|  |  |
| --- | --- |
| Internship Project Title | TCS iON RIO-125: HR Salary Dashboard - Train the Dataset and Predict Salary |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | B. K. Birla College of Arts, Science & Commerce (Autonomous), Kalyan |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 25-09-2023 | 30-09-2023 | | 24.0 hrs | | Google Colab | Python3 |
| Milestone #1 |  | Milestone: | | Create dataset, clean dataset, and sanitize dataset | | |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Charts, Table, Diagrams
* Outcome / Conclusion
* Link to code and executable file

**ACKNOWLEDGEMENTS**

I wish to express my heartfelt appreciation to my industry mentor, Debashis Roy, for his invaluable guidance and unwavering support throughout this project. His assistance and the opportunity he provided me to work on this remarkable platform have been indispensable in my successful completion of this endeavor. I am genuinely thankful for his willingness to address my queries at every stage of the project.

**OBJECTIVE**

The aim of this project is to develop a salary prediction dashboard specifically designed for human resource management. This dashboard will leverage the power of machine learning algorithms to forecast job candidates' salaries, taking into account various factors such as their experience, age, and qualifications. By providing this valuable information, the dashboard aims to assist HR managers in making more informed decisions during the candidate selection process and improve the overall effectiveness of hiring for job positions.

**INTRODUCTION**

The human resources department faces the complex task of evaluating job applicants and determining the best candidates for various positions. Salary considerations significantly influence a candidate's decision to accept a job offer, underscoring the importance of offering competitive compensation packages.

In this project, we have a substantial dataset with information on over 32,000 job candidates, including their experience levels and associated salaries. This dataset covers a wide range of job profiles and salary data, making it an invaluable resource. Our primary goal is to utilize this data to create a salary prediction dashboard, empowering HR managers to make well-informed decisions about salary offers for job candidates.

In the initial project phase, I successfully collected, cleaned, and refined the dataset, preparing it for the subsequent stage where we will build a robust salary prediction model. This project aims to streamline the hiring process, attract top talent, and contribute to the success of both job candidates and the organizations that employ them.

**INTERNSHIP ACTIVITIES**

* + Watched the welcome kit videos.
  + Done preparations for RIO – pre-assessment.
  + Attended the RIO – pre-assessment test.
  + Went through the day-wise plan.
  + Read the project reference material.
  + Read the industry project material.
  + Watched webinar 1.
  + Watched webinar 2.
  + Gone through all posts in the digital discussion room.
  + Watched lectures and other videos to gain a better understanding of the topic.
  + Created a GitHub account to store and share my project files.
  + Found a suitable data set for the project.
  + Wrote activity reports to document my progress.
  + Verified that the data set had enough data for the project.
  + Read articles and learned how to clean and sanitize the data.
  + Applied data cleaning and sanitization techniques to the data set.
  + Conducted exploratory data analysis to identify patterns and trends in the data.

**APPROACH / METHODOLOGY**

To reach the initial milestone of my internship project, I followed a well-defined methodology that can be summarized as follows:

Familiarization: I began my journey by immersing myself in the project's concepts and requirements. This involved extensive reading of relevant articles and watching informative videos. This initial step allowed me to grasp the fundamental aspects of the project, ensuring a comprehensive understanding of its scope.

Programming Environment: I chose Google Colab as my programming environment for its numerous advantages. This decision was primarily driven by its efficiency in writing and executing code swiftly. Leveraging the platform's features streamlined my workflow and expedited development tasks, allowing me to work more effectively.

Version Control: Recognizing the importance of collaboration and code management, I established a GitHub account. This step facilitated seamless code sharing and provided a platform for version control. It also enabled efficient collaboration with peers and potential reviewers, ensuring that the project's codebase remained organized and accessible.

By implementing this methodical approach, I not only acquired the necessary knowledge and skills required for project engagement but also made substantial progress toward achieving the initial milestone. This approach emphasizes thorough understanding, efficient coding, and collaborative development, setting a strong foundation for the project's success.

**OUTCOME**

Upon the successful completion of the first milestone in this internship project, I have gained valuable insights into the process of cleaning and preparing datasets for model construction. As part of this endeavor, I employed several techniques to enhance the cleanliness and accuracy of the dataset.

To begin with, I eliminated redundant columns such as "capital gain" and "capital loss" during the model building process. Additionally, I identified the "education-num" column as unnecessary and subsequently removed it from the dataset. These actions ensured that the dataset was streamlined, omitting irrelevant information and promoting data accuracy.

Furthermore, I conducted exploratory data analysis (EDA) to identify and subsequently remove records containing superfluous values. By doing so, I was able to improve the overall quality and precision of the data, eliminating any potential outliers or inconsistencies that may have hindered the model's performance.

Overall, these outcomes have equipped me with the necessary skills and knowledge to effectively clean and prepare datasets, setting a strong foundation for subsequent stages of the project.

**LINK TO CODE AND EXECUTABLE FILE**

https://colab.research.google.com/drive/1tsZggDi0iQt4nrLwS394103gBJcZbe4g#scrollTo=6v3wY5b1sncR