**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 | Prepinsta |
| Name of the Institute | Prepinsta |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 01/03/2023 | 30/05/2023 | | 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, 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. Then on the next 10 days of my project, I visualized all the features or attributes of salary prediction for HRs. After all the visualization part, I trained the dataset using logistic regression and random forest. Classification reports for each classifier were generated. At the end using a tuple of user defined data, a small prediction was made using the logistic regression classifier.

**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 the logistic regression and generated the classification report after the tuning process.
  + Did a prediction using the logistic regression classifier.
  + Watched the following videos for understanding about model training and hyperparameter tuning.
    - <https://www.youtube.com/watch?v=xXDoZLVjfbs&list=PLLy_2iUCG87CNafffzNZPVa9rW-QmOmEv&index=31>
    - <https://www.youtube.com/watch?v=Hy9tOyk33D0&list=PLLy_2iUCG87CNafffzNZPVa9rW-QmOmEv&index=32>
    - <https://www.youtube.com/watch?v=D7p5g8PMDvQ&list=PLLy_2iUCG87CNafffzNZPVa9rW-QmOmEv&index=41>
    - <https://www.youtube.com/watch?v=HdlDYng8g9s&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw&index=17>
  + Gone through the following articles
    - <https://www.pluralsight.com/guides/cleaning-data-using-python>
    - <https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/>
    - <https://www.geeksforgeeks.org/hyperparameter-tuning/>
    - <https://scikit-learn.org/stable/modules/grid_search.html>

**APPROACH / METHODOLOGY**

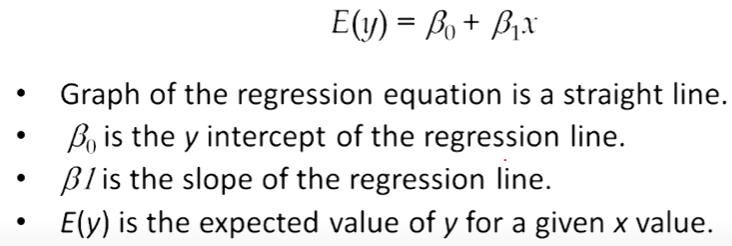
The approach I took for the internship project for completing the 2nd milestone is understanding the concepts of the requirements. Reading articles and watching videos helped in achieving knowledge about the requirements. Jupyter Notebook has been used for doing the programming. Google colab has also been used for much faster execution. A GitHub account has been created for publishing the codes.

**OUTCOME**

From these 15 days of learning and implementing, I have grasped much knowledge about the following:

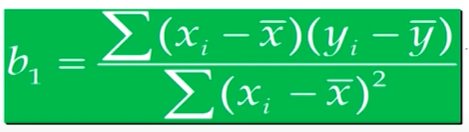
**Linear Regression**

Linear regression is perhaps one of the most well understood and well known algorithms in statistics and machine learning. Linear regression was borrowed from statistics to machine learning, which makes a statistical algorithm and machine learning algorithm. It is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). The equation that describes how y is related to x and an error term is called the regression model. The simple linear regression model is



Linear regression helps in finding the best line of fit through the data by searching for the regression coefficient (B1) that helps in minimizing the total error of the model.

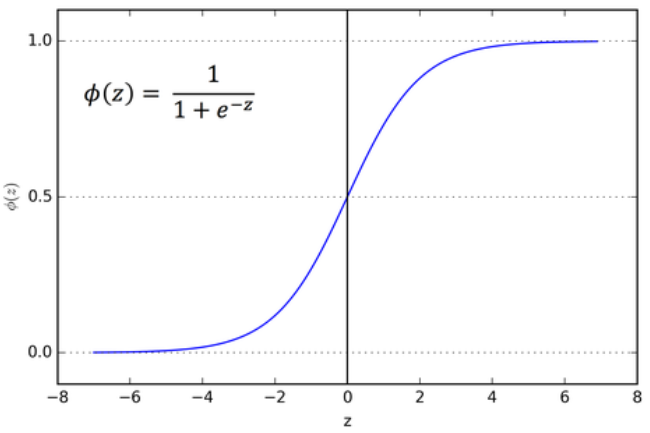
Now we can find or estimate the values for the parameters beta1 and beta2 using the equation for calculating the error. Then we define a model by minimizing the residual error.

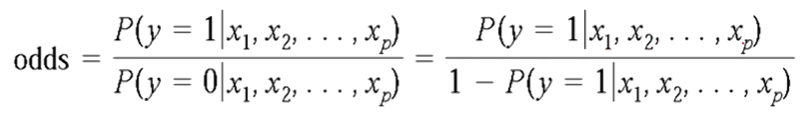
Gradient descent technique is an optimization algorithm that helps in finding the values of parameters of a function to minimize the cost function.

**Logistic regression**

This type of regression is used when the dependent variable is categorical. The funcion used here is a sigmoid function.



If ‘z’ tends to infinity, y tends to 1 and if ‘z’ tends to negative infinity, y tends to 0. So the outputs for a logistic regression will be 0 and 1. By using the function mentioned in the figure, we can define an odds ratio as



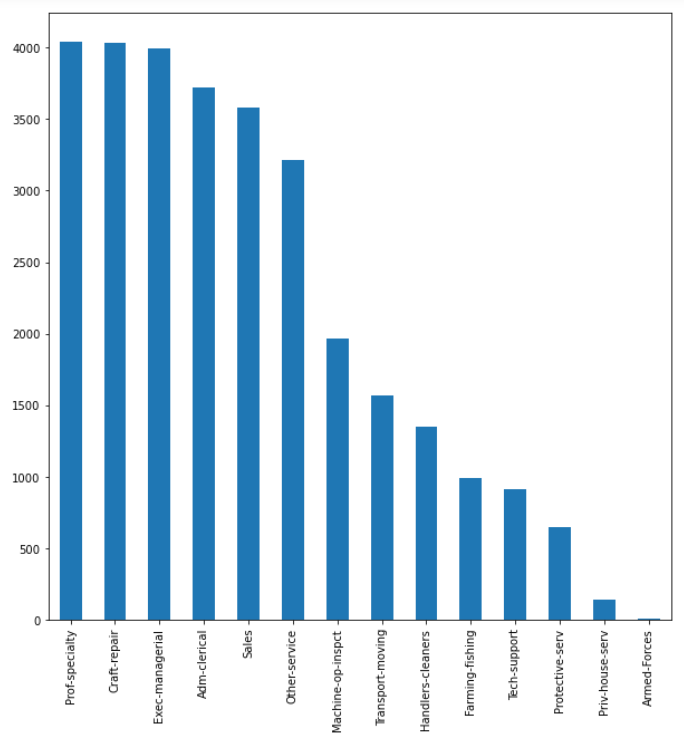
**PROJECT DEVELOPMENT**

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.

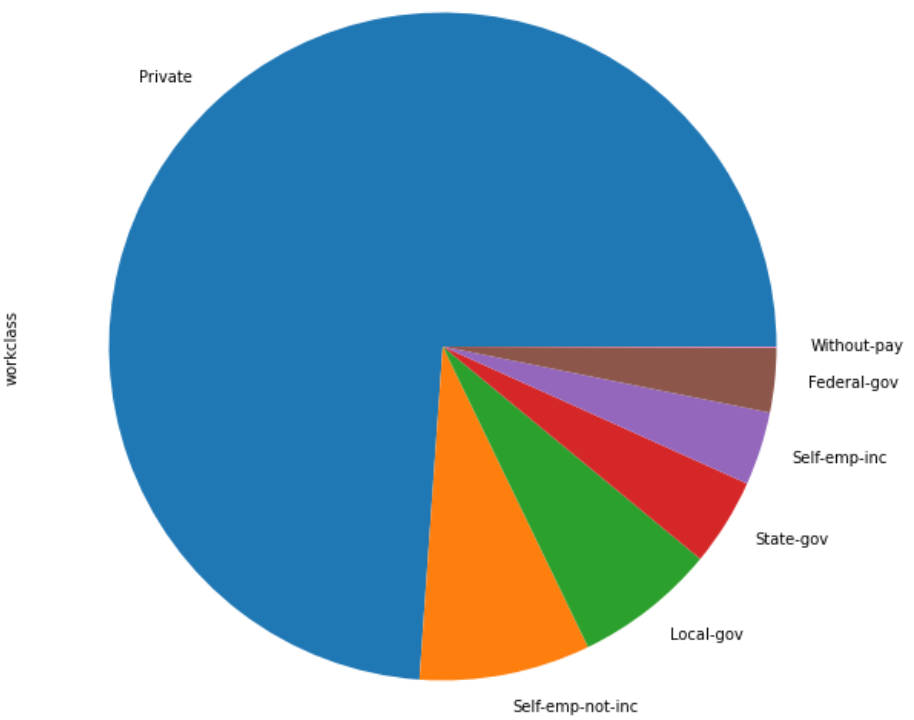
**PROJECT REFERENCE**

Throughout the project till the second milestone, I have done some EDA and visualising. They are as follows:

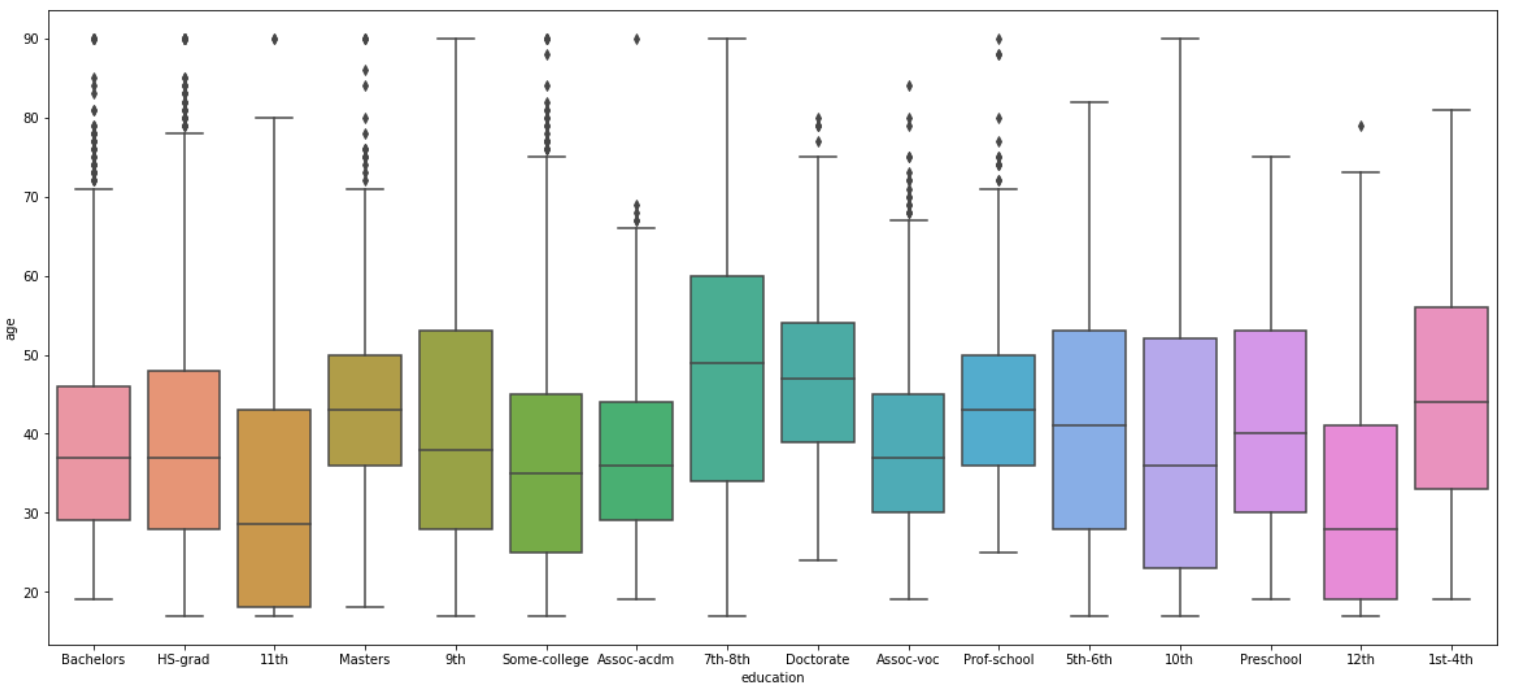
* This is a bar graph used for checking the variable ‘occupation’ in the dataset.



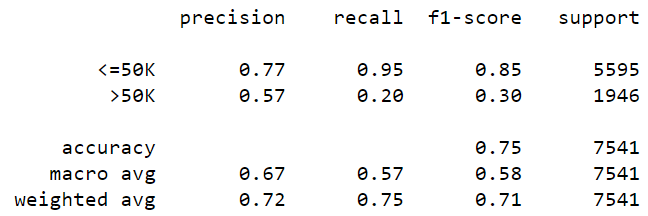
* This is a pie chart used for checking the variable ‘workclass’ in the dataset.



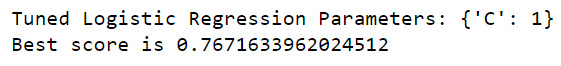
* This is a Box Plot used for checking the variables ‘education’ and ‘age’ in the dataset.



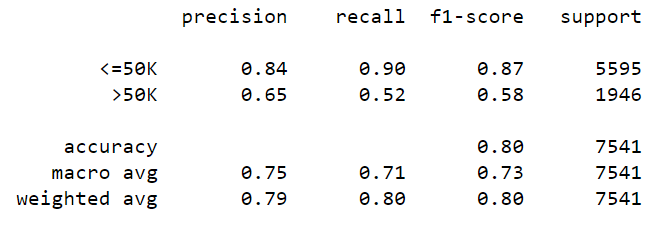
* This is a classification report of the logistic regression model before hyperparameter tuning. The accuracy score is almost 75.5%.



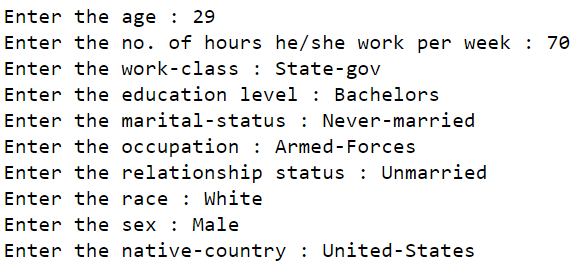
The next figure shows the best accuracy score of the tuned logistic regression model with the used parameter, which is 76.7% So some improvement has been shown when the model is gone through the tuning process.



* This is a classification report of the random forest classifier which shows a good accuracy score of 80% and also has an increase in the precision compared to the non tuned logistic regression model.



* This is the input given for achieving a prediction using the logistic regression model. The predicted output is given as the next figure.





**LINK TO CODE AND EXECUTABLE FILE**

* Link to the code: <https://colab.research.google.com/drive/1PXlic0zPslv423qobynzAV_5UPGFUFYn?usp=sharing>
* Executable file:

<https://github.com/Naveen-Bhargav/RIO-125-HR-Salary-Dashboard-Train-the-Dataset-and-Predict-Salary->