IMPLEMENTING DATA-DRIVEN HR ANALYTICS USING DATASCIENCE

Allaboina Manisha Yadav
Department of Information Technology
Sreenidhi Institute of Science and Technology, Hyderabad

Sunil Bhutada
Department of Information Technology
Sreenidhi Institute of Science and Technology, Hyderabad

Abstract- HR Analytics is a way of amassing and analyzing HR data to amend an organization's performance. HR analytics can perhaps carry incredible value to HR decision-making on employees and organizations by enhancing intuition and involvement. In this paper, we come up with an end to end approach to HR analytics that contributes to quality data for management's decisions. As a part of this approach, we focus on five inclusive challenges in the HR department; Consequently, we present techniques that can constrain these challenges. As has been noted, the accumulation of multiple fields yields authentically incipient intuitions: Subsequently, we prepare the data set that possibly is a coalescence of HR data, Survey data, and Management data. Further, we analyze the results such that HR can make decisions predicated on evidence and not on convictions; At first, we perform Text Analytics, by executing the Resume Parser which certainly screens the resumes and indeed spares the manual efforts while recruitment; Secondly, we cluster the data in a Wordcloud that helps in talent identification. Additionally, we transform the data to get more insight into the data. As a part of transformations, we convert Date of birth to age and salaries to other preferred currency rates. Next, we perform Sentimental Analytics on the data obtained from the employee survey and latter identify the key performance indicators. In the end, we visualize the data on an interactive dashboard so that HR makes the right decision based on evident data.

Keywords –Key performance indicators, Predictive power score, Resume Parser, Sentimental Analysis, Target Variable Analysis.

I. INTRODUCTION

HR analytics is the science that integrates all the data relating to HR functions like recruitment, talent identification, data collection, data analysis, and executing business strategies. For an organization to be lucrative and run prosperously, employees are a valuable asset [1]. The process of hiring the right person for the right job is profoundly paramount. A company spends its precious time and money on recruiting the employees; And if suddenly an employee leaves the company, it results in loss as the company loses a valuable asset; And now the company has to again invest its time in recruiting another employee in his position; And train the employee such that he is productive enough. An employee leaving the company is Attrition [2] and is one of the solemn quandaries in the HR. Sundry reasons for Employee attrition like voluntary, involuntary, and retirement. Voluntary attrition [3] is the one on which the HR has to focus while involuntary and retirement are inevitably ineluctable. The employee goes for voluntary retirement when he is not satisfied with the management and have any personal issues. If HR can visually perceive such quandaries, the company might preserve one of the valuable assets, a talented employee for his company. Moreover, the company should not leave the retired employees but, keep them as a component of their advisory board. Furthermore, their suggestions are predicated on their experiences and can become the best consultants [4] for the company. More the attrition rate more will be the loss to the company.

Recruiting and retaining a talented and adept person is one of the major tasks of HR. To recruit a person is one of the tough tasks for HR; This process involves talent hunting, magnetize the talent, invest time and money, and balance the business and the hiring in parallel. After the recruitment, it is also consequential that assigning the right person with the right job which he is adroit in; If not, the employee shows no interest in his job and will not be much productive and lucrative. For this purpose, it is significant that the management of the company takes measures to optically discern to designate the employee in the best designation as per his skills. The HR should go through the performance of every employee at customary intervals of time and as a component of inspiration, it should appreciate the best performing employees. As a consequence, the other employees withal get incentivized and endeavor their best. Therefore, we can understand that the functions, orchestrating, and strategies of HR are predicated on the data and from the analysis of that data. Thus, HR analytics is a data-driven process and if it can be executed well becomes an end to end solution to the HR beginning from the recruitment, development, and Retention of an employee.

II. LITERATURE REVIEW

In this corporate world, it is not an exaggeration to verbally express that HR Analytics is a game-changer. The only thing that makes a distinction in the corporate world is the efficacious and efficient strategies of HR. The way HR is addressing the strategies to ameliorate the overall development of the company has been varying over the decades. Still, some major issues are taking a different form and stand as challenges before the HR. Bhawna Gaur and Sadia Riaz have compared HR Analytics to Artificial Intelligence [5] which utilizes the test of cognitive faculties, adaptability, and productivity. Boston Consulting Group has put forward many implicative inferences for the future [6] by introducing AI at the workplace. This has amended the performance and the Managers could expect competitive advantages by deciphering the prominent features in their strategies. For this, companies need to understand how computers and humans can develop each other's strengths. Every company takes a look at digital transformation. Artificial Intelligence, Machine Learning, IoT, and other technologies have established a sizably voluminous transformation in HR practices. Wearable IoT devices [7] like wrist bands, biometrics, and real-time data sensing contrivances were utilized by the companies to track the data of the employees. With the avail of Biometrics, the management could capture the login and log out timings and the efficacious working hours of an employee to track the productivity of an employee. With the wearable IOT wrist bands [8] they endeavored to capture and monitor the heart rate and the ThermoSensors which monitor the body temperature. By this, the management could additionally monitor the health of the employees. All the data accumulated from the employees was uploaded into the HRMS implement accessible by both the employee and the management. In this way, IoT has brought a transformation [9] of Human Resource Management in the workplace. Regardless of taking effective measures there exists additionally earnest issues in HR. Some of the typical quandaries [10] in HR analytics include identify the cognizance and skills in the organization, estimate the churn rate, manage the data, prognostication of success. The next section of the paper discusses identically tantamount.

III. CHALLENGES IN HR

HR Analytics is a data-driven process. A company has variants of data like recruitment data, training data, personal data, productivity data, financial data, day to day evaluation data, and the traditional HR datasets. HR handles all these data and is incredibly valued [11]. In the past, this data was usually unused or just arranged in rows and columns or put in tables and charts. Today in this digital world, we have concepts like Big data Analytics, Artificial intelligence, and Data Science, and various techniques can be implemented on the HR data to get more out of the data collected. It all depends on the way we use the data which can add increased productivity and improve the performance of an individual. So, the data collected can answer many questions of HR and additionally assist with the challenges in HR. Correct business decisions are the result of quality data

The significant challenges in HR analytics include:

a) Data cleansing and quality data
b) Identifying the skills and knowledge of the employees
c) Estimating the Attrition Rate
d) Predicting success

Any organization depends on its employees and its customers. Comprehending the employees and then exploring the digital world is the best strategy for an organization. HR collects data from its employees since the time of recruitment until their attrition. Interpreting and utilizing this data is a significant concept focused on HR Analytics.

For the efficient functioning of HR, it is significant to use the Data wisely. Subsequently, HR should take a few steps in collecting more data at times by taking surveys at various levels of management. Consequently, through analysis, the collected data contributes a good insight into what an employee wants from his Administrators. Finally, the company records quality data. However, it’s a challenge to HR that how it should identify quality data.

Identifying the skills of an employee are also essential for HR Analytics. The company recruits a person primarily based on his talent required during the requirement. Additionally, he may have other complementary skills, but maybe of avail later. Suppose that there is any new project that demands specific skills, the Management should be ready to list the subsisting employees who can showcase their talents. Thus, the company can save time and money by identifying the skills of the organization.

Attrition, however, is one of the earnest issues in HR. Handling attrition wisely is a must, though concluded for an organization. If the company doesn't look at progression attrition, it results in a decline in productivity. To expel this, HR has to focus on the attributes that cause Attrition or the factors which may reduce the attrition rate.

Even the much intricate recruitment process is many times inefficient in predicting the success of the candidate. There can be many attributes that need to be verified for the success of a job. The company should incorporate success predictors based on the required skills along with cognitive assessments, skill tests, and previous profiles. These predictors should be used at regular intervals of time so that the company can assess itself. If there are undershoots, subsequently, Management can take the necessary measures. Now it's again a task for the HR to identify the attributes that may predict success.

IV. ARCHITECTURE DESIGN

The System design gives an end to end solution to HR. The below figure Fig.1 explains the flow of the process.

Initially, we perform Text Analytics in the context of Recruiting to screen the resumes and save the manual effort. Secondly, we perform target variable analysis in the context of Talent identification considering the identification of best-performing employees. Subsequently, we analyze the data obtained from the surveys conducted by the manager through Sentimental analysis. Latter we identify the key performance indicators by data clustering. Finally, we visualize the data such that HR can make decisions based on planning compensation to enhance productiveness.

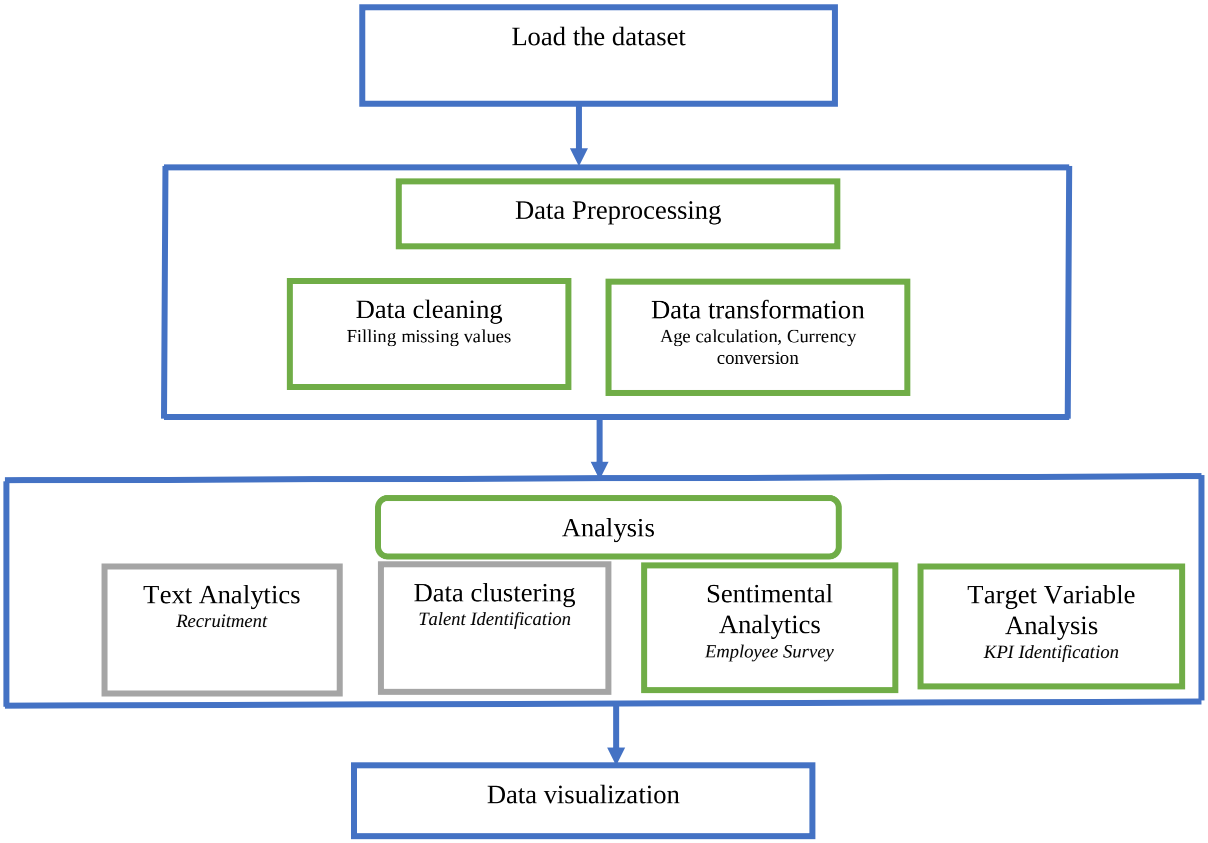


Figure 1. Image showing the architecture of the Data driven HR Analytics

4.1 Text Analytics

Resumes are unstructured documents that certainly enlist the personal and professional skills of a person. Briefly, a resume summarizes the professional and educational history of a person. A resume parser [12] is a program that significantly converts the unstructured data into structured data by extracting the knowledge and standardizing the information. This information indeed can be utilized for the evidence-based recommendation of the resume.

The input to the resume parser is primarily a resume, which can be PDF or Word formats. So, the resume parser design fortifies any of these formats. At first, we scan the documents. Subsequently, we have to extract the information from the resume. We do this by specifying keywords like name, education, email, skills. When the keywords match the words in the resume, the parser shall extract the data and exhibits the result as a data frame.

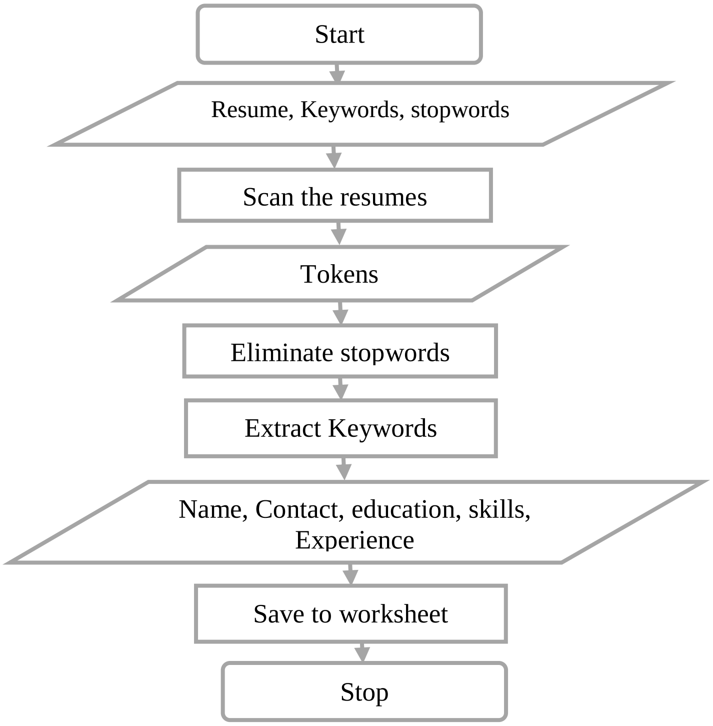


Figure 2. Flow chart of Resume Parser

4.2 Data clustering

Data clustering, a typical unsupervised learning technique for statistical analysis, is the process of creating clusters. A cluster is a group of similar and related things that indeed provides an incredible insight into the data. Several clustering algorithms are available based on the data clustered and the purpose of analysis.

Accordingly, Word Cloud is a data-clustering technique based on word frequency; The size of the words is proportional to their frequency in the given text. Word clouds can certainly help to identify the keywords. Moreover, the Word cloud visualization technique can reveal the patterns in a given text.

Figure3. Image showing a Wordcloud

4.3 Sentimental Analysis

Sentimental Analysis is a method of text analytics that significantly analyzes and interprets the sentiments in a given text. Sentimental Analytics is a straightforward technique. Initially, we split the given text into tokens and phrases, and identify the phrases holding sentiments. Latter we assign scores to these phrases based on a predefined scale. Consequently, based on the values we assign to the phrases, we identify the emotions. Sentimental Analytics extracts biased data through intent analysis. Furthermore, Employee sentiment [13] analysis can be the answer to many questions.

4.4 Target Variable Analysis

Target variable analysis determines how the feature variables are affected by the target variable; Feature variables form the building blocks for the analysis whereas, the Target variable is one of the feature variables on which we base our analysis. The target variable varies as per the business requirements, goals, and availability of the data.

In this specific instance, we have the feature variables like monthly salary, total working hours, job role, and attrition; If attrition is our target variable, we can analyze how the feature variable monthly income effect the target variable, attrition. Ultimately, the management can minimize the impact of the target variable on the other feature variables by planning compensations.

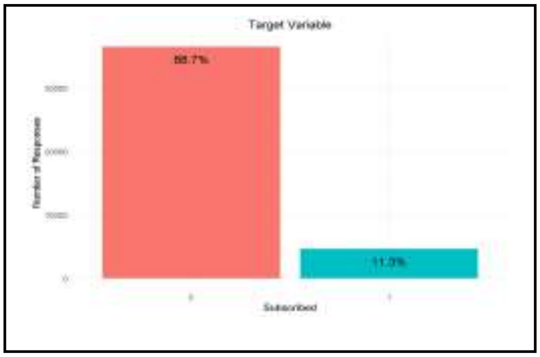


Figure 4. Target variable Analysis

V. RESEARCH METHOD

Meaningful insight into the data makes significant metrics. Consequently, it is necessary to have a keen comprehension of data amassed and the context of its collection. For this, we conduct analyses such that we get a good insight. At the beginning of the process, we start with Resume Parsing. Resume Parser is a code inscribed in Python that utilizes the inbuilt packages like, PDF miner and Spacy. The package PDF miner is a text extraction implement that converts PDF formats into HTML, XML, or Word Formats. As the resumes submitted may be of different forms, we convert them to an analyzable format. Spacy is a python library that helps in Natural Language

Processing. For the tokenization of the resumes, we make use of the library Spacy. These tokens help in comparison with the keywords. For comparison, we make use of the Matcher. We further download the list of stopwords. The parser eliminates the stopwords and extracts the keywords. The keywords are the words that help us gain insight from Resumes like education, skills, experience. The output of the Resume parser shall be the data frame that lists all the required data from the resume and eliminating the rest. Thus, we perform screening of the resumes and thereby preserve manual efforts.

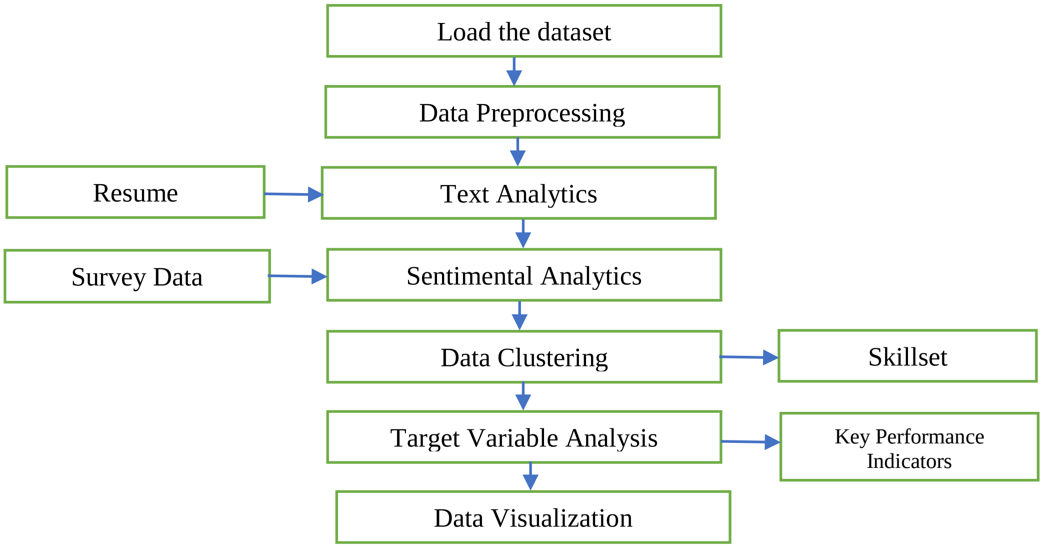


Figure 5. Image showing the process flow

Once the data frame gets saved into the worksheet, we have the additional skills of the employee in contrast to the skills he got selected in recruitment. Accordingly, we perform clustering on the available skills; For which we implement a Wordcloud. Wordcloud predication is on the groundwork of the word frequency. We can, therefore, get an idea of all the available skills in the organization. We can, furthermore, suggest the employees having specific skills at the onset of the new job description. Consequently, we formulate a code as per the requirement that lists the available skills. We make use of the python library Wordcloud and also counter to generate a Wordcloud.

If we are planning for analysis, we need to get ready with the dataset. For this, we need to clean and preprocess the data [14]. In the dataset used, there are some missing values that we have to handle. For making the best use of the data available, we transform the data such that we can get valuable insight. In this specific instance, we consider the attribute date of birth that, however, has nothing to do with analysis. In this case, we considerably transform the attribute date of birth to Age that preferably has a vital role to play. This process requires packages like DateTime and requests. As a consequence, we have a new field, Age added to our dataset. Currency conversion is also performed on the attribute salary, which is one of the key factors that influences the business. Eventually, our data is ready for analysis.

As a part of our process, we perform sentimental analysis on the Survey data. The survey conducted was at the manager level, which has the opinion of the employees on their job and management recorded on a five-point scale. This survey data is assumed to be the foundation of our analysis. Through the sentimental analysis, we can have an insight into the satisfaction levels of the employee, performance ratings, and decision on attrition. We can further make use of this analysis for predicting the attrition rate and compensation planning to preserve a talent.

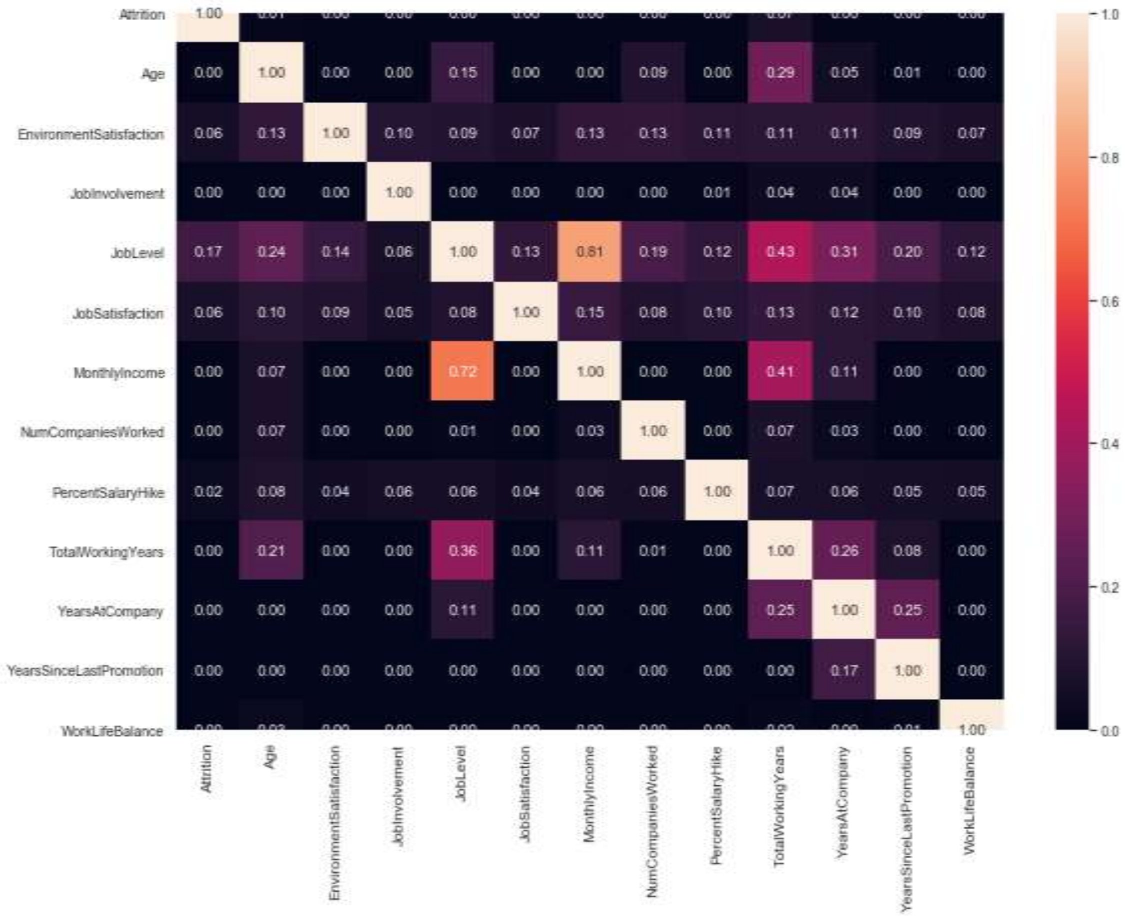


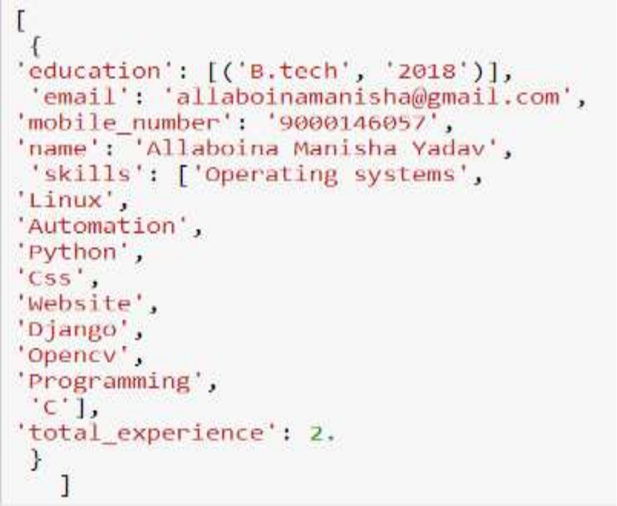
Figure 6. Image showing the Predictive Power Score for attrition

The next step in our process is to identify the key performance indicators. Key performance indicators [15] are, in fact, the features that influence the workflow. To order to obtain the Key performance indicators, we perform the predictive power score analysis as a part of Target variable analysis. The Predictive Power Score is significantly the quick data exploratory asymmetric technique that determines the linear and nonlinear relationships of the data correspondingly on a score ranging from 0 to 1. The representation of this score is on a symmetric matrix. For this, we acknowledge the most buzzing word, attrition as our target variable, and indeed obtain the relationship of the attribute attrition with the other attributes. We use the libraries' pandas, seaborn, and NumPy for the matrix. Thus, we perform the target variable analysis using the predictive power score and feature selection.

We finally visualize our findings on the interactive dashboard developed using the dash library and its components. The results of our analysis can not only be understood by the data scientist but also by the manager.

VI. Results

This section emphasizes the results obtained at each level of the process, Data-Driven HR Analytics. Starting with the resumes and recruitment, we have succeeded in analyzing the resumes using Text Analytics. The Resume Parser could extract useful data from the Resume and present the data frame. The output of the Resume Parser, the data frame is shown below in Fig.7.



```
[
 {
'education': [('B.tech', '2018')],
 'email': 'allaboinamanisha@gmail.com',
'mobile_number': '9000146057',
'name': 'Allaboina Manisha Yadav',
 'skills': ['Operating systems',
'Linux',
'Automation',
'Python',
'Css',
'Website',
'Django',
'Opencv',
'Programming',
 'C'],
'total_experience': 2.
 }
   ]
```

Figure 7. Output of Resume Parser

We have extracted Education, email, contact number, name, skills, and experience from the resume. This data frame is exported to the excel by using the python packages for ETL like xlrd and xlwt.

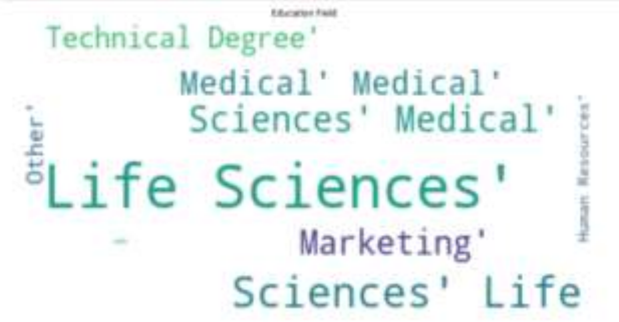


Figure 8. Word cloud of available Talents

Secondly, we identified the skills of the organization by clustering analysis and consequently generating the word cloud shown in Fig 8. We have also displayed the skills on the bar graph (Fig. 9). The results are shown below.

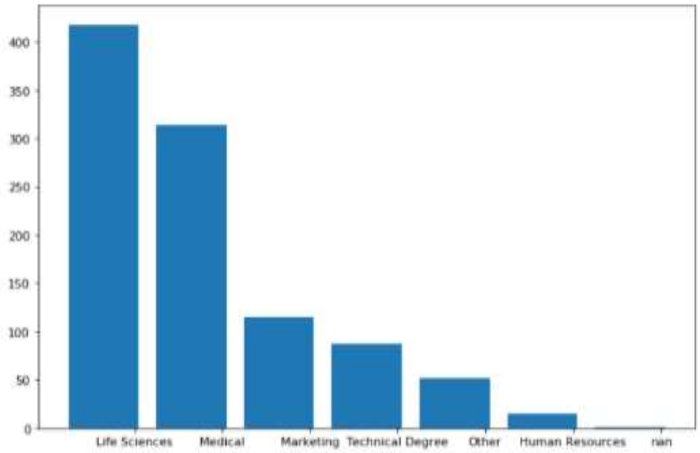
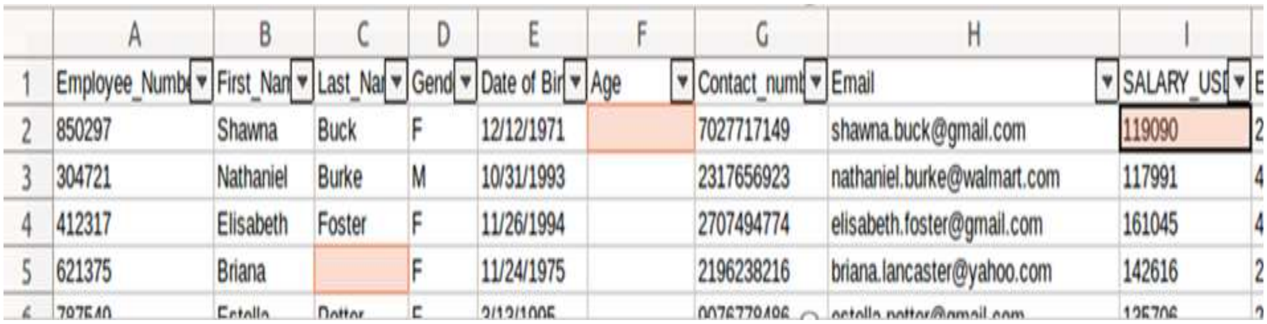


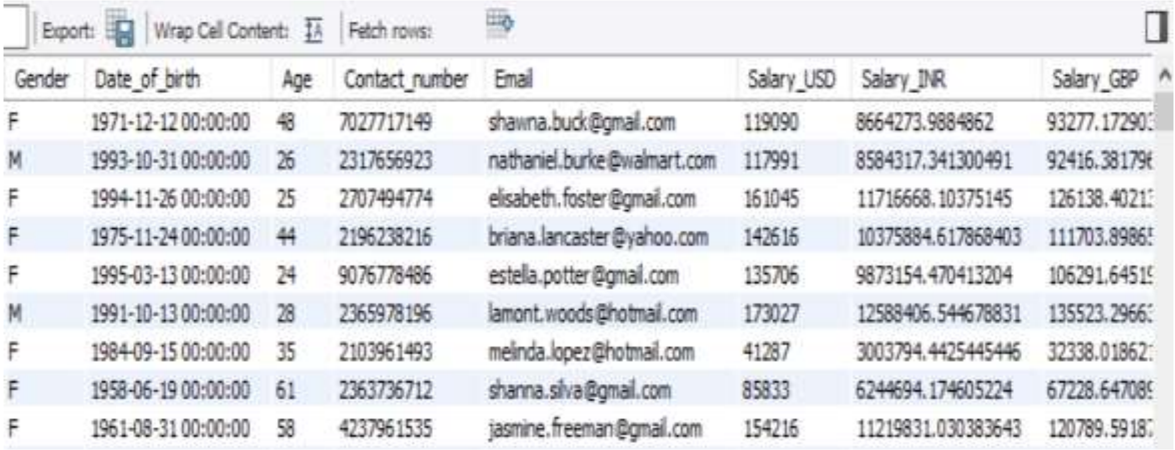
Figure 9: Image showing the frequency on talents based on which the word cloud is formed.

Consequently, we have preprocessed the data to make it ready for analysis. As we have said earlier, we have transformed the date of birth to age and filled the missing last names and first names of the employees. The results are displayed below;



	A	B	C	D	E	F	G	H	I	
1	Employee_Numb	First_Nan	Last_Na	Gend	Date of Bir	Age	Contact_numb	Email	SALARY_USI	E
2	850297	Shawna	Buck	F	12/12/1971		7027717149	shawna.buck@gmail.com	119090	2
3	304721	Nathaniel	Burke	M	10/31/1993		2317656923	nathaniel.burke@walmart.com	117991	4
4	412317	Elisabeth	Foster	F	11/26/1994		2707494774	elisabeth.foster@gmail.com	161045	4
5	621375	Briana		F	11/24/1975		2196238216	briana.lancaster@yahoo.com	142616	2
6	787540	Estella	Potter	F	3/13/1995		9076778486	estella.potter@gmail.com	135706	2

Figure 10. Image showing the datasheet before preprocessing and having missing values.



Gender	Date_of_birth	Age	Contact_number	Email	Salary_USD	Salary_INR	Salary_GBP
F	1971-12-12 00:00:00	48	7027717149	shawna.buck@gmail.com	119090	8664273.9884862	93277.17290
M	1993-10-31 00:00:00	26	2317656923	nathaniel.burke@walmart.com	117991	8584317.341300491	92416.38179
F	1994-11-26 00:00:00	25	2707494774	elisabeth.foster@gmail.com	161045	11716668.10375145	126138.4021
F	1975-11-24 00:00:00	44	2196238216	briana.lancaster@yahoo.com	142616	10375884.617868403	111703.8986
F	1995-03-13 00:00:00	24	9076778486	estella.potter@gmail.com	135706	9873154.470413204	106291.6451
M	1991-10-13 00:00:00	28	2365978196	lamont.woods@hotmail.com	173027	12588406.544678831	135523.2966
F	1984-09-15 00:00:00	35	2103961493	melinda.lopez@hotmail.com	41287	3003794.4425445446	32338.01862
F	1958-06-19 00:00:00	61	2363736712	shanna.silva@gmail.com	85833	6244694.174605224	67228.64708
F	1961-08-31 00:00:00	58	4237961535	jasmine.freeman@gmail.com	154216	11219831.030383643	120789.5918

Figure 11. Result of Preprocessing and data transformation

As discussed earlier, we considered the attribute attrition as the target variable and have performed the target variable analysis. As a part of this analysis, we have calculated the Predictive power score and analyzed how the other feature variables attribute influence attrition.

Table-1. Table displaying the Predictive Power Score

X: Attrition	Y: job satisfaction	Predictive Power Score
Monthly income	Job level	0.81
Total working Years	Monthly income	0.41
Age	Job level	0.24
Total working Years	Job level	0.43
Years since last promotion	Years at the company	0.25
Job level	Total working Years	0.36

From the above table (Table 1), it is clear that Monthly income and corresponding job-level have the highest PPS, symbolizing the influence of these attributes on attrition. Next, we have total working years and job level with high PPS. We latter predict the attrition rate using the Random Forest Classifier, and the F1 score is given below table 2.

Table-2. Table displaying the F1 score of Random Forest Classifier

	Precision	Recall	F1-score	support
0	0.90	0.93	0.91	245
1	0.57	0.49	0.53	49
accuracy			0.85	294
Macro average	0.74	0.71	0.72	294
Weighted average	0.85	0.85	0.85	294

With this Random forest classifier, we obtain the feature importance so that we can identify the Key performance indicators. From the figure, the top five performance indicators are no Over Time, StockOptionalLevel, Job Level, Monthly Income, and Job Satisfaction.

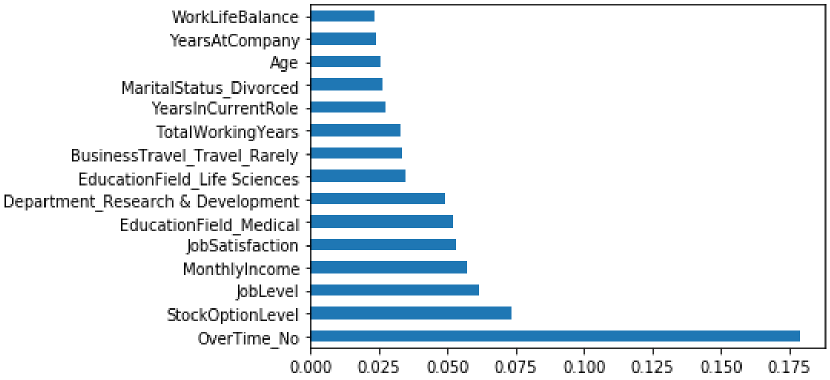


Figure 12. Image showing the key performance indicators

The following image (Fig. 13) shows the dashboard, which displays statistics of the company. This interactive dashboard assists the manager in identifying the factors that affect the performance and those influencing the attrition rate. A glimpse of the dashboard outfits a complete view of the business. We design the dashboard in such a way that almost every data field is involved in this visualization. The dashboard visualizes the comprehensive report of the Analytics.

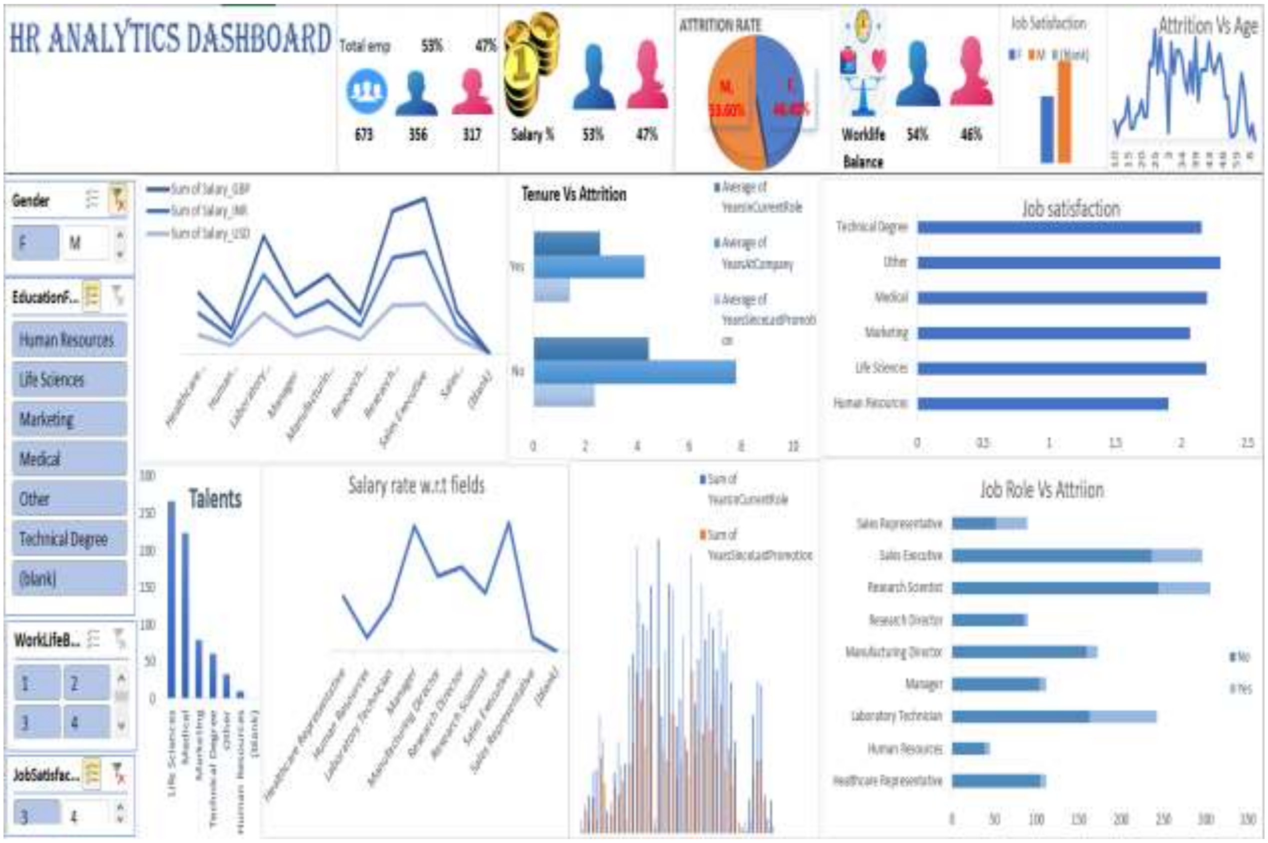


Figure 13. Image of the HR Analytics dashboard

VII. CONCLUSION

To conclude, Data Science is ruling this world with no bounds. Irrespective of the field of research, it can be made use of anywhere. The HR department is one that deals with more and more and data. We come up with a module which helps the HR in analyzing their data. This paper has focused on the significant issues and challenges in HR. The proposed system Data-Driven HR Analytics developed based on Data Science enhances and simplifies the performance of HR. The process was followed by analyzing the data and gaining incredible insight into the data. Correspondingly, the dashboard visualizes the results such that HR can make immediate decisions and implement strategies.

REFERENCES

[1] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari, "Evaluation of Machine Learning Models for Employee Churn Prediction" in Proceedings of the International Conference on Inventive Computing and Informatics, 2017.

[2] Neil Brockett, Catriona Clarke, Michele Berlingerio, Sourav Dutta, "A System for Analysis and Remediation of Attrition", IEEE International Conference on Big Data (Big Data), 2019.

[3] Moninder Singh, Kush R. Varshney, Jun Wang, Aleksandra Mojsilovic, Alisia R. Gill, Patricia I. Faur and Raphael Ezry, "An Analytics Approach for Proactively Combating Voluntary Attrition of Employees", IEEE 12th International Conference on Data Mining Workshops, 2012.

[4] Chiradeep BasuMallic, What Is Employee Attrition, HR Technologist.

[5] Bhawna Gaur, Sadia Riaz, "A Two-Tier Solution to Converge People Analytics into HR Practices" in 4th International Conference on Information Systems and Computer Networks (ISCON) GLA University, Mathura, UP, India, Nov 21-22, 2019.

[6] Reshaping Business with Artificial Intelligence, Findings from the 2017 Artificial Intelligence Global Executive Study and Research Project, (2017).

[7] Stefan Strohmeier, "Smart HRM – a Delphi study on the application and consequences of the Internet of Things in Human Resource Management", in The International Journal of Human Resource Management, 2018.

[8] Bhawna Gaur, Vinod Kumar Shukla and Amit Verma, “Strengthening People Analytics through Wearable IOT Device for Real-Time Data Collection” in International Conference on Automation, Computational and Technology Management (ICACTM) Amity University 2019.

[9] A Narasima Venkatesh, “Connecting the Dots: Internet of Things and Human Resource Management”, International Association of Scientific Innovation and Research (IASIR), USA, 2017.

[10] K. Simbeck, “ HR analytics and ethics”, IBM Journal of Research and Development, Volume: 63 , Issue: 4/5 , July-Sept. 1 2019.

[11] Bernard Marr, “Why Data Is HR's Most Important Asset” in Forbes 2018.

[12] Andrea De Mauro “Human Resources for Big Data Professions: A systematic Classification of Job Roles and Required Skill Sets”

[13] Rai B Shamantha, Sweekriti M Shetty, Prakhyath Rai, “Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance” in IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019.

[14] Warren R. Greiff, “The use of Exploratory Data Analysis in Information Retrieval Research” in Advances in Information Retrieval, pp 37-72, 2002.

[15] Adil Baykasoğlu, Zehra Nur Atalay, İlker Gölcük, “Analysis of key performance indicators in a manufacturing plant via fuzzy cognitive maps” in Innovations in Intelligent Systems and Applications Conference (ASYU), 2019.