Human Resource Analytics on Data Science Employment Based on Specialized Skill Sets with Salary Prediction

Tee Zhen Quan [a], Mafas Raheem [a,1,*]

[a] School of Computing, Asia Pacific University of Technology and Innovation, Technology Park Malaysia, Kuala Lumpur, 57000, Malaysia
[1] raheem@apu.edu.my.
* corresponding author

ARTICLE INFO

Article history
Received March 15, 2022
Revised August 7, 2022
Accepted October 17, 2022

Keywords
human resource analytics
specialized skill sets
data science
salary prediction

ABSTRACT

The research aims to perform meaningful human resource analysis on data science employment using the strong influences of specialized skills set with assisting salary prediction. With explosive big data development, a data science job shortage has occurred with high accurate recruitment demand to hire suitable professionals for specific data science roles. To achieve such outcomes, the current data science employment trends were analyzed based on a secondary dataset. Useful analytics insights for job securement and better career development were provided through the main dashboard. Besides, the significant in-demand data science skill variables were also identified for further effective model building. Particularly, certain data pre-processing techniques were performed extensively to prepare and optimize the dataset for the mentioned human resource analytics purposes. The ensemble model was selected as the most suitable salary prediction model with the lowest Average Squared Error (ASE) on validation. Despite the low prediction accuracy caused by numerous filtered skill variables, the salary prediction model's main objective was to interpret the relationships between input variables and the target salary levels variable. Overall, the results from both the human resource analytic dashboard and salary prediction model were tally where a detailed analytic report was provided to encourage different data science roles with specific and effective career development guidance, using salary as the motivation key.

1. Introduction

The era of IR 4.0 is underway and has gained widespread acceptance in numerous developed and developing countries [1], including Malaysia [2]. This technology and data-driven revolution are marked by high-speed and large-volume of data that affect almost all industries and internal operations of businesses. Companies implement digital transformations to enhance their operational efficiency, effectiveness, and maintainability [3]. Typically, a collaborative data ecosystem is developed to integrate different stakeholders such as customers, suppliers, production, and technical support to achieve a common organizational or business objective. One popular example is the optimization of manufacturing by using data to identify potential ways of improving specific manufacturing processes or collaborating with other departments. Additionally, prediction models are used to prevent excessive production and machine failure in product sales and maintenance. This approach has been successfully employed by Netflix and Amazon to predict customer preferences, leading to increased sales and growth.

On the other hand, the expansion of big data has led to a significant demand for data science jobs to transform these data into useful business values. These jobs entail precise scopes, ranging from data collection, processing, analysis, model building, and evaluation to a final business presentation, each requiring different levels of expertise. In 2016, a global consulting firm, McKinsey, conducted employment research that showed a shortage of 250,000 data science jobs in the United States [4], a remarkable employment shortage not encountered in other industries recently. Despite the introduction of many automatic and intuitive analytic software to the market, the 2020 McKinsey report showed similar data science job shortages [5], indicating the continuing importance of data science knowledge to fully utilize such software and interpret useful business insights. Moreover, the evidence suggests that big data expansion, especially, is still going on strongly.

Accurate recruitment is also essential for effective Human Resource (HR) budgeting and planning for an organization's position, directing recruitment costs, work productivity, and outcome. These represent internal capability and are responsible for promoting a company's market competitiveness and business expansion [6]. To achieve these goals, organizations must understand their recruitment requirements, the exact specialized skill sets, and knowledge required for the specified job scope, and the current recruitment trends and demand from an employee's perspective.

The challenge of job shortages in data science is not limited to developed countries like the United States, but also to developing and undeveloped countries, especially in the IR 4.0 transformation. Despite the availability of automated tools, the high demand for data science professionals still exists, and it is growing. Consequently, human resource analytics and salary predictions are not new in the general employment domain, and previous research has commonly aimed to discover essential employment trends for hiring organizations and job applicants. However, such analysis outcomes are too general for data science, a high-level employment domain that carries individual responsibilities and tasks for different job positions.

Therefore, accurate recruitment is the key component to resolving the data science employment problems in real life, which requires specific skills for a justified job position. This research performs extensive analysis and salary prediction specifically for the data science employment domain with a specialized skill set. The research produced outcomes that can effectively resolve real problems, providing insights into demanding data science skill sets and technical or soft knowledge that leads to higher salaries and employment opportunities.

2. Literature Review

2.1 Data Science Skills and Job Titles

1) Job Titles: The domain of data science employment is vast and encompasses various job scopes such as data collection, management, analytics, machine learning, and more. This necessitates the need for diverse job titles or positions that cater to each specialized task. To create an analytics dashboard and salary prediction model, it is imperative to have a comprehensive understanding of the entire data science employment domain, including the various job titles and specialized skill sets. By collating information from different credible sources, the top 10 most in-demand data science job titles are organized and presented in Table 1. While the top three job titles, namely data scientist, data analyst, and data engineer, are general roles, the remaining job titles have specific job scopes.

Table 1. Top 10 Most Currently Demanded Data Science Jobs

No	Job Title	Description
1	Data Scientist	The role of a data scientist is the most widespread and general, but it also carries the highest level of responsibility and skill set requirement. In particular, data scientists focus heavily on Research and Development (R&D) and possess an extensive set of skills to manage all project tasks effectively [7].

2	Data Analyst	Similar to a data scientist but differs in that it involves working with established algorithms rather than engaging in research. In general, the job entails handling all aspects of data management, including data collection, processing, and visualization, to convert data into valuable business insights [7].
3	Data Engineer	This role is responsible for designing, building, and maintaining the data pipeline. Its main function is to ensure the efficient flow of data throughout the system. Additionally, the data engineer is responsible for testing the entire pipeline ecosystem to ensure that it is optimized and ready for use by data scientists and data analysts [8].
4	Data Architect	The responsibilities of a data architect are quite similar to a data engineer, but the former is a specialized role that focuses specifically on designing and creating database systems based on a defined business model [8].
5	Data Storyteller	This is another specialized role that concentrates on data visualization in the final reporting phase. Therefore, it is a potent tool for converting technical information into business terms or stories that are more easily understood [8].
6	Business Intelligence Developer	Similar to a data storyteller, a Business Intelligence (BI) developer utilizes the transformed information to create and implement business strategies that facilitate significant decision-making. They should possess a business background with good BI tool knowledge, and be prepared for new business model deployment [8].
7	Machine Learning Scientist	This role is highly technical and similar to the data scientist role in its focus on research and development. However, it is specifically geared towards researching and programming new machine learning algorithms for predictive purposes [8].
8	Machine Learning Engineer	The job title is comparable to a machine learning scientist, but with less emphasis on technical skills. The main responsibility is to utilize machine learning algorithms for prediction tasks, including evaluating the performance of different models rather than developing new algorithms from the ground up [8].
9	Database Administrator	The role is responsible for the administration of an organization's database. Unlike data engineers and architects, this is a separate role that involves maintaining and monitoring the database rather than being involved in its design and creation process [8].
10	Technology Specialized Roles	As data science is an emerging field, it is expected that many more specialized roles will be in demand soon. In parallel with the growth of artificial intelligence, new roles in data science are emerging, such as Natural Language Processing (NLP) specialists, marketing storytellers, and many others yet to be developed [8].

2) Specialized Data Science Skill Sets: A study was conducted on 15,000 job listings related to data science to identify the most sought-after skills in 2021 [9] ranging from programming languages, tools, techniques, databases, and libraries. To address the issue of the vast array of skills and knowledge required in data science, this review provides a filter for the most in-demand data science skills of 2021, which include Python, SQL, R, Spark, AWS, Java, Tableau, Hadoop, TensorFlow, and Scala. Fig. 1 depicts the top three skills such as Python, SQL, and R, highlighting their critical importance in data science recruitment. Moreover, the remaining two skills, Java and Scala, also fall into the programming language category. Additionally, there is strong demand for cloud-based tools such as AWS and big data implementation tools like Spark and Hadoop in the current data science job market.

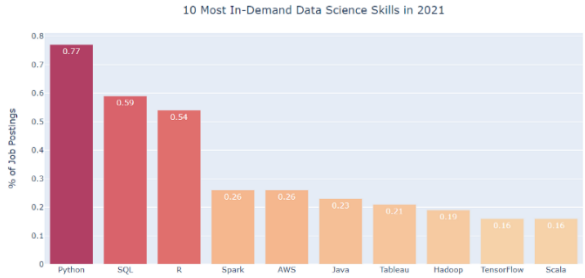

Figure 1. Top 10 Data Science Skills in 2021

The analysis of top growing skills from 2019 to 2021, as depicted in Fig. 2, provides further evidence of the increasing demand for cloud infrastructure. Amazon Web Services (AWS) demonstrated the highest growth rate and dominates the market, while its competitor Google Cloud Platform (GCP) also made the list, albeit with a lower growth rate. Additionally, machine learning or deep learning frameworks such as PyTorch, Scikit-learn, and TensorFlow exhibited significant growth and were not present in the previous analysis. In terms of programming languages, it can be observed that Python's dominance has decreased over the past two years. Surprisingly, the second most demanded language was SQL, which showed significant growth compared to Python, the most demanded functional programming language. This suggests that there is a strong possibility of SQL overtaking Python in demand if the same analysis is conducted in 2022.

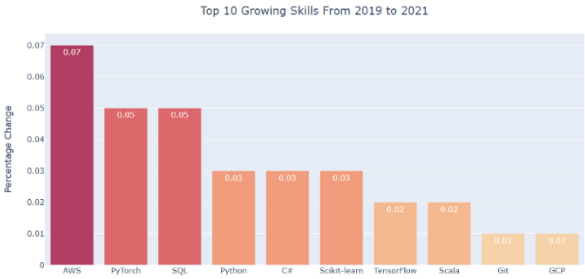

Figure 2. Top 10 Growing Data Science Skills between 2019 and 2021

2.2 Existing Human Resource Analytics

Human Resource (HR) analytics is a type of data analytics that focuses specifically on human resource data, involving data collection, analysis, and reporting. Its primary purpose is to measure the impact of human resource metrics on overall business performance and support informed decision-making based on relevant data [10]. By relying on a data-driven approach, HR analytics can provide more accurate and reliable insights compared to traditional human expertise and assumptions. This is especially important for organizations with limited experienced HR professionals or those with complex job roles. HR analytics can also help organizations predict trends and patterns early on, allowing for dynamic adjustments to maintain or improve workforce productivity. Despite its potential benefits, HR analytics is a relatively new field with much to explore in scientific research as more organizations are beginning to recognize its importance and invest in the proper collection and management of HR data.

Two of the most popular examples of HR analytics are employee turnover and recruitment [11]. In terms of employee turnover, HR analytics can analyze historical data to identify patterns and trends related to why employees quit. This includes analyzing employee behaviour data to better understand their work engagement and productivity. By correlating these data points, patterns can be identified that may lead to employee turnover. Predictive models can then be developed to forecast which employees may quit based on these identified patterns. This helps organizations develop effective HR strategies and decisions to improve the work environment and employee engagement levels.

Regarding recruitment, HR analytics can provide insights into job applicants by analyzing a range of variables such as education level, skills, and experience. This data-driven approach can avoid bias and ensure fair employment opportunities, as the measurement is based on reliable statistical data rather than personal opinion. For example, this research analyzes the most in-demand data science skills for specific roles, helping organizations hire applicants who are most suitable for the job and avoid the costs associated with over-hiring or under-hiring. Accurate recruitment and cost management are critical components of a long-term and effective hiring plan. [12] conducted a study on the use of HR analytics in the banking sector for decision-making purposes. The researchers collected data via questionnaires from various bank managers, focusing on several HR analytics use cases such as employment, training, employee performance, and turnover. The study concentrated on prescriptive analytics by providing decision-making recommendations on the continued use of HR analytics in the banking sector. Many managers agreed that HR analytics facilitated scientific and analytic recruitments with clearly defined policies, justifying the importance of technical and behavioural skills in bank employee recruitment. Additionally, HR analytics showed a positive influence on employee training and development supported by performance evaluation and management. The study found that employees with high satisfaction, competitive salary and benefits in similar job domains, fair treatment, and opportunities were more likely to stay. However, the research was limited to only 15 experienced branch managers which heavily relied on opinions.

Predictive analytics was conducted to identify employee churn issues and dashboards were also developed to perform diagnostic analytics to find the reasons for churn [13]. The Random Forest model performed better than the Logistic Regression model. The final model was used to provide real-time and accurate monitoring of the employee churn rate, and the dashboards presented employee performance and churn reasons. The implementation of this business intelligence is essential to prevent serious production problems caused by high employee turnover, especially in smaller organizations with limited human resources. The analysis also revealed important HR insights, such as employees with previous work accidents being more likely to quit, and only a small proportion of employees who were promoted in the past 5 years left the company. Furthermore, low job satisfaction contributed to a high churn rate. However, significant discrepancies were observed when comparing the actual and predicted turnovers, indicating that relying solely on prediction outcomes may lead to inaccurate interpretations, even with high-performing prediction models like Random Forest.

2.3 Existing Salary Prediction Methodologies

1) Statistical Methods: The term "statistical methods" generally refers to regression techniques that utilize predefined parameters. In a study by [14], a basic form of regression known as graph representation was employed to predict salaries. In general, linearity is expected to exist, as higher job levels are typically associated with higher salaries. However, the research findings indicated that the highest job level resulted in an elevated salary, but the increase was not consistent. This is consistent with real-world employment, as the salaries of top management positions are often incomparable with those of normal positions. Additionally, curve fitting was used to smooth the data and provide more accurate predictions. The polynomial regression was able to handle the exponential salary while allowing direct observation through the graph visualization. While the implementation was simple, data mining tools and programming could be used to explore wider models and improve performance evaluation.

In a similar approach, [15] developed a salary prediction model using linear regression. However, a stepwise method was employed where input variables were added one by one in each step. Initially, the researchers tested the working experience input variable against the salary target variable, which achieved 97% accuracy. After adding the remaining input variables, the baseline linear regression model's Mean Squared Error (MSE) increased significantly. Consequently, the researchers applied a polynomial transformation to the baseline model, which reduced the MSE and achieved a final accuracy of 76%. This experiment demonstrates that simple linear regression can be accurate when handling only one input variable. Multiple factors should be considered in a salary prediction

scenario. Nonetheless, the research did not use cross-validation to evaluate the model's performance, which raises doubts about the results.

In contrast to other studies, [16] utilized a statistical model called Multiple Linear Regression (MLR) where the researchers incorporated job applicants' examination scores, in addition to the typical salary prediction variables, although there was no linear correlation between the examination scores and the salary target variable. Nevertheless, the MLR model achieved a performance of 82% using Root Relative Squared Error (RRSE) evaluation, with other significant variables including logical, quantitative, programming, and language skills. This statistical model allowed the researchers to conclude that personal skill development is more critical than examination scores in the employment sector. Effective data cleaning and outlier treatments were necessary for the model to achieve such performance. However, a limitation was noted in that many engineers in the Indian dataset were unpaid. Thus, the salary prediction model may not be transferable to other locations where underemployment is illegal.

2) Machine Learning Methods: Modern machine learning methods differ from statistical methods where they do not rely on predefined parameters and are considered black-box techniques. However, it can achieve higher prediction accuracy, especially when tuned correctly with high-dimensional variables. [17] developed a popular tree-based machine learning salary prediction model using the built-in feature engineering function, variable importance without uncorrelated contract type and contract time variables. The researchers also discovered that the salary target variable was heavily skewed, with data concentrated at lower values. As a result, the data were transformed using the logarithmic function, which also reduced the noise in the dataset. By implementing an ensembled tree-based model called Random Forest (RF), the researchers concluded that RF achieved higher accuracy of 87.3% compared to the baseline decision model, which only achieved 84.8%. This study showed the reliability and accuracy advantage of ensembled methods by integrating multiple results instead of relying on one. However, machine learning methods heavily rely on the quality of the dataset, while statistical methods focus on the relationship between variables.

Typically, salary predictions involve predicting continuous numeric values. However, in a study by [18], salaries were grouped into different levels and converted into a classification problem using the K-Nearest Neighbors (KNN) algorithm. Similar to the present study's goals, the researchers developed a KNN classifier to predict salaries for Java back-end engineers based on specialized Java skills. This approach differs greatly from others as the optimal K value is initially determined based on distance measurement and K points. The best average accuracy of 88.1% was achieved when the K value was set at 7. This minimal parameter tuning approach demonstrated the advantage of fast model building, but with larger datasets, a higher K value may be necessary, requiring more tuning iterations and time. Evaluation through a confusion matrix revealed that the prediction accuracy of individual salary classes was unbalanced, with the top salary level achieving an exceptional result of 93.3%, while the bottom salary level only achieved 73.1%.

3) Deep Learning Neural Network Methods: Deep learning, a subset of neural networks, refers to a neural network algorithm with more than three layers [19]. For a similar problem, researchers [20] proposed an advanced cooperative neural network model. This two-stage model involves the first skill valuation model that provides a score for the skill, which is then used for the second cooperative salary prediction model. This framework allows for retraining the first model with feedback from the second model. Unlike previous statistical and machine learning methods that require strictly tabular data, the proposed deep learning method can extract information from unlabeled data and utilize a variety of models, including the baseline neural network, text-mining method, gradient boosting, Support Vector Machine (SVM), and linear regression. The proposed model achieved the least Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). However, since only the salary prediction model's performance was evaluated, it is difficult to accurately validate the existence of the skill score.

The same salary prediction problem was addressed by [21] through the proposal of a backpropagation neural network model (BPNN) that utilized gradient descent. The initial BPNN model had 14 input layers, 15 hidden layers, and 1 output layer, and the backpropagation update process was improved

using a gradient descent algorithm to enhance accuracy. Furthermore, the researcher experimented with a better gradient descent algorithm, Small-Batch Gradient Descent (SBGD), with Nadam optimizer for the same model's parameter update. Consequently, the final tuned BPNN achieved a high accuracy of 89.98%. However, the validation dataset consisted of only 100 records, which renders the model's performance evaluation unreliable.

In contrast, [22] proposed a novel approach using Bid-GRU-CNN (Bidirectional-Gated Recurrent Unit-Convolutional Neural Network) to tackle the same problem. Similar to previous studies, the proposed model utilized NLP techniques to process contextual or unlabeled data. However, in this case, the GRU was utilized to extract advanced information from single words and handle noisy data when combined into sentences. Additionally, the CNN layer employed ResNet to perform feature learning. The combined model outperformed the benchmark TextCNN algorithm, achieving the lowest error score according to the MAE evaluation. However, due to its high complexity, the proposed model requires significantly more training time than the ordinary TextCNN algorithm. [23] took a different approach by converting the job listing data into graph representations. A semi-supervised Graph Convolutional Network (GCN) was applied to manually label certain data for improving the accuracy of salary classification. The feature extraction process produced seven main classification features and an additional relationship feature to evaluate the similarity score of different jobs. Adding a spectral filter further improved the initial GCN model's accuracy to 77%. To compare with other supervised learning methods, another dataset was created with proper labels in tabular format. Surprisingly, deep learning neural network models, including the proposed GCN, underperformed with an average accuracy of 62%, while Random Forest achieved an accuracy of 76% with the same features. The GCN is considered a baseline model, and the advanced Region-based Convolutional Neural Network (RCNN) method was proposed to be applied for further performance improvement.

In general, data science jobs require specialized skills for specific job scopes, but there are some exceptions. For instance, data scientists and data analysts have multiple responsibilities, while data engineers may perform the job of a data architect. However, the demand for data science skills is not stable due to the high competition and availability of tools and techniques in the current market. Therefore, it is crucial to conduct up-to-date human resource analytics and salary predictions to reflect the latest data science employment trends. In the field of human resource analytics, employee turnover and recruitment are the most popular use cases. For instance, employee turnover predictive analytics research offers real-time monitoring and control. Many studies have demonstrated the importance of data-driven HR analytics, especially in employee performance monitoring, training, and development. However, before conducting salary predictive analytics, an HR visual analysis of the actual data is necessary to avoid conflicting results. By reviewing existing salary prediction methods, skill-based variables have demonstrated a significant positive influence. While regression models allow variable relationships to be interpreted, machine learning and deep learning models are more accurate in predicting salary on both linear and non-linear variables and require fewer data preprocessing steps. Deep learning models are particularly effective in processing unlabeled job listing data through NLP and text mining. However, machine learning methods overall achieve higher accuracy with less implementation complexity and easier parameter tuning using suitable techniques. A limitation of some past studies is the lack of sufficient records to reliably evaluate the model's performance. Overall, this paper fills a significant research gap in data-driven HR analytics, particularly in the accurate recruitment domain.

3. Methodology

CRoss Industry Standard Process for Data Mining (CRISP-DM) was chosen for its straightforward implementation flow and adaptability to various data science problems that involve understanding the business, preprocessing the data, and building the model. Additionally, the final deployment stage serves as a robust endpoint where the process concludes only after passing the evaluation stage through several iterations or cycles.

Figure 3. Vertical Slicing Strategy in CRISP-DM [24]

Referring to Fig. 3, the CRISP-DM methodology employs a vertical slicing strategy where a quick and narrow vertical slice is performed during each cycle. This involves addressing a specific problem piece by piece in each cycle, resulting in multiple vertical releases with frequent feedback updates. The benefits of this approach include increased validation of the research's aims and objectives before final deployment, thereby increasing the success rate while enabling strong implementation flexibility at a minimal cost.

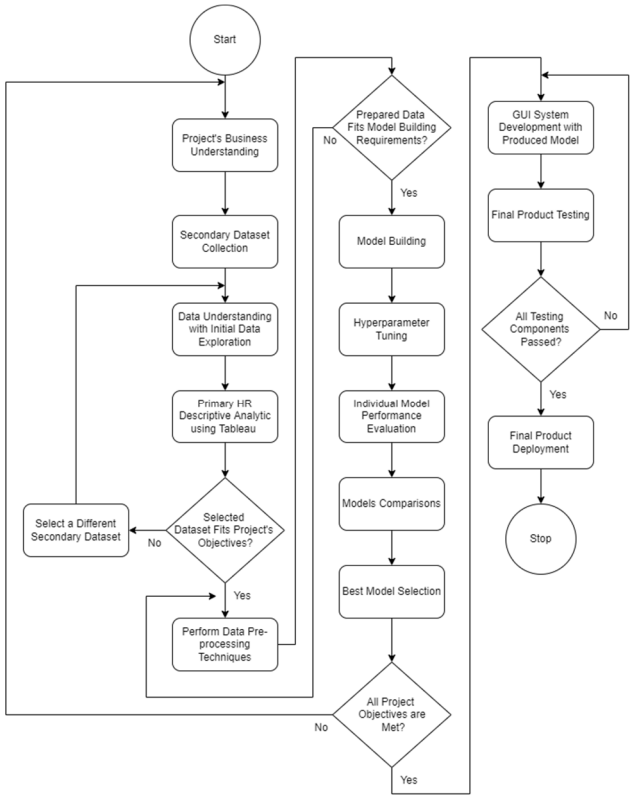

Figure 4. Detailed Research Design Flowchart Diagram

The business problem, as well as determining and comprehending the scope, objectives, and aim of the study was set to perform extensive analysis and salary prediction specifically for the data science employment domain. A secondary dataset was used with 1300 records and 400 variables. The data quality and integrity were ensured as the dataset was scrapped from actual data science job listings in Glassdoor. Both data understanding and business understanding were cross evaluated. The data understanding provided a necessary foundation for further analysis and ensured that the dataset is suitable for the business objectives.

Data pre-processing or preparation was involved to ensure the data is free from noise which is always considered the most important phase of any data science project. The dataset underwent feature selection, data imputation, and partitioning activities to make it more suitable for the intended analytics. Suitable machine learning algorithms were selected with interpretation capability to provide advanced career development recommendations. The statistical regression model and decision tree were built, and further tuning was done using a grid search algorithm to identify the best hypermeter combinations. The model evaluation was conducted using continuous prediction metrics to compare and determine the most appropriate prediction model. The deployment is often overlooked in many data science projects, despite its importance and carried out when required. Fig. 4 shows the detailed research design flowchart.

4. Analysis

4.1 Human Resource Analytics

An Exploratory Data Analysis (EDA) was conducted using interactive dashboards to present the findings in a coherent narrative that can be understood even by those with novice knowledge of data analysis as shown in Fig. 5, 6, 7 and 8.

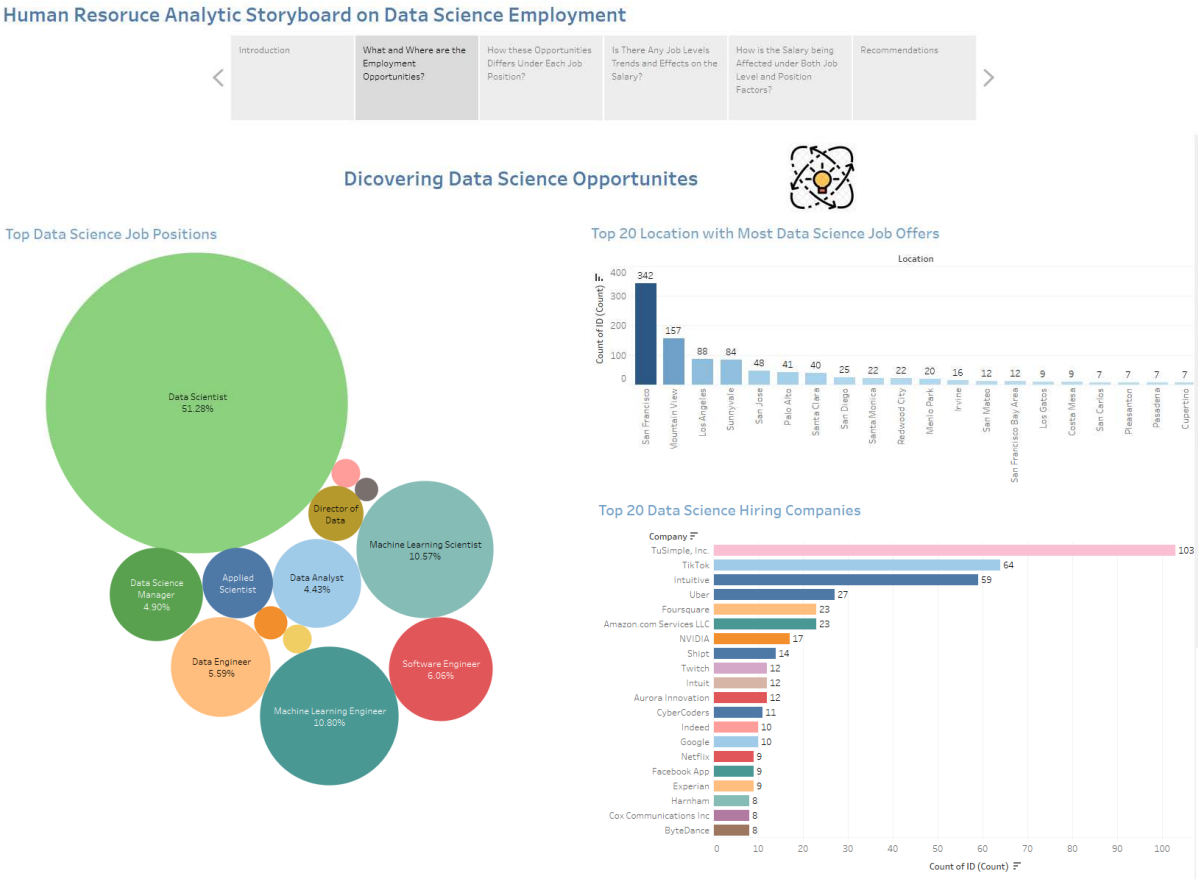

Figure 5. First HR Analytic Dashboard

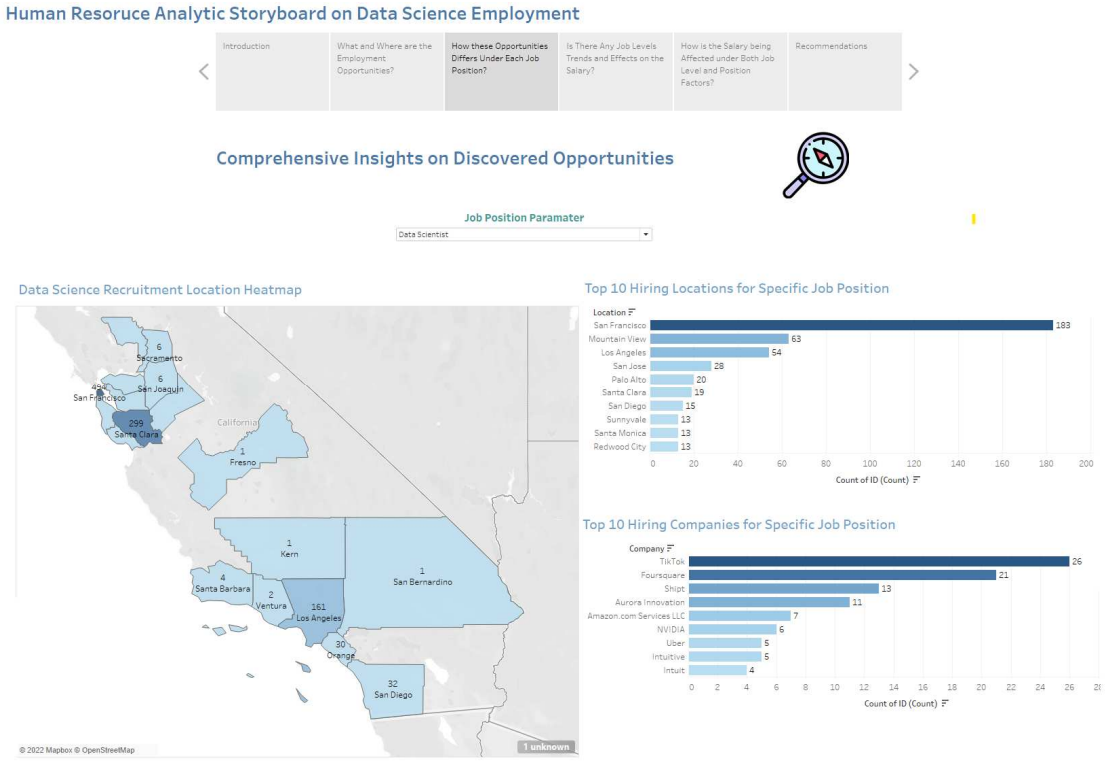

Figure 6. Second HR Analytic Dashboard

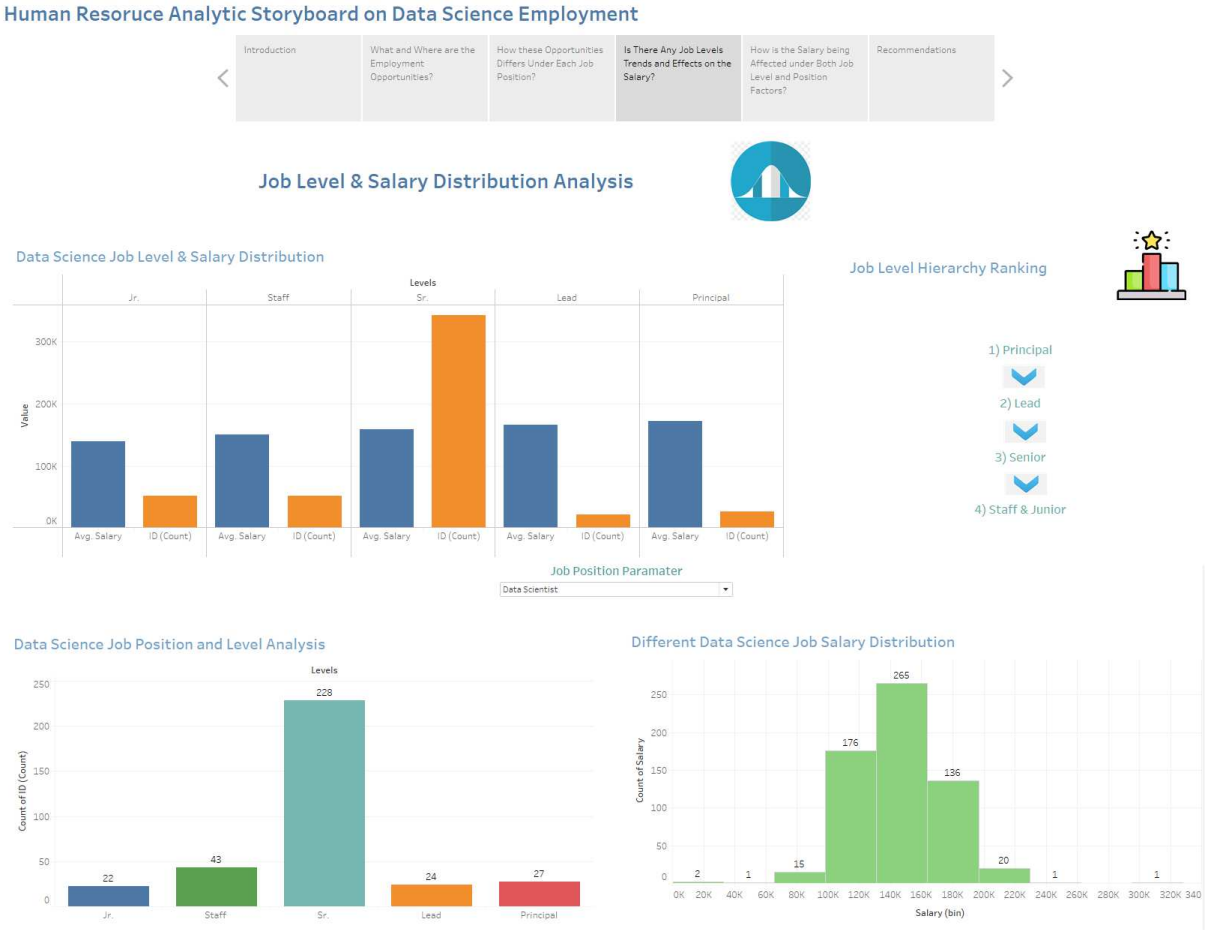

Figure 7. Third HR Analytic Dashboard

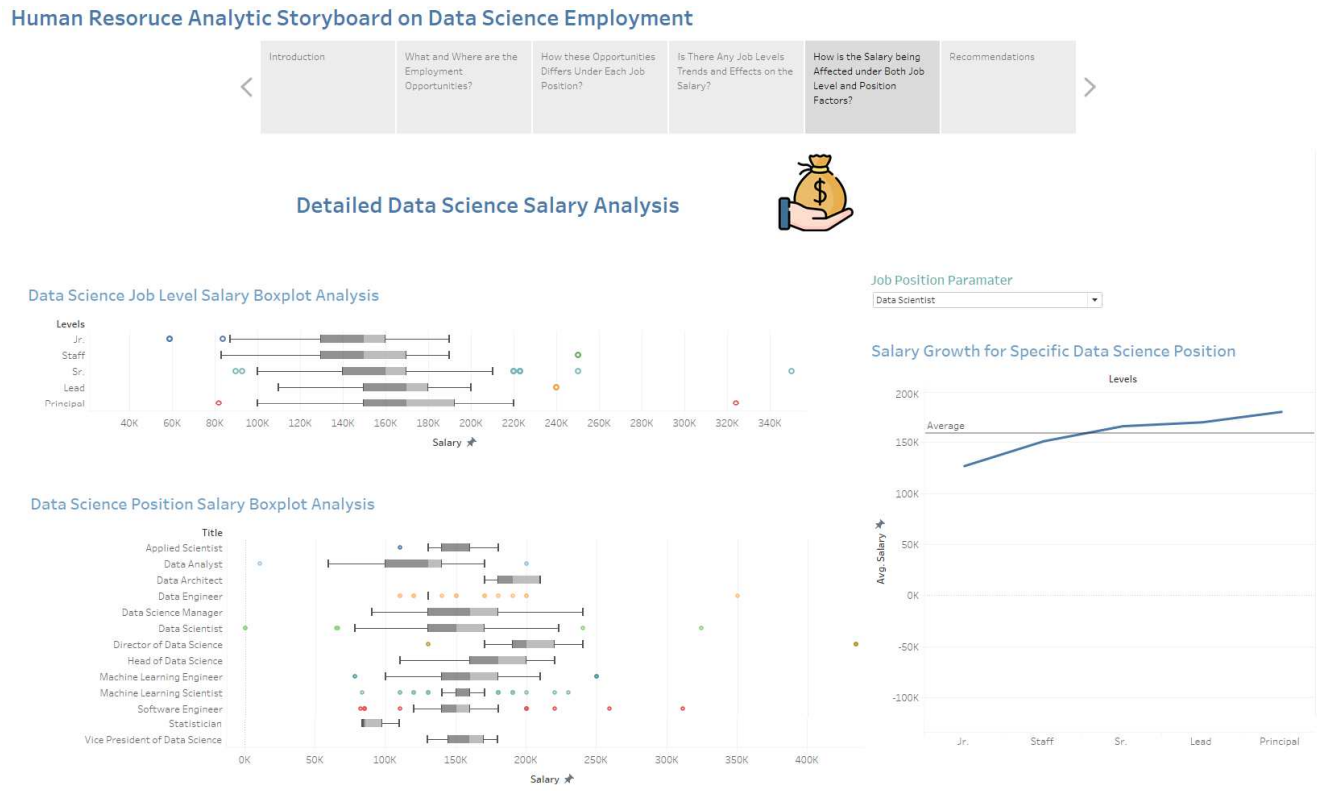

Figure 8. Fourth HR Analytic Dashboard

Skill variable visualization was designed to perform a follow-up HR analysis. Fig. 9 shows the top 10 general data science skill sets.

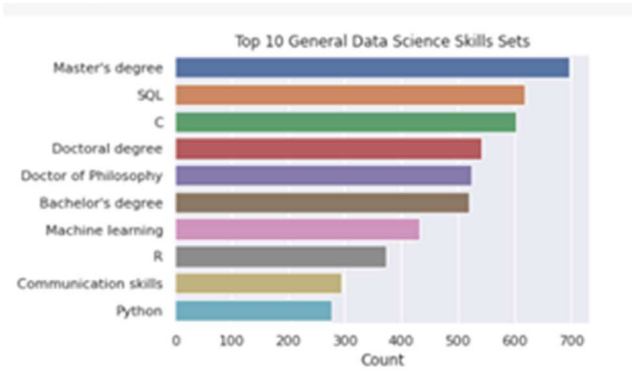

Figure 9. Top 10 General Data Science Skills Sets

The process was repeated to create a list of the top 10 most frequently occurring skill sets for various job positions, job levels, and salary levels. This was done to perform primary feature selection for more efficient model building in subsequent stages. The list of significant skill variables is identified and justified below.

a) Education Qualifications: Bachelor's degree, Bachelor of Science, Master's degree, Master of Science, Doctoral degree, Doctor of Philosophy.
b) Programming Skills: R, Python, C++, SQL, C, SAS, NoSQL.
c) Software of Tool Skills: Elasticsearch, Windows, New Relic, Tableau, Microsoft Excel, Torch, Sailpoint, Spark, Stata, TensorFlow, PyTorch, Azure, SSIS, DevOps, Keras.
d) Technical Skills: Business intelligence, Natural language processing, Machine Learning, Data visualization, Deep learning, System design, Marketing automation, Financial services, Editing, Computer networking, and Predictive analytics.

e) Soft Skills: Leadership, Research, Writing skills, Communication skills, Analysis skills, Analytics, Marketing, Continuous improvement, Mentoring, Ontology.

4.2 Salary Prediction Model

The SAS Enterprise Miner platform was used for the data preparation, model building, interpretation, and comparison processes as depicted in Fig. 10. The dataset was imported and subjected to various data pre-processing techniques, such as feature selection, data imputation, and partitioning, and the two selected prediction or classification algorithms were HP Regression and HP Tree, both of which are high-performance models that automatically perform parameter optimizations and use more complex structures to achieve higher prediction accuracy than baseline regression and decision tree models. Additionally, an ensembled model was built by averaging the classification outputs of both the HP Regression and HP Tree models to produce a more reliable prediction result. A "Surrogate Tree" was also added for interpretation purposes. Finally, the "Model Comparison" nodes evaluated the performance of the three identified models using classification metrics, and the results of model interpretation and comparison were further discussed in the results and discussion section.

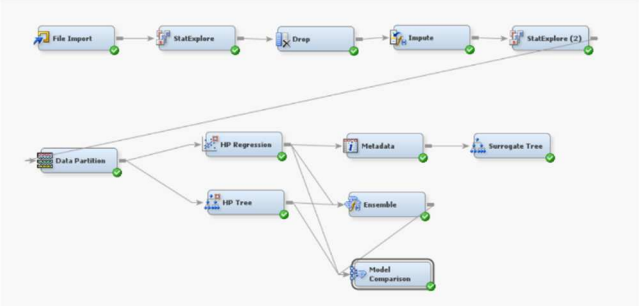

Figure 10. Implementation Flow Diagram in SAS Enterprise Miner

5. Results and Discussion

5.1 Human Resource Analytics

1) HR Analytic Dashboard – Fig. 5: The "Top Data Science Job Position" using a bubble chart indicates that data scientist is the dominant job position, comprising more than 50%. Machine learning scientists and engineers have a share of around 10%, followed by data engineers and data analysts come in 5th and 6th place, respectively. This result slightly differs from the literature review, which stated that data scientist, data engineer, and data analyst were the top three data science job titles. Nonetheless, all three positions are still popular options in this analysis. Other positions, such as data science manager, applied scientist, and director of data science, follow suit, while minority positions like data architect, statistician, head of data science, and vice president of data science contribute less than 1%.

The bar chart depicting "Top 20 Location with Most Data Science Job Offers" indicates that more than 25% of the data science job offers are from San Francisco which is the financial, technological, and commercial hub of Northern California [24]. Other popular locations include Mountain View, Los Angeles, Sunnyvale, San Jose, Palo Alto, and Santa Clara, which are major cities in California. The remaining 35% of job offers are distributed among 74 less popular cities.

The "Top 20 Data Science Hiring Companies" represented using a bar chart reveals that TuSimple, a truck company, has offered around 8% of the data science job opportunities in California, despite the dataset containing 527 different companies. TikTok and Intuitive are also prominent companies with significant shares. All of the top 20 companies belong to industries such as social media, transportation, information technology, entertainment, and e-commerce.

2) HR Analytic Dashboard – Fig. 6: The heatmap titled "Data Science Recruitment Location Heatmap" reveals that 65% of data science job opportunities are in the northern part of California, specifically in San Francisco and Santa Clara. However, there are very few or no job opportunities in the central region of California. The job offers only increase when moving down to Southern California, especially in the Los Angeles County region.

A drill-down analysis was conducted to examine the top hiring locations for specific data science job positions. The "Top 10 Hiring Locations for Specific Job Position" bar chart shows that most positions have the same geographic distribution pattern, with San Francisco continuing to be the state with the highest job offers for a particular position. However, data analyst and applied scientist positions have more evenly distributed job opportunities across different locations. Additionally, for certain job positions like machine learning engineer and data analyst, the location of Mountain View has surpassed San Francisco in terms of job positions offered. It was also found that over 80% of job offers for data engineers are located in Sunnyvale, which was only ranked fourth in general. Therefore, these locations offer significant opportunities for job applicants seeking such specific jobs.

Similarly, the "Top 10 Hiring Companies for Specific Job Position" bar chart performs a drill-down analysis of the top hiring companies for data science job positions. The top three companies identified, namely TuSimple, TikTok, and Intuitive, appear in almost all job positions, except for the director of data science position where there are no job offers due to missing values. TuSimple, a transportation company that specializes in autonomous trucking [25], has over 90% of its data science roles in machine learning scientist and software engineer positions. This is because they require machine learning skills to optimize their travelling route and autonomous driving feature, as well as software engineering skills to implement those algorithms in their work with human interactions. On the other hand, Intuitive, a company that offers robotic surgical systems, has over 90% of its data science roles in data engineer positions. This is because healthcare and medical companies usually have large and complex unstructured data such as screening results and heart diagnostic signals. Therefore, the company invests heavily in data engineers. Overall, TuSimple is an excellent opportunity for data science applicants or graduates who plan to work as machine learning scientists or software engineers, while Intuitive offers great opportunities for data engineers.

3) HR Analytic Dashboard - Fig. 7: The bar chart titled "Data Science Job Level & Salary Distribution" indicates that as the job hierarchy level increases from the lowest junior level to the highest principal level, the average salary gradually increases. This supports the expectation that salary should increase with promotion. Additionally, the chart shows that most job offers are for senior-level positions, with a minority of positions in the higher lead and principal levels. This suggests that data science professionals with only a few years of experience or fresh graduates may have more opportunities to find a new job compared to highly skilled professionals who are already at the lead or principal levels. As such, it is not recommended to frequently switch jobs when one has already attained these levels, as this could result in longer periods of unemployment.

The bar chart titled "Specific Data Science Job Position and Level Analysis," shows that most data science positions are on the senior job levels, consistent with the previous chart. Notably, the position of a software engineer has the most opportunities offered at the junior level, indicating a high demand for fresh graduates in this role. However, the data science manager and director of data science positions got missing values and no outputs.

Finally, the histogram titled "Specific Data Science Job Position's Salary Distribution" illustrates that most data science positions have normally distributed salaries. Exceptions to this are positions such as software engineer, data engineer, and director of data science, which got skewed salaries where lower salaries are more prevalent than high salaries. This suggests that salaries for normally distributed positions are more balanced and fairer.

4) HR Analytic Dashboard – Fig. 8: The “Data Science Job Level & Salary Boxplot Analysis” boxplot provides a more detailed statistical interpretation of the previous job level and salary

analysis. It shows that the interquartile range and median both increases as the job level hierarchy increases from low to high level. Although the median value for both the “Lead” and “Principal” levels is the same, the maximum salary value is still significantly higher in the “Principal” level. The boxplot for the “Staff” level and “Jr” level have similar median, maximum, and minimum values, confirming that both levels are the same. However, there is an unusual observation where the maximum value of “Lead” is smaller than the lower “Sr” level, which may be due to the insignificant record counts of “Lead” with high salary. The plot also reveals a weak outlier problem with only a few outliers occurring out of the 1287 records, which does not require outlier treatment.

The “Data Science Position & Salary Boxplot Analysis” boxplot shows that higher positions like “Head of Data Science” and “Director of Data Science” have a significantly higher salary range. The position “Vice President of Data Science” should also fall into this category, but the insignificant records resulted in the opposite trend. Surprisingly, the position “Data Architect” has a similar salary range despite not being a top management position, indicating that data architect is a high-paid job in the data science industry. On the other hand, “Statistician” and “Data Analyst” generally have a significantly lower salary range compared to other positions. The minimum value, first quartile, and median for “Statistician” are closely grouped, indicating that they can quickly start with a significant salary but have shorter salary growth. Similar observations occurred in other research-based positions like “Applied Scientist” and “Machine Learning Scientist,” while other positions have longer salary ranges and intermediate median values.

Finally, the “Salary Growth for Specific Data Science Position” line chart shows that “Data Scientist,” “Machine Learning Scientist,” and “Machine Learning Engineer” positions have a general upward trend, confirming that the average salary increases when the job level increases in such job positions. However, the “Applied Scientist” salary trend stays stationary at the “Staff” and “Sr” levels and decreases when moving to the highest “Principal” level, indicating that they start with a significant salary but have shorter salary growth. This observation may be due to the insignificant number of records for “Applied Scientist” with a principal job level. “Software Engineer” and “Data Analyst” reach their peak salary individually at “Sr” and “Lead” levels, respectively, indicating that it’s not necessary to be in top management to achieve high salaries for these particular jobs. Some positions have no visualization results to interpret because of missing values.

5) Skill Variables Visualization – Fig. 9: Academic qualifications are found to be necessary for various job positions, especially in research-focused roles such as "Data Scientist," "Machine Learning Scientist," and "Applied Scientist." These positions typically require a master's or doctoral degree and involve research and development, which necessitates strong academic skills and theories. In contrast, technical roles like "Machine Learning Engineer," "Software Engineer," "Data Engineer," and "Data Analyst" prioritize programming skills like C, R, and SQL. Other technical and software skills are required for specific domains, such as analytical skills and Tableau for "Data Analysts." Higher management positions like "Data Science Manager" and "Director of Data Science" require soft skills like communication and leadership, which are more important for managerial roles as their main focus is on management rather than technical tasks.

The importance of academic qualifications is evident across different job levels, with both "Doctoral degree" and "Doctor of Philosophy" as the top two most required qualifications at the "Principal" level. Programming skills are also in high demand, particularly at the "Senior" and "Lead" levels, with C, R, SQL, Python, and Spark being the core programming languages for data science positions. Technical skills like "Machine learning" and "Natural language processing" are also crucial for data science students and professionals to study and research for better job security.

Soft skills like communication are essential, appearing in the highest "Principal" level, and remain critical throughout different salary levels. "Deep Learning" is a new technical skill that has emerged with a high annual salary range between 150-170k, further emphasizing the interaction between data science and Artificial Intelligence (AI). Data science professionals need to acquire AI skill sets and knowledge for model building and predictions besides mathematics, statistics, visualization, and data analysis.

5.2 Salary Prediction Model

Human Resource (HR) analytics is a type of data analytics that focuses specifically on human resource data, involving data collection, analysis, and reporting. Its primary purpose is to measure the impact of human resource metrics

1) HP Regression Model Interpretation for Higher Salary Level: The model got a 64% likelihood of classifying a job in the highest salary range of over 170k and has a noteworthy probability of 24% of classified jobs in the second-highest salary range of 150-170k. This is particularly applicable for job positions such as Director of Data Science or Data Science Manager, where a Doctor of Philosophy degree is held. This suggests that top-tier academic qualifications and management positions play a crucial role in securing the highest-paying job positions in the field of data science. Further, the finding confirms that top salary levels are exclusively reserved for job positions with high hierarchy levels and that possessing C++ programming skills is crucial in attaining higher salaries.

2) HP Regression Model Interpretation for Lower Salary Level: In contrast, even though data analyst and data engineer positions may have high job levels, such as senior, lead, or principal, they tend to have particularly low salaries. Therefore, these roles are not recommended as they are lower paid compared to the positions mentioned in the previous section. In contrast to the higher salary level, the same job positions such as director of data science and data science manager were found to have a high probability of being classified as the lowest salary level of between 0 to 130k. However, the job level requirement was at the lowest with either junior or staff, and when the job offer was in the Los Angeles or Contra Costa regions. This led to the derivation of a new statement where the salary level for high managerial positions tends to decrease when moving to Southern California, as identified in the location heatmap in the HR analytic dashboard. The reason could be attributed to the lower demand for data science roles in such locations, making it more difficult for job applicants to negotiate for a higher salary. Therefore, data science job applicants and graduates should focus on applying for jobs in more competitive and highly demanded areas such as San Francisco and Santa Clara in Northern California.

3) HP Tree Model Interpretation for Higher Salary Level: The HP Tree model demonstrates that even though job positions like data science manager or director of data science have lower hierarchy levels such as junior or staff, there is still a significant chance of achieving the highest salary level if the applicant holds a Doctor of Philosophy degree. This finding reinforces the idea that having a doctoral degree or advanced academic qualifications can greatly influence the attainment of the highest salary level. Consequently, individuals pursuing a career in data science should continue to invest in their academic achievements, even after completing their undergraduate or master's degree programs.

Further, it demonstrates the immense impact of academic qualifications. If a candidate possesses computer networking skills and a Doctor of Philosophy, there is a probability of more than 95% receiving the second highest salary level of between 150 to 170k. Although this node has a simple set of rules, it is still significant with a considerable number of observations, considering that there are up to 58 nodes in the HP Tree model.

4) HP Tree Model Interpretation for Lower Salary Level: It indicates that if the job is in Southern California regions such as Los Angeles, Orange, San Diego, and San Mateo as shown in the location heatmap of Fig. 6, where there is a high possibility of receiving the lowest salary level. In addition, if the job level is low, such as junior or staff, and the applicant does not possess a Doctor of Philosophy, it also increases the likelihood of a lower salary. This emphasizes the lower salary pay trend in Southern California and suggests that applicants without doctoral degrees tend to achieve lower salaries.

5) Model Performance Comparisons: As mentioned previously, an additional ensemble model was built that combines the results of both the Optimized HP Tree and HP Regression models. By examining the performance measures of the three models in Fig. 10, it is evident that the Optimized HP Tree model achieved the highest testing accuracy of approximately 57%, followed by the Ensemble model at 56% and the Optimized HP Regression model at 52%. It is important to note that only testing measurements were considered as the objective was to make predictions on new data science employment data in the future. However, the Ensemble model achieved a significantly lower ASE value of approximately 0.138 compared to the other models, which had ASE values of over 0.144. This suggests that although the Ensemble model did not make the least number of misclassifications, its incorrect classifications were predicted more closely to the actual class. Therefore, the suitability of the Ensemble model can be further demonstrated with additional statistics in the following outcomes.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPTree	HP Tree	0.42740	0.12406	0.36461	0.14424
	Ensmbl	Ensemble	0.44110	0.11532	0.35986	0.13760
	HPReg	HP Regression	0.47397	0.11974	0.38361	0.14827

Figure 10. Fit Statistic between the Three Proposed Models

The combined confusion matrix shown in Fig. 11 reveals that the Optimized HP Regression model has a much higher rate of false negative classifications, while the Optimized HP Tree model has a higher rate of false positive classifications. Only the Ensemble model has a balanced rate of false negative and false positive misclassifications. False negative indicates a type 1 error, where the actual salary level is "Yes" but is wrongly classified as "No". In contrast, false positive indicates a type 2 error, where the actual salary level is "No" but is wrongly classified as "Yes". While false negatives may be preferred in Covid-19 detection, there is no preference for multiclass salary level classification. Therefore, it is advisable to select the more reliable Ensemble model with a balanced rate of type 1 and type 2 errors.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
HPReg	HP Regression	TRAIN	SalaryLevel		67	695	54	82
HPReg	HP Regression	VALIDATE	SalaryLevel		39	308	17	25
HPTree	HP Tree	TRAIN	SalaryLevel		51	668	81	98
HPTree	HP Tree	VALIDATE	SalaryLevel		19	291	34	45
Ensmbl	Ensemble	TRAIN	SalaryLevel		46	676	73	103
Ensmbl	Ensemble	VALIDATE	SalaryLevel		21	296	29	43

Figure 11. Event Classification Table between the Three Proposed Models

Judging the misclassifications of different salary levels in the proposed models helps to identify the model that is least biased in predicting all classes. In Fig. 12, it can be observed that the Optimized HP Regression model (on the left) is significantly weaker in correctly classifying the lowest and highest salary levels. On the other hand, the Optimized HP Tree model (in the middle) excels in correctly classifying the lowest and highest salary levels but has higher errors when classifying the two middle salary levels. Ultimately, the Ensemble model combines the characteristics of both models, mainly from the Optimized HP Tree model. This results in a more balanced incorrect classification across different salary levels while maintaining the high accuracy of the Optimized HP Tree model.

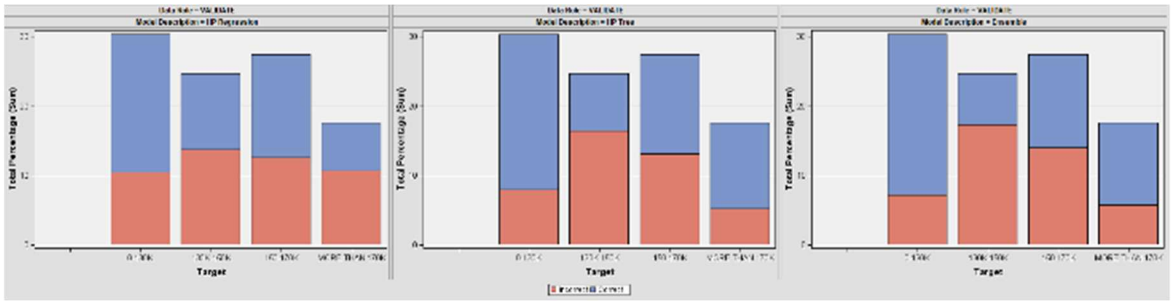

Figure 12. Testing Classification Charts between the Three Proposed Models

To select a final proposed model, a cumulative lift graph was used for all three models as shown in Fig. 13. The green line represents the centralized baseline cumulative lift, while the red line indicates the best cumulative lift across all the models. The blue line represents the actual cumulative lift for the particular model, with a closer alignment to the best cumulative lift (red) indicating better fitting. Upon comparison, the Ensemble model exhibits the best fitting, as its cumulative lift (blue) is smoother and closely follows the best cumulative lift (red) earlier, starting from a depth of 40.

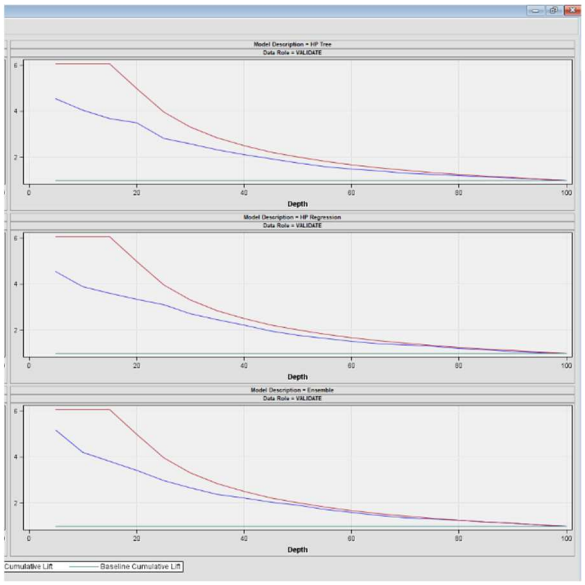

Figure 13. Testing Score Ranking Matrix between the Three Proposed Models

5.3 Analysis Summary Report

This analysis summary consolidates the outcomes of both the human resource visual analytics and salary prediction model interpretations. The data scientist position is shown to have the highest demand followed by machine learning scientist, machine learning engineer, software engineer, data engineer, and data analyst. Geographically, San Francisco offers strong data science job opportunities, while Mountain View is better for machine learning engineer and data analyst positions, and Sunnyvale is the dominant location for data engineer positions. The location heatmap highlights Northern California as the region with the most job opportunities, with social media, transportation, information technology, entertainment, and e-commerce companies being the top hiring companies, including TuSimple, TikTok, and Intuitive. TuSimple, in particular, offers strong opportunities for machine learning scientist and software engineer positions.

Furthermore, job hierarchy level affects salary, with higher job levels indicating higher salaries. Boxplot analyses further confirm this assumption as the interquartile range and median values increase with the job level hierarchy. The positions with higher ranking, such as head or director, have higher salary ranges, and the data architect position also has a similarly high salary range.

However, statistician and data analyst positions result in a lower salary range. Research-based positions, such as applied scientist and machine learning scientist, can receive a high salary at the beginning levels, but there is a smaller room for salary increments when moving to higher job levels.

Additionally, there are more job openings available for entry-level and senior-level professionals compared to those who are highly skilled. In particular, a majority of software engineer positions are offered at the junior level compared to other roles, making it an excellent starting position for fresh graduates pursuing a career in data science due to the high demand. Moreover, the detailed salary distribution analysis revealed that positions such as software engineer, data engineer, and director of data science have the majority of their salaries distributed on the lower side, indicating that achieving a high salary in these positions can be challenging but very rewarding if accomplished.

Academic qualifications are crucial for data science positions, levels, and salary levels, with all academic qualification variables remaining significant in the first feature selection. Research-based positions require particular academic qualifications, while technical positions require programming and software skills, and higher positions require soft skills. Programming skills are more important at the lead and senior levels, whereas software or tool skills are more critical at the lowest junior level. The highest principle level strongly requires top academic qualifications such as a doctoral degree, while the academic qualifications remain significant for salary levels. AI-related technical skills such as deep learning, machine learning, and natural language pre-processing are essential to career development domains for data science professionals.

The salary prediction model interpretations confirm the above assumptions, with high job levels, positions with a naturally high ranking, and a doctoral degree providing excessive possibilities for achieving the top salary level. Additionally, C++ programming and computer networking skills are significant in high-salary groups. Lower salary levels are generally associated with data analyst and data engineer positions, and the lowest salary level is often classified when a doctoral degree is not required, the job level is low and located in Southern California countries such as Los Angeles, Contra Costa, Orange, San Diego, and San Mateo. This suggests that locations with low data science job demand can also result in lower salaries, emphasizing the importance of applying for jobs in Northern California with more opportunities and competition.

6. Conclusion

The preliminary review of the literature provided essential technical knowledge for the implementation of both HR analytics and salary prediction modelling. It identified the most in-demand data science positions, such as data scientist, data engineer, and data analyst, and highlighted the significance of machine learning and deep learning models for accurate salary predictions. Moreover, the review pointed out that HR analytics are underdeveloped in the specialized data science employment field. To increase the accuracy of the reviewed works, this research followed an approach that converted salary prediction into a classification problem. The CRISP-DM methodology was selected to conduct the necessary data analysis and mining processes flexibly.

The research design flowchart provides a comprehensive overview of the entire research process, starting from data understanding in HR analytics, data preparation, modelling in salary prediction, evaluation, and final deployment. The cross-interpretation of HR visual analytics and salary prediction models showed that a doctoral degree is crucial for achieving higher salary levels and job hierarchy. Geographical analysis showed better job opportunities in Northern California, with a gradual reduction in salary levels and opportunities when moving down to Southern California. Moreover, higher job hierarchy levels significantly influence higher salary levels, with data architects having a much more significant salary. The proposed Ensembled model combining both Optimized HP Regression and HP Tree was determined as the final model with the best fitting, lowest ASE value, and classification accuracy of 56%. While this accuracy is lower than that of existing models reviewed in the literature, the interpretable rules are more significant in supporting the HR analytics component, with salary prediction serving as an assisting tool for the larger picture of HR analysis in the specialized data science employment field.

References

[1] Muratovski, G., 2020. Industry 4.0 Is Already Here, But Are You Ready?. [Online] Available: https://www.forbes.com/sites/forbesagencycouncil/2020/09/08/industry-40-is-already-here-but-are-you-ready/?sh=7187295144b5

[2] One, T., 2020. Malaysia is ready for Industry 4.0. [Online] Available: https://www.tmone.com.my/resources/think-tank/article/malaysia-is-ready-for-industry-4-0/.

[3] V.Sindhu, Anitha, G. & Geetha, R., 2021. Industry 4.0-A Breakthrough in artificial Intelligence the Internet of Things and Big Data towards the next digital revolution for high business outcome and delivery. Journal of Physics: Conference Series, Volume 1937, pp. 1-7.

[4] Jain, K., 2019. Big job opportunities in data science & machine learning. Express Computer, pp. 1-3.

[5] DuBois, J., 2020. The Data Scientist Shortage in 2020. [Online] Available: https://quanthub.com/data-scientist-shortage-2020/.

[6] More, A., Naik, A. & Rathod, S., 2021. PREDICT-NATION Skills Based Salary Prediction for Freshers. SSRN Electronic Journal.

[7] Olavsrud, T., 2020. What is a data analyst? A key role for data-driven business decisions. [Online] Available: https://www.cio.com/article/217583/what-is-a-data-analyst-a-key-role-for-data-driven-business-decisions.html#:~:text=Data%20analysts%20work%20with%20data,predict%2C%20and%20improve%20business%20performance.

[8] Metwalli, S. A., 2020. 10 Different Data Science Job Titles and What They Mean. [Online] Available: https://towardsdatascience.com/10-different-data-science-job-titles-and-what-they-mean-d385fc3c58ae.

[9] Shin, T., 2021. The Most In-Demand Skills for Data Scientists in 2021. [Online] Available: https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-in-2021-4b2a808f4005.

[10] Vulpen, E. v., 2021. aihr. [Online] Available: https://www.aihr.com/blog/what-is-hr-analytics/.

[11] Valamis, 2021. Human Resource (HR) Analytics. [Online] Available: https://www.valamis.com/hub/hr-analytics

[12] Nagpal, T. & Mishra, M., 2021. Analyzing Human Resource Practices For Decision Making in Banking Sector using HR analytics. Materials Today: Proceedings.

[13] Ameer, M., Rahul, S. P. & Manne, D., 2020. Human Resource Analytics using Power Bi Visualization Tool. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020), pp. 1184-1189.

[14] Das, S., Barik, R. & Mukherjee, A., 2020. Salary Prediction using Regression Techniques. International Conference on Industry Interative Innovations in Science and Engineering, pp. 1-5.

[15] Lothe, P. D. M. et al., 2021. Salary Prediction using Machine Learning. International Journal of Advance Sciencetic Research and Engineering Trends, 6(5), pp. 199-202.

[16] Pawha, A. & Kamthania, D., 2019. Quantitative analysis of historical data for prediction of job salary in India - A case study. Journal of Statistics & Management Systems, 22(2), pp. 187-198.

[17] Dutta, S., Halder, A. & Dasgupta, K., 2018. Design of a novel Prediction Engine for predicting suitable salary for a job. 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 275-279.

[18] Zhang, J. & Cheng, J., 2019. Study of Employment Salary Forecast using KNN Algorithm. International Conference on Modeling, Simulation and Big Data Analysis, pp. 166-170.

[19] Kavlakoglu, E., 2020. AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?. [Online]
Available: https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks#:~:text=Deep%20learning%20is%20a%20subfield,must%20have%20more%20than%20three.

[20] Sun, Y. et al., 2021. Market-oriented job skill valuation with cooperative composition neural network. Nature Communications, Volume 12, pp. 1-11.

[21] Zhu, H., 2021. Research on Human Resource Recommendation Algorithm Based on Machine Learning. Hindawi Scientific Programming, pp. 1-10.

[22] Wang, Z., Sugaya, S. & Nguyen, D. P., 2019. Salary Prediction using Bidirectional-GRU-CNN Model. 2019 The Association for Natural Language Processing, pp. 292-295.

[23] Chen, L., Sun, Y. & Thakuriah, P., 2020. Modelling and Predicting Individual Salaries in United Kingdom with Graph Convolutional Network. Hybrid Intelligent Systems 2018. Advances in Intelligent Systems and Computing, pp. 61-74.

[24] HOTZ, N., 2022. What is CRISP DM?. [Online] Available: https://www.datascience-pm.com/crisp-dm-2/

[25] Lamott, K., 2022. San Francisco. [Online] Available: https://www.britannica.com/place/San-Francisco-California

[26] tusimple, 2022. about us. [Online] Available: https://www.tusimple.com/