INTERNSHIP: INTERIM PROJECT REPORT

Internship Project Title	RIO-125: HR Salary Dashboard – Train the Dataset and Predict Salary
Name of the Company	TCS iON
Name of the Industry Mentor	Rushikesh Meharwade
Name of the Institute	Mepco Schlenk Engineering College,Sivakasi

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
7/09/2021	6/12/2021	25	Jupyter Notebook	Python 3
Milestone #	1	Milestone:	Create dataset, clean dataset, and sanitize dataset	

TABLE OF CONTENT

- Acknowledgements
- Objective
- Introduction
- Internship Activities
- Approach / Methodology
- Outcome
- Link to code and executable file

ACKNOWLEDGEMENTS

I am conveying my sincere gratitude towards my industry mentor, Rushikesh Meharwade, and academic mentor, Dr. Muthu Kumar for helping me throughout this project till now and providing me this wonderful platform to complete this project. I am also thankful for answering my queries at every phase of the project. I also want to thank all my friends who helped me with valuable suggestions during this project.

OBJECTIVE

The objective of this model is to make a salary prediction dashboard for human resource management.

INTRODUCTION

From the first 5 days of my project, I have collected the dataset. I also cleaned and sanitized the dataset. Now the dataset is ready for training which shall be used for salary prediction model building.

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.
- I went through the linear regression YouTube videos.
- Read the linear regression article.
- Watched the lectures provided and other videos for further understanding.
- Created a GitHub account.
- Searched and found out a proper data set for this project.
- Wrote activity reports.
- Checked and clarified the data set whether it has enough data for the project.
- Read articles and find out how to clean and sanitize the data.
- Cleaned the data set.
- Sanitized the data set.
- Done Exploratory Data Analysis(EDA)
- Watched videos on model training
- Used Logistic Regression and trained it
- Used Random Forest Classifier and trained it

APPROACH / METHODOLOGY

The approach I took for the internship project for completing the 1st milestone is firstly understanding the concepts of the requirements. Reading articles and watching videos helped in achieving knowledge about the requirements. Jupyter Notebook has been used for doing the programming. Google colab has also been used for much faster execution. A GitHub account has been created for publishing the codes.

OUTCOME

After the 1st milestone of this internship project, I have learned about regression models and understood how to clean and sanitize the dataset. I have removed 2 unnecessary columns,i.e capital gain, and capital loss. These two columns were to be used after the salary is provided. So using these variables is not have relevance to the model building. Also, there was another column education-num, in which this column is the numeric version of the education column. Since all the categorical variables are converted to numeric at the time of model building, thus I removed them.

I removed some of the records that had unnecessary values. These unnecessary values have been identified using the exploratory data analysis method(EDA).

I have checked all the columns using some graphs and bar plots. This gave an idea about what values were there in the corresponding columns. Then converted the categorical columns into numeric which was needed for model training. I trained two models; Logistic Regression and Random Forest Classifier. Both of their classification reports have been included in the Jupyter notebook. Among these to models, Random forest classifier showed the best score or accuracy in predicting the salary.