Employee Attrition Prediction Using Machine Learning

Aastha 2019224 | Aditi Sejal 2019228 | Samad Shahid 2019446 | Vaibhav Soni 2019454

Abstract

Predicting employee attrition can help organizations take the necessary steps to retain talent well within time. In this paper, several classification models, namely Logistic Regression, Naive Bayes, Decision Tree, Random Forest, AdaBoost, Support Vector Machine, Linear Discriminant Analysis, Multilayer Perceptron and K-Nearest Neighbors have been trained and tested on the IBM HR Dataset. Oversampled data with PCA had the best performances on which Random Forest, AdaBoost, SVM, and MLP achieved accuracy and F1 score above 90%. Based on our analysis, attrition rates were higher in younger employees, doing overtime, having lower monthly incomes and working for a shorter period of time.

1. Introduction

Employee attrition refers to an employee's voluntary or involuntary resignation from a workforce. Organizations spend many resources in hiring talented employees and training them. Every employee is critical to a company's success. Our goal is to predict employee attrition and identify the factors contributing to an employee leaving a workforce. We discuss various classification models on our dataset and assess their performance using different metrics such as accuracy, precision, recall and F1 score. We also analyze the dataset to identify key factors contributing to an employee leaving a workforce. Our project will assist organizations in gaining fresh insights into what drives attrition and thus enhance retention rate.

2. Literature Survey

There have been many machine learning works on employee attrition. Some of them are discussed in Table 1. Taking inspiration from the work done, we apply combinations of some of these models and techniques on our dataset and analyze them apart from training the base supervised models and testing on different evaluation metrics.

3. Dataset

3.1. Dataset Review

We used is the IBM Employee Attrition dataset from Kaggle. It contains 35 columns and 1470 rows and has a mix of numerical and categorical features. A sample row is shown in Fig. 1.


```
{'Age': 44,
 'BusinessTravel': 0,
 'DailyRate': 489,
 'Department': 1,
 'DistanceFromHome': 23,
 'Education': 3,
 'EducationField': 3,
 'EnvironmentSatisfaction': 2,
 'Gender': 1,
 'HourlyRate': 67,
 'JobInvolvement': 3,
 'JobLevel': 2,
 'JobRole': 2,
 'JobSatisfaction': 2,
 'MaritalStatus': 1,
 'MonthlyIncome': 2042,
 'MonthlyRate': 25043,
 'NumCompaniesWorked': 4,
 'OverTime': 0,
 'PercentSalaryHike': 12,
 'PerformanceRating': 3,
 'RelationshipSatisfaction': 3,
 'StockOptionLevel': 1,
 'TotalWorkingYears': 17,
 'TrainingTimesLastYear': 3,
 'WorkLifeBalance': 4,
 'YearsAtCompany': 3,
 'YearsInCurrentRole': 2,
 'YearsSinceLastPromotion': 1,
 'YearsWithCurrManager': 2}
```


Figure 1. Sample Employee Details

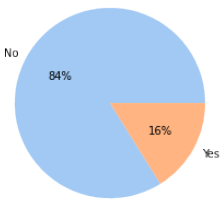

Figure 2. Employee Attrition Distribution

The target value Attrition is a categorical variable with

Paper and Dataset	Methodology	Results
USA bank data and IBM Watson Analytics HR dataset, Paper [**4**]	Analysed performance of DT, Random Forest, XGBoost, LR, SVM, Neural networks, Naive Bayes, LDA, KNN on datasets of small, medium, large employee population sizes across varied metrics	Tree-based algorithms performed better, XGBoost recommended along with trying different models and picking classifier which best fits data; Feature importance and rule sets are important for model interpretability
Global retailer's HRIS database, BLS (Bureau of Labor Statistics) data, Paper [**2**]	Trained and tested XGBoost, LR, Naive Bayes, Random Forest, SVM, LDA, KNN models on ROC-AUC metric	TXGBoost classifier is a superior algorithm in terms of significantly higher accuracy, relatively low runtimes and efficient memory utilization for predicting turnover
309 records from a Higher Institution in Nigeria between 1978-2006, Paper [1]	Data mining and classification tools (WEKA and See5) were used to generate classifiers (C4.5, CART, REPTree decision tree algorithms, boosted trees); if-then rule sets for an attrition prediction model were developed	The Boosted SeeTree performed best with accuracy 0.74. Identified employee salary and length of service as key factors in the attrition decision based on their high usage percentage across high-performing models
IBM HR Dataset, Paper [**3**]	SVM, Random Forest and KNN models trained and tested on original as well as balanced datasets (using ADASYN, undersampling, feature selection)	Improved performance for ADASYN and feature selected datasets (F1 scores between 0.90 - 0.93). Poor performance in undersampling (0.7 F1 score for SVM) due to important information being lost
IBM HR Dataset, Paper [**5**]	Analysis of base models (Decision Trees, Logistic Regression) and ensemble models (Random Forest, Adaboost, heterogeneous combinations of base models) on different datasets made using PCA, feature selection	Optimum results for the PCA algorithm dataset. Best base model: LR, best ensemble model: DT + LR (accuracy = 86.39%). Ensemble models will be more generalizable as compared to their base counterparts

Table 1. Related Work

the values "Yes" and "No." The dataset is highly imbalanced and contains significantly more examples for "No" than "Yes" as shown in Fig. 2.

3.2. Exploratory Data Analysis

Distribution graphs for features were analyzed. Some inferences are discussed below.

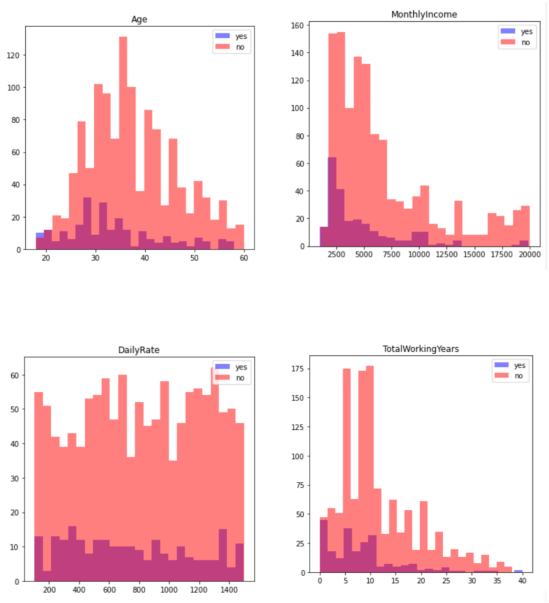

Figure 3. Graphs for Exploratory Data Analysis

From Fig. 3, we see that employees around 28 years of age seem more likely to leave the company. Low monthly income was associated with higher attrition rates. While attrition rates were higher among employees working for less than ten years, newer employees showed the highest attrition. Employees are more likely to leave if they work overtime. Attrition is more for employees who travel frequently. Sales executives are more likely to leave the company compared to other roles. No significant distinction in attrition based on gender was observed.

The correlation graph in Fig. 4 shows that:

1. Recent salary hikes are highly correlated with performance.
2. MonthlyIncome and JobLevel tend to be higher for employees who work longer hours.
3. YearsAtCompany, YearsWithCurrManager and YearsInCurrentRole are highly correlated, highlighting stagnant professional growth in the company.

3.3. Preprocessing

There are no missing/null values in the dataset. To visualize the distribution of different features, we plot bar graphs. Using these, we observe that the features 'EmployeeCount', 'Over18', and 'StandardHours' have only one

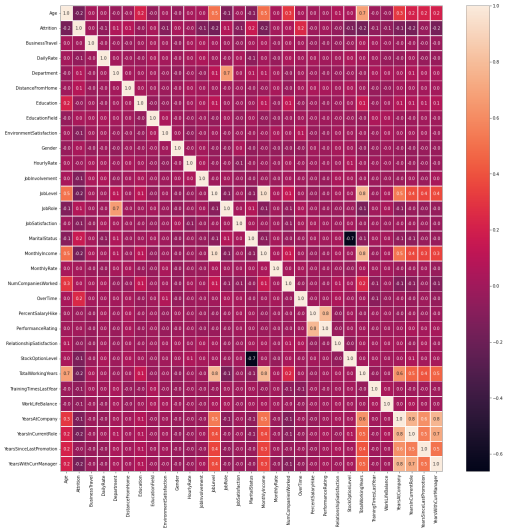

Figure 4. Correlation Matrix

unique value and hence add no value to attrition prediction. Thus, they are dropped. Employee number is varying for each row and is not related to the attrition column and is also dropped.

3.3.1 Encoding Categorical Columns

The values in the categorical columns are converted into numerical values using label encoding. Label encoding is the process of turning each of the n values in a column into a number ranging from 0 to n-1. This is done for BusinessTravel, Department, Gender, JobRole, MaritalStatus, OverTime, EducationField.

3.3.2 OverSampling and Undersampling

Random oversampling and undersampling was performed to handle class imbalance. Oversampling involves creating copies of data from the minority (No) class, while undersampling involves deleting data from the majority (Yes) class.

3.3.3 Feature scaling

After this, the data is standardized and normalized. Standardization is the process of scaling the features to have 0 mean and 1 variance, like in normal distribution. It is crucial not just for comparing measurements with various units, but it is also a basic criterion for many machine learning methods like Logistic regression. Normalization is the other approach for feature scaling. In this method, the data is scaled to a defined range - generally 0 to 1. This restricted range results in lower standard deviations, which reduces the influence of outliers.

3.3.4 Principal Component Analysis (PCA)

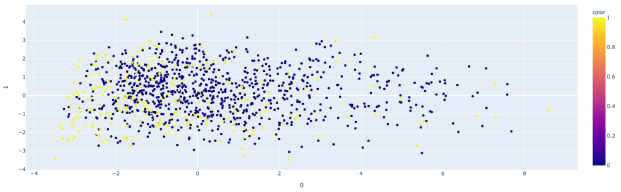

Figure 5. PCA Visualization

PCA is a dimensionality reduction technique which couples features based on relationships between them, Fig. 5. It improves interpretability while reducing information loss. As per Fig. 6, two principal components were retained with a total explained variance of 99.77%.

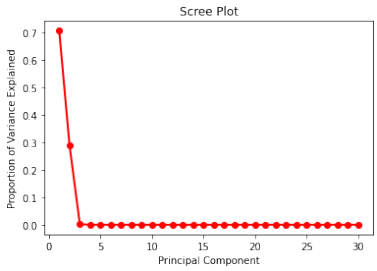

Figure 6. Scree Plot

4. Methodology

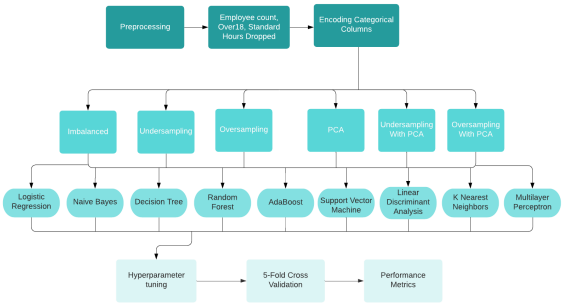

Figure 7. Project Methodology

We trained and evaluated nine supervised machine learning classification models. Simple supervised models like Logistic Regression (predicts binary outputs), Naive Bayes (maximises conditional probabilities for outputs), Decision Tree (branches on different feature values using entropy/information gain), Random Forest (ensemble of decision trees), Adaboost (adaptive boosting ensemble of

trees), Support Vector Machine (defines hyperplanes based on support vectors), Linear Discriminant Analysis (estimates probabilities using data statistics), Multilayer Perceptron (fully connected neural network) and K-Nearest Neighbors (minimises distance between points in k groups). We trained our models on six different datasets: imbalanced, undersampled, oversampled, PCA, undersampled with PCA and oversampled with PCA and evaluated their performance. Further, to get the best performance, hyperparameter tuning was carried out using RandomSearchCV and GridSearchCV. K-fold cross-validation with 5 folds was also performed on the training set. To handle model interpretability, appropriate graphs and figures were used. As suggested in [**4**] accuracy for the attrition decision is a biased metric and hence we evaluated the model on all the following classification metrics: accuracy, precision, recall and F1 score.

5. Results and Analysis

The results of various classification metrics for all models and imbalanced/balanced data are summarised in the tables 2, 3 and 4.

Model	Acc	Prec	Rec	F1
LR	**0.875**	0.753	0.346	**0.472**
NB	0.787	0.394	**0.586**	0.471
DT	0.836	0.498	0.279	0.351
RF	0.862	0.841	0.181	0.296
AdaBoost	0.858	0.769	0.181	0.293
SVM	0.867	0.741	0.283	0.406
LDA	0.867	0.683	0.325	0.439
KNN	0.848	**0.771**	0.089	0.157
MLP	0.866	0.677	0.338	0.447

Table 2. Classification Scores for Imbalanced Data

Model	Acc	Prec	Rec	F1
LR	0.706	0.714	0.733	0.723
NB	0.653	0.624	**0.776**	0.691
DT	0.664	0.673	0.645	0.655
RF	**0.724**	0.726	0.726	**0.724**
AdaBoost	0.722	0.730	0.708	0.718
SVM	0.719	0.729	0.696	0.712
LDA	0.550	0.539	0.696	0.607
KNN	0.704	**0.734**	0.646	0.686
MLP	0.715	0.720	0.705	0.712

Table 3. Classification Scores for Undersampled Data with PCA

Model	Acc	Prec	Rec	F1
LR	0.758	0.751	0.771	0.761
NB	0.680	0.649	0.782	0.709
DT	0.841	0.796	0.917	0.852
RF	**0.992**	**0.986**	**0.998**	**0.992**
AdaBoost	0.989	0.982	**0.998**	0.990
SVM	0.951	0.912	**0.998**	0.953
LDA	0.609	0.610	0.736	0.667
KNN	0.894	0.826	**0.998**	0.904
MLP	0.946	0.905	**0.998**	0.949

Table 4. Classification Scores for Oversampled Data with PCA

The logistic regression model performed best for imbalanced data with an accuracy of 87.5%. For undersampled data with PCA, Random Forest model had best metric values with 72.4% accuracy and F1 score and 72.6% precision and recall. In the case of oversampled data with PCA, tree based models performed best out of which Random Forest had the highest accuracy and F1 score of 99.2%, precision of 98.6%. As expected, the tree based models performed well as they are known to work with non linear data. They can make more complex decision boundaries that fit very well on non-linear data. Decision Tree was able to achieve an accuracy score of 84% and recall of 91%. We also tried other complex models such as the SVC and MLP. SVC with a non linear kernel 'rbf' and MLP also performed great on the testing data.

PCA, over and under sampling to balance data were also performed separately for these models. Among these, highest metric scores were observed for oversampled data while there was some improvement in performance for undersampled data. No considerable improvement in performance of models was obtained due to PCA alone.

Overall, all models performed better for oversampled data with PCA as compared to imbalanced data. The exceptions were LR and NB. Logistic regression didn't perform well as it assumes that the data is linearly separable which was not the case as was seen in the EDA. Naive Bayes also didn't perform well as many of the features are not conditionally independent such as the job role and the monthly income, education and job level as well as daily rate, hourly rate etc. This may also be because these classifiers were predicting the majority class most of the time and due to the imbalanced data scored high accuracies which was no longer the case for oversampled data.

There was no improvement in accuracy for any model for undersampling with PCA. Higher precision, recall and

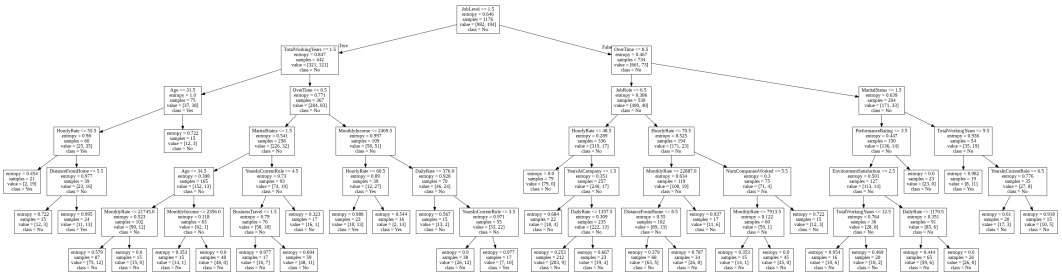

Figure 8. Decision Tree with tuned hyperparameters

F1 scores were obtained for some due to the balancing. This is because undersampling caused downsizing of data in the majority to around 16% from the earlier 84% leading to loss of valuable information on the way as proposed by [**5**]. The unsupervised model KNN had good metric scores yet there were many supervised models like RF, AdaBoost, SVM which performed better.

We also realize that in a real-world scenario, the data will inherently be imbalanced as employees leaving a workforce will generally be fewer than those staying in the organisation. Thus the above methods and results provide a good starting point for attrition prediction. Detailed, model-wise analysis is below.

5.1. Logistic Regression (LR)

The best performing logistic regression model was for oversampling with PCA; the hyperparamters obtained were: C as 0.1, penalty as l2 and solver as liblinear. It had higher accuracy for standardized data.

5.2. Naive Bayes (NB)

The Gaussian Naive Bayes achieved the best recall with imbalanced and undersampled data, 58.6% and 77.6% respectively. There was an increase in precision, recall and F1 scores in oversampled and undersampled data with PCA but a decrease in the accuracy.

5.3. Decision Tree (DT)

The decision Tree model trained on 30 features and unscaled data as shown in Fig. 8 had the following tuned parameters: criterion as gini, maximum depth as 13, maximum features as one-third of total features, the maximum number of leaf nodes as 100 and the minimum number of samples in leaf as 1. According to this tree, OverTime, JobLevel, HourlyRate, TotalWorkingYears, MaritalStatus, MonthlyIncome and Age had higher importance. The lack of stability of decision tress was responsible for lower accuracy, precision, recall, F1 score than other tree based counterparts.

5.4. Random Forest (RF)

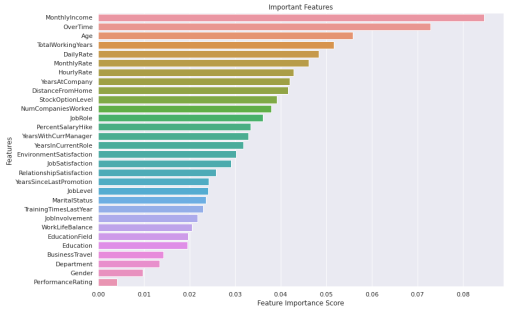

Figure 9. Feature Importance w.r.t. Random Forest with Oversampling

The best performance was obtained by setting the hyperparameters Bootstrap to False, max depth to 100, min samples leaf to 1, min sample required to split to 3 and the total number of decision trees in the random forest estimator to 250. From Fig. 9, we observe that the most important features were Monthly-Income followed by OverTime and Age, while the least important features were Performance Rating, Gender and BusinessTravel. This ensemble model offered stability, lower bias and variance and thus had the best performance.

5.5. AdaBoost

An improvement was seen compared to the decision tree results and the model achieved the best recall or 99.8% using under-sampled data with PCA. The best hyperparameters were the learning rate set to 1.0 and n_estimators set to 1000. It is also the second-best performing model with high accuracy, precision and F1 score values.

5.6. Support Vector Machine (SVM)

The best performance was obtained by setting the hyperparameter 'C' to 100 and kernel to 'rbf'. Its ROC-AUC curve is in Fig. 10. As expected, model trained on oversampled data performed the best and the model trained on the imbalanced data performed the worst.

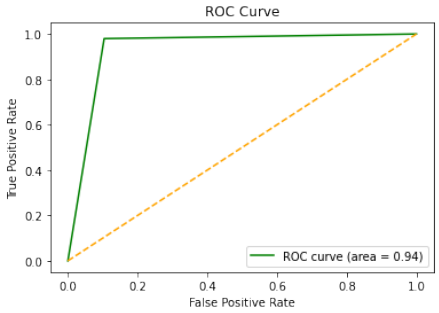

Figure 10. ROC-AUC Curve for SVC

5.7. Linear Discriminant Analysis (LDA)

The best performing LDA model was for oversampled data with PCA with hyper-parameters: shrinkage as auto and solver as 'lsqr'. LDA had higher accuracy and precision for imbalanced data but higher recall and F1 scores were observed for balanced data.

5.8. K-Nearest Neighbors (KNN)

KNN model performed best for oversampling with PCA and had hyperparameters: leaf size as 1, number of neighbors as 17 and weights as distance. It had a lower accuracy and precision for imbalanced data than undersampled data with PCA. This was the only unsupervised model and had the highest recall as well as good accuracy, precision, F1 Score.

5.9. MultiLayer Perceptron (MLP)

With hyperparameters: logistic activation function, alpha as 0.05 and lbfs solver, the best performance with highest metrics was obtained for oversampled data with PCA followed by imbalanced and undersampled data.

6. Conclusion

We trained various supervised classification models (LR, NB, DT, RF, AdaBoost, SVM, LDA, MLP and KNN) and summarised their results in this project. As observed from EDA and our previous analysis, each model performed significantly worse on the unprocessed dataset, due to its imbalanced nature. The best performance was obtained in Random Forest Model with PCA and Oversampling with accuracy of 99.2%, precision of 98.6%, recall of 99.8% and f1 score of 99.2%. Other models such as SVC and MLP also performed equally well with accuracies and F1 scores consistently more than 90%. Oversampling with PCA had better performances across models except LR and NB with tree based models having highest metric scores. In accordance to EDA, MonthlyIncome, Age, OverTime, TotalWorkingYears played major roles in the attrition decision and Gender did not impact attrition.

6.1. Learnings

We learnt Machine Learning in practice. We followed the ML pipeline starting from preprocessing and EDA on our dataset to get insights into the data, followed by training our models, hyper-parameter tuning, testing and cross-validation. Observing real-world intricacies in our project like imbalanced datasets, using combinations of techniques like undersampling, oversampling, PCA helped analyse the performance of different models. We learnt how to analyze and interpret models. Working on a dataset with a large number of features in a team was also a great learning experience.

6.2. Contribution

1. **Aastha** : Feature Scaling, Logistic Regression, SVM, MLP, KNN, Under/Over Sampling, Report writing
2. **Aditi** : Literature Review, Decision Tree, LDA, PCA, Report writing
3. **Samad** : EDA, Random Forest, AdaBoost
4. **Vaibhav** : Preprocessing, EDA, Naive Bayes, Sampling with PCA, Report Writing

References

[1] D. Alao and A. B. Adeyemo. "Analyzing Employee Attrition Using Decision Tree Algorithms". In: 2013.

[2] Rohit Punnoose and Pankaj Ajit. "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms". In: *International Journal of Advanced Research in Artificial Intelligence* 5.9 (2016). DOI: 10.14569/ijarai.2016.050904.

[3] Sarah S. Alduayj and K. Rajpoot. "Predicting Employee Attrition using Machine Learning". In: *2018 International Conference on Innovations in Information Technology (IIT)* (2018), pp. 93–98.

[4] Yue Zhao et al. "Employee Turnover Prediction with Machine Learning: A Reliable Approach". In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Nov. 2018, pp. 737–758. DOI: 10.1007/978-3-030-01057-7_56.

[5] Aseel Qutub et al. "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods". In: *International Journal of Machine Learning and Computing* 11 (2021), pp. 110–114.