least important. From the right side of the plot, we can see that as the value of a particular features increases it is marked in red or blue if it decreases. From the OverTime feature, we conclude that when the value of the OverTime feature is high (1) it makes the predictions towards Attrition Class and Non-Attrition class if it is low (0). Similarly, we can derive how each feature impacts model prediction. It can help the respective HR department determine the common factors causing attrition. Calculating SHAP values of a single example can find reasons driving employees to leave the firm at an individual level. It would even be more beneficial to the HR department. In figure 4, it can be seen which features may have caused for an employee to continue or leave the job.

5 Experimental Setup

We performed our proposed methodology extensively on Google Colaboratory. Our environment used Python (v3.7.12), scikit-learn (v1.0.2), CatBoost (v1.0.4), XGBoost (v0.90). Both CatBoost and XGBoost models have many hyperparameters to tune. To efficiently get the best results, we have used popular hyperparameter optimization frameworks Optuna (v2.1.0) and Hyperopt (v1.2.7). While training, the model that gave the best performance on the evaluation dataset is selected in all the experiments. For simulations of Federated Learning, experiments are carried out on a local machine having Python (v3.8.0) and Flower federated framework [17] (v0.17.0)

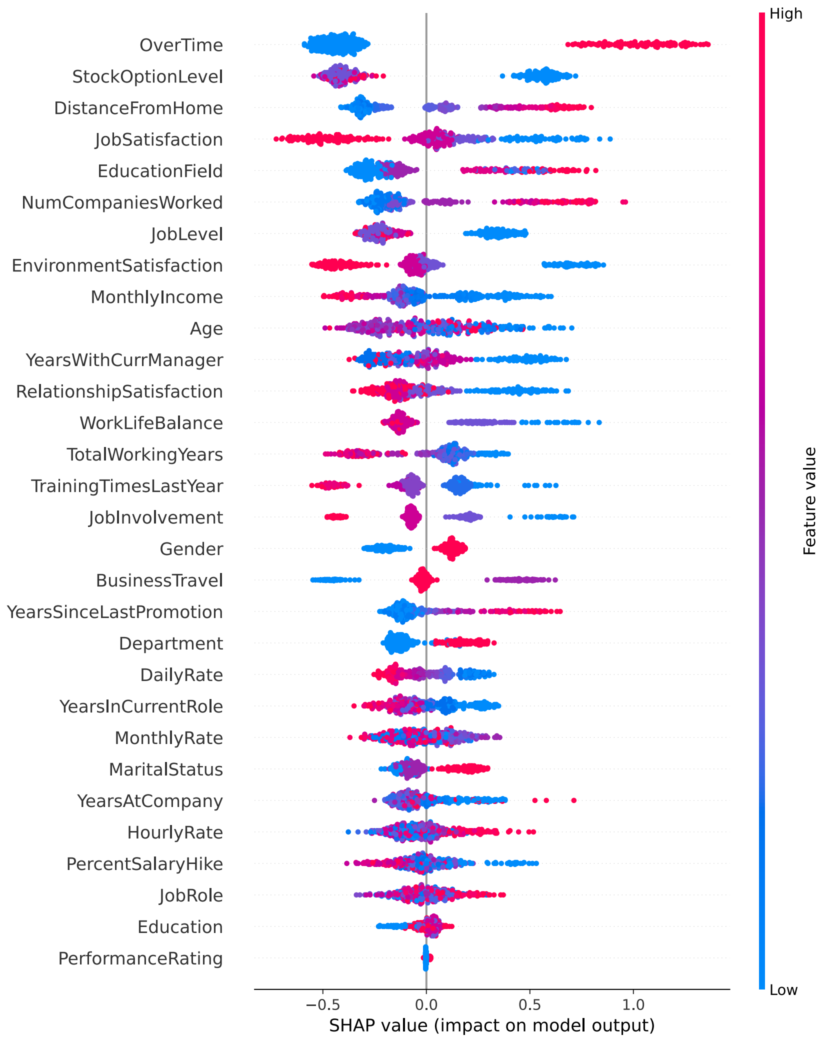

Fig. 3. Shap Values Summary Plot Imbalanced Dataset using CatBoost Model

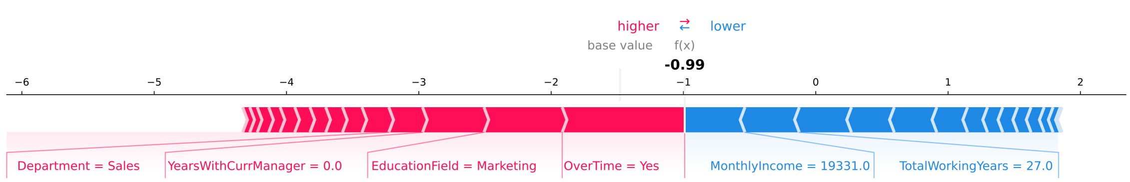

Fig. 4. Shap Values of Individual Test Sample

Table 2. Results and Comparison of F1-Score and Accuracy on Imbalanced Dataset

Paper	Model	Accuracy	F1-Score
[5]	Linear SVM	0.87	0.37
	Quadratic SVM	0.87	0.5
	Cubic SVM	0.84	0.46
	Gaussian SVM	0.87	0.34
	Random Forest	0.86	0.27
	KNN(K=1)	0.83	0.08
	KNN(K=3)	0.84	0.01
[7]	Gaussian NB	0.83	0.45
	Bernoulli NB	0.85	0.38
	Logistic Regression	0.88	0.45
	KNN	0.85	0.15
	Random Forest	0.86	0.19
	Decision Tree	0.82	0.35
	SVM	0.86	0.17
	Linear SVM	0.88	0.36
[2]	Logistic Regression	0.78	0.53
	Logistic Regression (feature selection)	0.81	0.56
[8]	Decision Tree	0.83	-
	Naive Bayes	0.81	-
[9]	XGBoost	0.89	0.60
[10]	Logistic Regression	0.79	0.52
	Naive Bayes	0.83	0.49
	Gradient Boosting	0.88	0.58
	Random Forest	0.89	0.59
Ours	**Catboost**	**0.90**	**0.69**
	XGboost	**0.90**	**0.68**

6 Results

In this section, we have discussed the performances of our models with previous work on both the balanced as well as the imbalanced dataset. We have compared the accuracy and F1 Score measures with the existing approaches. In both versions of the dataset, CatBoost performed slightly better than XGBoost. From table 2, we can say that our models have outperformed by a margin of 15% in the F1 Score on the imbalanced dataset. As shown in table 3, on the balanced dataset CatBoost gave an F1 Score of 0.94 and XGBoost gave 0.93 respectively. Authors of [5] got an F1 Score of 0.97 with KNN (K=1) but they have mentioned that their model was overfitted.

Table 3. Results and Comparison of F1-Score and Accuracy on Balanced Dataset

Paper	Model	F1-Score	Accuracy
[5]	Linear SVM	0.78	0.78
	Quadratic SVM	0.88	0.88
	Cubic SVM	0.93	0.93
	Gaussian SVM	0.91	0.91
	#KNN(K=1)	0.97	0.97
	KNN(K=3)	0.93	0.93
	Random Forest	0.93	0.92
Ours	**Catboost**	**0.94**	**0.94**
	XGboost	**0.93**	**0.93**

[5] have mentioned that their model was overfitted.

7 Conclusion and Future Scope

Employee attrition is one of the biggest obstacle for companies. We have performed 3 experiments as a part of our methodology. Firstly, we have trained our Catboost and XGBoost classifier models on an imbalanced dataset. The results obtained are an F1-score of 0.69 and an accuracy of 90%. Secondly, we trained our XGBoost and CatBoost classifier models on the balanced dataset. The results obtained are an F1-score of 0.94 and an accuracy of 94%. Finally, we applied the federated learning technique to the balanced dataset. We hope that this solution creates a win-win situation for employers and employees. As a part of future work, we would like to create a more robust model trained on a large, realistic, and balanced dataset. We can add new features to the dataset after taking feedback from employees who left.

References

1. Bamboohr Blog, https://www.bamboohr.com/blog/onboarding-infographic/. Last accessed 7 Feb 2022
2. S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection," Mathematics, vol. 9, no. 11, p. 1226, May 2021, https://doi.org/10.3390/math9111226.

3. Khera, Shikha N. “Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques.” Vision 23, no. 1 (March 2019): 12–21. https://doi.org/10.1177/0972262918821221.
4. P. Sadana and D. Munnuru, "Machine Learning Model to Predict Work Force Attrition," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-6, https://doi.org/10.1109/I2CT51068.2021.9418140.
5. S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), 2018, pp. 93-98, https://doi.org/10.1109/INNOVATIONS.2018.8605976.
6. Jain, P.K., Jain, M. Pamula, R. Explaining and predicting employees' attrition: a machine learning approach. SN Appl. Sci. 2, 757 (2020). https://doi.org/10.1007/s42452-020-2519-4
7. Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. Computers 2020, 9, 86. https://doi.org/10.3390/computers9040086
8. Usha, P.M.; Balaji, N. Analysing employee attrition using machine learning. Karpagam J. Comput. Sci. 2019, 13, 277–282.
9. R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," 2018 International Conference on System Modeling Advancement in Research Trends (SMART), 2018, pp. 113-120, doi: https://doi.org/10.1109/SYSMART.2018.8746940.
10. Predicting Employee Attrition with Machine Learning, Knime Blog, https://www.knime.com/blog/predicting-employee-attrition-with-machine-learning. Last accessed 14 Mar 2022
11. IBM HR Dataset, Liu: Attrition, (2020). https://doi.org/10.5281/zenodo.4323396
12. Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.
13. Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics data analysis, 38(4), 367-378.
14. Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785
15. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.
16. McMahan, B., Moore, E., Ramage, D., Hampson, S. amp; Arcas, B.A.y.. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 54:1273-1282 Available from https://proceedings.mlr.press/v54/mcmahan17a.html.
17. Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P. P., Lane, N. D. (2020). Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390.
18. Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.