A

PROJECT REPORT ON

"*HR ANALYTICS TO* IMPROVE DECISION MAKING PROCESS"

&

USE CASE USING MACHINE LEARNING ALGORITHMS TO UNDERSTAND RELIABLE WAYS TO FIGURE OUT, IF AND WHY THE BEST AND MOST EXPERIENCED EMPLOYEES ARE LEAVING PREMATURELY

Submitted to

SCDL, Pune

SYMBIOSIS
Centre for Distance Learning

Submitted By:
NAME : **Santosh Kumar**
REGISTRATION NO. : **201613110**
UNDER SUPERVISION OF: **Mr Jainendra Udai Singh**

Submitted in partial fulfilment of the requirements for qualifying
POST GRADUATE DIPLOMA IN HUMAN RESOURCE MANAGEMENT (PGDHRM)

ACADEMIC YEAR 2016-17

NO OBJECTION CERTIFICATE

This is to certify that **CDR SANTOH KUMAR** is an employee of this organization for the past **02 Yrs. W**e have no objection for him to carry out a project work titled ***HR ANALYTICS TO* IMPROVE DECISION MAKING PROCESS & USE CASE USING MACHINE LEARNING ALGORITHMS TO UNDERSTAND RELIABLE WAYS TO FIGURE OUT, IF AND WHY THE BEST AND MOST EXPERIENCED EMPLOYEES ARE LEAVING PREMATURELY"** in our organization and for submitting the same to the Director, SCDL as a part of fulfillment of the PGDBA (HR) Program. We wish him/her all the success. Seal of the company Signature of the competent authority of the Institute / Organization

Place: Mumbai

Date: Jun 18

DECLARATION BY THE LEARNER

This is to declare that I have carried out this project work myself in part fulfilment of the PGDHRM Program of SCDL. The work is original, has not been copied from anywhere else and has not been submitted to any other University/Institute for an award of any degree/diploma.

Date: Jun 18

Place: Mumbai

Name: Santosh Kumar(Reg No. 201613110)

DECLARATION OF GUIDE

Certified that the work incorporated in this Project Report **"*HR ANALYTICS TO* IMPROVE DECISION MAKING PROCESS & USE CASE USING MACHINE LEARNING ALGORITHMS TO UNDERSTAND RELIABLE WAYS TO FIGURE OUT, IF AND WHY THE BEST AND MOST EXPERIENCED EMPLOYEES ARE LEAVING PREMATURELY"** submitted by **SANTOSH KUMAR** is her original work and completed under my supervision. Material obtained from other sources has been duly acknowledged in the Project Report

Date: Jun 18

Place: Mumbai

Index

SECTION I

Introduction

1. **Human resource analytics (HR analytics).** Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes. [1]

2. **Machine learning**. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.[2] Machine learning can greatly assist the HR function. By using machine learning, many traditional activities like 'Talent Acquisition' and 'Employee engagement' can be greatly improved. Machine learning can help quickly sift through thousands of job applications and shortlist candidates who have the credentials that are most likely to achieve success. This, while also helping HR managers, have access to continual insights into how their employees are feeling about their workplace and how engaged are they. Machine learning is already efficiently handle the following HR aspects[3] :-

(a) Scheduling of HR functions such as interviews, performance appraisals, group meetings and a host of other regular HR tasks.

(b) Analytics and reporting on relevant HR data

(c) Streamlining workflows

(d) Improve recruitment procedures

(e) Reducing staff-turnover

(f) Personalize training

(g) Measure and manage engagement

(h) Enhance rewards and recognition programs

---

[1] *https://www.techopedia.com/definition/28334/human-resources-analytics-hr-analytics*
[2] *https://www.expertsystem.com/machine-learning-definition/*
[3] *https://www.techemergence.com/machine-learning-in-human-resources/*

3. As machine learning algorithm gains a deeper understanding of the company and has can absorbed all relevant information. In future, machine learning capable to do the following[4]:-

(a) Identify knowledge gaps or weakness in training

(b) Fine-tune and personalize training to make it more relevant and accessible to the employee

(c) Become a resource for information and questions related to policies, benefits, procedures and basic conflict resolution

(d) Aid in performance reviews

(e) Track, guide and enhance employee growth and development

4. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

5. **Aim**. The paper aims to understand the insight into HR Analytics and the established guidelines for solving HR problems using statistics and Machine Learning Algorithms. The scope has been divided as follows:-

(a) Chapter II deals with usage of **Data Analytics and Role of Data Scientists in HR Analytics.**

(b) Chapter III deals with **Cross-industry standard process for data mining**, known as **CRISP-DM**,[5] is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems.

(c) Chapter IV, has case study on **"Data Analytics and Machine Learning algorithms to understand reliable ways to figure out, if and why the best and most experienced employees are leaving prematurely".**

---

[4] https://www.peoplematters.in/article/techhr16/scope-of-robotics-and-artificial-intelligence-in-hr-13806
[5] Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

(d) In, Chapter V, **Machine Learning algorithms** have been used to solve the case study enumerated in chapter IV.

6. The use case used in the project is a **randomly generated data set of 15000 employees generated by the author and is nowhere related to any organisation/institution.** It is merely to understanding the concepts of HR Analytics and Machine Learning advancement in the field of HR Analytics. The data set is composed of both present employees and people who have already left the organisation.

7. **Tools Used**. The programming language used for data manipulation/cleaning is 'R'. Python has been used for coding and Visualization. SQL for querying. Also, Business Intelligence (BI) tools like Tableau have been used for generating graphs. Codes for Python used by the author for the use case has been provided as motivation for personnel.

8. **Limitations of the Paper**. The use case used in the paper is based on the randomly generated data set by the author. It is a general paper on industries/ organisations, wherein profits are measurable and accounting of HR in terms of CTC (Cost to Company) is measurable. The analysis can't be directly applied to Government organisations, as the dynamics/ requirements of such organisations are different.

SECTION II

HR Analytics Usage

1. **Background**. One of the most important functions of human resource (HR) professionals is to evaluate talent management and development techniques and identify opportunities to more effectively manage human capital. As human behaviour is much more complex and much less predictable than that of machinery or other tangible assets, the optimization of human capital allocation has, historically, been a difficult undertaking.[6] The use of HR analytics has noticed a recent rise in popularity in response to this challenge. By using data and metrics to design, evaluate, and implement new management policies, the "tried and true" method of using of experience, intuition, and guesswork to guide HR strategy is beginning to fall by the wayside[7].

2. **Classification.** The relationship between an organization's investment in human capital and its performance was first evaluated more than 50 years ago[8]. Only in the past two decades, however, has the application of data analytics to HR really taken off. Data analytics has been described as a merging of art and science[9]. While statistics are obviously a major component of any analytical exercise, analytics also involve a mental framework and logical understanding of the information at hand and the problems that need to be solved. In this way, analytics may be viewed as a "communications device," bringing together information from multiple sources to provide an actionable representation of a current state and a likely future[10]. By providing an evidence-based approach to decision making, analytics is a logical method that enables technological manipulation of information to provide insight on relevant issues. Classification of different levels of analysis are as follows:-

(a) **Descriptive Analysis**. Most commonly employed by organizations, descriptive analysis gathers data on past events or trends. This could include such measures as turnover rates or cost to hire a new employee.

---

[6] *Fitz-enz & Mattox, 2014*
[7] *Pfeffer & Sutton, 2006; Schwarz & Murphy, 2008*
[8] *Becker, 1964*
[9] *Fitz-enz & Mattox, 2014*
[10] *Fitz-enz & Mattox, 2014*

(b) **Predictive Analysis**. Predictive analysis evaluates why past trends have occurred and how they will change or continue without intervention. An example of predictive analysis would be the use of a model to increase the probability of selecting the right candidate for a job.

(c) **Prescriptive Analysis**. . Prescriptive analysis, designs treatments for fixing current issues. This could entail creating a model to understand how alternative investments in employee training affect the organisation's profits.

3. **Application of HR Analytics**. The application of HR analytics within a organisation may be a one-time effort or may coincide with a newly overhauled approach to organizational management. It is not uncommon, however, for one-time efforts to inspire more broad-reaching organizational change. It is important for organizational or HR leaders driving the incorporation of analytical methods to consider the purpose behind these efforts. According, analytics must be rooted in an understanding of the data to be used and the context under which that data were collected if any meaningful insight is to be gained[11]. This understanding will help determine the resources that are required and the form that the analysis will eventually take.

4. **HR Strategy**. HR professionals and management must develop a strategic understanding of how human capital contributes to organizational success prior to incorporating HR analytics. If the nature of the issue to be tackled using analytical tools is not explicitly defined, the likelihood of adding any value to the organization is extremely low. Before solutions are "fired at" the perceived issue, it is important to understand the potential causes behind the problem at hand[12]. Once the purposes behind analytical efforts are realized and an understanding of how human capital may contribute to organizational success is obtained, it is important to consider the data that will be used in subsequent analyses. Thanks to the increasing popularity of HR information systems, data are now regularly stored in one place which makes the gathering of information

[11] *Angrave et al. (2016*
[12] *Fitz-enz & Mattox, 2014*

relatively fast and painless[13]. However, even if data are in one place, this does not mean those data are ready to be plugged in to your statistical analysis program of choice. Data quality must be considered in terms of both missing data and possible errors in data entry. An analysis of the data may require descriptive or inferential statistics (or both) and may involve descriptive, predictive, or prescriptive analyses.

5. When analyses begin to move beyond summaries of the current state and inspire "what-if" questions, an organization moves from descriptive to predictive analysis. These analyses may involve evaluations of correlation, regression models, or structural equation modeling techniques; however, more advanced data-driven decision making extend past these methods to experimental studies that identify how human capital inputs affect organizational performance[14]. It may be argued that HR analytics should facilitate experimentation to identify the causes of performance improvement and quantify the return on investment that such efforts may provide. This involves complex projects that begin with question formulation, specify a logical research design, organize data in a meaningful way, and use appropriate statistical modelling, including a variety of techniques requiring different levels of mathematical complexity. By measuring the overall impact or "lift" of an intervention, these results may then be applied more broadly to provide further improvement in different areas[15].

6. In considering these recommendations about implementing an analytical approach, it is important to remember that support from the top of an organization is usually required to achieve success. Management support provides those driving analytical efforts with resources to allow such efforts to begin in the first place, and also with support when data are difficult to access or when such efforts are met with resistance (as is nearly always the case!).

7. **Intuition Versus Data**. There is a possibility of argument between HR professionals, about the balance of using intuition and data in making decisions. According, evidence-based management requires managerial decisions that are based on

[13] *Fitz-enz & Mattox, 2014*
[14] *Angrave et al., 2016*
[15] *Davenport, 2006*

hard facts to avoid "dangerous half-truths and total nonsense" that result from a reliance on past experience, benchmarking, or commonly accepted beliefs[16]. However, it can be argued that not all decisions should be wholly grounded in analytics and that instinct and anecdote should be used in decisions involving human capital, pointing to research that most people are able to make fast and accurate judgments of personality and character. There needs to be a balance of relying on numbers and trusting common sense.

8. Analytical techniques, in and of themselves, will not provide limitless rewards to those organizations who want to seek their use. It is believed that the application of HR analytics, combined with human judgment and managerial expertise, will allow better conclusions to be reached and practices to be realized than could have resulted from following the status quo of intuition and gut reaction alone.

9. **Drawbacks of HR Analytics**. The recent and dramatic rise in the popularity of HR analytics should be accompanied by a certain skepticism over the value of these efforts. There is substantial variability in the measurement maturity of organizations. Approximately 75% of HR departments do not have usable base metrics[17]. This means that, for many organizations, there is a big leap from their current state to the appropriate use of analytics.

10. **Role of Data Scientists/Analytics and HR Professionals.** Data scientists can help mitigate the problems with initiating successful HR analytics programs, assisting the HR profession in entering a "new world of strategic analytics-driven HR"[18]. This is true in terms of both the application of content and the use of quantitative methods. Many HR professionals lack a detailed understanding of analytical approaches. This hinders their ability to have meaningful interactions with data. Similarly, many analytics experts do not understand HR. This lack of overlapping skill and expertise leads to a mismatch between what HR information systems can do and what HR departments need. This calls for a different approach to HR analytics. Data Scientists, with an understanding of both the field of HR as well as quantitative methods, should play an important role in this

[16] *Schwarz and Murphy (2008),*
[17] *Fitz-enz and Mattox (2014),*
[18] *Angrave et al., 2016*

approach. Also problematic to the implementation of HR analytics, the most commonly used HR information system packages typically lack the statistical functionality to enable the types of analyses that are required to solve the problems at hand (e.g., longitudinal and multivariate analysis methods). There is an opportunity for data scientists to facilitate the application of quantitative methods, developed in other contexts, to the realm of HR management and development. One such example of this application is discussed below.

11. Organizations should not implement HR analytics programs because they are trendy or because their competitors use these approaches. HR analytics should be adopted because their use can drive widespread firm improvement in the present and for years to come.

12. Most of the foreign companies and private sector have already integrated the benefits of using their data to understand which customers are most likely to churn and are using that information to engage special efforts to retain the key employee. The public sector in India and government organisations still have some progress to make to reach that level of analytics.

13. The Personnel Departments generate a huge amount of data on a daily basis; wages and benefits, recruitment, leaves, departures, social conflicts, annual evaluations, career evolution, etc. Big Data combined to predictive analytics can open a tremendous potential for HR professionals and may generate huge benefits for all stakeholders in the organization: mainly managers and employees. Some non-exhaustive list of applications of predictive analytics for HR Usage are enumerated in succeeding paragraphs.

14. **Recruitment Optimization.** Recruitment is the first time the employer gets in contact with its future employee. Past datasets combined to applications received and publicly available information can contain opportunities to make the right decisions in short listing of the candidates for future and facilitate the employee's recruitment. It may aid into the following:-

(a) Detect talents and high potentials.

(b) Increase hiring success rates.

(c) Predict recruitment channel effectiveness.

(d) Predict employer brand strength.

(e) Find the right balance between contingent and fixed workforce.

15. **Employee Performance.** Efficiently applying analytics and prediction to HR data can offer new insights into current and future performance optimization opportunities at several moments of the employee life cycle:-

(a) Predict absenteeism and work accident risks.

(b) Analyze employee engagement.

(f) Analyze and predict best on-boarding processes to reduce time-to performance.

16. **Employee Retention.** HR Analytics allow to understand very accurately the employees' motivations and what makes them stay longer in the organisation or decide to leave. Based on a few data sets and, algorithmic models are able to extract powerful analysis that aid to the management to decide about the right actions and may immediately improve employee retention rates and management decision for VR(Voluntary Retirement):-

(a) Detect potential leavers and take preventive action

(b) Analyze employee attrition by business unit or department

(c) Predict turnover evolution on the short, mid and long run

HR Reporting & Analytics Study

17. Since its introduction, HR analytics has been food for thought among HR practitioners, experts and researchers. Early adopters and fans praised the big potential of data driven HR. Some even say we have entered the age where HR analytics is 'the new normal'. Others remain reserved or have second thoughts before they want to embark into the world of data driven HR.

18. What does HR analytics success look like in terms of organisation, people and data? There are four levers that set HR analytical front runners apart from the rest. HR

analytics is successful when it is positioned strategically, focuses on business impact with support by key roles and capabilities, captured in clear processes and responsibilities to make optimal use of available, high quality data.

19. It is clearly seen that organisations that have made significant progress in the field of HR analytics have claimed a place for it at the core of strategic decision making. In these companies, HR moves in the same space as marketing, operations and finance in terms of demonstrable strategic impact on the organisation. In most cases, it does not mean they started their HR analytics efforts from this strong strategic position supported by senior management, but they did start at some point with the data and resources available at the time and are therefore now front runners.

20. It is no surprise that maturity in HR reporting is strongly correlated with maturity in HR analytics. In practice, many organisations are not spending enough energy on developing HR analytics though. Taking the first steps in HR analytics requires a mix of knowledge, capabilities and most of all dedication. However, the immediate need for improving HR data and metrics is always present, using employee data for operational management and as a basic hygiene factor to support your workforce. HR needs to claim space for HR analytics, preferably in a dedicated role. Through HR analytics it is possible to show the strategic value of people insights, go beyond knowing and really act to increase people's impact on the organisation.

21. Business impact value and growth HR analytics is a strategic tool to analyse and predict HR's effects on key organisational outcomes. Front runners position it as such. This means they organise all HR analytics efforts to optimise the impact of HR interventions on their business and increase the chance of successfully attaining business objectives. Finding out how HR can have a positive impact on realising important objectives for the organisation is the essence of HR analytics. This means it is crucial to know what issues business management are facing and for them to know how HR can help. You need senior management support to be able to position HR analytics strategically. Usually, a scoped but appealing pilot with a clear link between HR and business results will get the support from senior management. The next challenge is to

gain momentum and start building capability to grow and add sustained value, by integration of a relevant HR analysis in every important business decision.

22. To keep the appetite for analytical insights fresh and be able to work on challenging cases for HR analytics, specific roles and capabilities are needed. Analytical capabilities, IT capabilities and management capabilities are important, but front runners also acknowledge the importance of the HR business partner capabilities as linking pin between business and HR. Having the capabilities to collect important issues business leaders have in realising their objectives is key in order to work on useful HR analytics projects. Once you have gathered those burning issues, the next important role and capability needed is the deep analytical kind: the data scientist. Many organisations are struggling to find one, especially in HR. Therefore some organisations are opting for a centralised analytics function, working on deep analyses for other functions, including HR. However it is organised, being able to work together with data scientists and understand and interpret analytical outcomes is an important ability for any HRBP or HR consultant in the HR analytics process.

23. A successful execution of HR analytics requires good governance: all key players and other stakeholders are on board and involved, know what to do, and understand the relations and interdependencies between roles. Good cooperation within the HR analytics process is key for the delivery of data driven HR insight. Clear and well-documented processes will help, as well as including the analytics responsibilities within the HR job descriptions. Role clarity is a big differentiator between starters and front runners and contributes to creating a more data driven mindset in HR. Creating a data driven culture is not easy in most organisations, it is a long-term change process. For HR, this is especially the case, as HR people are not used to it.

24. IT issues in relation to HR analytics are common for most companies. The difference between starters and front runners in this respect was relatively the smallest from all areas in our research. One of the reasons front runners are at the front is because, regardless of technical shortcomings, they started anyway, built their database and improved HR data along the way based on business priorities. You do not need all HR

data to be available and full access to business data to get started. And as research shows, most organisations are far from achieving both holy data grails anyway. Despite the fact support from IT resources and IT capabilities does help to achieve these goals, unfortunately, many HR organisations still cannot fully rely on their IT function and are lacking support from the CIO.

25. A vast majority of the organisations still can be categorised as starters (or willing to start). Successful organisations organise their HR analytics in order to make a business impact. At the same time, practical directions for those who are still in an earlier phase of HR analytics maturity. There is a lot of opportunity for growth and development in data driven HR to be exploited, and thus organisational success to be harvested in the near future.

Strategic Positioning of HR Analytics aimed at Business Impact

26. The first key factor for HR analytics success in our research is to have a clear strategy for the use of HR analytics in line with business strategy. HR analytics is a strategic tool to analyse and predict HR's effects on key organisation outcomes. Front runners position it as such. This means they organise all HR analytics efforts to optimise the impact of HR interventions on their business and increase the chance of successfully attaining business objectives.

27. Just like a sound people strategy should follow from your business strategy, HR analytics efforts should always be in line with business strategy if you want to make business impact with HR. Front runners indicate their people strategy explicitly mentions that data driven HR is a key factor for creating business impact. Moreover, the HR analytical projects can be clearly linked to their organisation's strategic objectives.

28. Organisations that build their HR analytics strategy on the foundation of their organisation strategy started with one question: what are our business objectives and how can we use analytics to optimise the effect of our people practices on attaining these objectives? An effective way to further align with business strategy in your HR analytics efforts is to talk to business leaders and ask them what their biggest hurdles are in realising their objectives. Indeed, most front runners and practitioners have also identified and

prioritised issues in the organisation where HR analytics can provide the most valuable insights.

29. Business impact, the ultimate goal required for business management to act on insights provided by HR. For that to happen, they need to see the business value of a proposed HR intervention. The power of HR analytics is that statistical analyses of HR combined with business data allows you to predict possible business improvements to a certain extent. In practice, a start to harness this power could be to conduct an inventory of key business processes that could benefit from specific HR insights. However, on average, organisations score relatively low on systematically assessing key business processes that could benefit from specific HR reports and insights from HR analytics.

30. The most important major differentiator between low mature HR analytics organisations and higher mature HR analytics organisations is having HR analytics projects with measurable business impact.

31. A formal strategy and execution plan for clear direction In order to gain serious momentum with HR analytics, it is important to formulate a formal strategy and strategy execution plan. Actually, our research demonstrates this is one of the biggest differences between HR analytics front runners, practitioners and starters. This means it needs to be clear to all how HR analytics is positioned and used within the organisation and how this affects the organisation. When this strategy and plan is initially created, it is wise to involve all key stakeholders with at least one business sponsor and develop a short-term and long term ambition for HR analytics. A short term ambition helps in creating a pilot case for HR analytics, with business impact to create appetite in the organisation. The long-term ambition should be focused on creating a solid process for the delivery of key insights, working on important areas such as data quality, people capabilities, governance, and a data driven mindset.

32. Three ‘strategic positioning’ factors that clearly standout between starting analytical organisations and developed analytical organisation:-

(a) Having strategy for HR analytics

(b) Performing HR analytics with proven business impact

(c) Actively involving senior management in HR analytics project

33. All the former key elements cannot be accomplished if senior business management doesn't support it. It is key to help business leaders understand the potential of HR analytics and involve them to help gain focus on the right business issues. Overall, over 60% of organisations indicate senior management actively promotes the use of analytics and data driven decision-making in general. The only convincing they need is why it is also relevant for HR, which is all about ROI. It would be wise to gain insights on what investment in their people generates in terms of organisation outcomes. A simple calculation is usually enough to get their attention. To create further buy-in, front runners indicate they always have active involvement of at least one key member of senior management in HR analytics projects

34. Dedicated budget helps a lot In order to be successful in HR analytics, there needs to be budget in order to get organised and have the means to do useful projects. This also means investments in data & IT architecture, capabilities, hiring of experts and putting together a dedicated team in order to create impact.

Support by Key Roles and Capabilities

35. The second key factor for HR analytics success in our research is to have key roles and capabilities in place to be able to execute all necessary tasks in HR analytics projects. Analytical capabilities, IT capabilities and management capabilities are important, but front runners also acknowledge the importance of the business partner capabilities as linking pin between business and HR.

36. In practice, we see that in starter organisations, HR analytics activities are often assigned to the same team or person as HR reporting tasks. Time is mostly spent on HR reporting due to immediate urgency and HR analytics is put on a slow side-track or on hold completely. Our research indicates that HR analytics front runners and practitioners have more dedicated management and execution roles for HR analytics projects and active support from senior management for these projects compared to HR analytics starters.

37. Not only should senior management be involved in HR analytics projects in order to link to strategy and provide business support, it is also important that senior management is supportive of HR analytics in general. Key roles for senior management support are CEO, CFO, CHRO and CIO, and each one of them could be convinced of the added value of HR analytics in a different way.

38. The HR business partner role is very important in the context of data driven HR. Their role is to understand both business and HR and figure out how HR can help solve business issues through analytics. For HR analytics to work, you need HR people who understand how analytical techniques and technologies work and how to use these techniques and technologies. Moreover, HR business partners should be able to understand, interpret and advise on analytical outcomes. This helps them fulfil their role as linking pin. In our research, nearly all front runners indicate that their HR business partners and/or HR managers are able to explain analytical outcomes and translate them into clear actions for the organisation.

39. In order to create relevant insights from HR analytics, you need people who can understand a research question, create a research model and perform the necessary analysis. In this function data science meets HR and although it is very helpful to have HR experience and knowledge in this function, it isn't a top priority. Still, many organisations are struggling to tap into analytical capabilities in-house; almost half of respondents indicate they have no analytical professionals with statistical knowledge, while front runners seem to have relatively less issues with finding the right analytical talent and experts and often employ a HR analytics manager with consultancy skills and statistical background.

40. Without proper support from IT any effort to leverage data driven HR will be strained. We need to work together closely with the IT department to collect and improve HR data, gain access to other data sources and the latest statistics tools. In practice, we often see that IT capabilities are a bottleneck. There are often many different systems for gathering an ever-growing diversity of data.

41. A proper data warehouse can be the solution to link all data, but in practice we see that relatively few organisations have the services to simply link data and create easy access. The results of this research show that the biggest challenge cited by HR analytics front runners, is IT support in terms of adequate capabilities to stimulate the use of analytics within the organisation. Although there is a significant difference between starters and front runners in this area, the difference is not particularly high. It seems that support from IT is not optimised in most organisations. This is stressed by the lack of CIO support mentioned earlier in this chapter.

42. Successful execution of HR analytics requires good governance. In other words, all key players and other stakeholders are on board and involved, know what to do, and understand the relations and interdependencies between roles. Plus, all the activities related to the execution of HR analytics should not only be evident and unambiguous, but also well-documented, accessible and communicated to all people involved. Involvement of key stakeholders (for example when designing processes) means involving decision- makers and experts from business and other departments such as Finance and Operations. Creating a data driven culture is not easy. It is a long-term change process.

43. A data driven mindset starts with clarity on roles and responsibilities Implementing and executing data driven HR means transforming to a more data driven mindset. Creating this mindset is not easy for organisations nor for the people involved. For HR professionals and specifically HR Business partners, a lot will change when it comes to the required and needed capabilities to do their work. One of the key success factors in change management is to make sure everyone knows how the change will affect the daily operations of one's role and responsibilities. It will be no surprise that while implementing a new process like HR analytics, it is equally important that everyone knows their position within the (HR analytics) value chain in order to achieve an optimised outcome.

44. Organising for an effective HR analytics pipeline Good cooperation also means having clear and well-documented processes to deliver data driven HR insights. This forces you to have a clear vision on the process as a whole and the understanding of one's

role within the process. This is a must for a well-oiled working HR analytics machine. Just as important is the communication and training of people working with the process.

45. Next factor for HR analytics success in our research is easy accessible and reliable data. With accessible, we mean that the needed data field is available for analytics when records of all data are obtained in an editable format by the data specialist with only one or few actions. And with reliable data we mean that the data has a certain accuracy (correct, reliable, up-to-date and free of errors). HR analytics practitioners and front runners view having readily available data with high quality standards as an important predisposition for HR reporting and analytics.

46. Most organisations have created difficult infrastructures to link data to each other, requiring a specific IT skillset as mentioned in the roles and capabilities chapter. When access to data is restrained, a lot of time and effort needs to be put into creating a dataset ready for analysis. Eighty percent of front runners state that HR data is easily accessible to those who need it within their organisation.

47. The accessibility of HR data (to anyone who needs it) poses a challenge for most organisations however, and scores relatively low. For HR analytics practitioners, it is their biggest challenge, ranking this aspect as number 1 of the items with lowest average scores. Moreover, in HR analytics, preferably also business data is used and linked to HR data. Access to business data by HR for the benefit of HR analytics is, although relatively speaking, one of the biggest challenges for HR analytics front runners (stating it as number 2 on the bottom of the ranking). Overall, half of the organisations cite HR has access to business data for analysis purposes. At the same time, the number of HR analytics practitioners and front runners that agree with this statement is not overwhelmingly high.

48. To realise business improving HR analytics results, HR and business data should not only be available for analysis, but also of a certain quality. Without good data quality, HR analytics fails to produce reliable insights. Still, it is still common in most current HR analytics projects to spend a significant amount of time on data selection and data cleaning.

49. It is important that the use of HR data meets specific privacy standards since the used data is about people. There are strict ethical and legal standards to comply to and meeting those standards is often considered a rocky road for many organisations. It means dealing with works councils, compliance and legal in order to have a unified and documented set of agreements on how to deal with people data. The key to success is to start identifying the total landscape and specific requirements through data governance. Also, involving the relevant stakeholders and consulting with legal in time can make your analytics journey a lot easier.

50. Organising people for organisational success has grown more complex as the world of work diversified and became more global over the past century. HR has grown into a range of specialisms to serve specific needs for different cultures. This has made HR organisation complex in itself, resulting in all kinds of governance and operating models but the HR discipline still boils down to one common goal: getting people organised to maximise organisation success.

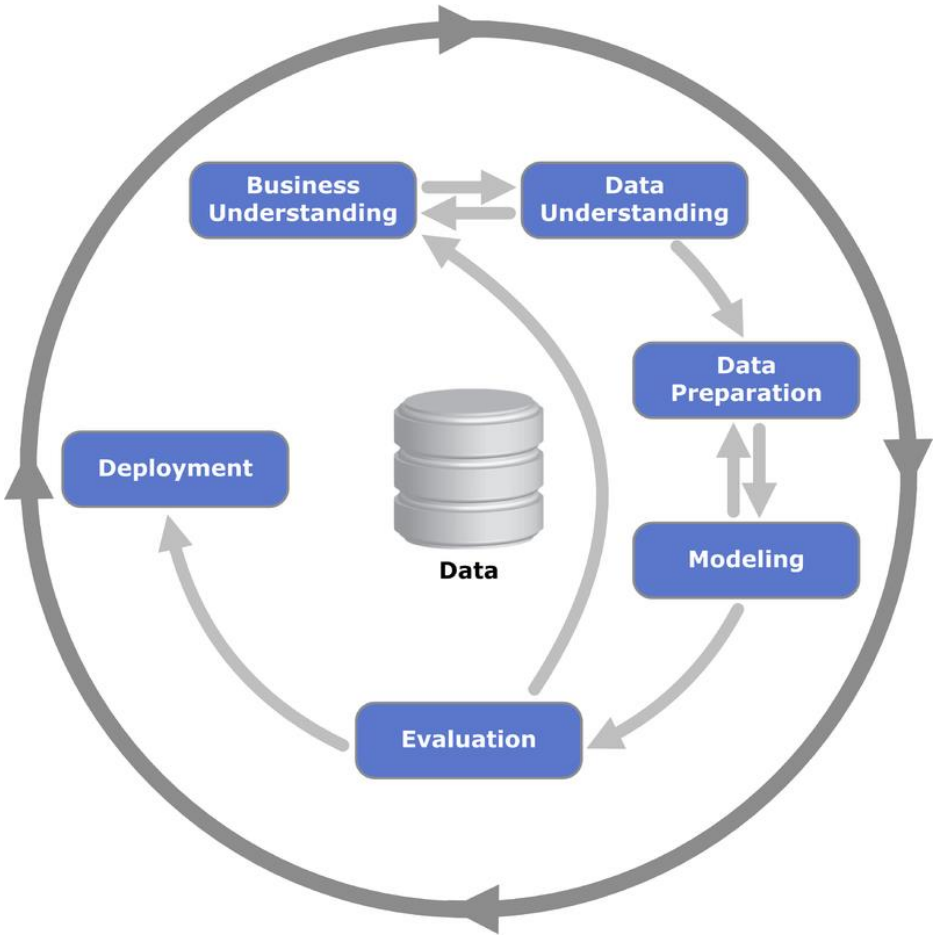
51. As it turns out, HR analytics is a great method to contribute to this goal. It uses data and statistics to create clarity on relationships between the way we organise our people and organisational success. It can therefore be a valuable strategic tool for an organisation and this is starting to dawn with organisations on a global scale. As our study shows, positioning of HR analytics within the strategic realm of the organisation is crucial to make progress.

52. Although it is a small group that is already integrating HR analytics as an extension of business strategy (and as part of daily operations), we see HR analytics is gaining traction in the market. This means many organisations would like to get started with this methodology but do not know where to begin, do not have the time and support or are hesitant to begin due to technology immaturity

SECTION III

Cross Industry Standard Process for Data Mining Framework (CRISP-DM)

1. **Analytics Problem**. Analytics problem solving involves multiple steps like data cleaning, preparation, modelling, model evaluation etc. Completing a typical analytics project may take several months, and thus it is important to have a structure for it. The structure for analytics problem solving is called the CRISP-DM framework - Cross Industry Standard Process for Data Mining.. It involves a series of steps which are not necessarily sequential in nature. But it all starts with business understanding. Business Objectives can be articulated from the broad problem statement, so that the next steps in the CRISP-DM framework can be performed effectively. There are six steps involved in CRISP-DM framework, which are as follows:-



Process Diagram Different Phases of CRISP-DM [19]

[19] *https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining*

Step 1: Business understanding

2. Understanding the business and its specific problems is of utmost importance for data analyst. The problem needs to be clearly understood and needs to be converted into a well-defined analytics problem. Only then a brilliant strategy can be laid to solve it. In summary, to understand the business problem, one has to undertake the following steps of analysis:-

(a) Determine the business objectives clearly

(b) Determine the goals of data analysis

Step 2: Data Understanding

3. After business understanding, the next step is data understanding. The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. It is important to understand the data structure (number of files, rows, columns etc.), understand how they are related to each other and whether something looks fishy–like a date column having negative values. Broadly, we are interested in the following:-

(a) The type of data sets that are available for analysis.

(b) The information extracted from the datasets.

(c) Exploring the data (by plotting graphs and observing them).

(d) Performing quality checks on the data sets.

4. To summarize, data understanding has the following steps:-

(a) Collect relevant data and classify them as:

  (i) Structured

  (ii) Unstructured

(b) Describe datasets

  (i) Create data dictionary

  (ii) Summarize data

(c) Explore data by plotting graphs

(d) Check data quality in terms of:

(i) Completeness

(ii) Correctness

(iii) Types of error/Missing values

5. **Plotting Charts**. A critical part of data understanding is exploring the data though plotting charts. While a line chart can be used to present a time-dependent trend, bar graphs and histograms are best used for categorical and continuous data respectively. A pie chart best summarises the share of different components in an aggregate whole, while a stacked bar chart is used to compare the share and contribution of categories across different sectors. A box plot is suitable to represent the quartile, percentile and outliers values, whereas a scatter plot summarises the variation of data points across two dimensions or two parameters. A grouped bar chart is best suited to present different sub-groups among the main categories.

Step 3: Data Preparation

6. The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modelling tools. Across projects, data analysts spend around 50-80% of the time on data cleaning and preparation, and therefore data preparation becomes one of the most crucial steps. Data is usually spread across different files. Collating those files together and selecting the required rows and columns based on business understanding is a major step in data preparation. After collating the data set, missing values and outliers, needs to be addressed.

7. Data preparation is considered the most crucial step because the model for analysis would be built on the data sets created. If the data set is erroneous, the solution to the problem after building a model would be erroneous too-no matter how the model is being created. To summarise, data preparation is one of the most time-consuming steps of the entire analysis. It consists of the following steps:-

(a) Select relevant data

(b) Integrate Data- one merge file is essential

(c) Clean data

(d) Construct Data: Derive new features- to reduce the no of variables

(e) Format Data

Step 4: Data Modelling

8. It is well said that, **"If you torture the data long enough, it will confess."** In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. Data Modelling is the heart of data analytics. One can think of a model as a black box which takes relevant data as input and gives an output you are interested in. The following points are important for Data Modelling. The Models selected should be Succinct, Mathematically sound, Efficient and Easy to use

Step 5: Model Evaluation

9. At this stage in the project a model has been built, that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached. In data analytics, evaluation is when you put everything you have done to litmus tests. If the results obtained from model evaluation are not satisfactory, the whole process needs to be reiterated. If the model performs well and gives accurate results, model can be implementation. Evaluation is necessary to ensure that the model is robust and effective. Finally, implementation is the natural fruition of a project life-cycle.

10. One interesting insight is that the whole process is iterative in nature. The intelligence of a model has to evolve continuously. Model evaluation is done to verify or validate that the model developed is correct and conforms to reality. After the model is built, we need to check if the model works well on the actual data and not just the data from which it was built. Multiple models can be built for a certain phenomenon, but a lot of them may be imperfect. Model evaluation helps us choose and build a perfect model. Model evaluation helps us choose the best model among a given set of possible models that can be built. Comparison should be done to assess the performance of different models in various situations and then decide the best model class or algorithm to use for the required business problem.

Step 6: Model Deployment

11. Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process.

SECTION IV

Use Case- Reduce Employee Attrition and Make Talents Stay Longer

1. **Aim/ Business understanding.** Retaining key employees is a major stake for each organization. But are there **reliable ways to figure out if and why the best and most experienced employees are leaving prematurely**?

2. Personnel Departments generate a huge amount of data on a daily basis: wages and benefits, recruitment, leaves, departures, social conflicts, annual evaluations, career evolution, etc. Big Data combined to predictive analytics can open a tremendous potential for HR professionals and may generate huge benefits for all stakeholders in the organization.

3. The data set in the use case is a randomly generated data set of 15000 employees generated by the author and is nowhere related to any organisation/institution. It is composed of both currently working employees and people who have already left the organisation.

4. **Data Understanding/ Data Discovery.** 'R' programming language has been used to clean the data and to dig into the different columns of the data set. Fields in the data set include:-

(a) **name**. The name of the employee.

(b) **satisfaction_level**. Employee satisfaction level. Ranges between 0 and 1. 1- indicates highly satisfied.

(c) **last_evaluation**. The grade the employee got at their last evaluation. Ranges between 0 and 1.

(d) **number_appointments**. The number of simultaneous appointments (duties) /projects the employee has worked on.

(e) **average_monthly_hours**. The number of monthly hours the employee is working.

(f) **time_spent_org**. The number of years the employee has been working for the company

(g) **work_accident./ Medical Issues**. Whether the employee has already had a work accident in the past (1 for yes, 0 for no)

(h) **promotion_last_5_years**. Whether the employee has got promoted during last 5 years. (1 for yes, 0 for no)

(i) **department**. The department the employee is working,

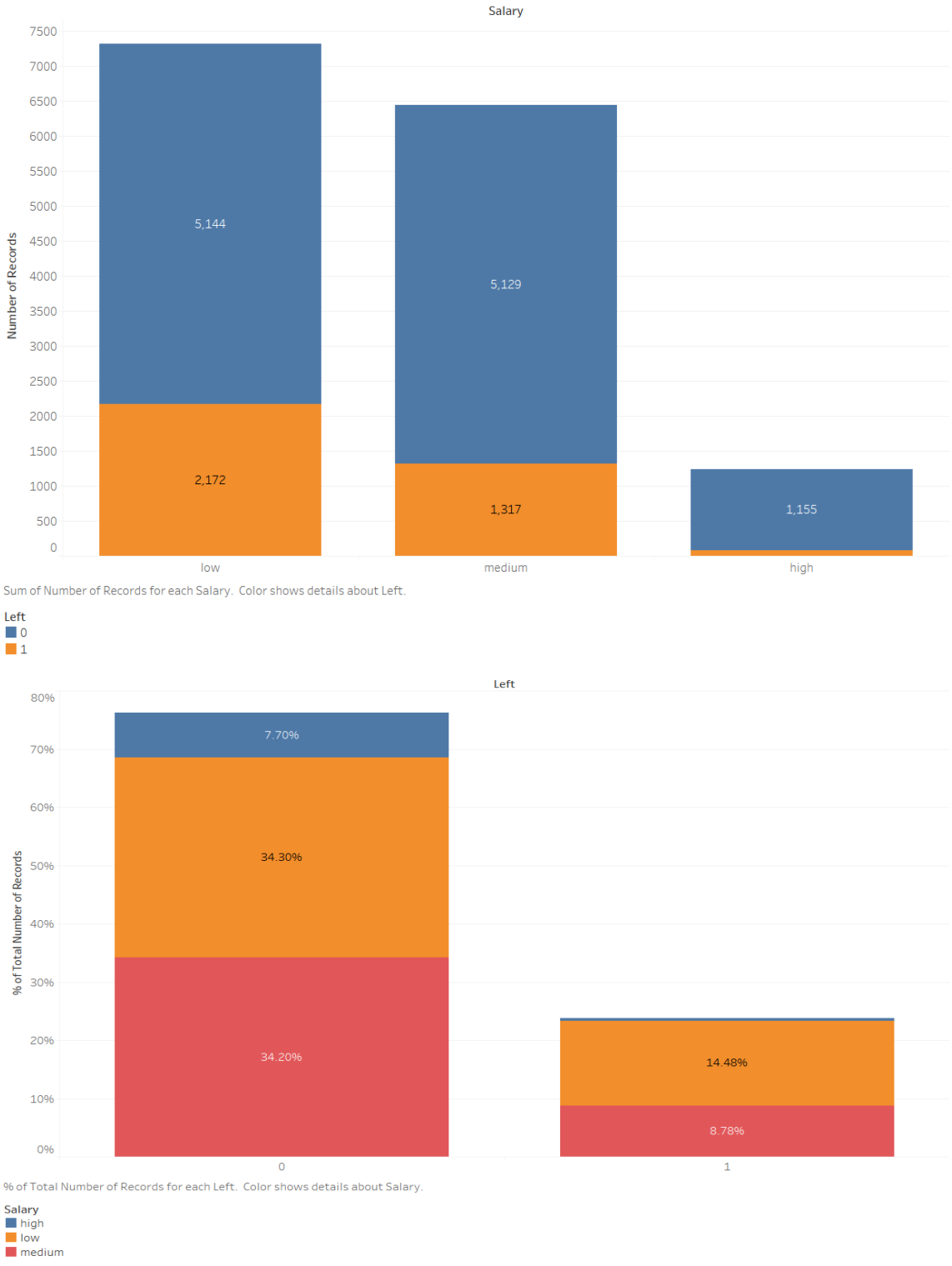
(j) **salary**. The current level of salary of the employee (3 categories : high, medium, low)

(k) **left**. Whether the employee has left. 0 if the employee is still working for the company, 1 if not.

N13

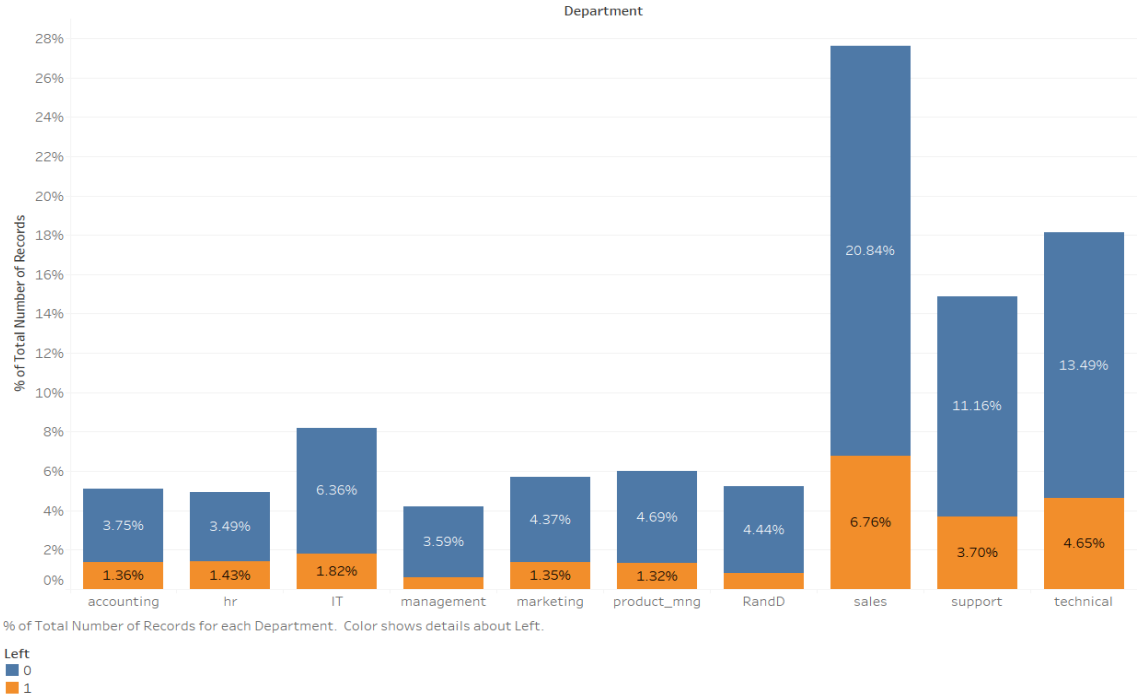
	A	B	C	D	E	F	G	H	I	J	K	L
1	name	satisfaction_level	last_evaluation	number_appointments	average_monthly_hours	time_spent_org	work_accident	left	promotion_last_5_year	department	salary	salary_level
2	AMAR	0.38	0.53	2	157	3	0	1	0	sales	low	1
3	AKHBAR	0.8	0.86	5	262	6	0	1	0	sales	medium	2
4	ANTONY	0.11	0.88	7	272	4	0	1	0	sales	medium	2
5	RAJA	0.72	0.87	5	223	5	0	1	0	sales	low	1
6	RANI	0.37	0.52	2	159	3	0	1	0	sales	low	1
7	MANTRI	0.41	0.5	2	153	3	0	1	0	sales	low	1
8	WAKIL	0.1	0.77	6	247	4	0	1	0	sales	low	1
9	GARCIA	0.92	0.85	5	259	5	0	1	0	sales	low	1
10	RODRIGUEZ	0.89	1	5	224	5	0	1	0	sales	low	1

5. Prior to doing any specific analytics, it important to understand if how the dataset might be correlated to each other.



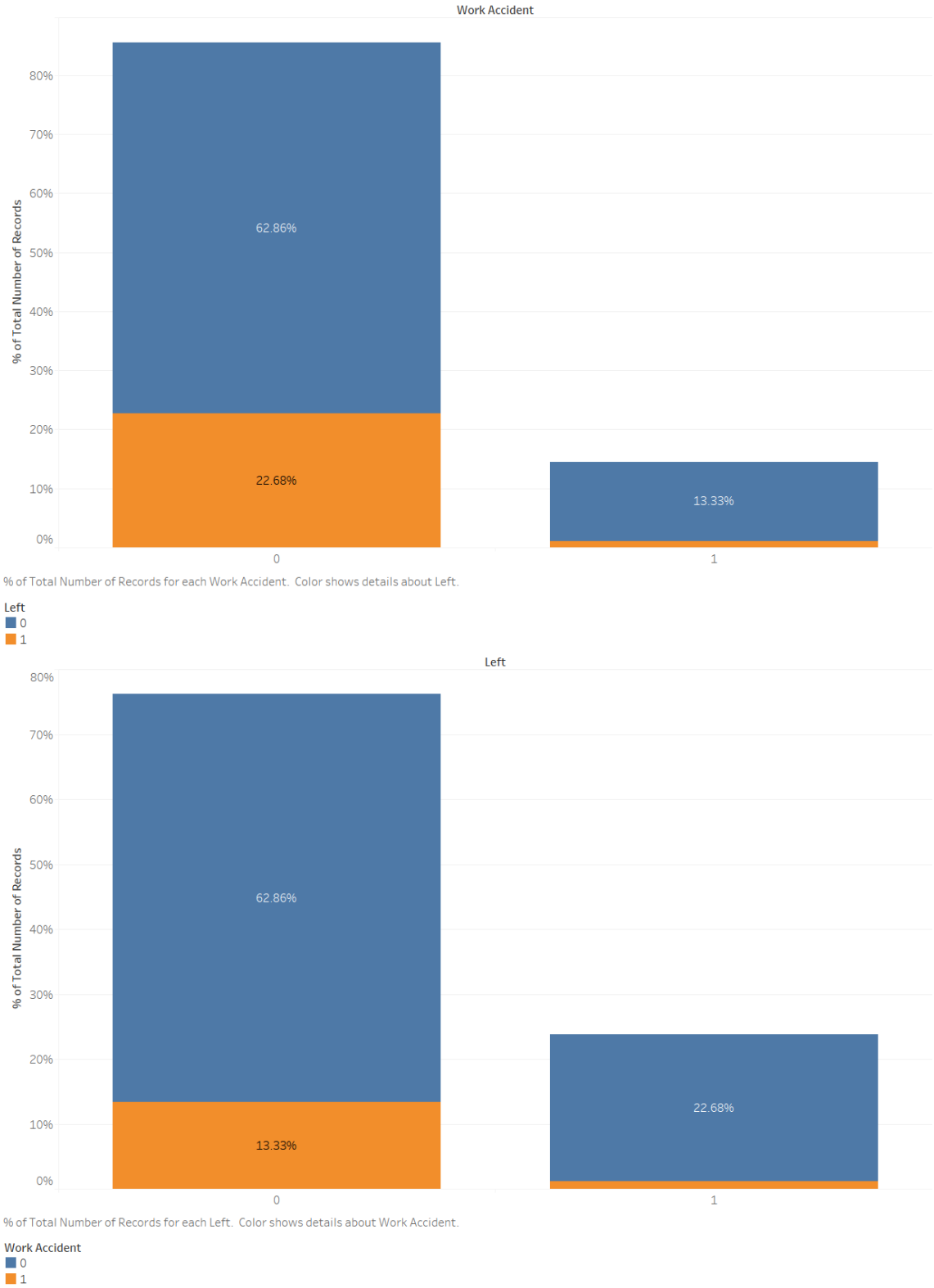
RELATIONSHIP BETWEEN PERSONNEL LEFT(0-WORKING, 1-LEFT) AND SALARY(3- HIGH, 2-MEDIUM, 1LOW)

*(**Interpretation**- The number of personnel leaving are highest from the Low Salary group. It accounts for 14% of the total dataset)*



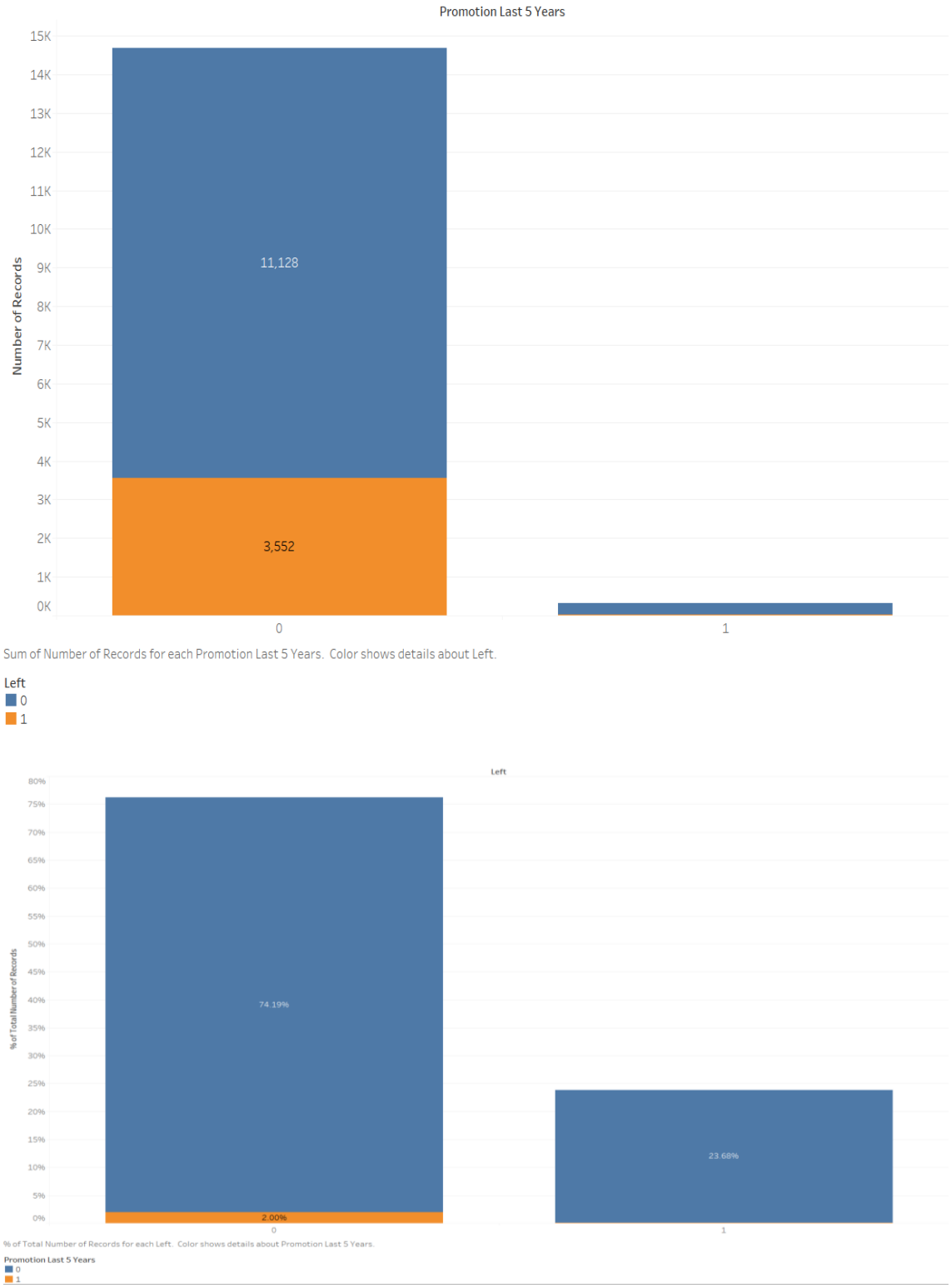
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND DEPARTMENTS)

*(**Interpretation**- The number of personnel leaving are highest from Sales department followed by technical and support (% of the total dataset)*



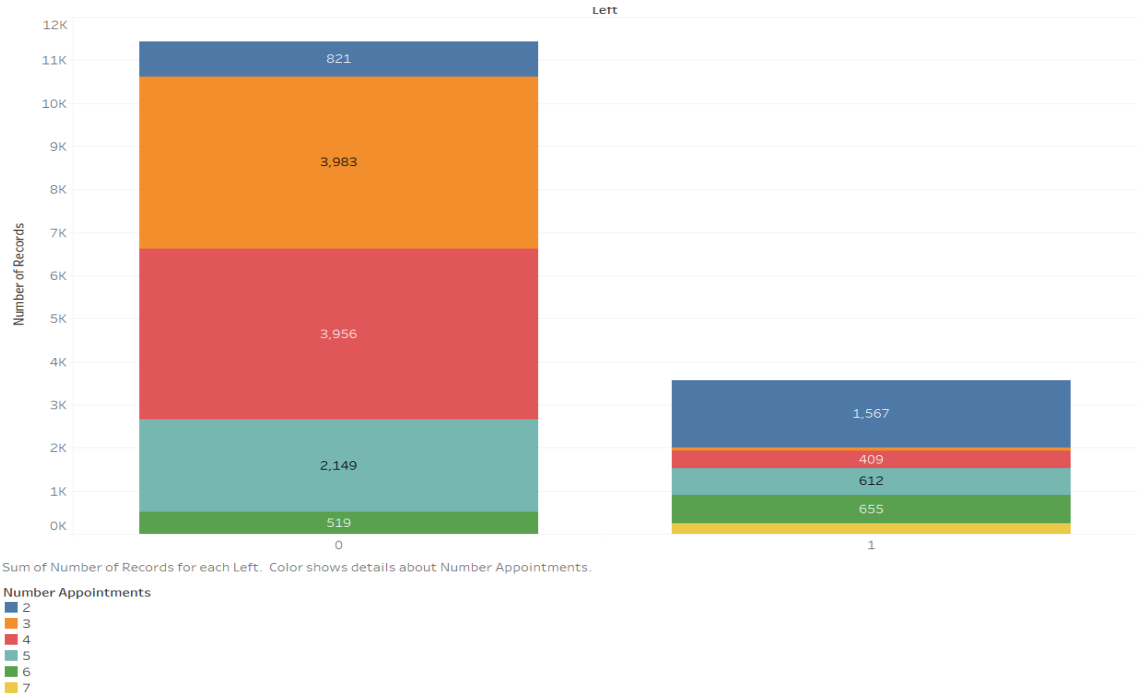
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND WORK ACCIDENT (1- YES, 0- NO)

*(**Interpretation**- The number of personnel who met with any accident, normally do not leave the organisation)*

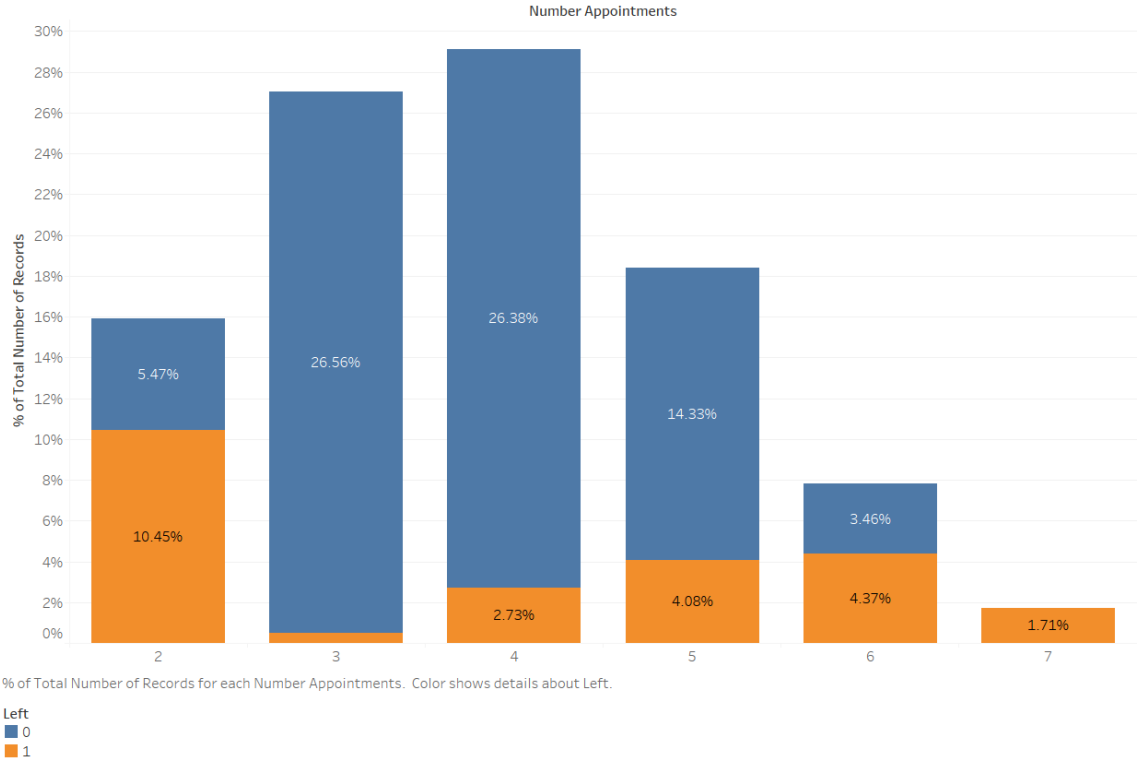


<u>RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND PROMOTION IN LAST FIVE YEAR (1- YES, 0- NO)</u>

*(**Interpretation**- <u>Personnel who have left the organisation, have not been promoted in last five year</u>*

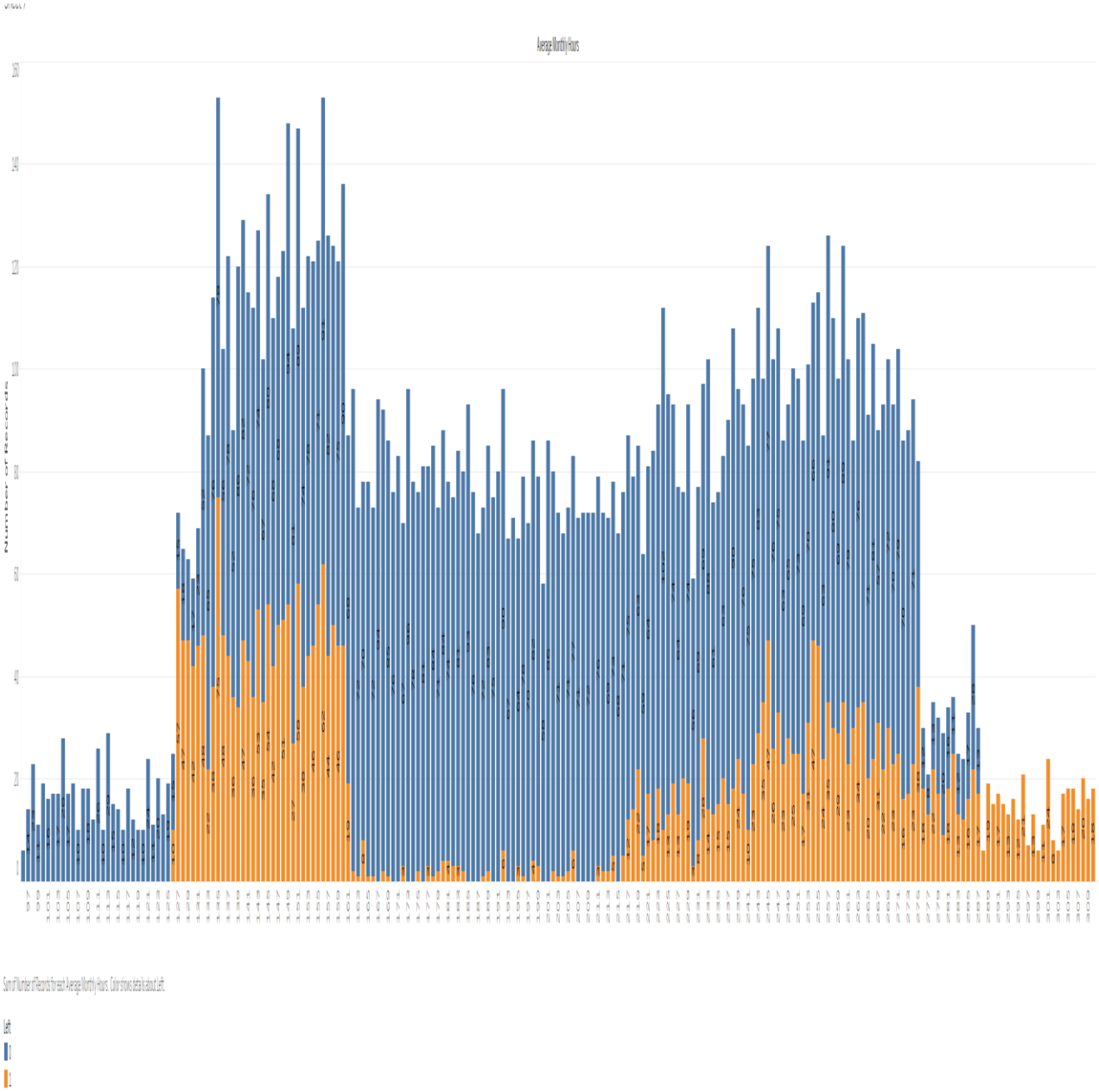


Sheet 6



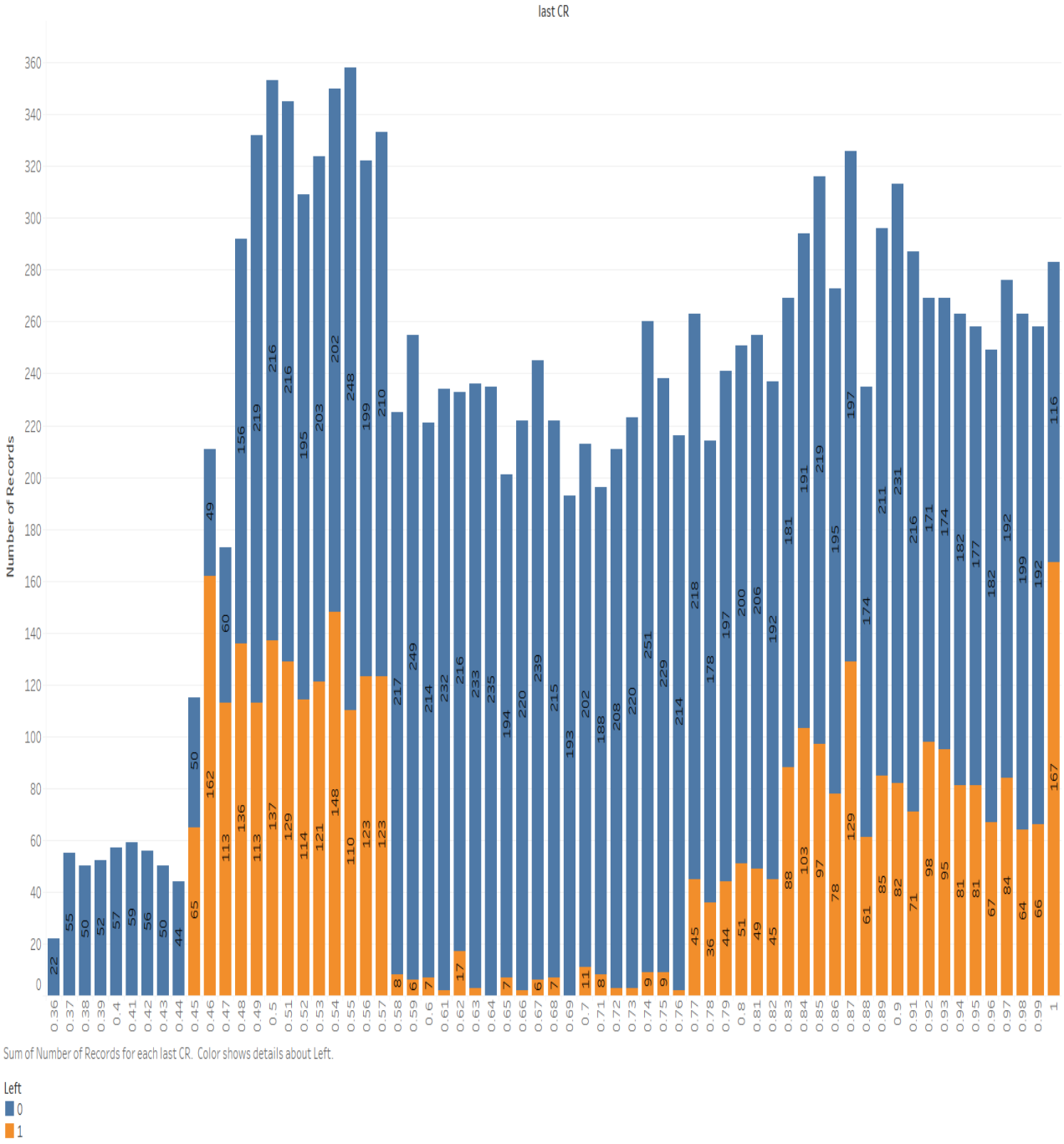
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT), AND NUMBER OF SIMULTANEOUS APPOINTMENTS/PROJECTS

*(**Interpretation**- Maximum number of Personnel have left large number of simultaneous appointments/projects(>6 .*



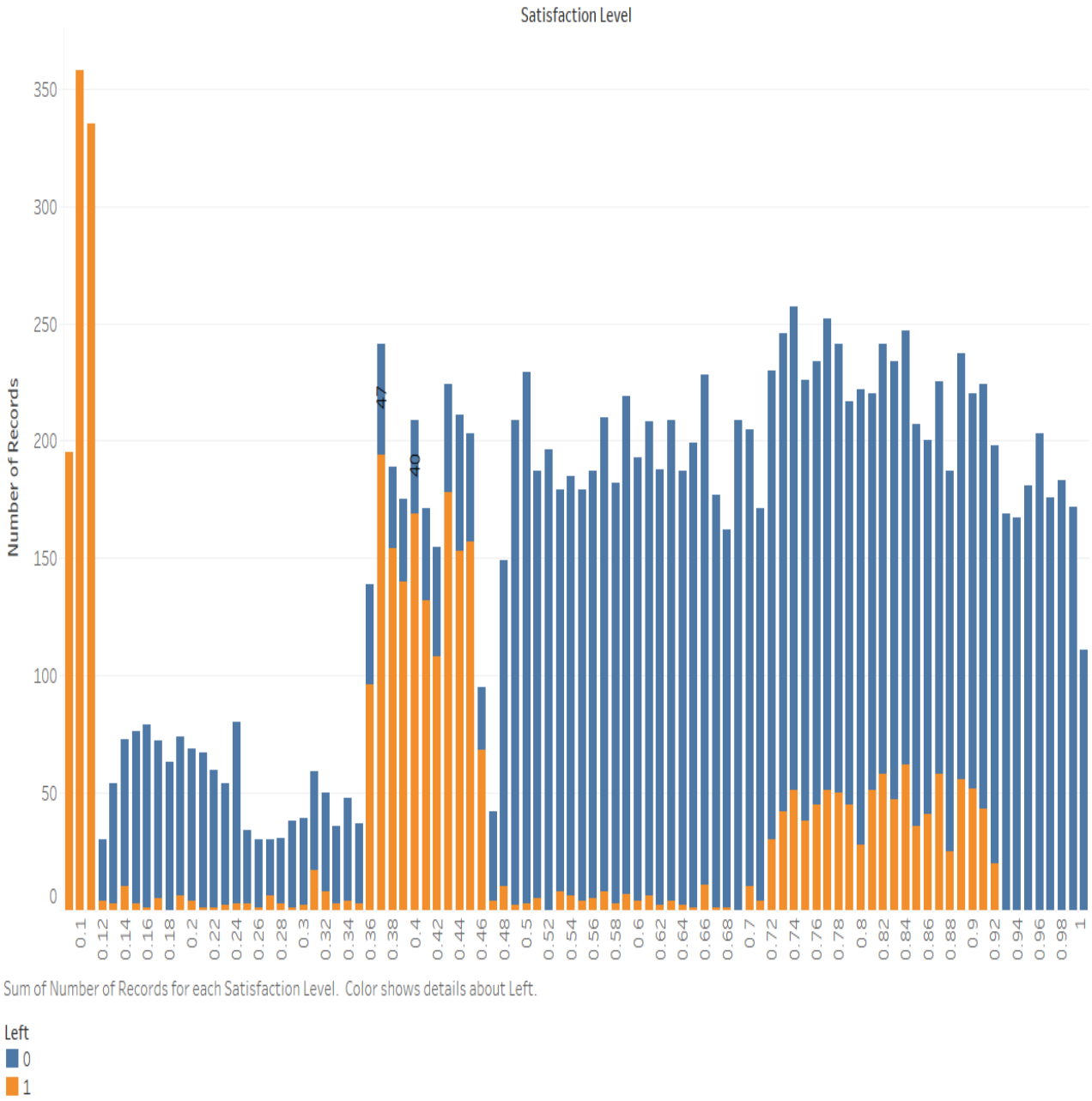
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND NUMBER OF WORKING HOURS/MONTH)

*(**Interpretation**- As the number of hours/month has increased, the resignation rate has increased. All employees who have put more than 287 hrs/month have left)*



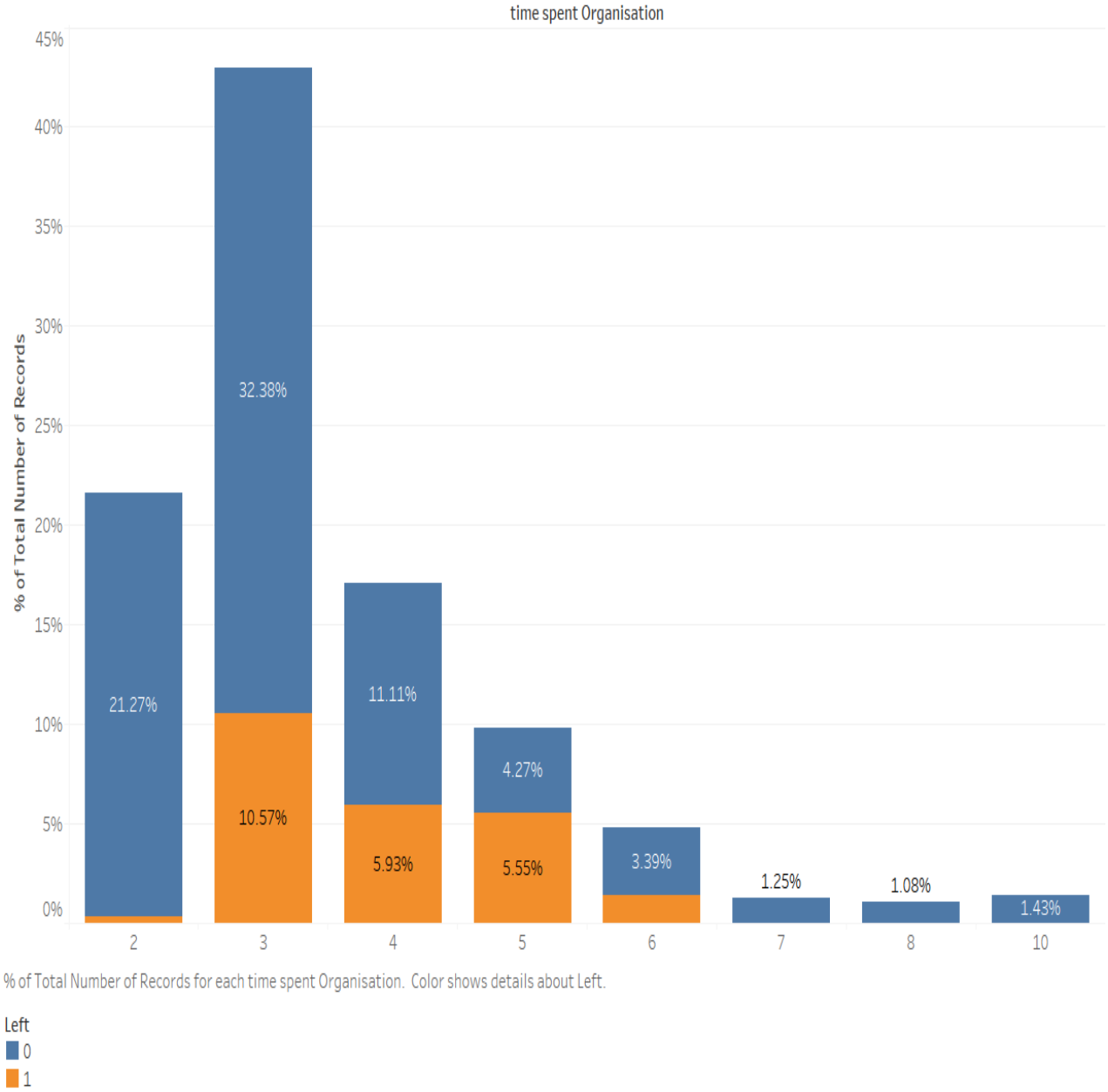
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND LAST CR)

*(**Interpretation**- Personnel leaving are from various CR range. Mostly from .45-.57 and .77-1. 167 out of 283 have left, who have been given 100% CR)*



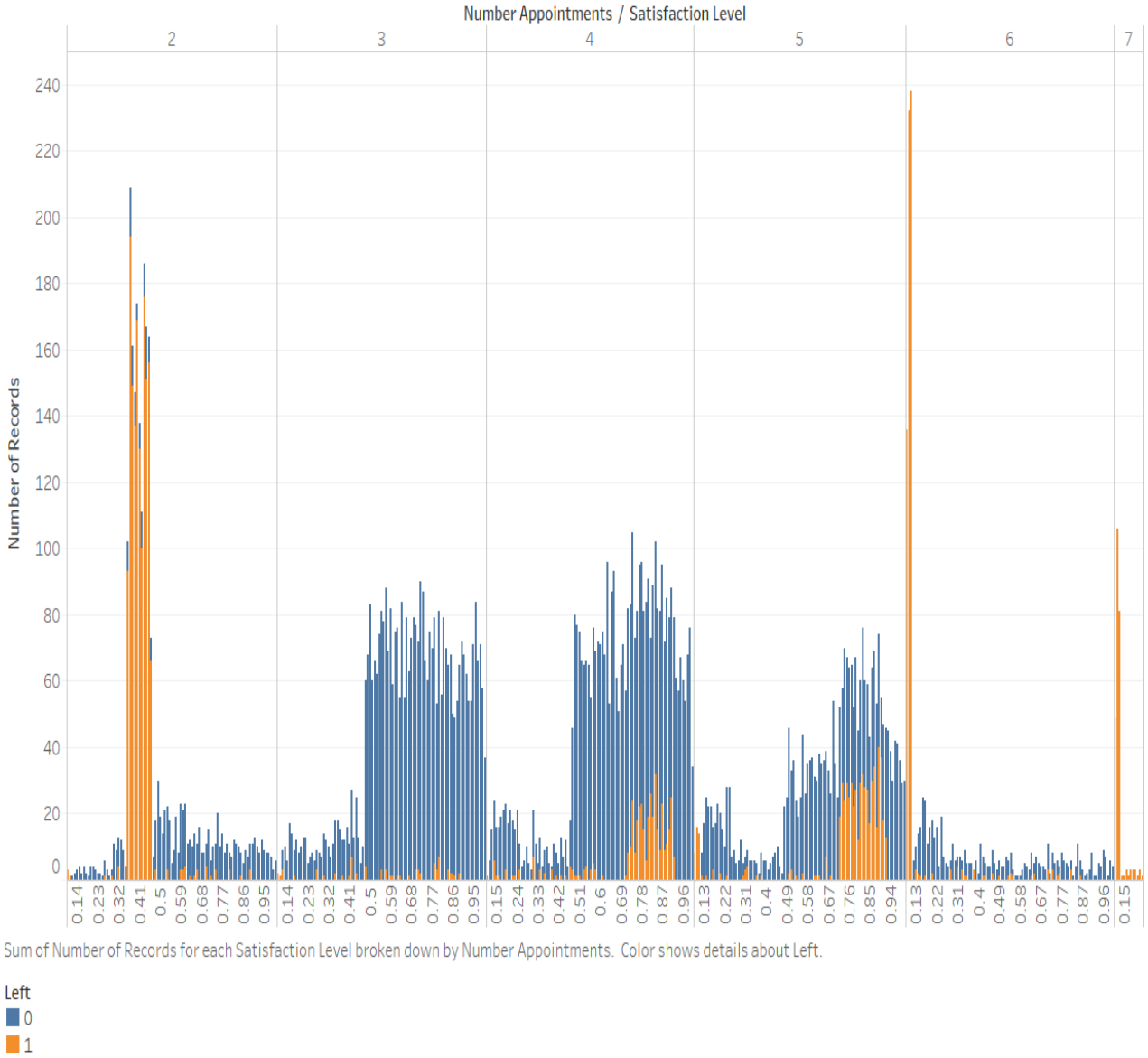
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND SATISFACTION LEVEL)

*(**Interpretation**- Most of the unsatisfied personnel have left. Also, there are large number of personnel left who have satisfaction level between .36-.46)*



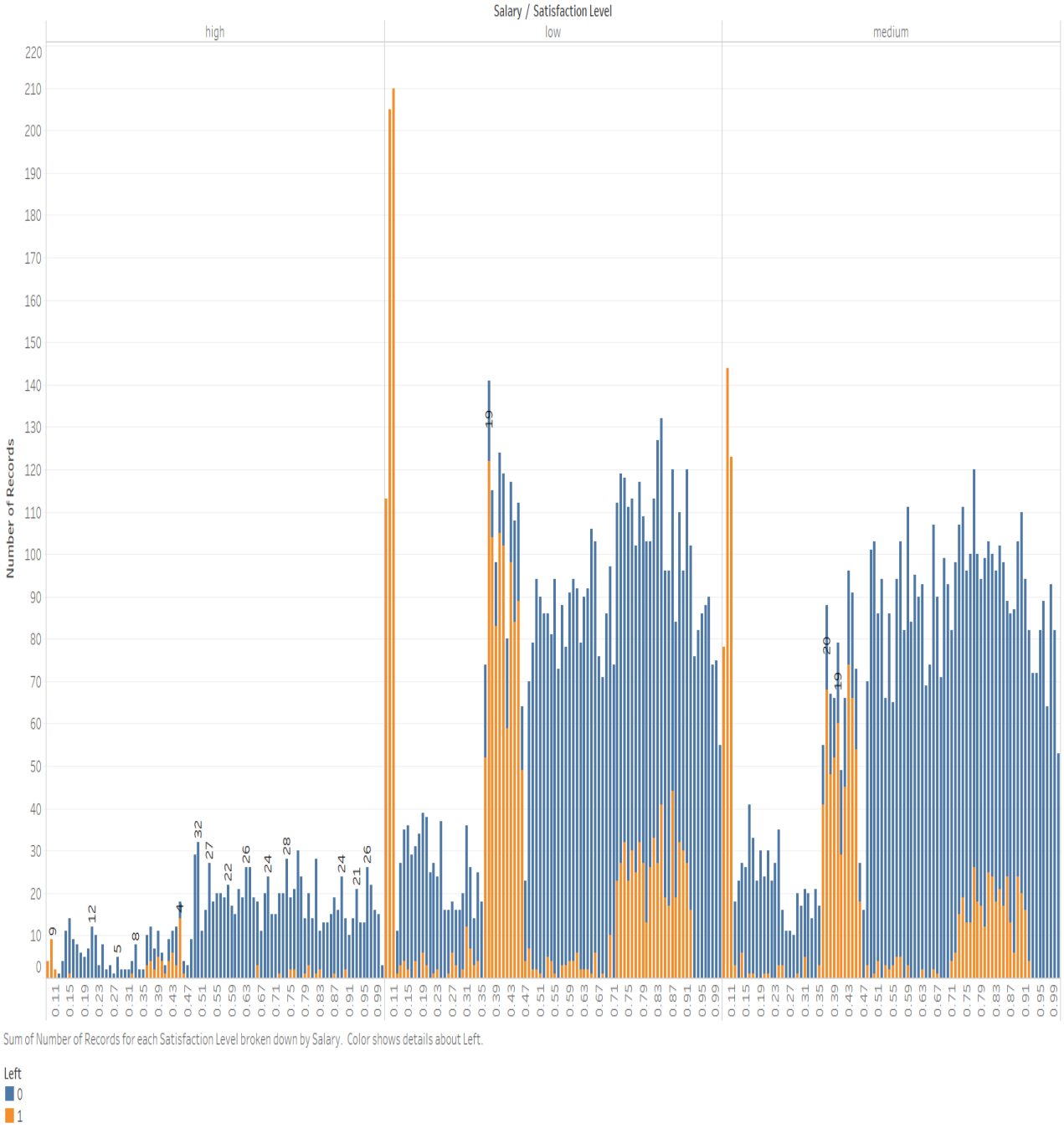
RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND TIME SPENT IN THE ORGANISATION)

*(***Interpretation**- *There are less number of personnel leaving within 2Yrs of service. Also, personnel who have stayed in the organisation beyond 6 yrs, hardly leave. There are 6,123 employees having spent more than 3 years within the company and evaluations higher than 0.7 and 30.44% (1,864 employees) of them have left the company)*



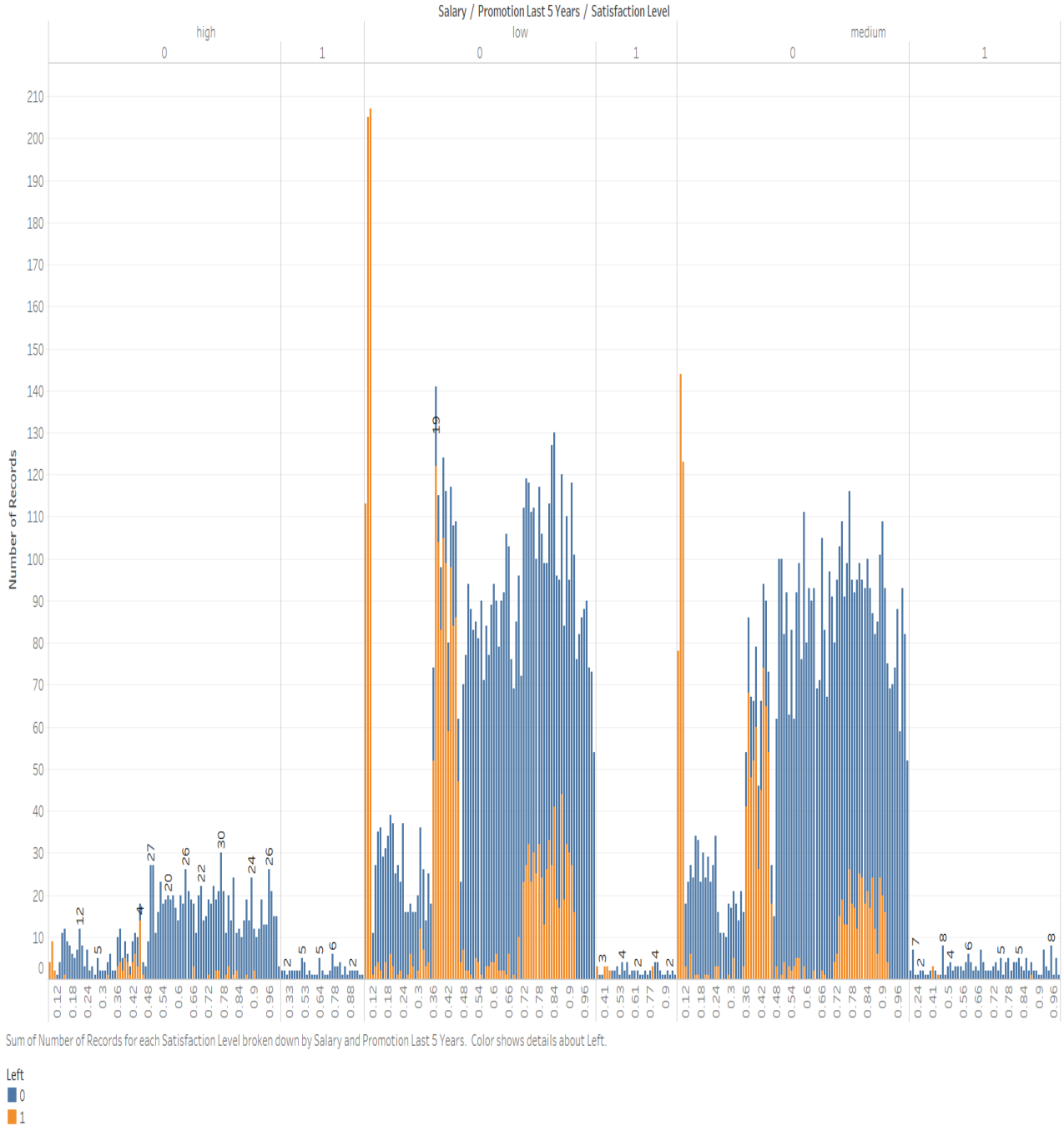
<u>RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT) AND NO OF SIMULTANEOUS APPOINTMENTS(PROFILES) / PROJECTS AND SATISFACTION(0-1)</u>

*(**Interpretation**- Personnel who have less or very large(>6) no of simultaneous projects, seems to have less satisfaction and have left the organisation. The other clusters of orange in the graph may have been due to some policy decision of the firm and may be considered to be the outliers)*



RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT), SALARY(3- HIGH, 2-MEDIUM, 1-LOW) AND SATISFACTION (0-1)

*(**Interpretation**- Personnel in low and medium salary group with low satisfaction level have left the organisation. The other clusters of orange in the graph may have been due to some policy decision of the firm)*

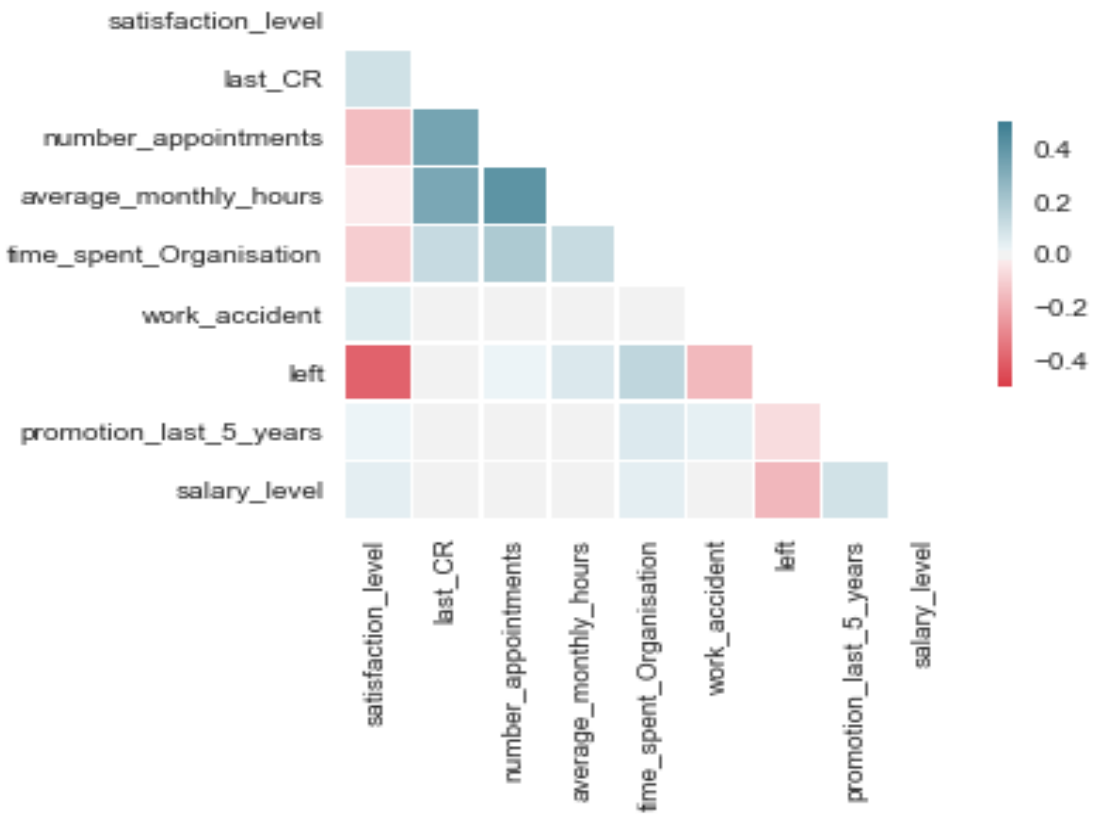


RELATIONSHIP BETWEEN PERSONNEL LEFT (0-WORKING, 1-LEFT), SALARY(3- HIGH, 2-MEDIUM, 1-LOW) AND PROMOTION IN LAST 5 YRS

*(**Interpretation**- Personnel in low and medium salary group with low/medium satisfaction level have left the organisation, if not promoted in last five years.)*

Data Preparation

6. **Analyse correlations.**Calculating the correlations between all different combinations of data allows us to get first hints on why people leave in order to orient our analysis into the right perspective. Red fields mean negative correlations, blue fields indicate positive correlations ie. The field on the crossing point between “left” and “satisfaction_level” is dark red which means that when the satisfaction level of employees goes down, the value of “left” goes up (which means that employees are leaving, as satisfaction_level can only be 0 or 1).

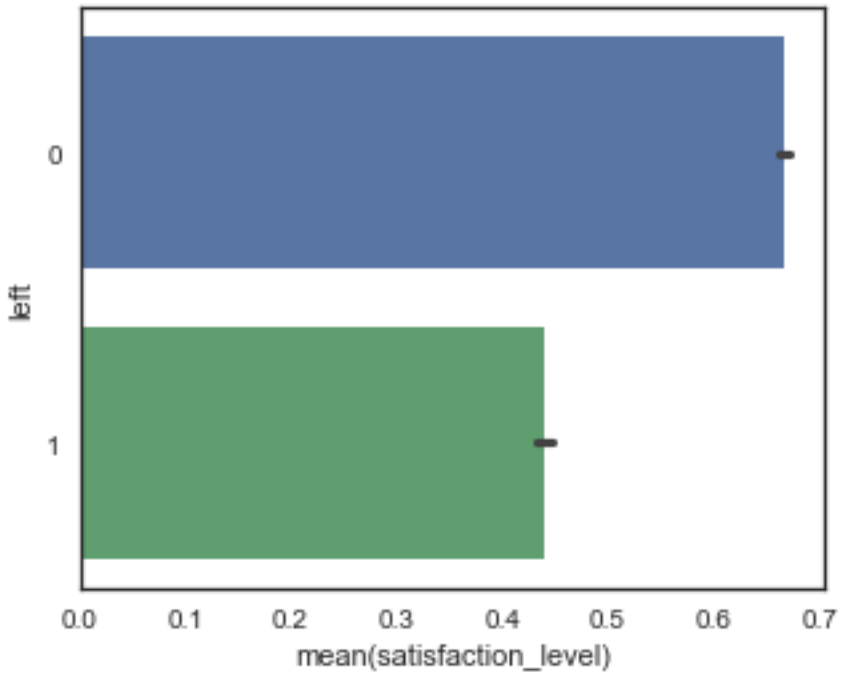



```
#Cdr Santosh Kumar"@2018
# Analyze correlations
sns.set(style="white")

# Compute the correlation matrix
corr = data.corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(5, 4))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(10, 220, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.5,
            square=True, xticklabels=True, yticklabels=True,
            linewidths=.5, cbar_kws={"shrink": .5}, ax=ax)
```

7. From the plot, it can be seen very clearly that the **satisfaction level of the employees is strongly related to the fact that they leave the organisation**. Other significant factors making people leave are the **salary level, the work accidents and if they have got a promotion during the last 5 years**. Regarding the correlation between the satisfaction level and the other dimensions, it can be understood, **that satisfaction mainly decreases when the number of simultaneous appointments(profiles)/ projects and the time spent in the company increase.**

8. **<u>Focus on employee satisfaction</u>.** The mean satisfaction level of personnel leaving is less(.45) as compared to personnel still in the organisation(.65).

```
#Cdr Santosh Kumar"@2018
#Analyze features
sns.set(style="white")
f, ax = plt.subplots(figsize=(5, 4))
sns.barplot(x=data.satisfaction_level,y=data.left,orient="h", ax=ax)
```

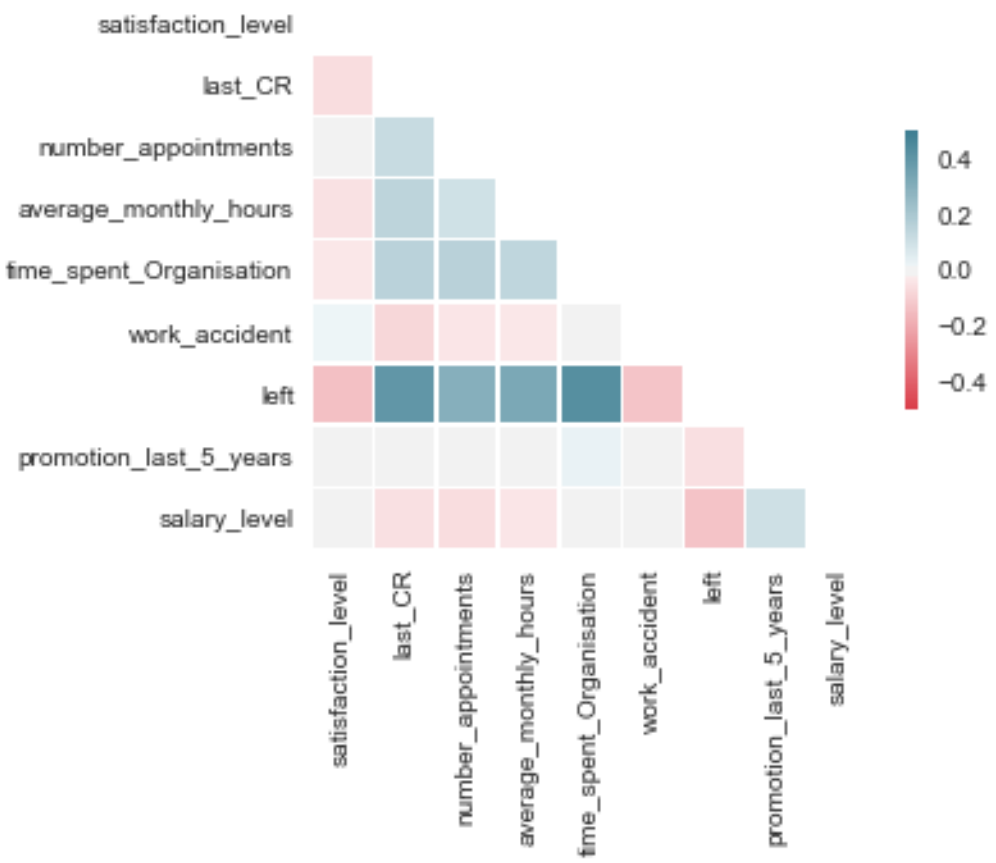


9. We plot histogram for how employee satisfaction looks like for the different departments. In the left range is the charts for employees still in the company (left=0), in the right range is the the employees that have already left the company (left=1).

```
#Cdr Santosh Kumar"@2018
sns.set(style="darkgrid")
g = sns.FacetGrid(data, row="department", col="left", margin_titles=True)
bins = np.linspace(0, 1, 13)
g.map(plt.hist, "satisfaction_level", color="steelblue", bins=bins, lw=0)
```

10. Employees that have left can be split up into 3 distinct groups; those who were unsatisfied, those who were very satisfied and those in between. There is no smooth transition between those groups like there is for employees still in the company. It appears quite clearly why unsatisfied people leave the company, but it could be interesting to explore why satisfied employees left. Therefore, we plot the correlation chart including satisfied employees only (satisfaction_level > 0.7).

```
#Cdr Santosh Kumar"@2018
# Analyze correlations
sns.set(style="white")

# Compute the correlation matrix
corr = data.corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(5, 4))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(10, 220, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.5,
            square=True, xticklabels=True, yticklabels=True,
            linewidths=.5, cbar_kws={"shrink": .5}, ax=ax)
```



11. This chart shows that **satisfied employees leave the company when they work on a high number of appointments/ projects or a high number of hours each month and when they have already spent a long time in the company**. Leave decisions are also influenced by a low salary level and when employees haven't got a promotion during the last 5 years.

```
#Cdr Santosh Kumar"@2018
#Count key employees
#all key employees
key_employees = data.loc[data['last_CR'] > 0.7].loc[data['time_spent_Organisation'] >= 3]
key_employees.describe()
```

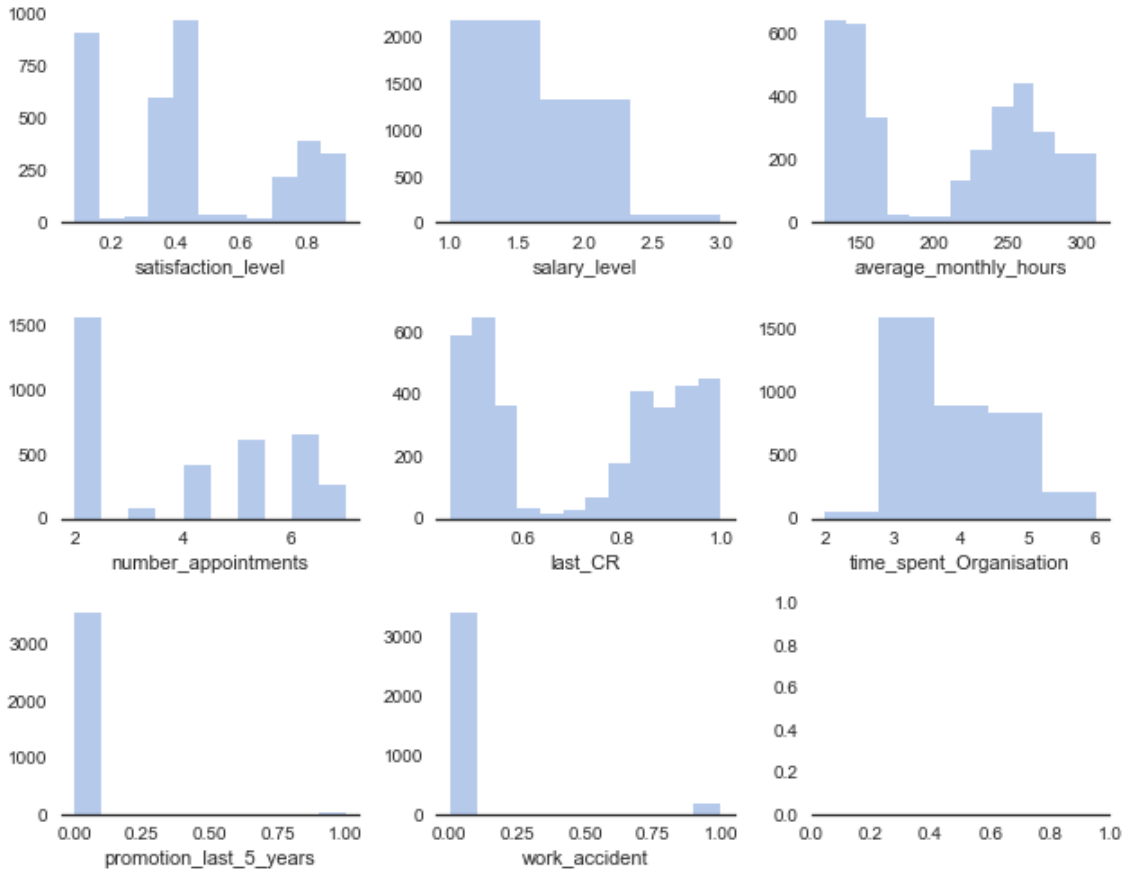
	satisfaction_level	last_CR	number_appointments	average_monthly_hours	time_spent_Organisation	work_accident	left	promotion_last_5_ye
count	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000	6123.000(
mean	0.603059	0.864467	4.301813	219.332027	4.127225	0.138984	0.304426	0.022
std	0.287024	0.083265	1.215323	48.552356	1.383378	0.345958	0.460201	0.149
min	0.090000	0.710000	2.000000	96.000000	3.000000	0.000000	0.000000	0.000(
25%	0.430000	0.800000	3.000000	180.000000	3.000000	0.000000	0.000000	0.000(
50%	0.690000	0.870000	4.000000	229.000000	4.000000	0.000000	0.000000	0.000(
75%	0.830000	0.930000	5.000000	258.000000	5.000000	0.000000	1.000000	0.000(
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000(

```
#Cdr Santosh Kumar"@2018
#lost key employees
lost_key_employees = key_employees.loc[data['left']==1]
lost_key_employees.describe()
```

	satisfaction_level	last_CR	number_appointments	average_monthly_hours	time_spent_Organisation	work_accident	left	promotion_last_5_years
count	1864.000000	1864.000000	1864.000000	1864.000000	1864.000000	1864.000000	1864.0	1864.000000
mean	0.462328	0.896357	5.325107	257.935622	4.622318	0.047747	1.0	0.002146
std	0.354372	0.067570	1.061447	30.686214	0.695091	0.213287	0.0	0.046287
min	0.090000	0.710000	2.000000	130.000000	3.000000	0.000000	1.0	0.000000
25%	0.100000	0.840000	5.000000	243.000000	4.000000	0.000000	1.0	0.000000
50%	0.505000	0.900000	5.000000	258.000000	5.000000	0.000000	1.0	0.000000
75%	0.820000	0.950000	6.000000	278.000000	5.000000	0.000000	1.0	0.000000
max	0.920000	1.000000	7.000000	310.000000	6.000000	1.000000	1.0	1.000000

12. **Other factors**. We plot histogram for various parameters to analyse on other factors that describe leaving employees.

```
#Cdr Santosh Kumar"@2018
#Leavers analysis
sns.set(style="white", palette="muted", color_codes=True)

# Set up the matplotlib figure
f, axes = plt.subplots(3, 3, figsize=(9,7))
sns.despine(left=True)

#people that left
leavers = data.loc[data['left'] == 1]

# Plot a simple histogram with binsize determined automatically
sns.distplot(leavers['satisfaction_level'], kde=False, color="b", ax=axes[0,0])
sns.distplot(leavers['salary_level'], bins=3, kde=False, color="b", ax=axes[0, 1])
sns.distplot(leavers['average_monthly_hours'], kde=False, color="b", ax=axes[0, 2])
sns.distplot(leavers['number_appointments'], kde=False, color="b", ax=axes[1,0])
sns.distplot(leavers['last_CR'], kde=False, color="b", ax=axes[1, 1])
sns.distplot(leavers['time_spent_Organisation'], kde=False, bins=5, color="b", ax=axes[1, 2])
sns.distplot(leavers['promotion_last_5_years'],bins=10, kde=False, color="b", ax=axes[2,0])
sns.distplot(leavers['work_accident'], bins=10,kde=False, color="b", ax=axes[2, 1])
plt.tight_layout()
```



13. It can be seen that leaving employees tend to have lower salaries, a higher number of projects, higher monthly working hours and fewer promotions. All this sounds logic as the satisfaction analysis provides the same conclusions and satisfaction is closely related to the leave decision. It is to be noted that a **large number of the employees leaving the company are people with a high evaluation and several years spent in the company**. These employees are highly valuable assets that should not be lost.There are 6,123

employees having spent more than 3 years within the company and evaluations higher than 0.7 and 30.44% (1,864 employees) of them have left the company.

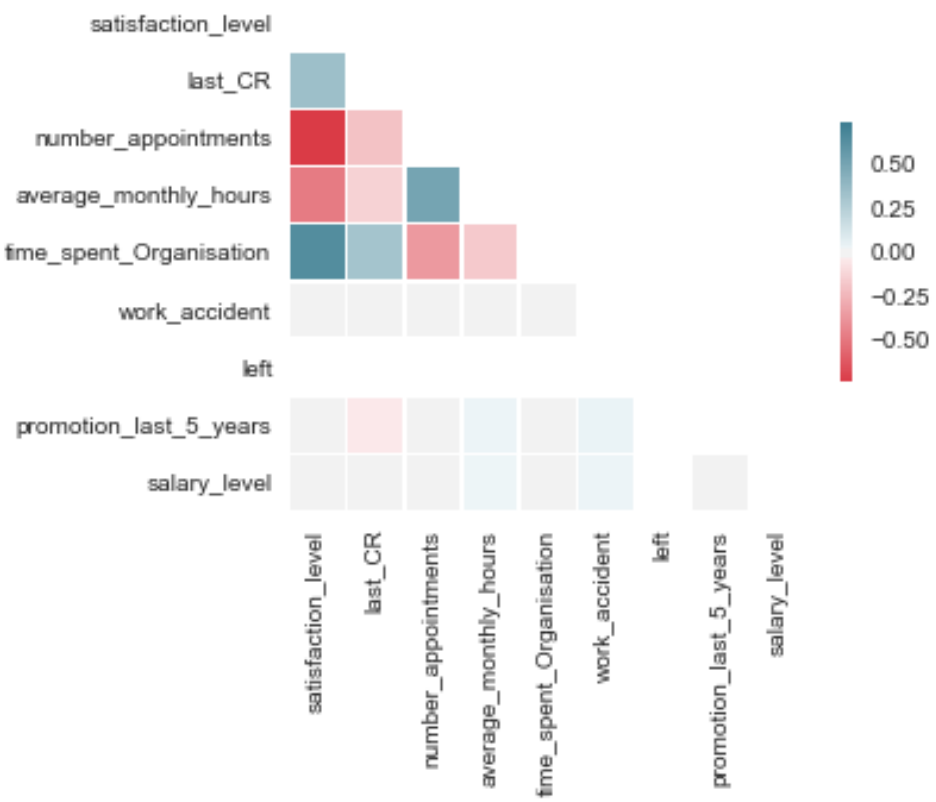
```
#Cdr Santosh Kumar"@2018
print ("Number of key employees: ", len(key_employees))
print ("Number of lost key employees: ", len(lost_key_employees))
print ("Percentage of lost key employees: ", round((float(len(lost_key_employees))/float(len(key_employees))*100),2),"%")
```

```
Number of key employees:  6123
Number of lost key employees:  1864
Percentage of lost key employees:  30.44 %
```

14. **<u>Why do good employees leaving</u>?** Before trying to predict which people are most likely to leave the organisation, it is important to understand what makes high performers leave.

```
#Cdr Santosh Kumar"@2018
# Why do performing emploees leave ?
#filter out people with a good last evaluation
leaving_performers = leavers.loc[leavers['last_CR'] > 0.7]
```

```
#Cdr Santosh Kumar"@2018
sns.set(style="white")

# Compute the correlation matrix
corr = leaving_performers.corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(5, 4))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(10, 220, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.5,
            square=True, xticklabels=True, yticklabels=True,
            linewidths=.5, cbar_kws={"shrink": .5}, ax=ax)
```



15. This correlation matrix shows that good employees have left mainly because of a high number or simultaneous projects and a high amount of working hours.

SECTION V

Use Case- Using Machine Learning Models Predict which Employees would Leave the Company

1. In the first part of our analysis, some plots have been generated to get some basic insights about data set and the features have showed quite good correlation rates. In this part data prepared would be used to predict which employees will leave the company. A model needs to be created to predict extremely accurately which employees will leave the company and who will stay. The precision level of the model has to be high.

```
#Cdr Santosh Kumar"@2018
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
#disable warnings to make notebook smoother
import warnings
warnings.filterwarnings('ignore')
```

2. Loading all 15000 lines of the data into a data frame and display the column names as well as the first five rows of the data set.

```
#Cdr Santosh Kumar"@2018
# Load data frame
data = pd.read_csv('C:/Users/Ranjita/Desktop/HR Analytics/raw_data.csv')
data.head(10)

```

	name	satisfaction_level	last_CR	number_appointments	average_monthly_hours	time_spent_Organisation	work_accident	left	promotion_last_5_years
0	AMAR	0.38	0.53	2	157	3	0	1	0
1	AKBHAR	0.80	0.86	5	262	6	0	1	0
2	ANTONY	0.11	0.88	7	272	4	0	1	0
3	RAM	0.72	0.87	5	223	5	0	1	0
4	LAKSHMAN	0.37	0.52	2	159	3	0	1	0
5	SITA	0.41	0.50	2	153	3	0	1	0
6	BAZIGAAR	0.10	0.77	6	247	4	0	1	0
7	HAMRAZ	0.92	0.85	5	259	5	0	1	0
8	KULVEER	0.89	1.00	5	224	5	0	1	0
9	WILSON	0.42	0.53	2	142	3	0	1	0

3. **Feature Preparation**.Feature preparation is required for feature standardization. After having transformed the department labels into integers and put all values into a feature matrix, feature standardization converts all values into floats ranging between -1.0 and 1.0. This procedure contributes considerably to the accuracy of the predictions.

```
#Cdr Santosh Kumar"@2018
#Feature preparation

leave_df = pd.read_csv('C:/Users/Ranjita/Desktop/HR Analytics/raw_data.csv')
col_names = leave_df.columns.tolist()

# Isolate target data
y = leave_df['left']

# We don't need these columns
to_drop = ['name', 'salary', 'left']
leave_feat_space = leave_df.drop(to_drop,axis=1)

# Pull out features for future use
features = leave_feat_space.columns

# convert label features to integers
from sklearn import preprocessing
le_sales = preprocessing.LabelEncoder()
le_sales.fit(leave_feat_space["department"])
leave_feat_space["department"] = le_sales.transform(leave_feat_space.loc[:,('department')])

# transforme the whole feature space into a matrix
X = leave_feat_space.as_matrix().astype(np.float)

# standardize all features
scaler = preprocessing.StandardScaler()
X = scaler.fit_transform(X)

print ("Feature space holds %d observations and %d features" % X.shape)
print ("Unique target labels:", np.unique(y))
```

```
Feature space holds 14999 observations and 9 features
Unique target labels: [0 1]
```

Data Modelling

4. **Prediction function.** The prediction function uses a 3-fold cross-validation and it takes several prediction algorithms as input (we would use it to compare 3 different algorithms). We call this function both for comparing the algorithms and predicting the leave probabilities.

```
#Cdr Santosh Kumar"@2018
# prediction function
def run_cv(X,y,clf_class, method, **kwargs):

    from sklearn.model_selection import cross_val_predict
    # Initialize a classifier with key word arguments
    clf = clf_class(**kwargs)
    predicted = cross_val_predict(clf, X, y, cv=3, method=method)
    return predicted
```

5. **Prediction Algorithms.** The following classification techniques would be evaluated:-

(b) **Support Vector Machine (SVM).** It is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is

number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.[20]

(c) **Random Forest Classifier.** It creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.[21]

(d) ***K*-Nearest Neighbors Algorithm (*k*-NN)**. It is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. [22]

6. **Compare prediction algorithms.** In order to chose the best classification algorithm we need to compared a Support Vector Classifier (SVC), a Random Forest Classifier (RF) and a K-Nearest-Neighbors classifier (KNN). All models produce quite satisfying results (between 95% and 99% accuracy) without even tuning the algorithms. We would retain the Random Forest Classifier for the prediction as it produces the best results (98.8% accuracy).

```
#Cdr Santosh Kumar"@2018
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier as RF
from sklearn.neighbors import KNeighborsClassifier as KNN
from sklearn import metrics

def accuracy(y, predicted):
    # NumPy interprets True and False as 1. and 0.
    return metrics.accuracy_score(y, predicted)

print ("Support vector machines:")
print ("%.3f" % accuracy(y, run_cv(X,y,SVC, method='predict')))
print ("Random forest:")
print ("%.3f" % accuracy(y, run_cv(X,y,RF, method='predict')))
print ("K-nearest-neighbors:")
print ("%.3f" % accuracy(y, run_cv(X,y,KNN, method='predict')))
```

```
Support vector machines:
0.958
Random forest:
0.990
K-nearest-neighbors:
0.952
```

7. **Calculate confusion matrices.** A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test

[20] *https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/*
[21] *https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1*
[22] *https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm*

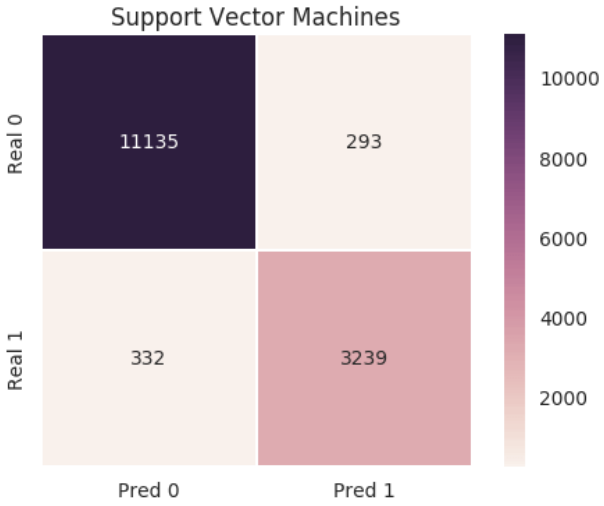
data for which the true values are known**.** Confusion matrices are a very useful tool to get an overview of the accuracy of a prediction. The matrix provides a value for each crossing point between predicted and realized classes. The confusion matrix has 4 fields: Left-Left, Left-Not left, Not left-Left and Not left-Not left. The Left-Left and Not left-Not left fields contain by far the largest amount of values. It indicates high quality of the prediction.

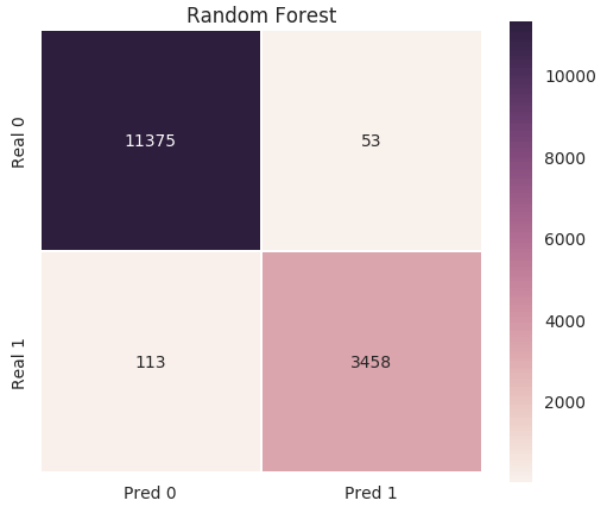
```
#Cdr Santosh Kumar"@2018
from sklearn.metrics import confusion_matrix

y = np.array(y)
class_names = np.unique(y)

# calculate confusion matrices
confusion_matrices = [
    ( "Support Vector Machines", confusion_matrix(y,run_cv(X,y,SVC, method='predict')) ),
    ( "Random Forest", confusion_matrix(y,run_cv(X,y,RF, method='predict')) ),
    ( "K-Nearest-Neighbors", confusion_matrix(y,run_cv(X,y,KNN, method='predict')) ),
]

# show confusion matrix values
print (confusion_matrices)
```

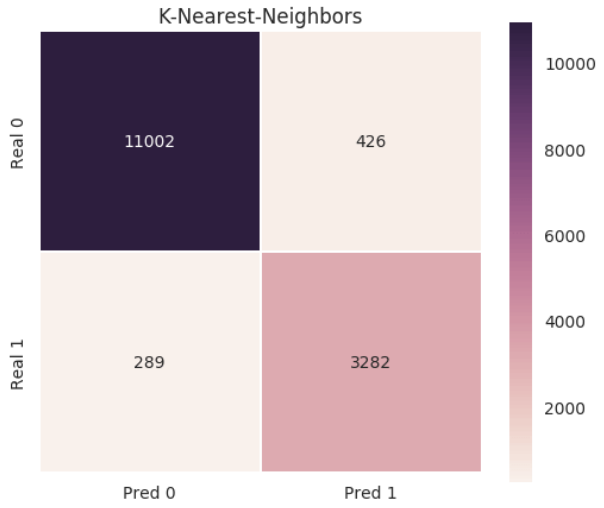
```
[('Support Vector Machines', array([[11135,   293],
       [  332,  3239]])), ('Random Forest', array([[11374,    54],
       [  127,  3444]])), ('K-Nearest-Neighbors', array([[11002,   426],
       [  289,  3282]]))]
```

8. **Plotting confusion matrices.** Using Seaborn visualization library of Python we plot the confusion matrices for 3 prediction algorithms. The Random Forest Classifier model predict 11,375 times correctly that an employee will stay.

```
#Cdr Santosh Kumar"@2018
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline

# draw confusion matrices
for cf in confusion_matrices:

    ax = plt.axes()
    ax.set_title(cf[0])

    df_cm = pd.DataFrame(cf[1], index = ["Real 0", "Real 1"], columns = ["Pred 0", "Pred 1"])
    plt.figure(figsize = (6,5))
    sn.heatmap(df_cm, annot=True, ax = ax, square=True, fmt="d",linewidths=.5)
```


Support Vector Machines
11135
293
332
3239
Real 0
Real 1
Pred 0
Pred 1
10000
8000
6000
4000
2000


Random Forest
11375
53
113
3458
Real 0
Real 1
Pred 0
Pred 1
10000
8000
6000
4000
2000


K-Nearest-Neighbors
11002
426
289
3282
Real 0
Real 1
Pred 0
Pred 1
10000
8000
6000
4000
2000

9. **<u>Calculating Prediction Probabilities for all Employees</u>.** We use the predict function to calculate the probabilities for staying and leaving (left=0, left=1) for all 15000 employees in our data set. As every predictor makes several hundreds or even thousands of predictions we can compare our probabilities with the actual outcome of each class.For example, for the group of employees for which we predicted a 60% probability of leaving, the actual leaving percentage is 78%. As we can see the predicted probabilities for the two main classes (pred_prob=0% and pred_prob=100%) are very close to the real probabilities which shows another time that our model is extremely accurate.

```
#Cdr Santosh Kumar"@2018
# Use 10 estimators so predictions are all multiples of 0.1
pred_prob = run_cv(X, y, RF, n_estimators=10,  method='predict_proba',)

pred_leave = pred_prob[:,1]
is_leave = y == 1

# Number of times a predicted probability is assigned to an observation
counts = pd.value_counts(pred_leave)

# calculate true probabilities
true_prob = {}
for prob in counts.index:
    true_prob[prob] = np.mean(is_leave[pred_leave == prob])
    true_prob = pd.Series(true_prob)

# pandas-fu
counts = pd.concat([counts,true_prob], axis=1).reset_index()
counts.columns = ['pred_prob', 'count', 'true_prob']
counts
```

	pred_prob	count	true_prob
0	0.0	9448	0.004234
1	1.0	3139	0.993629
2	0.1	1363	0.019809
3	0.2	427	0.030445
4	0.9	184	0.951087
5	0.3	145	0.068966
6	0.4	88	0.204545
7	0.8	76	1.000000
8	0.7	53	0.792453
9	0.5	39	0.564103
10	0.6	37	0.783784

10. **Generate key employees with leaving/ staying probabilities.** In the last step we filter out all key employees (employees that have a last evaluation higher than 0.7) that are still in the company. This gives us a table of about 6000 employees. In order to be able to alert managers about employees that are most likely to leave we order the employee list by their leaving probability and save the whole list as a CSV file.

```
#Cdr Santosh Kumar"@2018
#create a dataframe containing prob values
pred_prob_df = pd.DataFrame(pred_prob)
pred_prob_df.columns = ['prob_not_leaving','prob_leaving']

#merge dataframes to get the name of employees
all_employees_pred_prob_df = pd.concat([leave_df, pred_prob_df], axis=1)

#filter out employees still in the company and having a good evaluation
good_employees_still_working_df = all_employees_pred_prob_df[(all_employees_pred_prob_df["left"] == 0)
                                                              & (all_employees_pred_prob_df["last_evaluation"] >= 0.7)]

good_employees_still_working_df.sort_values(by='prob_leaving', ascending=False, inplace=True)

#write to csv
good_employees_still_working_df.to_csv("C:/Users/Ranjita/Desktop/HR Analytics/good_employees_leaving_prob1.csv")
```