

DATA-DRIVEN HR ANALYTICS DEPLOYMENT USING DATA SCIENCE

Allaboina Manisha Yadav ¹, Sunil Bhutada ²

¹Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Email: allaboinamanishayadav@gmail.com

²Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Email: sunilb@sreenidhi.edu.in

Abstract: HR Analytics is a way to amass and evaluate HR data and change the efficiency of an enterprise. Through developing insight and engagement, HR analytics will maybe add incredible benefit to HR decision-making for workers and organizations. In this post, we come up with an end-to-end approach to HR analytics that leads to quality data for the decisions of management. We concentrate on five inclusive issues in the HR department as part of this approach; accordingly, we present strategies that can restrict these challenges. As noted, the aggregation of several fields creates truly incipient intuitions: we then plan the data collection that could be a coalescence of data from HR, survey data, and management. In addition, we interpret the findings in such a manner that HR can make decisions based on facts and not on convictions; First by running the Resume Parser, we conduct Text Analytics, which definitely screens the resumes and potentially spares the manual efforts during recruitment; Second, we cluster the data in a Word cloud that helps to recognize talent. Additionally, to get more insight into the data, we convert the data. We translate the date of birth to age as part of transformations and wages to other preferred currency values. Next on the data collected from the employee survey, we conduct Nostalgic Analytics and later determine the main success metrics. In the end, on an interactive dashboard, we visualize the information so that HR makes the right decision based on obvious 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 information related to HR functions such as hiring, identifying talent, collecting data, analyzing data, and executing business strategies. Employees are a valuable asset for an organization to be lucrative and run prosperously [1]. The process of hiring the right individual for the right job is deeply important. A company spends its valuable time and money on recruiting employees; and if an employee leaves the company suddenly, it results in loss as the company loses a valuable asset; and now the company has to invest its time again in recruiting another employee in its position; and train the employee to be sufficiently productive. Attrition [2] is an employee leaving the company and is one of the solemn dilemmas in HR. There are various reasons for employee turnover, such as voluntary, involuntary, and retirement.

The HR has to concentrate on voluntary attrition [3], while involuntary and retirement is inevitably ineluctable. When he is not satisfied with the management and has any personal problems, the worker goes for voluntary retirement. If HR can perceive such dilemmas visually, one of the valuable assets, a talented worker for his company, could be preserved by the company. In addition the company should not leave retired employees, but should keep them as part of its advisory board. In addition, their suggestions are based on their experiences and may become the company's best advisors [4]. The greater the attrition rate, the greater the loss to the organization. One of HR's major tasks is to recruit and retain a talented and adept individual.

One of the tough tasks for HR is to recruit a person; this process involves hunting for talent, magnetizing talent, investing time and money, and balancing the company and hiring in parallel. It is also consequential after the recruitment to assign the right person to the right job in which he is proficient; if not the employee shows no interest in his job and will not be very productive and lucrative. For this purpose, it is important that the management of the company takes steps to optically discern the best designation of the employee according to his abilities. At normal intervals of time, the HR should go through the performance of every employee and as a component of inspiration; it should appreciate the best performing employees. As a result, the other employees get incentivized and try their best. We can therefore understand that HR's functions, orchestration, and strategies are based on the data and analysis of that information. HR analytics is therefore a data-driven process and becomes an end-to-end solution to HR starting from the recruitment, development and retention of an employee if it can be executed well.

II. LITERATURE REVIEW

It is not an exaggeration in this corporate world to express verbally that HR Analytics is a game-changer. In the corporate world, the only thing that makes a distinction is HR's effective and effective strategies. Over the decades, the way HR deals with strategies to improve the overall development of the company have varied. Nonetheless, some major problems take a different form and stand before the HR as challenges. Bhawna Gaur and Sadia Riaz compared HR Analytics with Artificial Intelligence [5], which uses cognitive skills, adaptability, and productivity testing. By introducing AI in the workplace, the Boston Consulting Group has put forward many implicit inferences for the future [6]. This has altered performance and by deciphering the prominent features in their strategies managers could expect competitive advantages. For this, companies need to understand how the strengths of each other can be developed by computers and humans. Each organization takes a look at digital transformation.

A significant voluminous transformation of HR practices has been established by Artificial Intelligence, Machine Learning, IoT and other technologies. Companies have used wearable IoT devices [7] such as wrist bands, bio-metrics, and real-time data sensing contrivances to track employee data. The management could capture the login and log out timings and the effective working hours of an employee with the use of Bio-metrics to track an employee's productivity. They attempted to capture and track the heart rate and the ThermoSensors that monitor the body temperature with the wearable IoT wrist bands [8]. Through this the management could also track the health of the staff. All the data collected from the staff was uploaded into the implementation of the HRMS accessible by both the employee and the management. In this way, IoT has introduced a transformation [9] of the management of human resources in the workplace. In addition, serious problems in HR exist regardless of taking effective measures. In HR analytics, some of the typical dilemmas [10] include identifying the organization's knowledge and skills, estimating the churn rate, managing the data, and forecasting success. Identical amounts are discussed in the next part of the paper.

III. CHALLENGES IN HR

HR Analytics is a process driven by data. A business has data variants such as recruitment data, training data, personal data, and data on productivity, financial data, day-to-day assessment data, and traditional HR datasets. All this information is handled by HR and is incredibly valued [11]. Usually, this information was unused in the past or just arranged in rows and columns or placed in tables and charts. Today, we have concepts like Big Data Analytics, Artificial Intelligence, and Data Science in this digital world, and different techniques can be applied to HR data to get more out of the collected data. It all depends on how we use the data, which can increase productivity and improve an individual's performance. Thus the collected data can answer many HR questions and also help with HR challenges. Correct business choices are the consequence of quality data.

In HR analytics, the important challenges include:

- a) Data cleansing and data quality
- b) Identifying the abilities and knowledge of the staff

- c) Estimating the rate of attrition
- d) Predicting achievement

Any organization is dependent on its staff and its clients. The best strategy for an organization is to understand the staff and then explore the digital world. From the time of recruitment until their attrition, HR collects data from their employees. A significant concept focused on HR Analytics is the interpretation and use of this information. It is important to use the data wisely to ensure the effective functioning of HR. Subsequently, by taking surveys at different levels of management, HR should take a few steps in gathering more information at times.

As a result, the collected data gives a good insight into what an employee wants from his administrators through analysis. Finally, quality data is recorded by the company. However it's a challenge for HR to determine how quality data should be identified. It is also important for HR Analytics to identify the skills of an employee. A person is primarily recruited by the company based on his talent required during the requirement. In addition, he may have other complementary skills, but later he may be of use. The management should be prepared to list the surviving employees who can showcase their talents if there is any new project that requires specific skills. Thus by identifying the organization's skills, the company can save time and money.

Attrition, however, is one of HR's serious problems. Handling attrition wisely is a must for an organization, even if concluded. If the business does not look at the attrition of progression, it results in a productivity decrease. HR must concentrate on the attributes that cause attrition or the variables that may decrease the rate of attrition in order to expel this. In predicting the success of the candidate, even the very intricate recruitment process is often inefficient. For the success of a job, there can be many attributes that need to be verified. Based on the required skills, the company should incorporate success predictors along with cognitive evaluations, skill tests, and previous profiles. In order for the company to assess itself these predictors should be used at regular time intervals. If undershoots exist, then the necessary measures can be taken by management. Now it is again a task for the HR to identify the characteristics that can predict achievement.

IV. ARCHITECTURE DESIGN

The architecture of the device provides HR with an end to end approach. Fig.1 in the diagram below describes the phase flow.

Initially, in the sense of recruitment, we conduct text analytics to screen the resumes and archive the manual effort. Secondly, in the sense of skill recognition, we execute goal variable research regarding the selection of the best-performing workers. Subsequently, by way of nostalgic research, we analyze the data collected from the surveys carried out by the manager. Later, by data clustering, we determine the main performance metrics. Finally, we visualize the data in such a way that HR can make decisions to increase efficiency dependent on planning compensation.

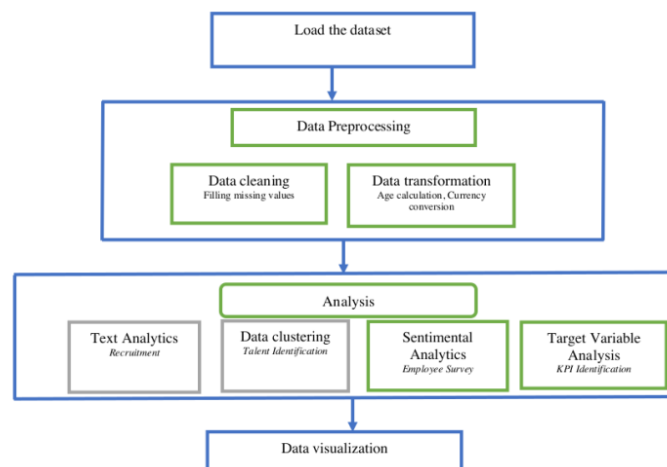


Fig 1: Image showing the architecture of the Data driven HR Analytics

4.1 *Text Analytics*

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

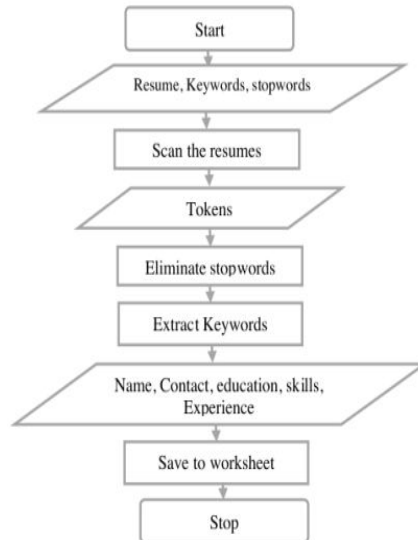


Fig 2: Flow chart of Resume Parser

The resume parser input is primarily a resume that can be in PDF or Word format. So, any of these formats is fortified by the resume parser design. We'll scan the documents first. Subsequently, from the resume, we must extract the data. By specifying keywords such as name, education, email, skills, we do this. The parser must extract the data and display the result as a data frame when the keywords match the words in the resume.

4.2 Data clustering

The process of creating clusters is data clustering; a typical unsupervised learning technique for statistical analysis. A cluster is a group of things that are similar and related that actually provide an incredible insight into the data. Based on the clustered data and the purpose of the analysis, several clustering algorithms are available. Therefore, Word Cloud is a word frequency based data-clustering technique; the size of the words in the given text is proportional to their frequency. To identify the keywords, Word clouds can definitely help. In addition, in a given text, the Word cloud visualization technique can disclose the patterns.

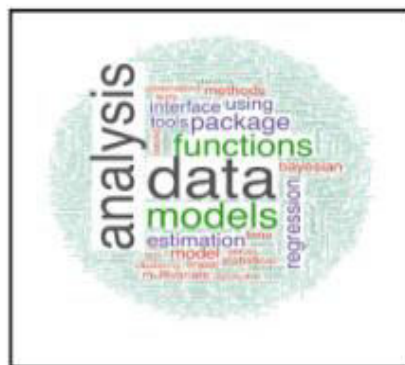


Fig 3: Image showing a Wordcloud

4.3 Sentimental Analysis

Sentimental Analysis is a text analytics approach that analyses and interprets the emotions in a given text significantly. A simple approach is Romantic Analytics. We break the given text into tokens and phrases initially, and define the phrases that contain feelings. Later, based on a predefined scale, we allocate ratings to these sentences. Consequently, we define the feelings depending on the values we attribute to the words. By analysis of intent, Sentimental Analytics extracts biased data. In addition, the analysis of employee sentiment [13] can be the answer to many questions.

4.4 Target Variable Analysis

The interpretation of the target variable specifies how the target variable influences the function variables; feature variables form the building blocks for the analysis, while the target variable is one of the function variables on which our analysis is based. The goal variable varies according to the data's business conditions, priorities, and availability. In this particular case, we have the function variables such as monthly wage, total working hours, job position, and attrition; if attrition is our target variable, we will examine how the monthly income variable factor affects the target variable, attrition. Ultimately, through scheduling rewards, management can mitigate the effect of the goal variable on the other function variables.

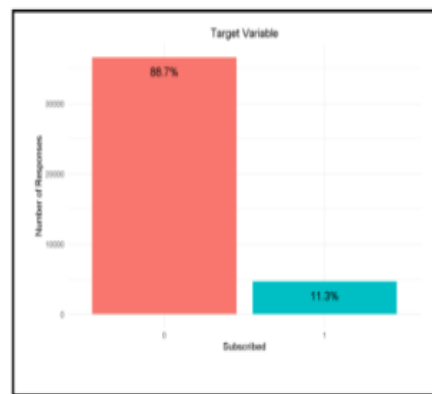


Fig 4: Target variable Analysis

V. RESEARCH METHOD

Major data insight generates valuable indicators. Consequently, a keen knowledge of the cumulative data and the sense of its compilation are required. For this, we administer experiments in such a way that we get a clear interpretation. We start with Resume Parsing at the start of the process. Resume Parser is a Python-inscribed code that utilizes embedded packages such as PDF miner and Spacy. The PDF Miner kit is an application for text extraction that transforms PDF formats to HTML, XML, or Word formats. We transform them to an analyzable format as the resumes sent could be of various types. In natural language processing, Spacy is a python library that supports. We use the Spacy library for the tokenization of resumes. In contrast with the keywords, these tokens support.

We allow use of Matcher for contrast. We will continue to download a list of stopwords. The stopwords are replaced by the parser and the keywords are retrieved. The keywords are the terms that allow us to gain knowledge from CVs such as schooling, expertise, experience. The Resume Parser output is the data frame that lists all the data needed by the resume and excludes the rest of the data. We then execute scanning of the resumes and thereby maintain manual efforts.

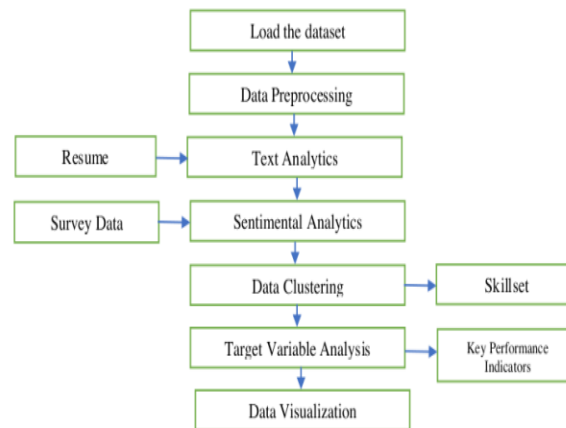


Fig 5: Image showing the process flow

When the data frame is saved in the worksheet, in comparison to the skills he has chosen in recruitment, we have the employee's extra skills. Therefore, we conduct clustering on the available skills on which a Wordcloud is deployed. Wordcloud forecasts are based on the frequency of the word. Therefore, we will get an understanding of all the abilities in the organizations that are available. In addition, at the outset of the current job description, we should recommend that the workers have unique expertise. Consequently, according to the criterion that lists the available talents, we formulate a code. We allow use of the Wordcloud python library and counter to create a word cloud as well. We need to get ready for the data set if we are preparing for research. We need to clean and preprocess the data [14] for this. There are some missed values we have to handle in the dataset used.

We transform the data and make the most use of the data accessible, so that we can obtain useful insight. In this particular case, we consider the date of birth variable, which however, has little to do with research. In this case, the date of birth attribute is substantially changed to age, which ideally has a crucial role to play. Packages such as Date Time and requests involve this procedure. As a result, we've added a new area, Era, to our dataset. Currency translation is also carried out on the wage attribute, which is one of the main factors affecting the business. Finally, our material is ready for review. We conduct sentimental research on the survey details as part of our operation. The survey was performed at the level of the boss, who reported the employees' opinions on their jobs and management on a five-point scale. The basis of our research is supposed to be this survey results. Via the sentimental study, we will get an insight into the employee's productivity rate, success scores, and attrition decisions. In order to forecast the turnover rate and salary scheduling to retain a talent, we will further use this research.

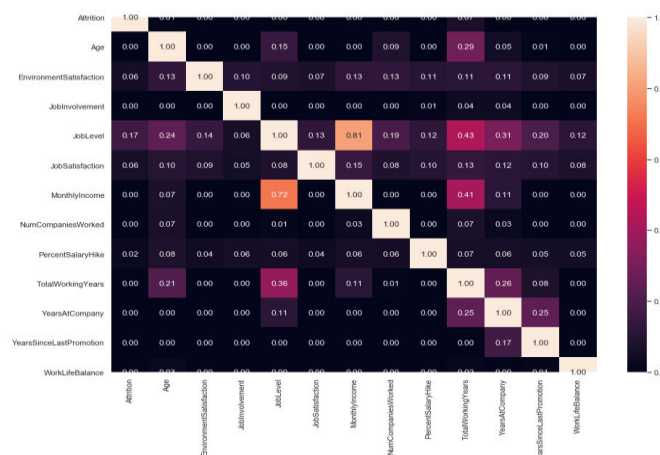


Fig 6: Image showing the Predictive Power Score for attrition

Identifying the main success metrics is the next step in our process. In reality, main performance indicators [15] are the features which affect the workflow. The predictive power score analysis is conducted as part of the Goal variable analysis in order to achieve the main success metrics. The Predictive Power Score is the quick exploratory asymmetric data approach that correspondingly calculates the linear and nonlinear data relationships on a score ranging from 0 to 1. On a symmetric matrix, the representation of this score is. For this, we consider as our target variable the most buzzing term, attrition, and indeed obtain the relationship of the attribute attrition with the other attributes. We use the Pandas, Seaborn, and NumPy libraries for the matrix. Thus, using the predictive power score and function collection, we conduct the target variable analysis. Finally, on the virtual dashboard created using the dash library and its elements, we simulate our results. Not only can the data analyst, but also the boss, grasps the consequences of our study.

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.
  }
]
```

Fig 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.



Fig 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.

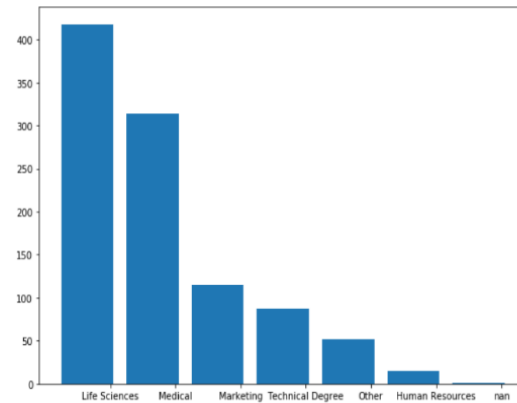





Fig 9: Image showing the frequency on talents based on which the word cloud is formed.

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_Nam	Last_Nam	Gender	Date of Bir	Age	Contact_num	Email	SALARY_USD
2	850297	Shawna	Buck	F	12/12/1971		7027717149	shawna.buck@gmail.com	119090
3	304721	Nathaniel	Burke	M	10/31/1993		2317656923	nathaniel.burke@walmart.com	117991
4	412317	Elisabeth	Foster	F	11/26/1994		2707494774	elisabeth.foster@gmail.com	161045
5	621375	Briana		F	11/24/1975		2196238216	briana.lancaster@yahoo.com	142616
6	707540	Estella	Potter	F	11/12/1995		0076770406	estella.potter@gmail.com	135706

Fig 10: Image showing the datasheet before preprocessing and having missing values

Export:  Wrap Cell Content:  Fetch rows: 

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	93277.172903
M	1993-10-31 00:00:00	26	2317656923	nathaniel.burke@walmart.com	117991	8584317.341300491	92416.381796
F	1994-11-26 00:00:00	25	2707494774	elisabeth.foster@gmail.com	161045	11716668.10375145	126138.40211
F	1975-11-24 00:00:00	44	2196238216	briana.lancaster@yahoo.com	142616	10375884.617868403	111703.89869
F	1995-03-13 00:00:00	24	9076778486	estella.potter@gmail.com	135706	9873154.470413204	106291.64519
M	1991-10-13 00:00:00	28	2365978196	lamont.woods@hotmail.com	173027	12588406.544678831	135523.29663
F	1984-09-15 00:00:00	35	2103961493	melinda.lopez@hotmail.com	41287	3003794.4425445446	32338.018621
F	1958-06-19 00:00:00	61	2363736712	shanna.silva@gmail.com	85833	6244694.174605224	67228.647089
F	1961-08-31 00:00:00	58	4237961535	jasmine.freeman@gmail.com	154216	11219831.030383643	120789.59187

Fig 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	Precision	Recall	F1-score	support
	0	0.90	0.93	0.91	245
Total working Years	1	0.37	0.49	0.53	49
Years since last promotion	Years at the company			0.85	294
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

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.

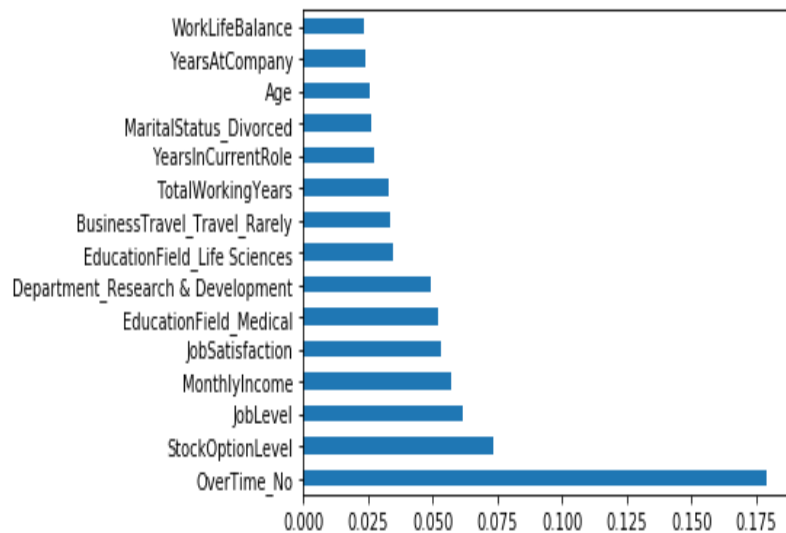


Fig 12: Image showing the key performance indicators

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.

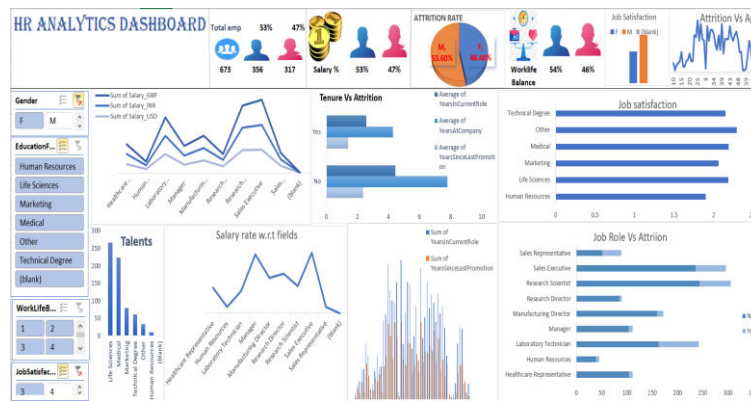


Fig 13: Image of the HR Analytics dashboard

VII. CONCLUSION

In end, with no restrictions, Data Science dominates this globe. It can be made use of anywhere, regardless of the study area. The HR branch is one that discusses details and more and more. We have a module that lets the HR evaluate their results. The focus of this paper was on the essential problems and difficulties in HR. The proposed Data-Driven HR Analytics approach, built on the basis of Data Science, increases and simplifies HR efficiency. The approach was accompanied by an interpretation of the data and an amazing insight into the data. The dashboard then visualizes the consequences in such a way that HR can make quick choices and execute policies.

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. NeilBrockett, Catriona Clarke, Michele Berlingiero, 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 Baykasoglu, Zehra Nur Atalay, İlker Golcuk, “Analysis of key performance indicators in a manufacturing plant via fuzzy cognitive maps” in Innovations in Intelligent Systems and Applications Conference (ASYU), 2019.