**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 | Debashis Roy |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 13/02/2022 | 14/05/2022 | | 75 | | Jupyter Notebook | Python 3 |
| Milestone # | 2 | Milestone: | | Train the dataset and predict the salary of particular HR based on the dataset | | |

**TABLE OF CONTENT**

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

**ACKNOWLEDGEMENTS**

I am conveying my sincere gratitude towards my industry mentor, Debashis Roy and academic mentor, Aswathy P, Project coordinator, ICT Academy of Kerala 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. In order to select a right candidate for a job position, along with assessing the credentials and qualifications of the candidates, HR should consider the previous salary offered to similar candidates based on their experience, age, qualifications etc. Since the salary is affected by many factors, using algorithms to predict salary is a good idea and machine learning models are very useful.

**INTRODUCTION**

The human resource department is facing a tough job nowadays to process a huge number of applicants and at the same time, they need to select the right candidates for the right position. Candidates usually considering salary as one of the defining factors while accepting a job offer. Hence offering the right and comparable salary in the market to the candidates are very important.There are different datasets available over the internet. However, by considering the prerequisites, this work analyses a well-respected dataset having 32562 salary data with 14 columns. It represents a wide range of job profiles, experience and salary hence it is a good dataset for our analysis.The target salary in the dataset I have selected contains only two classes(<=50K and >50K ). Hence, the model needs to predict the salary to be in one of these two classes. So, our model converges to a binary classification model. There are several methods to make a binary classification, which include SVM, logistic regression, random forest, etc.

The objective of this model is to make a salary prediction dashboard for human resource management. I have collected the dataset, cleaned and sanitized the dataset. Once the dataset is ready for training which shall be used for salary prediction model building, visualized all the features or attributes of salary prediction for HRs. After all the visualization part, trained the dataset using Logistic regression, KNN, Decision Tree, Random forest, Gradient boost trainings. Classification reports for each classifier were generated which showed hypertuned Gradient boosting classifier showed much accuracy.

**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.
  + Generated Classification reports for both classifiers.
  + Did Hyperparameter tuning for gradient boosting and generated the classification report after the tuning process.
  + Did a prediction using the gradient boosting classifier.

**APPROACH / METHODOLOGY**

The approach I took for the internship project for completing the 2nd milestone is 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**

The current stage of the project is that I have cleaned and sanitized the data, i.e data has been preprocessed. A logistic regression model has been trained and tested at the end of the milestone. A random forest classifier has also been implemented to understand the difference between certain models. Classification reports have been generated for both the models. The parameters of the logistic regression model have been tuned for showing better performance. Also for the tuned model, a classification report has been generated.

│ Model │ Accuracy │

╞═════════════════════╪════════════╡

│ Logistic Regression │ 0.773244 │

├─────────────────────┼────────────┤

│ KNN │ 0.817297 │

├─────────────────────┼────────────┤

│ Decision tree │ 0.832135 │

├─────────────────────┼────────────┤

│ Random Forest │ 0.836541 │

├─────────────────────┼────────────┤

│ Gradient Boost │ 0.863668 │

╘═════════════════════╧════════════╛

**LINK TO CODE AND EXECUTABLE FILE**

* Link to the code:

<https://colab.research.google.com/drive/1vIP6kaQIX0EwWLRa18TZ82DCDAeOEHCj?usp=sharing>