Employee Attrition Prediction Using Machine Learning

Dhruvin Dhaduk

Abstract

Predicting employee attrition can help organizations take the necessary steps to retain talent well within time. In this paper, several classification models, namely Logis tic Regression, Naive Bayes, Decision Tree, Random Forest, AdaBoost, Support Vector Machine, Linear Discriminant Analysis, Multilayer Perceptron and K-Nearest Neighbour's have been trained and tested on the dataset. Oversampled data with PCA had the best performances on which Random Forest, AdaBoost, SVM, and MLP achieved accuracy and F1 score above 95%. 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 analyse the dataset to identify key factors contributing to an employee leaving a workforce. Our project will as sist 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 analyse them apart from training the base supervised models and testing on different evaluation metrics

3. Dataset

3.1. Dataset Review

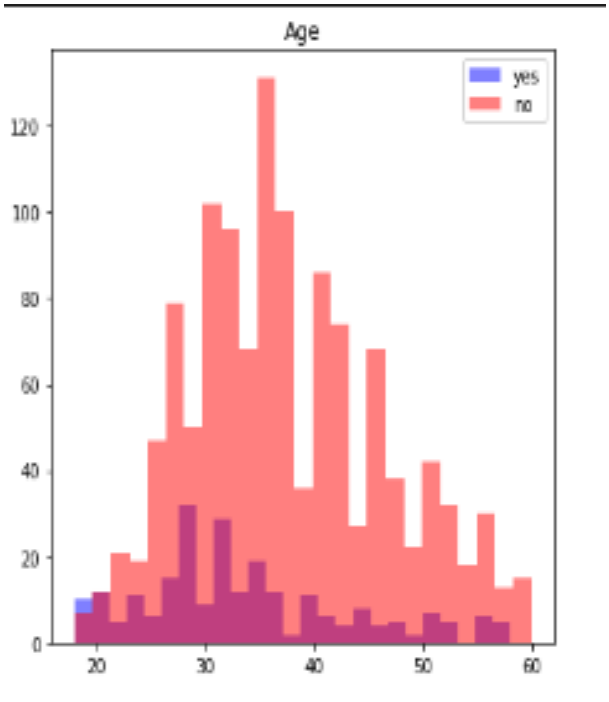
We used the IBM Employee Attrition dataset from Kaggle. It contains 26 columns and 445 rows and has a mix of numerical and categorical features. A sample row is shown in Fig. 1.

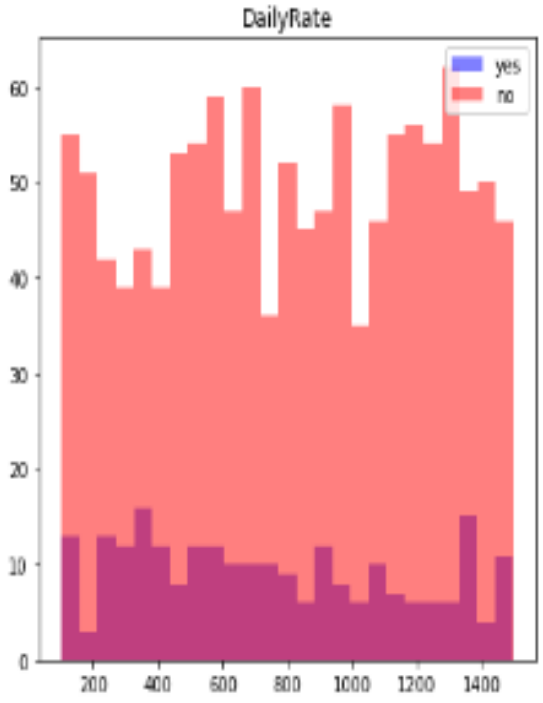
	Id	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeNumber	EnvironmentSatisfa
0	1	30	0	Non-Travel	Research & Development	2	3	Medical	571	
1	2	36	0	Travel_Rarely	Research & Development	12	4	Life Sciences	1614	
2	3	55	1	Travel_Rarely	Sales	2	1	Medical	842	
3	4	39	0	Travel_Rarely	Research & Development	24	1	Life Sciences	2014	
4	5	37	0	Travel_Rarely	Research & Development	3	3	Other	689	

5 rows × 29 columns

3.2. Exploratory Data Analysis

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

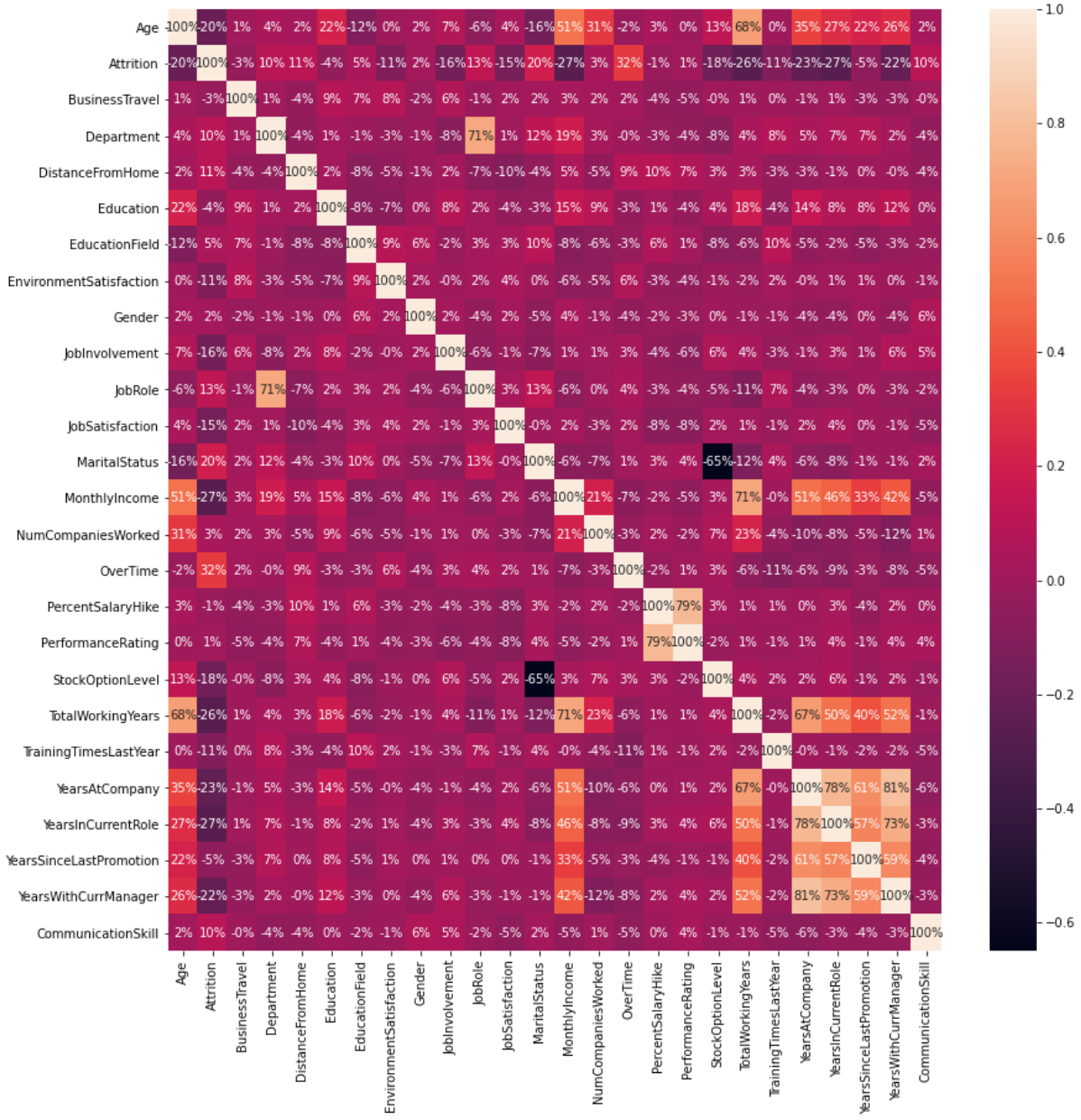




we see that employees around 28 years of age seem more likely to leave the company. Low monthly income was associated with higher attrition rates. While at triton rates were higher among employees working for less than ten years, newer employees showed the highest attrition. Employees are more likely to leave if they work over time. 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.

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



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 Business Travel, Department, Gender, JobRole, MaritalStatus, Overtime, Education Field

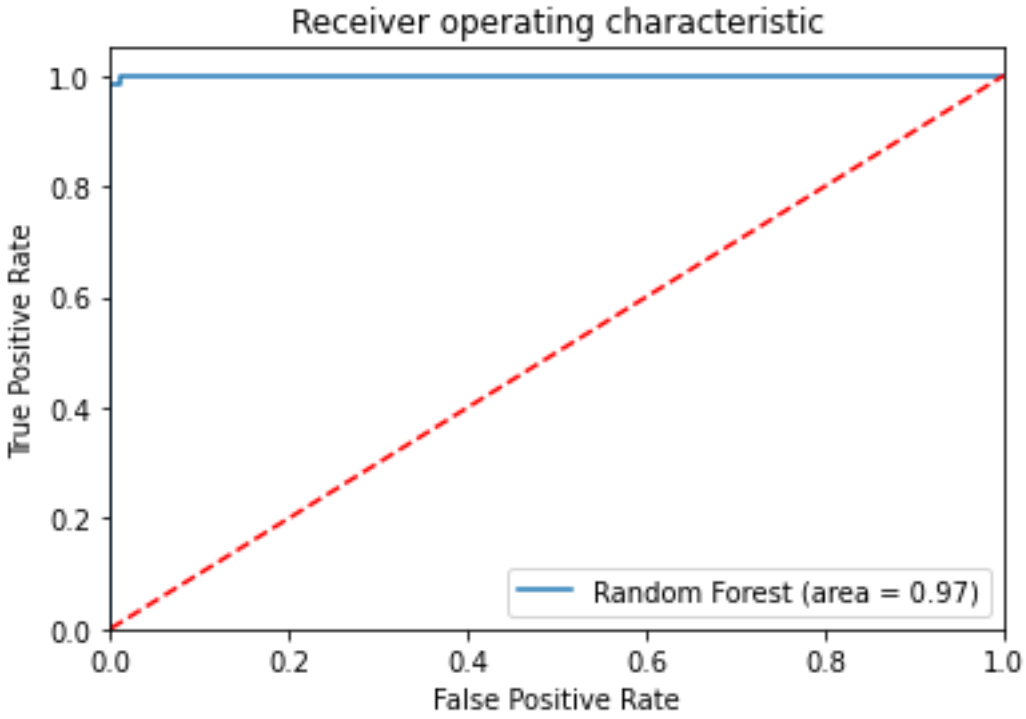
3.3.2 Oversampling and Under sampling

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

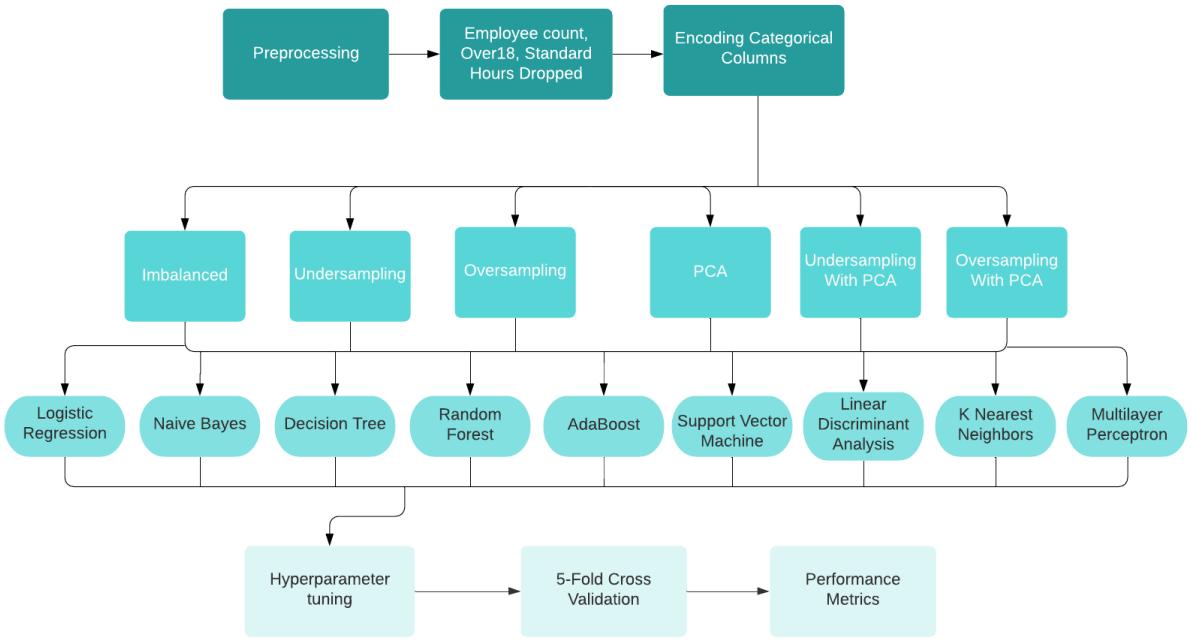
3.3.3 Feature scaling

After this, the data is standardized and normalized. Standardisation is the process of scaling the features to have 0 meanand1variance, like in normal distribution. It is crucial not just for comparing measurements with various units, but it is also a basic criterion for many machines 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 in fluence of outliers

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%.



4. 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 Neighbour's (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, hyper parameter 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 inter pretability, appropriate graphs and figures were used. As suggested in [4] accuracy for the attrition decision is a biased metric and hence

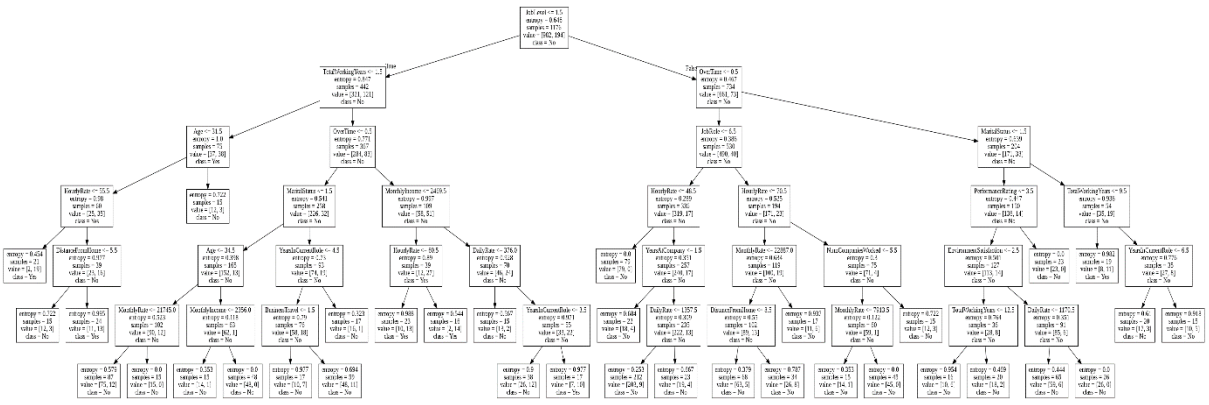
we evaluated the model on all the following classification metrics: accuracy, precision, recall and F1 score.

5. Results and Analysis

The 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 MLPalso 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 ac curacies which was no longer the case for oversampled data.



Well-organized 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

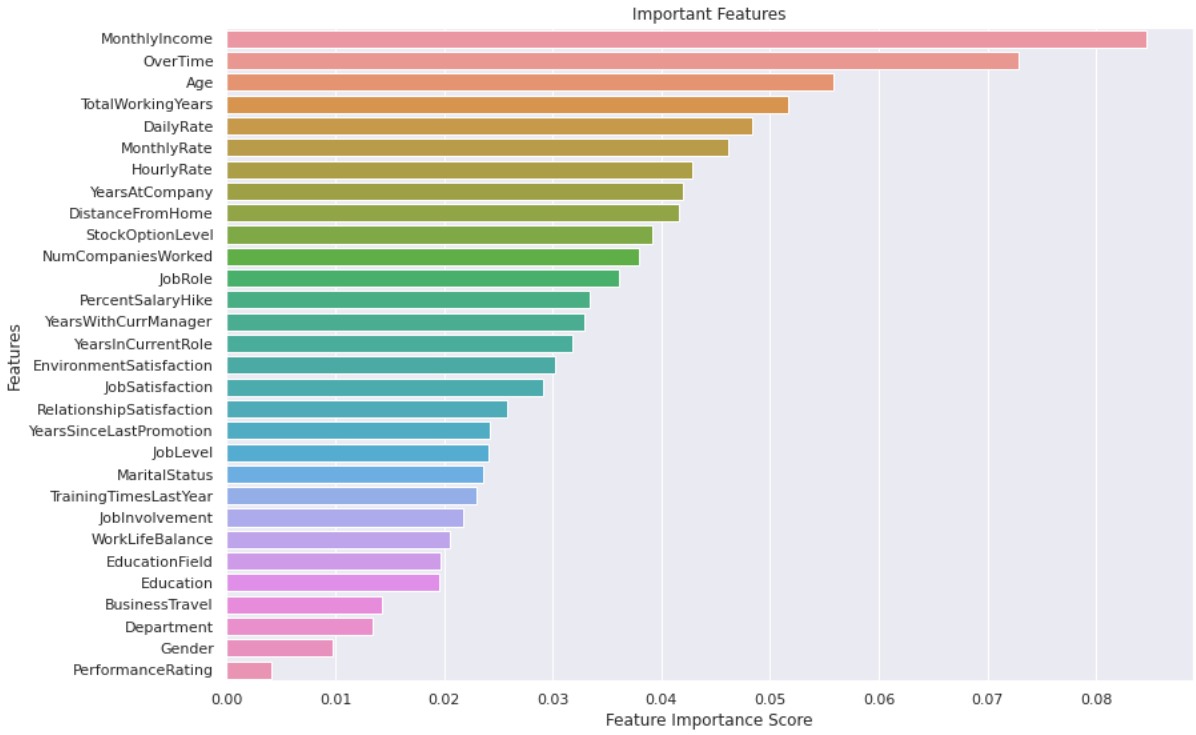
5.2. Naive Bayes (NB)

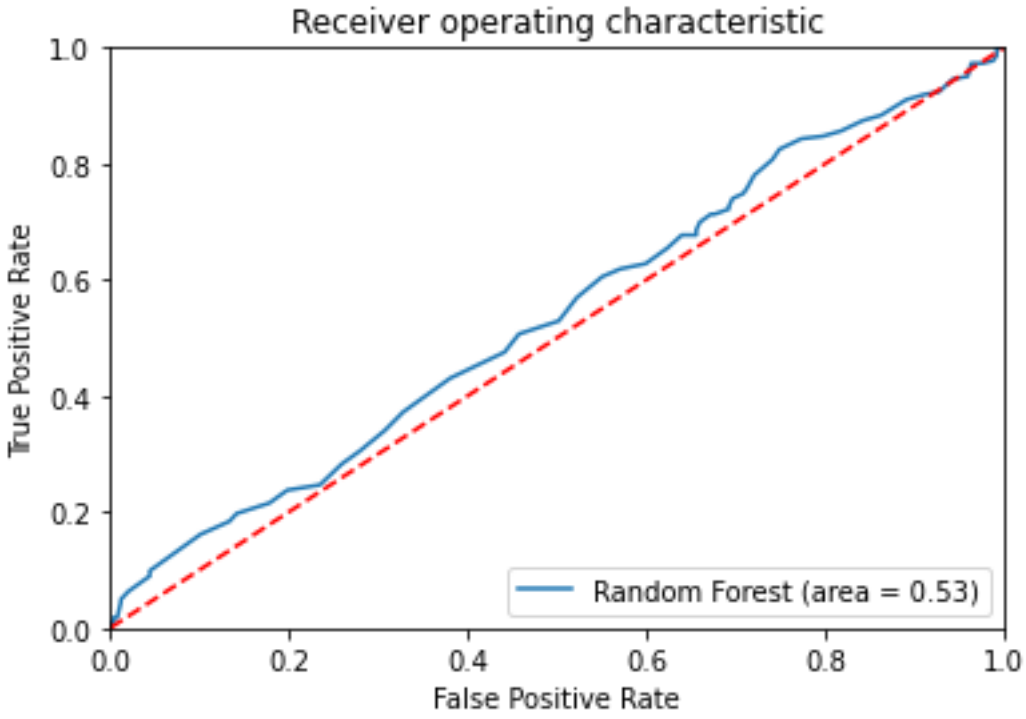
The Gaussian Naive Bayes achieved the best recall with imbalanced and undersampled data, 58.6% and 77.6% re spectively. There was an increase in precision, recall and F1 scores in oversampled and undersampled data with PCA but a decrease in the accuracy.

5.3. Decision Tree (DT)

The decision Tree model trained on 30 features and un scaled data as shown in Fig. 8 had the following tuned parameters: criterion as gain, maximum depth as 13, maxi mum 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)





The best performance was obtained by setting the hyper parameters Bootstrap to False, max depth to 100, min samples leaf to 1, min sample required to split to 3 and the to tal 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

6. Conclusion

Wetrained various supervised classification models (LR, NB, DT, RF, AdaBoost, SVM, LDA, MLP and KNN) and summarised their results in this project. As observed from EDAand 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, Total WorkingYears played major roles in the attrition decision and Gender did not impact attrition

7. 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.