|  |  |
| --- | --- |
| 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 | Harish Kumar |
| Name of the Institute | ICT Academy of Kerala,Thiruvananthapuram |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 17-02-2023 | 07-03-2023 | | 32 | | Jupyter | Python3 |
| Milestone # | 2 | Milestone: | | Train the Dataset and Predict Salary | | |

**TABLE OF CONTENT**

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

**ACKNOWLEDGEMENTS**

I am conveying my sincere gratitude towards my industry mentor, Harish 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 is to Train the Dataset and Predict Salary of Employee.

**INTRODUCTION**

The next 10 days of my project, I visualized all the features using EDA. After all the visualization part, I trained the dataset using Logistic Regression and Random Forest. Classification reports were generated.

**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.
  + Watched few of the linear regression YouTube videos.
  + Read the linear regression article.
  + Went through the linear regression & Random Forest YouTube videos.
  + Read the linear regression article.
  + Watched the lectures provided and other videos for further understanding.
  + 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.
  + Watched videos for understanding about model training and hyperparameter tuning.

**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 coding.

**OUTCOME**

After the 2nd milestone of this internship project, I have learned about regression models and lazy predict.

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

**Classification Vs. Regression**

There are four main categories of Machine Learning algorithms: supervised, unsupervised, semi-supervised, and reinforcement learning.

Even though classification and regression are both from the category of supervised learning, they are not the same.

The prediction task is a classification when the target variable is discrete. An application is the identification of the underlying sentiment of a piece of text.

The prediction task is a regression when the target variable is continuous. An example can be the prediction of the salary of a person given their education degree, previous work experience, geographical location, and level of seniority.

**Logistic Regression:**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



**Linear Regression:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Mathematically, we can represent a linear regression as:

**y= a0+a1x+ ε**

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Consider the below image:



When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

**Metrics to Evaluate Machine Learning Classification Algorithms**

**1.Accuracy:**

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

****

It works well only if there are equal number of samples belonging to each class.

**2.** **Confusion Matrix**

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

There are 4 important terms:

* True Positives: The cases in which we predicted YES and the actual output was also YES.
* True Negatives: The cases in which we predicted NO and the actual output was NO.
* False Positives: The cases in which we predicted YES and the actual output was NO.
* False Negatives: The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the “main diagonal” i.e



**3.Area Under Curve**

Area Under Curve (AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

**4.F1 Score**

F1 Score is used to measure a test’s accuracy

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).



Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.



Recall: It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

**Introduction to Lazy Predict**

When working on a regression or classification-based machine learning problem where you want to compare which model will work best on your dataset, you should train and test the performance of various machine learning models one by one on your dataset. This is where the Lazy Predict library in Python comes in. It allows you to compare the performance of all classification and regression models in a few lines of code.

So, by using Python’s lazy predict library, you can easily compare the performance of all machine learning models so that you can choose the best performing model for your problem. If you have never used this library before, you can easily install it in your system using the pip command:

*pip install lazypredict*

**LINK TO CODE AND EXECUTABLE FILE**

* Executable file: [http://localhost:8888/notebooks/HR%20Salary%20Dashboard.ipynb#](http://localhost:8888/notebooks/HR%20Salary%20Dashboard.ipynb%23)