Churn Prediction using Machine Learning Models

*

Soham Padmakar Mohire
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x19225491@student@ncirl.ie

***Abstract*—The process of decision making forms an essential aspect in managing a company and thereby its employees are considered to be a valuable form of asset to the organization. Hence, the process of hiring them by taking right decisions is widely accepted as a well-known hurdle by administrative officials. Since, hiring of freshers demands not only extra time, but also a large amount of capital investment; employee attrition becomes a tedious process. Apart from this, there are several other factors as well that contributes in selecting and hiring a good candidate that would in turn generate profitable returns of an organization. Therefore in this report, I propose to develop a model that could predict employee churn rate on the basis of certain attributes selected from three datasets, obtained from Kaggle repository. The summarized work so presented, involves implementation using machine learning methods along with pre-processing techniques that could analyze employee attributes and thereby predict their turnover and churn rate. The implementation of the report is experimented on extracted datasets using machine learning algorithms such as logistic regression, AdaBoost, XGBoost, KNN, decision trees and Naïve Bayes. After conducting extensive research and training selected attributes, they are finally evaluated against evaluation parameters such as accuracy and precision factors.**

***Index Terms*—Churn Rate, Employee Attrition, Employee Retention Strategy, Employment Features, Machine Learning Algorithms**

I. Introduction

The definition of employee attrition or churn rate generally refers to the reduction of manpower and human employment in an organization wherein the employees may leave the organization without prior notice or when retired. This turnover of manpower and employees leads to attrition cause for a specific period of time and may result into expenditure of human resources [1]. This expenditure is later caused due to new recruitment process and financial investment required in training and development of newly hired individuals. However, this attrition or churn rate heavily depends on the termination criteria of the company and an employee's reason to leave the job. Hence, to analyze and measure the involved manpower, terminologies such as abdication, occupational relinquishment and termination are used.

The department of HR generally forms the backbone of any company, as the members of HR are responsible to invest and select individuals and resources that could compete in the market. This process is generally known as the "hiring process" and involves the company's time and money and results into company's loss if any such resource leaves the organization.

This occurrence of unpredictable loss is often termed as "employee attrition" and tends to have a direct impact towards the functioning of the organization. The presence of this limitation has led to the rise of HR Analytics which is widely adopted to gain insights of an employee's needs and behavior, so that the department can come up with strategies and overcome the limitation using data mining techniques such as classification and regression models so that the entire situation of an employee attrition is resolved on a larger scale [2].

Apart from time and monetary investment, the employee also needs to be managed by some higher authority that might guide them throughout their service period until they get acquaintance of their work. Hence, the process of attrition is said to be voluntarily unavoidable. Therefore, this problem of employee morale needs to be solved, so that a desirable working environment is achieved and the issue of employee turnover and churn rate is reduced significantly [3]. The usage of computer aided technologies such as data mining and machine learning has helped various HR departments to judge the historical patterns of an employee and predict the attrition rate in a company. However, until now this process of attrition has been prevailing manually by domain experts on the basis of gender, cause of termination etc. Thereby motivated by the same, the proposed study aims to develop a machine learning based model that could predict the probability of such attrition occurring in an organization on the basis of certain factors such as:

- Gender of an Employee
- Joining year of Employee
- Cause of Termination
- Cause of Resignation

The above mentioned points are important to analyze so that the model could optimally predict and assist the domain experts with reliable outcomes. For this purpose, I put forward to create a model that could adopt machine learning based algorithms such as logistic regression, KNN and decision trees to implement the same and achieve real time analysis of employee attrition.

The main contributions of this work can be briefed as follows:

- Implementation of machine learning techniques with pre-processing steps to accurately predict churn rate.
- Utilization of three datasets to analyze attribute features and balance the noisy dataset.
- CGeneration of optimized results using desired algorithms.
- Evaluation of work using hyperparameters such as confusion matrix and classification report.

Further, the report is organized as follows. Section II focuses on the literature survey of the topic that highlights similar work being done by previous research scholars. Section III describes the methodologies used and the system design of the model required for implementation, followed by section IV; wherein a detailed experimental analysis is being provided with results obtained from three datasets so used. The report finally ends with a conclusion in section V preceded by references.

II. Related Work

This occurrence of unpredictable loss is often termed as "employee attrition" and tends to have a direct impact towards the functioning of the organization.

One of the major issues that companies face is the loss of talented employees. Hence in order to overcome this a model was implemented by Sarah.et.al in [4], wherein she made use of machine learning algorithms to determine on what factors did the attrition depend on. These factors were determined and machine learning based models were trained using KNN and SVM. In another approach by D. Vanden [5] the authors proposed a voice based attrition detection technique that were used in call centers to support their decision in attrition detection. For this research, authors developed a model based on textual formats wherein the attrition factors were manually entered by department professionals.

In another work by Chiu.et.al [6], he proposed attrition detection using data mining approach wherein the model was implemented based on telecommunication factors such as subscriber data and time period of the call. Singh.et.al [7] aimed to focus how employee attrition affected the normal functioning of an organization and highlighted the key factor that could determine further occurrence of this problem. He mentioned detailed statistics based on various reports that mentioned how a company's productivity was dependent on its resources and employees.

Another study, highlighting the features of employee retention was mentioned in [8] by Zhou.et.al, wherein the author emphasized how the overall efficiency of an employee could be enhanced. The author also gave a detailed explanation on various unpleasant reasons that resulted into the termination of an employee.

Authors in [9] aimed to develop a model to predict primary reasons for employee attrition. For this purpose, they made use of six machine learning algorithms and compared all the algorithms against basic evaluation parameters to achieve optimal accuracy. In a similar work [10] authors made us of the IBM-HR dataset to predict employee attrition. Since many features from the dataset were unstructured, the methods of feature selection and feature extraction were employed. This method was later followed by PCA to reduce the dimensionality of the feature.

The authors of [11] successfully implemented prediction of employee attrition using logistic regression as the ML technique. For this purpose, they gathered the dataset from Kaggle repository and established relationship between input features and output vectors. The model obtained an optimal accuracy and was further extended to deploy in various organizations. However, the process of feature selection was based on random forest and logistic regression was used in the later stages to finally predict the final output of the model. This work was highly appreciated in the domain of machine learning.

In another study by authors in [12] they implemented their concept on a dataset using machine learning algorithms such as decision trees and KNN for predicting the churn rate of employees. They conducted their validation using 10-fold cross validation technique and further had the dataset split as 70:30 for training and testing purpose respectively. However, their resulted accuracy was less when compared to other research works and hence, their contribution was just limited to the pre-processing stage and could not be used further.

In another work contributed by authors in [13] they analyzed the methods and limitations of paper [12] and later adopted different sets classifiers to solve the problem of employee attrition. The work in [13] also involved Naïve Bayes and SVM that were validated using train-test split. However, the authors in this also divided the dataset as 70:30 for training and testing purpose which was similar to the work proposed in [12]

Authors in [14] also proposed to predict churn rate of employees by presenting various stages involved in the framework. They briefed various stages of the model by making use of feature selection method and further combined with data reduction techniques. The next stage of the implementation involved training the model using machine learning methods such as logistic regression and finally proposed to predict the model in the third stage by performing confidence analysis. This system was witnessed to provide the best accuracy as compared to other techniques. A research work in [15] was experimented using classification trees and random forests that aimed in prediction of employee churn rate. The authors in this utilized the concepts of Pearson correlation pre-processed the data by eliminating non-desirable features. This work also contributed in comparing various algorithms used to accomplish optimized accuracy.

III. Methodology

This section of the report mentions machine learning based methodologies that are required for successful implementation of the system model.

A. KDD Method

I first explored the dataset, and checked the dataset for null and missing values and cleaned the data for any noisy data due to variance error. Next I transformed data, dropped columns that were not useful, renamed some columns, segregated column as per their class, transformed data type and also created additional columns when required. Machine learning algorithms were then preformed on this transformed data to evaluate patterns. Output of this patterns are then represented in visual formats to make them more understandable to anyone.

B. Logistic Regression

In literal terms, the usage of a logistic model occurs by assigning the probability of a certain class and labelling it as pass/fail, healthy /sick, win/lose etc. Hence, the same concept is extended towards several events such as employee attrition in my case. Since, the final output of the predictive model for attrition detection would either be a “yes” or “no”. In such a scenario of categorical dependent variable, we can’t implement linear regression and thereby move towards the implementation of logistic regression. Logistic regression is a machine learning based model that is to be used when the output is expected to be a binary dependent variable. Mathematically, logistic regressions are representations of binary variables as “0” and “1”. Hence, the labelling takes place only between this range and final parameters are estimated using a sigmoid function.

C. KNN

The implementation of a KNN classifier is generally referred to as a lazy learner as its concept follows of never learning from the data models provided to it, rather it tends to execute by finding similarities from the dataset which it uses to train its model. However, the finding of similar values under the trained dataset is done on neighboring samples and its closest value is determined based on the values drawn from these samples. In the later stages, prediction is followed on a new dataset and the value of ‘k’ is determined with the closest data points from ‘k’.

D. Decision Trees

The working concept of a decision tree takes place by accumulating various sub-classifiers together and generating the best result amongst them. Hence, this method is applicable for classification as well as regression methods. Decision tree corrects the habit of relying too much on the training set and improves the accuracy of the model. Initially, sub-datasets are formed from original dataset and instances are randomly selected from these sub-sets. Since the element and the content of each sub-dataset are different, multiple trees from these sub-sets are created. In the later stage, these trees are randomly selected based on a selection criteria and forms multiple decision trees known as random forests.

E. Naïve Bayes

Majority of the classification problem associated with machine learning are resolved using Naïve Bayes. Its implementation occurs by considering certain features of a dataset that are completely independent with the occurrence of other features. Hence, it is believed to classify problems on the basis of occurrence probability. Therefore, it is very much likely and suitable to perform on a complex dataset as conventional machine learning algorithms are difficult to implement on datasets containing larger and imbalanced values.

F. AdaBoost

One of the most simple and widely accepted machine learning algorithms is the boosting method that executes by assigning weights to every sample and conducting its respective observation in the training phase. One of its major advantages is the usage of weak algorithms and later assigning them higher weights. By this process, it tends to boost the overall performance of the system by combining multiple weak classifiers and increasing their initial weights. This process therefore contributes in discarding all the results so generated during the misclassification phase. In the later stages, iterations are performed so that the previous weights of the weak classifiers are increased. The creation of these iterations helps to develop boundaries and improve the overall performance of the system by boosting initial weights.

G. XGBoost

This machine learning algorithm is majorly used in classification and regression problems of a system model and has its concepts similar to that of the AdaBoost algorithm. However, this algorithm tends to estimate weak set of values that are prevalent to decision tress and thereby tends to generate a large set of weak classifiers in order to produce stronger and larger dataset. This process proves to enhance the overall efficiency of the system by boosting lower accuracy algorithms to create superior precision factors. In this process, the caused error in the system is minimized to a great extent. Therefore this algorithm is known as a boosting algorithm as it teaches all the weak learners to generate accurate predictions.

IV. SYSTEM DESIGN

The implementation of the model highly depends on the features selected from the dataset. This selected feature tends to impact and influence the performance accuracy of prediction. The overall implementation of the model can be summarized as follows:

- Gathering dataset from three repositories.
- Pre-processing the obtained data.
- Analyzing the data by selecting relevant features and employee attributes.
- Balancing the data to eliminate errors.
- Splitting the data into train-test phase.
- Predicting the model using machine learning based algorithms.

A. Dataset

The dataset used to implement the proposed work is obtained from Kaggle repository. A total of three datasets are used to achieve an improved level efficiency in the system model and develop an experimental verification. However, the characters from the dataset are initially converted to numerical values to provide efficiency of the system. The datasets used are briefed as follows:

- Employee Attrition Dataset
  This dataset gives information on employee terminations that have taken place in the last 10 years and thereby gives information on active employees working in the organization, it was obtained from kaggle[1]. The primary intent of this dataset is to analyze employee terminations and predict its possibility in the future from the data provided. The dataset includes a total of 18 columns and 49653 rows with employee attributes such as city name, job title, ID of the employee and gender.
- Telco Customer Churn Dataset
  The dataset on kaggle[2] is used to predict of retention customers and analyze customer relevance by focusing on retention programs. This dataset includes informational attributes such as customer who left in the previous month (churn), all the available services for which new customer subscribed for, all required account information of the customer, gender, age and demographic data of the customers. The raw data of this dataset contains 7043 rows and 21 columns with the “churn” column as my targeted feature.
- HR Dataset
  This dataset from kaggle[3] is used to gain insights of an employee working in an organization so that the probability of employee attrition is avoided in the future. The dataset includes a total of 10 columns with selected attribute features such as last evaluation, promotion in last 5 years, number of projects undertaken, sales and salary.

B. Pre - Processing

This step is followed by cleaning the raw data obtained from the repository by eliminating missing values and thereby getting effective results. The stage of pre-processing also involves transforming one form of data to another form and smoothing numerical values through integration. Hence, the obtained data is now reduced to a larger extent and if therefor formatted to achieve effective results.

C. Dataset Analysis

In the data analysis phase, the obtained categorical values from the dataset are converted to their respective numerical values so that the overall efficiency of the classification algorithms increases. The figure below represents a correlation matrix which is generated to establish relationship amongst dataset features.

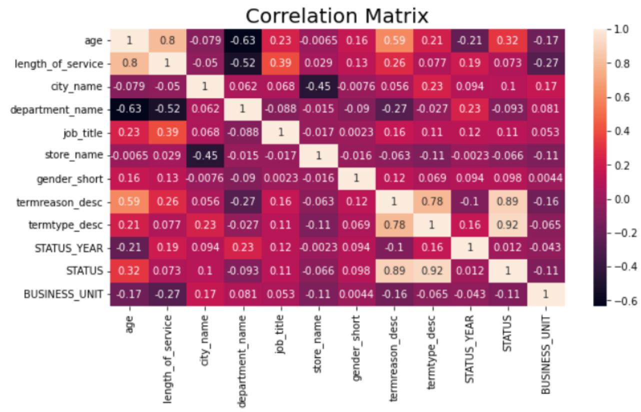

Fig. 1. Attrition Analysis Steps

Initially, all the categorical attributes such as “salary” that contains values such as low, medium and high are converted to numerical values such as 0, 1, and 2 respectively. Hence, in such a case a correlation matrix helps to identify such attributes depicting strong and weak correlation.

D. Workflow

The implementation phase of the proposed model begins by collecting raw data from three repositories containing multiple information of employees that might help to calculate the turnover rate. In the next step, the data undergoes a phase of pre-processing wherein all irrelevant features are eliminated and relevant features are worked upon. Next, the process of EDA is conducted where the data is visually analyzed through correlation matrix to gain insights on numerical values of the dataset. Once, the process of EDA is finished the dataset is split into train-test phase and respective machine learning algorithms are imposed on the data. After the implementation of respective algorithms, its accuracy score is measured to determine the highest predicting algorithm.

V. EXPERIMENTAL ANALYSIS

A. Experimental Analysis on Employee Dataset

As mentioned in the above datasets, employee attributes such as salary, city name, gender etc are given. Based on the values derived from this data machine learning algorithms are applied and churn rate is predicted to gain insights that whether an employee would quit the organization or not. Since a total of three datasets are used, this section is divided into three chunks of experimental analysis with respective descriptions of performance analysis.

This dataset contains a total of 18 columns and 49653 rows and is used to gain insights on employee termination which have been taken place in the last 10 years. This dataset is precisely used to further predict the rate of employee attrition that might occur in the future. The model is trained on this dataset using decision trees and Naïve Bayes as machine

[1]https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data
[2]https://www.kaggle.com/datasets/blastchar/telco-customer-churn
[3]https://www.kaggle.com/code/jacksonchou/hr-analytics/data?select=HR_comma_sep.csv

learning classifiers. The table in Fig 3. provides information and numerical values obtained from confusion matrix.

Name Of Algorithm	Train\Test	Values	Accuracy
Decision Tree	Train Set	[37880,625] [00038,563]	99.18
	Test Set	[9477,186] [0009,605]	99.03
Gaussian Naive Bayes	Train Set	[349413,564] [2235116,212]	65.84
	Test Set	[8751,912] [566939,36]	65.84

Fig. 2. Obtained values from Confusion Matrix

The graph below illustrates the generated confusion matrix and provides numerical values on false positive and true negative cases. However it can be observed that the implementation of decision trees produced 9477 true positive and 186 false positives cases of employee attrition prediction. Similarly, GNB produced 8751 true positive and 912 false positive cases for the same. Therefore, it can be observed that the testing accuracy of decision trees proved to be better than that of GNB.

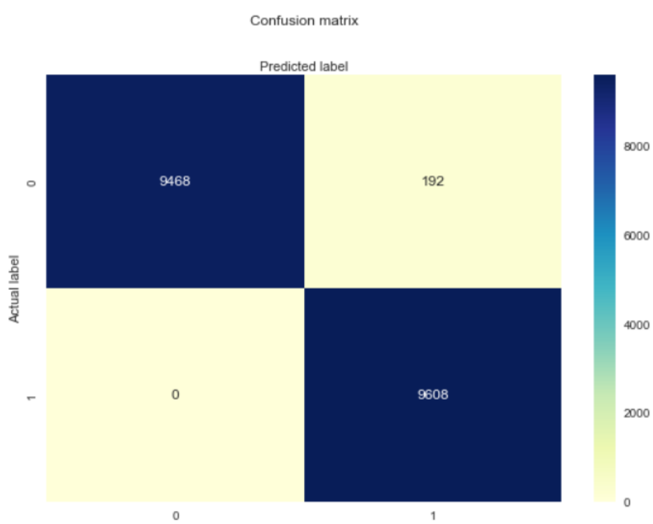

Fig. 3. Decision Tree

The table in Fig 5. depicts numerical values obtained from classification reports. The numerical data obtained from table 2 gives a detailed summary on the values obtained through classification report in the form of accuracy and precision. The overall results are thereby represented with their respective accuracy values. However, it has been observed that the highest accuracy has been generated when the model is experimented using decision trees and thereby provides an accuracy of 99%.

B. Experimental Analysis on Telco Churn Dataset

This dataset contains a total number of 7043 rows and 21 columns and is developed to focus on predicting customer

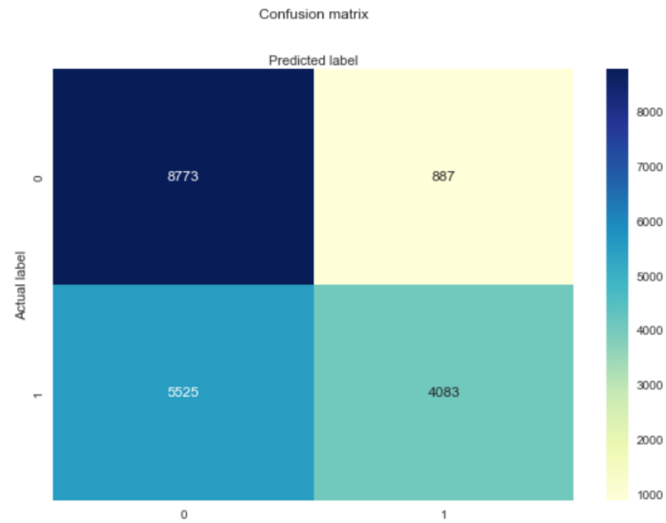

Fig. 4. Decision Tree

Name Of Algorithm	Precision		Recall		FI Score	
	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
Decision Tree	1.00	0.98	0.98	1.00	0.99	0.99
Accuracy			0.99			
Gaussian Naive Bayes	0.61	0.81	0.91	0.41	0.73	0.54
Accuracy			0.66			

Fig. 5. Obtained values from Classification Report

behavior on their respective retentions. The model is trained on the dataset using AdaBoost and XGBoost algorithms as machine learning classifiers. The table below provides information and numerical values obtained from confusion matrix.

Name Of Algorithm	Train\Test	Values	Accuracy
ADA Boost	Train Set	[3775,377] [689,784]	81.04
	Test Set	[908,103] [191,205]	79.10
XG Boost	Train Set	[3836,316] [674,799]	82.39
	Test Set	[924,87] [187,209]	80.52

Fig. 6. Obtained values from Confusion Matrix

The graph below illustrates the generated confusion matrix and provides numerical values on false positive and true negative cases. However it can be observed that the implementation of ADaBoost produced 908 true positive and 103 false positives cases of employee attrition prediction. Similarly, XGBoost produced 924 true positive and 87 false positive cases for the same. Therefore, it can be observed that the testing accuracy of XGBoost proved to be better than that of AdaBoost.

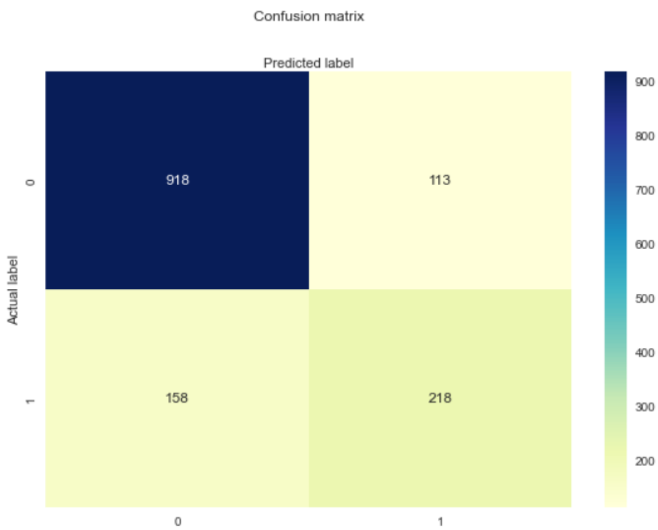

Fig. 7. ADa Boost

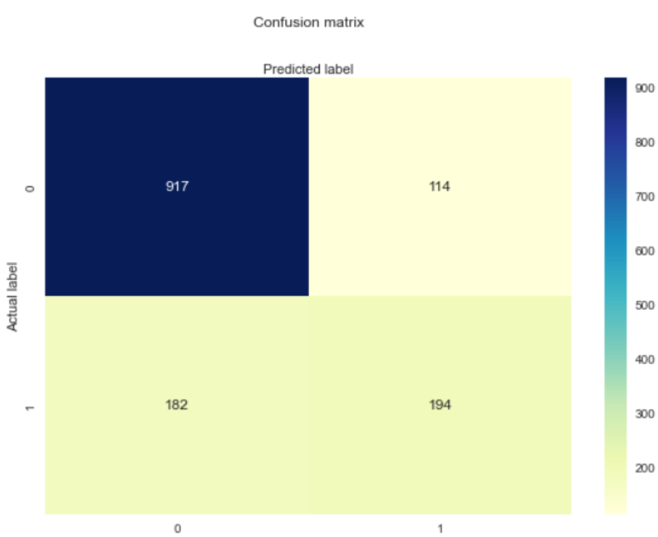

Fig. 8. XG Boost

The table in Fig 9. depicts numerical values obtained from classification reports. The numerical data obtained from table 4 gives a detailed summary on the values obtained through classification report in the form of accuracy and precision. The overall results are thereby represented with their respective accuracy values. However, it has been observed that the highest

Name Of Algorithm	Precision		Recall		FI Score	
ADA Boost	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
	0.83	0.67	0.90	0.52	0.56	0.58
Accuracy			0.79			
XG Boost	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
	0.83	0.71	0.91	0.53	0.87	0.60
Accuracy			0.81			

Fig. 9. Obtained values from Classification Report

accuracy has been generated when the model is experimented using XGBoost and thereby provides an accuracy of 81%.

C. HR Dataset

This dataset contains a total number of 10 rows and 10 columns and is developed to take wise decisions by the HR in hiring employees for a company based on employee attributes such as number of project undertaken, time spend in a company, sales, salary etc. The model is trained on the dataset using Logistic Regression and KNN algorithms as machine learning classifiers. The table below provides information and numerical values obtained from confusion matrix.

Name Of Algorithm	Train\Test	Values	Accuracy
Logistic Regression	Train Set	[8442,703] [2130,724]	76.38
	Test Set	[2104,179] [523,194]	76.60
KNN	Train Set	[8882,263] [1812,673]	96.29
	Test Set	[218,895] [74,643]	94.36

Fig. 10. Obtained values from Confusion Matrix

The graph below illustrates the generated confusion matrix and provides numerical values on false positive and true negative cases. However it can be observed that the implementation of Logistic Regression produced 2104 true positive and 179 false positives cases of employee attrition prediction. Similarly, KNN produced 2188 true positive and 96 false positive cases for the same. Therefore, it can be observed that the testing accuracy of KNN proved to be better than that of Logistic Regression.

The table in Fig 13. depicts numerical values obtained from classification reports. The numerical data obtained from table 6 gives a detailed summary on the values obtained through classification report in the form of accuracy and precision. The overall results are thereby represented with their respective

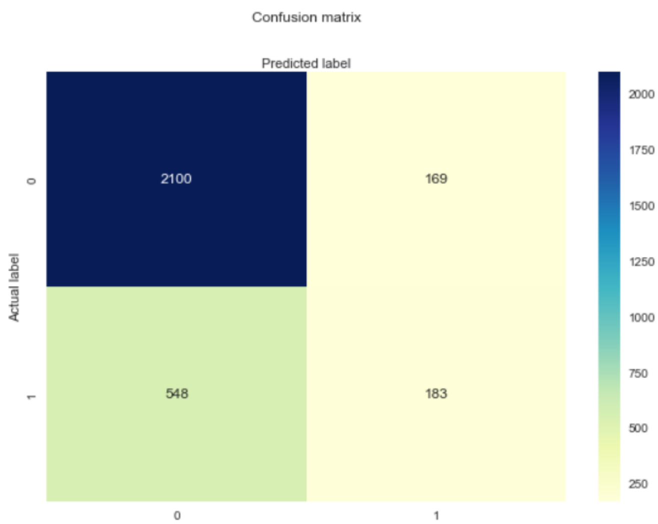

Fig. 11. Logistic Regression

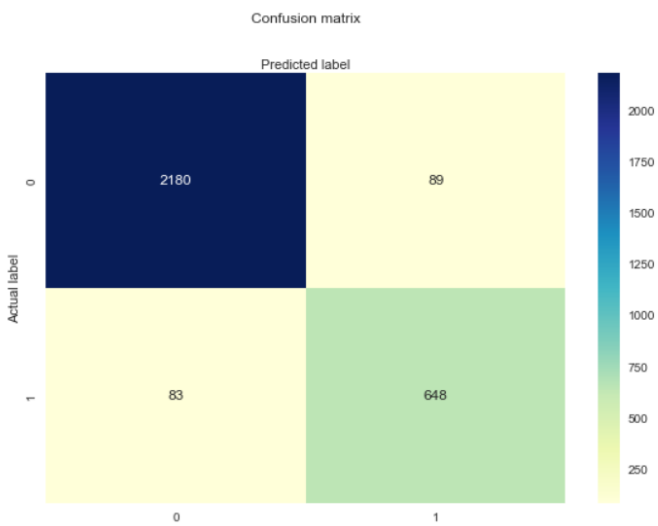

Fig. 12. KNN

Name Of Algorithm	Precision		Recall		FI Score	
Logistic Regression	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
	0.80	0.52	0.92	0.27	0.86	0.36
Accuracy			**0.77**			
KNN	Positive Cases	Negative Cases	Positive Cases	Negative Cases	Positive Cases	Negative Cases
	0.97	0.87	0.96	0.90	0.96	0.88
Accuracy			**0.94**			

Fig. 13. Obtained values from Confusion Matrix

accuracy values. However, it has been observed that the highest accuracy has been generated when the model is experimented using KNN and thereby provides an accuracy of 94%.

VI. Conclusion

The primary aim of the submitted report is to implement machine learning algorithms and predict the associated churn rate of employees. Since this churn rate and turnover has a direct impact on the profits of an organization, it becomes a very critical decision to hire and terminate employees and individuals to maintain the working balance in the environment. Therefore, I have proposed to predict the same scenario using three different datasets and various ML based methods. From experimental analysis it has been observed that KNN, XGBoost and decision trees generated optimized accuracies of 94%, 81% and 99% respectively.

References

[1] Jarrahi, M.H. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Bus. Horiz. 2018, 61, 577–586

[2] Duan, Y.; Edwards, J.S.; Dwivedi, Y.K. Artificial intelligence for decision making in the era of Big Data—Evolution, challenges and research agenda. Int. J. Inf. Manag. 2019, 48, 63–71

[3] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. Computers 2020, 9, 86

[4] Sarah. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Vaesens,"New insights into a churn prediction in the telecommunication sector. An profit driven datamining approach," European journal of operational research, vol. 218, no. 1, pp. 211-229, 2012

[5] K.Coussement and D. VandenPoel, "Integrating the voice of customers through call centre emails into a decision support system for attrition prediction," Information & Management, vol. 45, no. 3, pp. 164–174, 2008

[6] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to attrition prediction: a data mining approach," Expert systems with applications, vol. 23, no. 2, pp. 103–112, 2002

[7] Singh, M., Varshney, K. R., Wang, J., Mojsilovic, A., Gill, A. R., Faur, P. I. and Ezry, R. (2012). An analytics approach for proactively combating voluntary attrition of employees, Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, IEEE, pp. 317–323

[8] Zhou, N., Gifford, W. M., Yan, J. and Li, H. (2016). End-to-end solution with clustering method for attrition analysis, Services Computing (SCC), 2016 IEEE International Conference on, IEEE, pp. 363–370

[9] Mohbey, K.K. Employee's Attrition Prediction Using Machine Learning Approaches. In Machine Learning and Deep Learning in Real-Time Applications; IGI Global: Hershey, PA, USA, 2020; pp. 121–128

[10] Ponnuru, S.; Merugumala, G.; Padigala, S.; Vanga, R.; Kantapalli, B. Employee Attrition Prediction using Logistic Regression. Int. J. Res. Appl. Sci. Eng. Technol. 2020, 8, 2871–2875

[11] Yang, S.; Ravikumar, P.; Shi, T. IBM Employee Attrition Analysis. arXiv 2020, arXiv:2012.01286

[12] Usha, P.M.; Balaji, N. Analysing employee attrition using machine learning. Karpagam J. Comput. Sci. 2019, 13, 277–282

[13] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. Computers 2020, 9, 86

[14] Najafi-Zangeneh, S.; Shams-Gharneh, N.; Arjomandi-Nezhad, A.; Hashemkhani Zolfani, S. An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. Mathematics 2021, 9, 1226

[15] Pratt, M.; Boudhane, M.; Cakula, S. Employee Attrition Estimation Using Random Forest Algorithm. Balt. J. Mod. Comput. 2021, 9, 49–66