This article intends to investigate by means of Data Science, Human Resource (HR) managers to retain talent and avoid attrition in the organizations as it costs companies huge waste of time and money.

The following are the adverse effects of employee attrition:

The remaining employees have to take over the job that the person who left was in charge of, which translates into more unsatisfied employees, who are more likely to leave the company.   
Lost experience and knowledge as all the experience and knowledge that the former employee acquired through the years are no longer available for the company.   
The selection/recruitment process to replace the employee costs the HR department both money and time.   
In addition to the training courses the new employee might need, further costs might be uncured as the worker will take some time to be fully able to do his/her job and he/she will need help from other co-workers, which leads to a significant drop in productivity.

The actual costs have been estimated and published in a study by the Center of American Progress, which reveals that, in average, replacing a lost employee costs businesses one fifth of the employee’s yearly salary (Boushey & Glynn, 2012).

All of the above show the importance of retaining talented employees in order to reduce costs and increase productivity. However, avoiding employee churn is not an easy task, since the factors that drive attrition might not always be foreseeable and thus, preventable. Such task were traditionally carried out by HR; who would manually go over exit interviews of former employees; if any, trying to extract conclusions or patterns that could explain their exit, which implies a significant waste of time and money for the company.

This is where Data Science plays a vital role. With Data Science, the analysis of the available exit interviews at a company can be done systematically, and predictions for the future can be obtained based on the data from previous employees, minimizing costs and increasing the chances of retaining good employees and preventing their attrition. Data Science could be   
defined as the exploration and quantitative analysis of all available structured and   
unstructured data from a domain to develop understanding, extract knowledge, and formulate actionable results (Rudin & Elston, 2015)

Data Science is closely related to widely-known terms such as Big Data, Data Mining or Business Intelligence. However, those terms are neither equivalent nor mutually exclusive.

Big Data is a collection of very huge data sets with a great diversity of types so that it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms. It is characterized by what has been called The 3Vs: Volume, which is the size of the data set, Velocity, which indicates the speed of data in and out, and Variety, which describes the range of data types and sources (Zhang & Chen, 2014)

Data Mining involves the inferring of algorithms that explore the data contained in large and complex datasets, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction (Maimon & Rokach 2010).

Business Intelligence (often referred to as BI) is a process that includes two primary activities: getting data in, i.e. moving data from a set of source systems into an integrated data warehouse, and getting data out, i.e. the access to data from the data warehouse done by business users and applications to perform enterprise reporting, OLAP (On-Line Analytical Processing), querying and predictive analytics (Watson & Mixon, 2007).

People Analytics, sometimes also referred to as HR Analytics, is the use of data and analytic tools to inform decisions about how to manage people. It represents a data-driven approach to managing people at work, instead of using traditional methods of personal relationships, decision making based on experience, and risk avoidance (Massey, Haas & Bidwell 2016).

CRoss Industry Standard Process for Data Mining (CRISP-DM)

The de facto standard methodology for data mining: CRISP-DM. CRISP-DM stands for CRoss Industry Standard Process for Data Mining (Shearer, 2000), and is a data mining process model that describes the most commonly used approaches that data mining experts use to tackle problems. This model encourages best practices and offers organizations the structure needed to achieve better, faster results from data mining. Polls carried out by KDNuggets (leading site that covers the news in the field of Business Analytics, Big Data, Data Mining and Data Science) in 2002, 2004, 2007 and 2014 (KDNuggets, 2016) revealed that CRISP-DM was the most used methodology among the users.

Business Understanding

The first phase in the CRISP-DM process model is Business Understanding. This phase focuses on understanding the project objectives and requirements from a business perspective and gathering all the details about the resources, assumptions, constraints, specific data mining goals and data mining success criteria. Finally, a preliminary project plan is developed.

The business goal of this study is to predict whether an employee is going to   
voluntarily leave the company or stay. The benefits of this project will be minimizing   
employee turnover costs for the company and retaining valuable workers. The hardware available for the execution of this project is a Lenovo Idea pad Z510 laptop, along with an Asus desktop computer with Windows 10, an i5-6600 microprocessor and 32GB RAM.   
  
The data that will be analyzed has been obtained from the Watson Analytics website   
and is static, meaning that no further collection of data will take place. Metadata for   
the variables in the dataset is available.

The data mining goal of this study is to extract insights from the data that can explain   
factors that drive attrition in the company, and to predict attrition for a particular   
employee, given some personal and professional information. For this purpose,   
several prediction algorithms will be tested in order to obtain the attrition prediction.   
  
The output model of the data mining process will be required to have a minimum   
value of accuracy of 70% for the project to be successful. If such value is not achieved,   
the whole process must be repeated and corrected in order to obtain the desired   
accuracy.

Data Understanding

The next stage in the CRISP-DM process model is Data Understanding. This stage begins with the data collection, followed by an initial description and exploration of the data, where the distribution of the variables or relationships between pairs of them may be included. Finally, a data quality report is generated, listing the results of the data quality verification, and explaining if the data is complete, that is if it contains errors, or if any missing values are present.

Data Preparation   
The third phase in the CRISP-DM process model is Data Preparation. The main goal   
of this phase is to construct the final dataset that will be fed into the modelling   
algorithms. In order to achieve that, a data cleansing process must be carried out,   
along with constructive data preparation operations to obtain derived attributes or   
create generated records. Besides, the data has to be formatted to meet every   
algorithm’s requirements.

Modeling   
The fourth phase in the CRISP-DM process model is Modeling where the modeling techniques are selected, applied and their parameters calibrated to optimal values. Finally, the models are assessed in terms of accuracy and the results are collected and ranked.

Evaluation   
The fifth phase in the CRISP-DM process model is Evaluation. In this phase, the model that has been built and evaluated in the Modeling phase is assessed with respect to business success criteria. Besides, the whole process is reviewed to determine if any important factors or tasks have been missed.

Deployment   
The sixth and last phase in the CRISP-DM process model is Deployment. In this   
phase, the deployment strategy is carried out. This strategy could be very simple,   
like generating a report, or very complex, like implementing a repeatable data   
mining process.

Reference

1. Clara Cabañas Pujadas, “Application Of Data Science To Reduce Employee Attrition” 2016. [Online]. Available: https://e-archivo.uc3m.es/bitstream/handle/10016/27189/tfg\_clara\_cabanas\_pujadas\_2016.pdf?sequence=1&isallowed=y
2. H. Boushey and S. J. Glynn, "Center for American Progress," 16 11 2012. [Online]. Available: https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf.
3. C. Rudin and S. Elston, "edX," 2015. [Online]. Available: https://courses.edx.org/courses/course-v1:Microsoft+DAT203x+3T2015/courseware/5bbcfcf04e6d49bda1eecb1e1c0bfc  
   24/a57949170b154fdd87057996c87717c7/.
4. C.-Y. Zhang and C. P. Chen, "Data-intensive applications, challenges,   
   techniques and technologies: A survey on Big Data," Information Sciences, vol.   
   275, pp. 314-347, 2014.
5. O. Maimon and L. Rokach, "Introduction to Knowledge Discovery in   
   Databases," in Data Mining and Knowledge Discovery Handbook, New York,   
   Springer, 2010, p. 1
6. H. J. Watson and B. H. Mixon, "The Current State of Business Intelligence,"   
   Computer, vol. 40, no. 9, pp. 96-97, 2007
7. C. Massey, M. Haas and M. Bidwell, "Coursera," 2016. [Online]. Available:   
   <https://www.coursera.org/learn/wharton-people-analytics>.
8. C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining,"   
   Journal of Data Warehousing, vol. 5, no. 4, p. 13, 2000.
9. "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2016.   
   [Online]. Available: http://www.kdnuggets.com/polls/2014/analytics-data-  
   mining-data-science-methodology.html. [