|  |  |
| --- | --- |
| 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 |
| Name of the Institute | Indian Institution of Information Technology and Management of Kerala |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 12/03/2021 | 10/06/2021 | | 70 | | 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 / Description of Internship
* Internship Activities
* Outcome / Conclusion
* Link to code and executable file

**Acknowledgements**

First of all, I am expressing my sincere gratitude towards my industry mentor, Rushikesh and academic mentor, Dr. Manoj Kumar T.K for helping me throughout this project until 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. You kept me going until now. 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.

**Introduction**

During the first 5 days of my project I have collected the dataset. I have also cleaned and sanitized the dataset. Now the dataset is ready for training which shall be used for salary prediction. Then on the next 10 days of my project I have visualized the dependence of every attributes on salary of HRs’. Then I trained the dataset using logistic regression since our target column only has two classes. Therefore a binary classifier using logistic regression is appropriate. Then I have generated the classification report. At the end I have predicted the result using a user defined data tuple.

**Internship Activities**

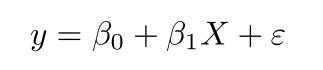
* Watched all welcome kit videos
* Did preparation for RIO – pre assessment test
* Attended the RIO – pre assessment test
* Gone through the day wise plan
* Read the project reference material
* Read the industry project material
* Gone through all posts in the digital discussion room
* Posted queries and doubts about the project and webinar 1 in DDR
* I went through the linear regression you tube video
* Read the linear regression article
* Made myself clear with the math behind the model
* Implemented a linear regression model on my own
* Did some participation in the digital discussion room
* 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 2 articles and find out how to clean and sanitize the data
* Again, checked and clarified the data set whether it has enough data for the project
* Cleaned the data set
* Sanitized the data set
* Did visualization of the dataset using all the attributes.
* Pre-processed the dataset
* Trained and tested the dataset using logistic regression
* Created classification report
* Did the hyper parameter tuning and generated classification report using the new model.
* Tested the data against a user defined data tuple.
* Watched the following videos regarding training a model
  + - <https://youtu.be/3YWJAooQTZQ>
    - <https://www.youtube.com/watch?v=32o0DnuRjfg>
    - <https://youtu.be/3YWJAooQTZQ>
* Went through the following articles
  + - <https://appen.com/blog/training-data/>
    - <https://www.cloudfactory.com/training-data-guide>
    - <https://www.techopedia.com/definition/33181/training-data>
    - <https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac>
    - <https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86>
    - <https://machinelearningmastery.com/linear-regression-for-machine-learning/>
    - <https://www.excelr.com/blog/data-science/regression/simple-linear-regression>

**Outcome**

The following are some of the things I've learned during the last 15 days as a result of my activities.

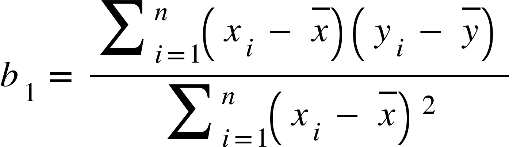
**Linear Regression**

Linear regression is perhaps one of the most well known and well understood algorithms in statistics and machine learning. It is a kind of regression in which we try to fit a line onto a set of data points. It is one of the most common algorithms that’s used for predictive analysis. The base of the model is the relation between a dependent and independent variable basically represented as



* **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
* **B0** is the **intercept**, the predicted value of **y** when the **x** is 0.
* **B1** is the regression coefficient – how much we expect **y** to change as **x** increases.
* **x** is the independent variable ( the variable we expect is influencing **y**).
* **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B1) that minimizes the total error (e) of the model.

By differentiating the above formula, we can obtain an equation for beta1 and beta2 using which we can define the equation for error. Then we will try to define the model by minimizing the residual error.



[Gradient descent is an optimization algorithm that finds the values of parameters (coefficients) of a function (f) to minimize the cost function (cost)](https://machinelearningmastery.com/gradient-descent-for-machine-learning/).

**Logistic Regression**

 Logistic Regression is used when the dependent variable(target) is categorical.

The type of function used here is a sigmoid function.

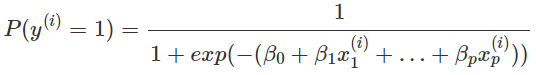
If ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0. So, the only outputs of a logistic regression model are ‘0’ and ‘1’.

logistic(η) =

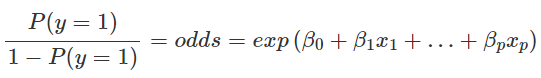
The step from linear regression to logistic regression is kind of straightforward. In the linear regression model, we have modelled the relationship between outcome and features with a linear equation:



For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.



From the above equation, in the end we can define the odds ratio as

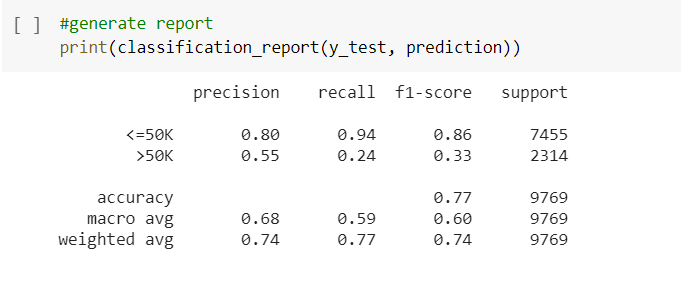
****

**Project Development**

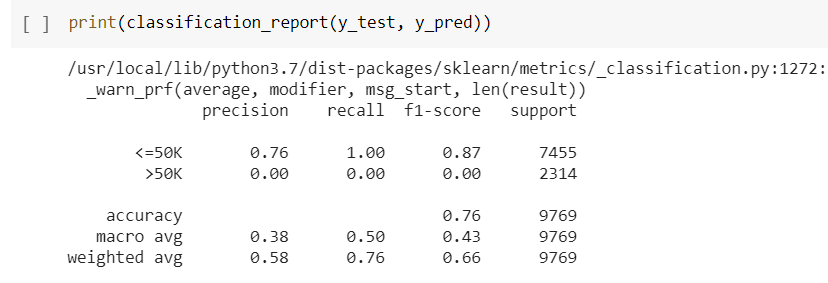
Now at the end of 2nd milestone, I have preprocessed the dataset. I have also trained the model using logistic regression and tested the model. Then I have created classification report. Then I have tuned the parameters of the model and chose the best ones. And again, I have tested and printed the classification report. Even though there is not much change in the precision, we can see a small increase in the ‘<=50k’ class. At the end I have tested the model against a user defined data tuple also.

**Project Inference**

So, I have tried a logistic regression model here for the HR salary dashboard. And my classification report was like this.

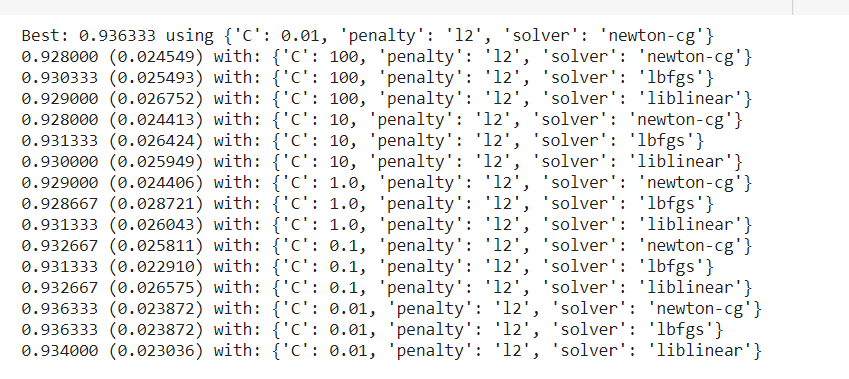


We can see that the model could achieve a 77% accuracy on predicting the salary. Then I tried it with the SVM model and my classification report was like this

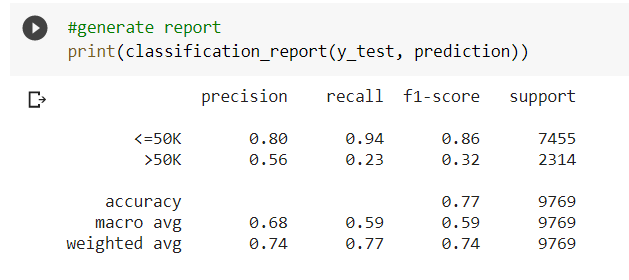


I found a decrease in the accuracy, so I continued with my logistic regression model.

Now I went on to adjust the parameters of my model even though there was not much biasness in prediction of 2 classes. So I did the hyper parameter tuning and the results were as following.

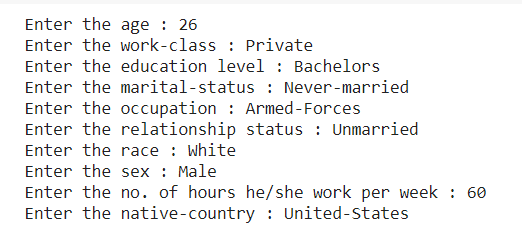


Hence I found the best parameters, and then I trained the model using them, and unfortunately I couldn’t find much difference in result. The classification report is as follows.



There is a small increase in the precision of ‘>50k’ class.

Then I went onto test my model on a particular user defined tuple of data.



The output was as following.



**Link to code and executable file**

Link to the colab file:

<https://colab.research.google.com/drive/1KOhRZ8AzNSDxyZ1Llr3yopWI1p9DGwEs?usp=sharing>

Link to the GitHub file:

<https://github.com/Rithik-Alias/Salary-prediction-dashboard>