Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor’s feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

|  |  |
| --- | --- |
| 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 | ICT ACADEMY OF KERALA |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 22-04-2023 | 22-04-2023 | | 20 | | Remote Internship | Jupyter Notebook |
| Milestone # | 1 | Milestone: | | Cleaned and sanitized the data and performed EDA | | |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**ACKNOWLEDGEMENT**

I would like to express my deepest gratitude to **Mr. Debashis Roy**, my industry mentor, TCS **iON** and **ICT Academy of Kerala** for providing me with the necessary facilities for the completion of this project. I am thankful for the valuable discussion I had at each phase of the project and for being a very supportive and encouraging project mentor. I would like to express my sincere thanks to all my friends who were actively part of the discussion room in this project and gave valuable suggestions.

To achieve my chosen career goals, I will work to apply my newly acquired skills and information as effectively as possible and to further develop them.

**OBJECTIVE**

The objective of HR salary prediction is to use data analytics and machine learning techniques to accurately predict the salary of employees. This project is to develop a dashboard for predicting the salary of employees using various machine learning algorithms. The goal of this process is to provide HR departments with the information they need to make informed decisions about employee compensation and benefits. This is a classification problem.

**INTRODUCTION**

Predictive analytics is an upcoming trend in Human Resources (HR). Human resource (HR) salary prediction is the process of using data analytics and machine learning techniques to estimate the salary of employees within an organization. The goal of this process is to provide HR departments with accurate, data-driven insights into employee salaries, enabling them to make informed decisions about compensation and benefits.

The process of HR salary prediction involves collecting data on employee salaries and relevant factors, cleaning and pre-processing the data, and then building and training machine learning models to make predictions. In this project the main aim to train the dataset and predict the salary of new employee. For developing this, we are using Classification of supervised learning in machine learning.

Data with 32561 rows and 14 features is gathered to do a predictive study of HR compensation. A training set is taken from the dataset, and a testing set is taken from it, with the training set being used to train the model and the testing set being used to assess its accuracy. Once trained, the model can be used to forecast new hire salaries of employees.

**INTERNSHIP ACTIVITIES**

TCS iON Remote Internships has two major learning components spread across 125 hours and 30 days of engagement for successfully completion of program. TCS iON Remote Internship comprises of various activities including self learning sessions, recorded webinars, activity report, Day wise plan, Pre-test, Digital discussion room. What I have done so far is:

* Finalized the dataset salarydata.csv.
* Started project in Jupyter notebook.
* Imported the libraries, loaded the dataset into the Python environment.
* Studied about the data.

Checked shape of data(32561 rows x 14 columns), info of the data, statistical

summary, value\_counts, unique and nunique.

* Cleaned the data.
* Checked null values, duplicates.
* Find the unique values and found some ? in workclass, occupation and native-country and replaced ‘?’ with nan.
* Filled the null values in workclass, occupation and native-country using mode.
* Found the correlation matrix and dropped irrelevant columns like education\_num, capital gain, capital loss and native country.
* Performed univariate, bivariate, and multivariate analysis.
* Derived meaningful insights from the data.
* Found the outliers present in continuous numerical columns and handled it using clip method.
* Cleaned and sanitized the data successfully.

1. Exploratory Data Analysis

|  |  |  |
| --- | --- | --- |
| Univariate Analysis | | |
| Feature | Plot | Insights |
| Age | Hist plot | Most peoples are in the range of 20- 50 years of age. |
| workclass | Count plot | We can see that 75% people are privately employed. |
| education | Count plot | Looking at the education, majority of people have completed high graduation, have attended college, or have at least completed masters. |
| Marital-status | Count plot | Considering the marital status, 45.9% are married-civ-spouse and 32.8% are never married. |
| occupation | Count plot | Most people are working as professional-specialty |
| relationship | Count plot | Husbands having 40% than other relationships |
| race | Count plot | considering race, 85.43% of people are white. |
| sex | Count plot | 66.92% of people are male and 33.08% are female. |
| native-country | Count plot | Most people belong to US. |
| salary | Count plot | 25% people have salary above 50k and 75% below 50k, this indicated that we have imbalanced data. |

|  |  |  |
| --- | --- | --- |
| Bivariate Analysis | | |
| Feature | plot | Insights |
| salary Vs age | Bar plot | Age above 40 having salary more than 50k. |
| salary Vs hours-per-week | Bar Plot | People with salary >=50 works for more than 40 hrs per week. |
| salary Vs workclass | Count plot | In workclass, People who are privately employed, also self employed individuals having salary more than 50k |
| salary Vs education | Count plot | In education, masters and bachelors, also High school and college grads are having salary more than 50k. |
| salary Vs marital-status | Count plot | In marital status, people who are married are having salary more than 50k |
| salary Vs occupation | Count plot | Exec-managerial having salary more than 50k. |
| salary Vs relationship | Count plot | In relationship husbands having salary more then 50k. |
| salary Vs race | Count plot | In race White people having salary more than 50k. |
| salary Vs sex | Count plot | In sex, males have salary more than 50k. |
| salary Vs native-country | Count plot | In native-country people living in the United states having salary more than 50k |

**LINK TO CODE AND EXECUTABLE FILE**

https://github.com/SWATHYE/TCS-iON/blob/main/SALARY%20PREDICTION%20CLEANING%20DATA.ipynb