Workforce Analytics using Machine learning

1st Dhwani Dharmesh Hingu
Dept. of Computing
National College of Ireland
Dublin, Ireland
x19216742@ncirl.student.ie

***Abstract*—In order to gain profits of the organization as well as employees, every organization tries their best to make prime utilization of their employees. However, they face a lot of issues identifying while their finest employees, this is where People Analytics comes into the picture. HR Analytics looks after day-to-day operations, Efficiencies in procedure or other strategic operational problems. Thus, HR Analytics is concerned about all aspects of organization at large level, whereas workforce analytics focusses on employee's data such as employee's engagement, job satisfaction and employee's success. Workforce analytics falls under the broad umbrella of HR Analytics. Employees quit for a number of reasons, including salary dissatisfaction, stagnant career growth, and so on. A great loss in company can be not only in terms of money but also in terms of losing their skilled workers. If the company identifies whether which employee should get promotion, to predict how much salary an employee should get and also to predict whether an employee will leave the company or no in near future, then the company can work on employee retention in advance so that they can save their valuable and hardworking employees. This prediction of employee's attrition and retention can be done through machine learning techniques. Every modern organization collects massive number of employee's data, we will use this data, analyse this data, extract insights from this data, so that company can take better decisions to handle their workforce.**

I. Introduction

There has been much interest in the topic of HR Analytics in recent years. A simple online search of scholarly articles will bring up over thousands of results for 'HR Analytics'. HR Analytics is a broad term, a lot of things are incorporated. Under HR Analytics one of the most important sub domain is employee engagement and management which is now prominently known as workforce analytics, because with growing technology now we are able to analyse and identify our finest employees as well as our employees engagement through technology. In this project, supervised machine learning techniques are used to build models and analyse factors which affect employees attrition and engagement in order to help recruiting team and company to know their finest employees.

A. Dataset 1: Employee Salary - Adult Census Income

Wage is one of the important factor in employees job satisfaction. An employee works for fulfilling their basic necessities such as food, clothes and house. Employees productivity is directly promotional to their salary. If an employer is happy and satisfied for what he is paid for his work than he'll be happy with his work and satisfied with his job. However, this indirectly helps for the company's growth as well also there are lot of chances that the employee will stay for along in the company. On the other hand, it benefits company also not just in term of growth but also in terms of knowing their optimal employee's. If an employee is not paid much according to employee's performance than there are high chances that an employee will leave the company. Thus, salary is one of the importance aspect a company should identify. Also if the company is paying too much on the employee and the employee doesn't work and doesn't meet the company's expectation, they tend to waste their lots of time and money on an employee who doesn't deserve to be in the company. Thus using Adult Census Income dataset, will try to identify What are the factors that mostly impact on an employee salary greater than $50k?

B. Dataset 2: HR Analytics : Employee promotion data

Whenever an employee performs good the employee is usually awarded with a promotion. This majorly helps in knowing employees effectiveness and efficiency, also promotion contributes to job satisfaction of an employee. It takes time to train employee for a specific role specially for Team Lead or Manager, as these roles have to lead and manage team, if a wrong promotion is done this will result as massive loss for company to gain the profits. Thus, under work force analytics identifying the right people for promotion is one of the biggest problem. The final promotions are only revealed after the assessment and this leads to delay in transfer to new positions. Hence, company needs help in determining the deserving applicants at a specific checkpoint so that they can expedite the entire promotion cycle. With HR Analytics : Employee promotion data will try identify following research questions like:
what are the factors that affects employee's promotion?
What are the factors that can classify which employee deserves promotion and employee should not?

C. Dataset 2: HR Analytics : Job change of Data Scientist

Be it employee turnover or employee voluntarily leaving company, in both condition a company faces major loss. When company hires an employee it invest a lot of time and money to train that employee so that, that employee knows

how the company works what he has to do for the company. In most of MNC's, when an employee is hired they are given 3-6 month training, the company bares all cost to train the employee, most of the times some employees simply leaves after taking training. However this leads to loss for the company. On the other hand when an employee is not satisfied with the job or company, the employee leaves the company. The research says most of the employee leaves the company if their work life and personal life is not balanced, and this affects into their job productivity also. These days no body wants wo work extra hours and or overtime if they are not respected for their work in the company or if they do not love their job. So by using HR Analytics : Job change of Data Scientist, will try to solve following research question: What are the factors that affects an employee to leave the company? What are the factor that helps to identify which employee is not looking for job change and will stay in the company? Are the employee's satisfied with their job and is their work life balanced?

II. Related Work

In the paper "Salary Prediction Using Regression Techniques", the author proposed to predict the salary for an employee in an organisation. The author used regression models to predict the salary and opted to use graphical representation which would be easily interpreted by organization member. The aim was to help an employee where they can check their growth similarly they can be plotted to a particular field based on their salary. For implementation Linear Regression and Polynomial Regression were used based on the data collected from the organization. Although, the prediction was near correct, author suggests to go with K-nearest regression for better accuracy [1]

Similarly, in the paper "Salary Prediction in It Job Market" the author had predicted the salary of an employee using other regression models like Random Forest, Decision Tree, and SVM where they got better accuracy of 97% by using Random forest. The paper shows, how author implemented salary prediction on dataset on 10,000 records where the experience of employee along with CTC per annum influenced the output. It also, opened way for investigation how a job post influenced an employee salary and also discovered data-driven profiles obtained via skill based aggregation [2].

In the paper "Turnover of Employees within the Irish Hospitality Sector" explains about Irish hospital market and its employees and their relation with the employee. Employee's job satisfaction is the most important factor. The project was related to research and investigation, a research methodology was used. This paper was helpful for the project to understand business understanding of employee turnover. One limitation of this paper was there were no machine learning techniques used only data gathering and understanding was done [3].

On the other hand, in the paper "Wages and Employees Performance: The Quality of Work Life as Moderator" [4] authors focus on how wages impacts work life and performance of an employee. The data collected was based on questionnaire form and random sampling was applied to get data of around 100 employee's. Linear regression and moderated regression model was implemented for the analysis which showed negative effect on wages of an employee. Also, the author gave analysis of how quality of work is affected by wages. Quality of work life and less salary was the reason for negative impact. The analysis showed, all the independent variable had significant relation with the dependent variable. Based in the F Statistics, the significance values was 0.021 which is less than 0.05. The adjusted R-Square was 0.044 which suggests that only 4% increase in employee's performance was based on wages of the employee whereas, 95.6% is influenced by other variables. After implementation of Moderated Regression Analysis the adjusted R-Square value increased to 15.9%. Also, P values obtained for wages was 0.008 which was less than 0.05 similarly, for quality of work life was 0.016.

The purpose of this paper [5] was to evaluate the effects of promotion practice on job satisfaction. Primary and Secondary data were used in this case study. Multi stage sampling and Simple random sampling were implemented on the employee's data. The data was analyzed by SPSS and also with the help of correlation and regression models. Correlation helped in predicting positive & negative relationship between independent and dependent variables. In this case study, output for the regression model indicated that independent variables had accounted for 44.5% of variance on the target variable. Based on the results of this research paper, the author concluded that the total assessment of perception of employees towards the promotion practice of the bank as being irregular and dissatisfying and further,the results showed that the results that employees were not satisfied with the current promotion opportunities in the bank. Also,the result of this case study suggests that promotion practice is a very key factor in keeping an employee satisfied.

Also, in the paper "A Study on Employee Attrition: Inevitable yet Manageable" [6] the author examined the employee attrition which means what influences an employee resign or retire. For analysis, the author gathered random sample of 100 data records analysed using SPSS software. The paper also gave statistical analysis by computing correlation analysis, T test, Chi-Square, Annova test and Multiple Regression. The output from correlation analysis was meaningful although, multiple regression helped to check the impact of employee attrition. The R-square was 0.218 where as Adjusted R-Square value was 0.47 which were obtained using Multiple Regression model. The study suggests that there was no meaningful association of employee searching for new job comparing to lower attrition hence, null hypothesis was accepted.

In the research paper"Employee Satisfaction as an Important Tool in Human Resources Management" [7], the author has discussed the companies and their employees. This research paper constructs a casual model which shows the impact of employee perceived value on employee satisfaction, as well as the impact of employee satisfaction, affective commitment and on employee loyalty. The conclusion of this paper was that

by building a casual model between employee and employee satisfaction, it will help the companies make their employees more valuable and enjoyable the the working environment.

A similar analysis was conducted in paper "Prediction of Employee Promotion Based on Personal Basic Features and Post Features" [8] where it focused on HR management. They managed to get data of a Chinese company and constructed various features and applied machine learning model for the same. As per the result random forest model was better and was verified based on the validity features. Also Gini importance was calculated for every feature where it showed how it influences the staff promotion. Their final analysis depicted that job post based features had higher impact on promotion where as personal details had low impact. The other features like working year, different posts and higher department level had high impact on employee promotion.

The author of this research paper [9] helps HR team in making strategies over employees and managerial decisions. The contribution of this paper work is to develop a framework for a company which will help them in predicting various factors such as employee turnover, employee satisfaction. The author used various analysis models such as descriptive analysis, predictive analysis, and entity sentiment analysis to build this framework. To predict employee's turnover, employee satisfaction, various machine learning algorithms were used in this paper by the author. The conclusion was that by using these models, the HR team can get more insights on the company's employees like salary expectations, employee promotions and may other risk factors regarding the employees

In this paper [10], the author is helping the HR team in predicting why the employees tend to leave the company and the possible reasons behind it. This research analysis was performed by using various ML techniques such as SMOTE, SVM & Linear Regression. The author calculated the accuracy of the model by focusing on True Positives. The author concluded the paper by giving precise results, and indicating that the management team should select significance features which will help in different employee attributes. From this paper, it was observed that the author performed SA-SVM with Bayesian Optimization, which gave the best accuracy percentage as compared to other performed models.

The given paper, ""Exploiting linkedin to predict employee resignation likelihood" [11] has implement another way to to predict employee resignation which is different from others. They have generated dataset from Linkedln where they identified features which influence an employee to resign. Also, they implemented different Machine Learning techniques like Decision Tree, Back Propagation, Self-Organizing Maps where supervised learning algorithm like cross validation score was applied with ten folds. Decision Tree was the best out of others which gave accuracy of 88.4% where as other gave around 50% & 56% respectively. Mean of Kappa score for Decision Tree was 0.34, Back Propagation was 0 whereas SOM was 0.02

In the given paper [12], the author discusses how the company's success or failure depends on the performance of their employees. Employee performance matrix for the next year was calculated by using Decision tree algorithm and it also elaborates that how the decision clustering algorithm helps in calculating employee's performance. K-means clustering algorithm was implemented by the author to classify employees based on their performances. The author's conclusion in this paper predicts the number of employees that are selected for promotion, designation, employees performance.

III. Methodology

For this project (C) Cross (I)Industry (S) Standard (P) Process for (D) Data (M) mining (CRISP-DM) methodology is used as this model helps in workforce analytics to manage their employees. For Data Mining this methodology was developed by IBM, apart from data mining this methodology it is also useful for other projects. This methodology works like a revision model with datasets which can add new features and delete features which was are not useful. Moreover, this models is helpful for Human Resource and Business Analysts as the business understanding explains the problems and solutions for the evaluation and development. Thus CRISP-DM is used in solving employee turnover, employee job satisfaction and employee productivity using the dataset. This project undergoes six steps of CRISP-DM as shown bellow [13]

Fig. 1. CRISP-DM process flow

A. Business Understanding

For an organisations to gain profits and trust from their customers they have to build a good reputation in the market. The company needs to keep track of all positive and negative

feedbacks of customers as well as the company needs to keep track of their workforce, the well-being of their employees. If your workforce is happy and motivated than the company can grow exponentially and will also save from loss of their valuable employees. If employees of the company are not satisfied and if employees are leaving the job abruptly than this affects not only to company by loosing their employee but it indirectly affects company by loosing their customers also. So with the help of dataset this project aims to solve research questions. By using machine learning techniques this project will try to predict who are the company's prime employees and based of their features will try to identify who deserves a promotion, this project also tries to classify that what are the factors that helps an employee earn $50K a year.

B. Data Understanding

If someone is not aware what the data is which attributes are useful for modelling than it leads to bad modelling and bad results, thus data understanding is very crucial step after business understanding Following are three datasets are chosen from Kaggle [14], as Kaggle is world's largest data science community which allow users to publish datasets and explore datasets and create models in a web-based data-science environment, collaborate with other data scientists and machine learning engineers, and compete in data science competitions.

1) Dataset 1- Adult Census income: This dataset [15] consists of employee features, based on the features it helps to identifies which are the factors that affects employee income. It has 15 columns and 325262 rows. Independent Variables – age, workclass, fnlwgt, education, education.num, marital.status, occupation, relationship, race, sex, capital.gain, capital.loss, hours.per.week, native.country. Dependent Variable : income ($\leq$50K, $>$50K) Based on independent variables will try to analyse what are the factors that helps an employee to earn $>$50K a year or $\leq$50K

2) Dataset 2- HR Analytics : Employee promotion data: This dataset [16] consists of employee features, based on the features it helps to identifies which are the factors that affects employee's promotion. The dataset consists of 13 columns and 54809 rows. Independent variables: employee_id, department¸ region, education, gender¸ recruitment_channel¸ no_of_trainings, age, previous_year_rating, length_of_service, awards_won?, avg_training_score. Dependent variable : is_promoted (0-not promoted, 1- is promoted) Based on independent variables will try to analyse what are the factors that helps an employee to gain promotion.

3) Dataset 3 - HR Analytics : Job change of Data Scientist: This dataset [17] consists of employee features, based on the features it helps to identifies what are the factors that leads an employee leave the company or stay in the company. The dataset consist of 14 columns and 19159 rows. Independent variables: enrollee_id, city, city_development_index, gender, relevent_experience, enrolled_university, education_level, major_discipline¸ experience, company_size, company_type, last_new_job, training_hours. Dependent variable: target (0- Not looking for job change, 1- Looking for a job change). Based on independent variables will try to analyse what are the factors that helps to analyse whether the employee is looking for a job change or no.

C. Data Preparation:

Before using machine learning models it is important to pre-process data, clean data, and make the data in a way that helps machine to understand and provide best results as per the machines understanding. Thus, Data pre- processing is very vital step in building a model. There are many things involved in data pre-processing like understanding whether the data consists of numeric data or categorical data, followed by how many missing values or redundant values are as well as if any outliers are, once these things are identified data needs to be cleaned and transformed appropriately.

1) Dataset 1: Adult Census income: Initially how many null values are present in the dataset was checked. The null values were present in form of "?". So in order to understand how many of these "?" present used relace function to replace them with NAN. Once the values were replaced, the dataset was checked for NAN values and fig.2 shows which column consists of null values.

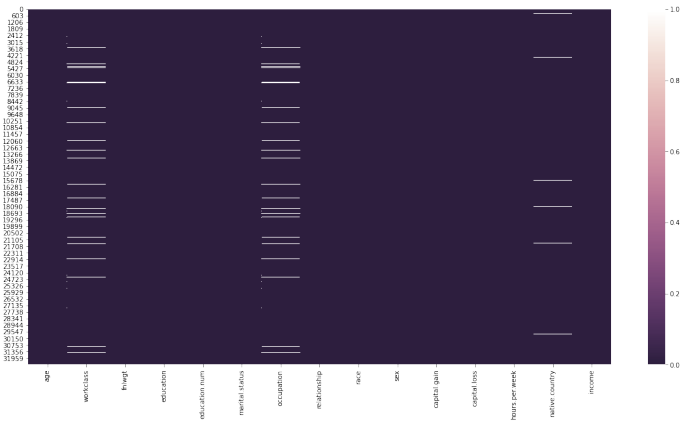

Fig. 2. Null values in Salary Dataset

As there were not much null values the null values were dropped,as the dataset had enough number of data that can help in building model. Also, found out that columns like fnlwgt which is continuous variable was not helpful for building model as well as columns like education and education.num are helping for same thing to understand what is employee's education, thus dropped education.num and tried to group education categories into 1 category as there were lot of categories eg. 'Preschool', '1st-4th', '5th-6th', '7th-8th', '9th','10th', '11th', '12th' group as school. Similar done for other categories in the education column. Also for feature engineering, did ordinal encoding on education

column as ranking is given.

For other categorical columns ordinal encoding was not possible so did OneHotEnconding by creating dummy variables using pd.get_dummies function of python's pandas. The dataset was cleaned and transformed and was ready for building model. Before model building, the dependent variable was checked whether the data is balanced or no and it was observed that the data is not balanced hence SMOTE resample technique was used to balanced the data. The fig.3 shows before and after sampling.

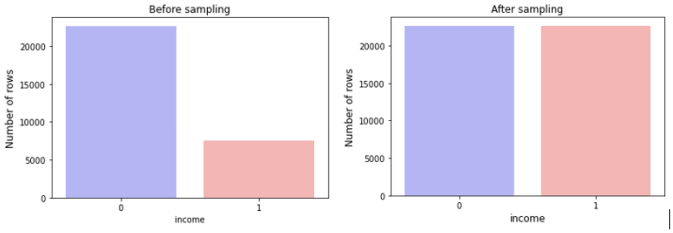

Fig. 3. Balncing Data with SMOTE

2) Dataset 2: HR Analytics : Employee promotion data: The dataset was checked if there are any null values, and it was observed that that two columns had 4% and 7% missing values. However, the dataset had 54k records as it was large enough to build a model thus decided to drop null values. Once the dataset was free from null values, the dataset was checked whether how many categorical features it consisted, 5 out 13 columns were categorical. So, for feature engineering, the education column was handled with ordinal encoding that is ranking was given based on the education lowest being "Below Secondary" and highest being "Master's & above" category. Other columns like department, region, recruitment_channel cannot be given ranking so dummy variables were created and dummy trap was handled so that the machine do not get confused. Once the data got converted to numeric, tried to identify which columns are useful building model and thus observed that employee_id doesn't help as a factor to identify whether an employee should get promotion or no. Also looked for outliers and found that there were outliers in the some columns also did feature selection so got to know inspite of having outlier these outlier were not harmful for model building.The fig.4 shows outliers and fig.5 shows the best features for model building

3) Dataset 3 - HR Analytics : Job change of Data Scientist: Starting with in this dataset missing values were checked and fig.6 is heatmap which demonstrates the tile with dark colour had correlation between columns having missing values. There is strong correlation hence dropping null values was not a good option as well as if null values were dropped than the structure of the dataset would have been reduced a lot which would not help in building model. Thus, Data imputation was done statistically by using mode of every column. Columns which had lowest number of null values were dropped eg. Experience column had 0.34% missing values so null values were dropped. Column "enrollee_id" was not helpful for analysing whether the employee will stay or leave company so dropped it. For Feature engineering ordinal encoding was done on column like "relevent_experience", "enrolled_university", "education_level" and "company_size" as ranking these column helps in building and analysing the model. For columns like "gender", "major_discipline" and "company_type" ranking cannot be given thus dummy variables were made and dummy trap was taken care of by dropping first colum. Outliers were detected in only one column and were not very harmful for building models.The fig.6 shows outliers in the dataset 3

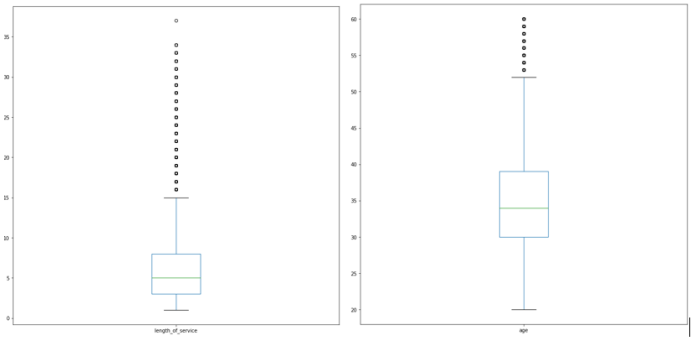

Fig. 4. Balncing Data with SMOTE

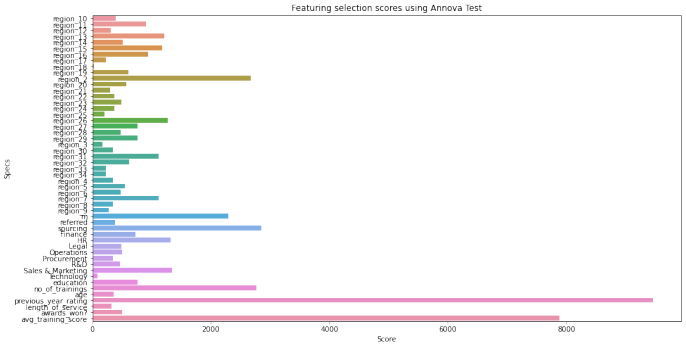

Fig. 5. Feature Selection on Dataset 2 using ANOVA Score

D. Modeling:

After the data is cleaned and analysed, modelling takes place, modelling is the 4th process step in CRISP-DM for this project only classification machine learning techniques are used as all three data sets have features where classification done. For every dataset at least two machine learning classification techniques are used and compared for the best results.

1) Dataset 1: Adult Census income: For this dataset, two classification algorithms are used which are naïve bayes and K-Nearest Neighbour(KNN). Naïve bayse: with the presumption of predictor independence A Naive Bayes classifier, in simple terms, assumes that the existence of one function in a class is unrelated to the presence of any other feature

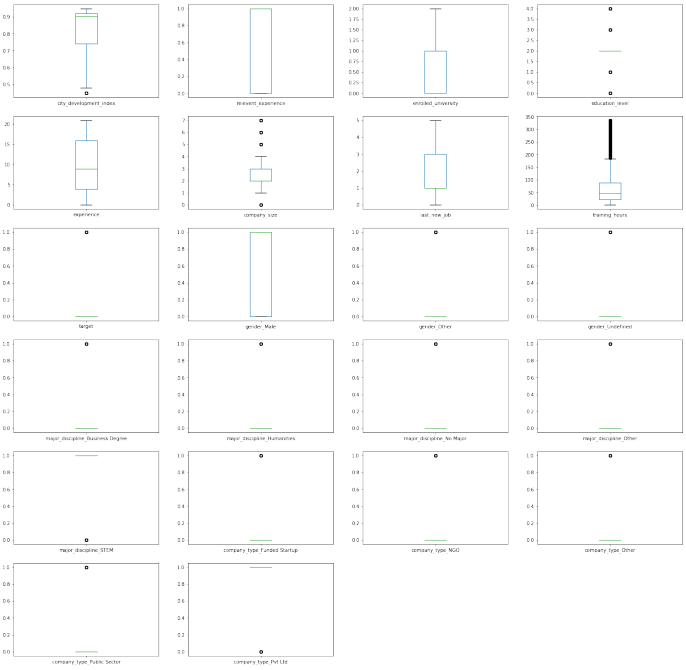

Fig. 6. Outliers in Dataset 3

Rational for choosing Naïve bayse:
- As the dataset had lots of categorical variables and naïve bayes performs well with categorical varibales.
- A Naive Bayes classifier outperforms other models such as logistic regression and needs less training data when the assumption of independence is met.
Rational for choosing KNN:
- KNN is lazy learner algorithm, and it is one of the simplest machine learning algorithm based on supervised machine learning
- During the training process, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a group that is very close to the new data.

2) Dataset 2: HR Analytics : Employee promotion data: For this dataset model was build using Logistic Regression and naïve bayes, Following are the rational for choosing the algorithms
Rational for Logistic Regression:
- It is used for both classification and regression. Logistic regression, like linear regression, uses an equation as its representation.
- The objective or dependent variable is diverging in nature, implying that there are only two possible groups.

3) Dataset 3 - HR Analytics : Job change of Data Scientist: For this dataset once the dataset was cleaned Decision Tree and Random Forest Classifier was chosen to build models. Following are the rational for choosing them:
Rational for choosing Decision Tree :
- As this dataset consists of lots of categorical values, a decision tree is a tree structure that looks like a flowchart, but each internal node represents a function.
- In machine learning Decision tree is like a white box, internal
decision-making logic is shared, which is not present in black box algorithms like Neural Networks [19]

Rational for choosing Random Forest Classifier:
- It is used for both classification and regression as well as it is most flexible and easy to use. The trees make up a forest. A forest is said to be more durable the more trees it has. However, in this dataset there are lots of categorical values thus it helps producing best results as when trees are form.
- Compared to decision tree it model is difficult to interpret also it is slow in generating prediction as it has multiple decisions but however gives the best results.

E. Evaluation

- Confusion metrics : Confusion metrics helps to check whether when and model is given something to predict whether the model rightly predicts and classify it or no. The confusion metrics consist of following terms:
  True positive(TP) : Should predict positive and actually positive
  False positives (FP): Should predict positive and are actually negative.
  True negatives (TN): Should predict negative and are actually negative.
  False negatives (FN): Should predict negative and are actually positive.

- Accuracy : Accuracy is a statistic that sums up how well a model performs in all groups. It's helpful when all of the classes are equally relevant. The ratio between the number of accurate predictions and the total number of predictions is used to calculate it.

- Precision : It helps to evaluate that out of total predicted positive occurrence how many positive instances are. It solves as how much the model is right. TP/ (TP + FP)

- Recall : It helps to evaluate that out of total actual positive occurrences how many are positive instances. Here the equation is like TP / (TP + FN) , where (TP + FN) are the actual number of positive instances.

- F1 Score : F1 is mean of precision and recall, higher the F1score it is better.

1) Dataset 1: Adult Census income: For this dataset models were trained before and after feature selection, however it was observed when the model was trained without feature selection the model performed good. However, after feature selection the model was performed average.fig shows results before feature selection and fig. shows results after feature selection.

	Algorithm	Accuracy	Train Score	f1_Score	Recall_Score	Precision_Score
1	KNeighborsClassifier	0.845509	0.886581	0.862286	0.943595	0.793878
0	GaussianNB	0.839219	0.832478	0.848245	0.876642	0.821630

Fig. 7. Dataset 1 results before feature selection

	Algorithm	Accuracy	Train Score	f1_Score	Recall_Score	Precision_Score
1	KNeighborsClassifier	0.820900	0.862881	0.836770	0.895587	0.785202
0	GaussianNB	0.818252	0.813055	0.827232	0.848870	0.806669

Fig. 8. Dataset 1 results after feature selection

2) Dataset 2: HR Analytics : Employee promotion data: For this dataset the model did not perform well when naïve bayes was trained. However the model was able to classify though logistic regression. In order to check whether the model's performance was accurate or no. Confusion metrics was used to clarify the actual positive and predictive occurrences out of total positive occurrences

As seen in fig while the model was train through naïve bayes it failed to predict false negative, when the model had to predict 0 occurrence it predicted it as 1. However, on the other been observed that recall result was good.

The model was successfully able to classify true positive and true negatives. However just for false negative again the model was not accurate to predict it. Overall, for this dataset, logistic regression was able classify prime employees who should get promoted.

3) Dataset 3: HR Analytics : Job change of Data Scientist: After building model using Decision Tree and Random forest Classifier, evaluation methods like confusion matric, F1 score, Precision, recall, AUC etc are used to check machine learning model performance

Fig.7 is the result after training the dataset with decision tree and random forest. As seen, Random forest has good accuracy compared of to decision tree. However, random forest performed well in all aspect, The F1 score and precision is also good. As seen the fig. when model was given to predict to 0 it predicted all 0, also when model was given to predict 1 it

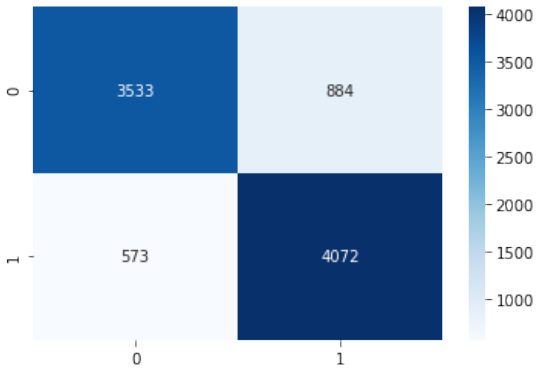

Fig. 9. Confusion metrics of Naive Bayes before feature selection

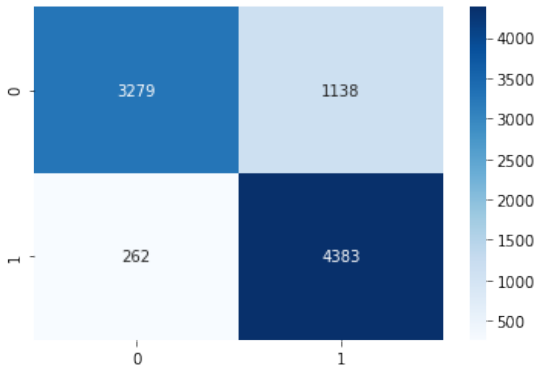

Fig. 10. Confusion metrics of KNN before feature selection

	Algorithm	Accuracy	Train Score	f1_Score	Recall_Score	Precision_Score
1	LogisticRegression	0.780104	0.769709	0.787285	0.803913	0.771331
0	GaussianNB	0.674769	0.769709	0.743817	0.932748	0.618532

Fig. 11. Dataset 2 results

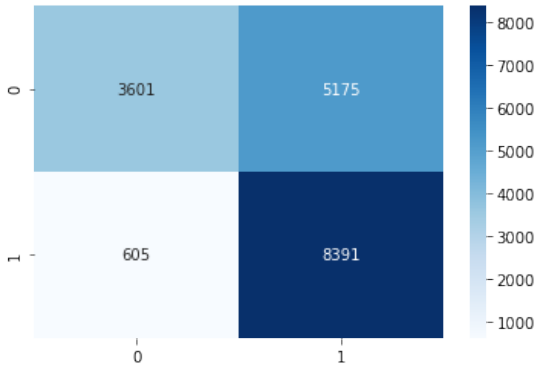

Fig. 12. Confusion Metrics for naive bayes classifier

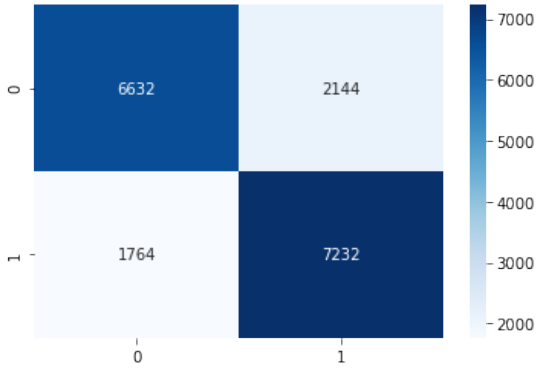

Fig. 13. Confusion Metrics for Logistic Regression

	Algorithm	Accuracy	Train Score	f1_Score	Recall_Score	Precision_Score
1	RandomForestClassifier	0.809100	0.998213	0.810913	0.822129	0.800000
0	DecisionTreeClassifier	0.757322	0.807035	0.765182	0.794118	0.738281

Fig. 14. Dataset 3 results

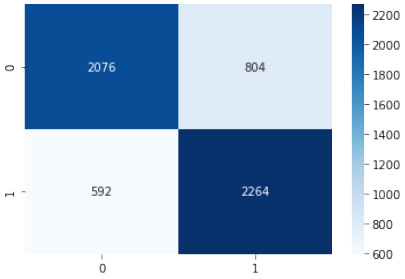

Fig. 15. Confusion Metrics for Decision tree

predicted 1. However, when model was said to predict 0 It predicted 1 as 804 number of time, on the other hand when model was given to predict 1 it predicted 0 as 592 number of times.

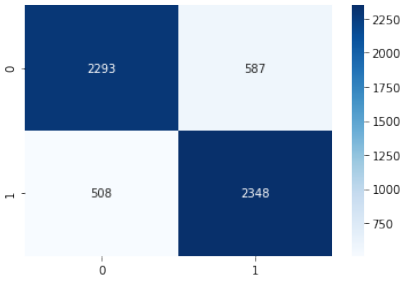

Fig. 16. Confusion Metrics for Random Forest Classifier

Also when the model was train using random forest classifier, it predict the actual positive very well as seen in the fig. Confusion metrics demonstrates that the model has performed good enough and producing the right results. Thus, it can be said that model is able to classify number of employee who will stay in company and number of people who are looking for job change.

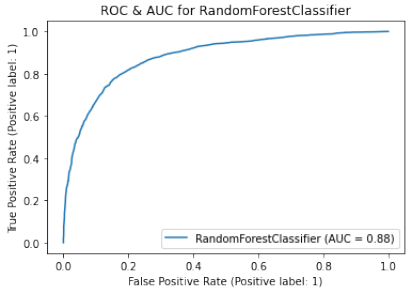

Fig. 17. ROC & AUC for DecisionTreeClassifier

IV. Conclusions and Future Work

Thus, through machine learning techniques the project can help workforce analytic to manage their employees. Various machine learning algorithms were used to build the model for all three detests. All research questions were solved as the model was able to classify the occurrences. By using machine learning technique and CRISP-DM methodology companies can identify who are their prime employees, also company can check who deseves the promotion and which employees are looking for job change. During building models, Support Vector Classifier did not run properly, so it can be done in the future work. Also, Adaptive boosting was not used due to time constraint,can be used in future work to get the best results.

References

[1] S. Das, R. Barik, and A. Mukherjee, "(PDF) Salary Prediction Using Regression Techniques," ResearchGate, 01-Jan-2020. [Online].[Accessed: 01-May-2021].

[2] Navyashree M, Navyashree M K, Neetu M, Pooja G R, Arun Biradar, "Salary Prediction in It Job Market," International Journal of Computer Sciences and Engineering, Vol.07, Issue.15, pp.78-84, 2019.

[3] V. Jagun, "An Investigation into the High Turnover of Employees within the Irish Hospitality Sector, Identifying What Methods of Retention Should Be Adopted," NORMA@NCI Library, 02-Sep-2015. [Online]. Available: http://norma.ncirl.ie/2096/. [Accessed: 01-May-2021].

[4] Gunawan H, Amalia. R, "Wages and Employees Performance: The Quality of Work Life as Moderator," International Journal of Economics and Financial Issues. 5. 349-353.

[5] A. Tadesse, "The Effect of Employee Promotion Practice on Job Satisfaction: The Case of Dashen Bank S.C.," AAU-IR. [Online] [Accessed: 02-May-2021].

[6] Dr. Latha L," A Study on Employee Attrition: Inevitable yet Manageable," International Journal of Business and Management Invention, ISSN (Online): 2319 – 8028, Volume 6 Issue 9

[7] N. An, J. Liu, L. Wang and Y. Bai, "Employee Satisfaction as an Important Tool in Human Resources Management," 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008, pp. 1-4, doi: 10.1109/WiCom.2008.1677.

[8] Long, Yuxi & Liu, Jm & Fang, Ming & Wang, Tao & Jiang, Wei, "Prediction of Employee Promotion Based on Personal Basic Features and Post Features," ICDPA 2018: Proceedings of the International Conference on Data Processing and Applications. 5-10. 10.1145/3224207.3224210.

[9] L. Liu, S. Akkineni, S. Paul,D. Clay," Using HR Analytics to Support Managerial Decisions: A Case Study," Proceedings of the 2020 ACM Southeast Conference, 01-Apr-2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3374135.3385281. [Accessed: 01-May-2021].

[10] T. Attri, "Why an Employee Leaves: Predicting using Data Mining Techniques," NORMA@NCI Library, 13-Aug-2018. [Online]. Available: http://norma.ncirl.ie/3434/. [Accessed: 01-May-2021].

[11] A. C. C. de Jesus, M. E. G. D. Júnior, and W. C. Brandão, "Exploiting linkedin to predict employee resignation likelihood," in Proceedings of the 33rd Annual ACM Symposium on Applied Computing - SAC '18, 2018.

[12] B. A. Sarker, S. M. Shamim, M. S. Zama, and M. M. Rahman, "Employee's performance analysis and prediction using K-means clustering & decision tree algorithm," Core.ac.uk. [Online]. Available: https://core.ac.uk/download/pdf/231150638.pdf. [Accessed: 01-May-2021].

[13] Quantum, "Data science project management methodologies," 20-Aug-2019. [Online]. Available: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb [Accessed: 01-May-2021]

[14] "Your Machine Learning and Data Science Community," Kaggle. [Online]. Available: https://www.kaggle.com/. [Accessed: 01-May-2021].

[15] U. C. I. M. Learning, "Adult Census Income," Kaggle, 07-Oct-2016. [Online]. Available: https://www.kaggle.com/uciml/adult-census-income. [Accessed: 01-May-2021].

[16] Möbius, "HR Analytics: Employee Promotion Data," Kaggle, 23-Dec-2020. [Online]. Available: https://www.kaggle.com/arashnic/hr-ana. [Accessed: 01-May-2021].

[17] Möbius, "HR Analytics: Job Change of Data Scientists," Kaggle, 07-Dec-2020. [Online]. Available: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists. [Accessed: 01-May-2021].

[18] Sunil RayI am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years., "Learn Naive Bayes Algorithm: Naive Bayes Classifier Examples," Analytics Vidhya, 18-Oct-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. [Accessed: 01-May-2021].)

[19] "Tutorials for data scientists: DataCamp," DataCamp Community. [Online]. Available: https://www.datacamp.com/community/tutorials/decision-tree-classification-python). [Accessed: 01-May-2021].