Sri Lanka Institute of Information Technology

Fundamentals of Data Mining

[IT3051]

Mini Project – Statement of Work Document

2023

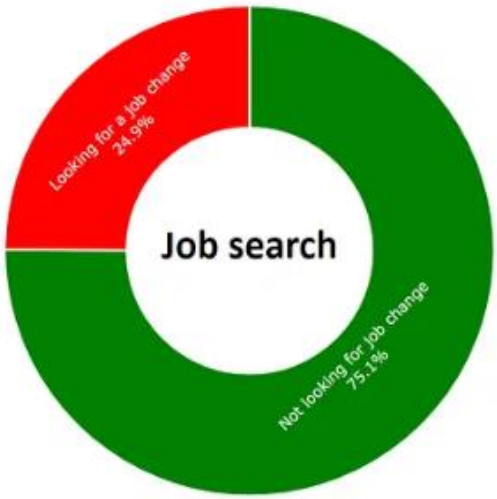
Group – G09

Dissanayake D.J.R	**-**	**IT21313370**
Sulakkana H.D.S.R	**-**	**IT21224348**
Kuhananth C	**-**	**IT21302244**
Manathunga M.A.O.S	**-**	**IT21224652**

Table Of Content

Background

The company specialize in the execution of substantial data science and big data project and, as part of their recruitment strategy, offers training courses to potential employees. The primary objective is to ascertain the genuine commitment of candidates towards long-term employment with the organization to subsequent to their training. This determination holds considerable strategic value as it continues to cost reduction, time efficiency, enhance the quality of training programme and facilitates targeted categorization of candidates.



The graph illustrates the proportion of employee contemplating a job change vs not.

Following a throughout analysis of the job changing issue, we have determined that the most practical solution is to devise a method for identifying or predicting the likelihood of employees changing their data science job.

As students in group G09, we have comprehended the problem and have chosen to devise a solution. Our approach centers on addressing this challenge through the lens of Human Recourses (HR) Team, as they are the primary stakeholders significantly impacted by this issue.

- **Problem –** The company invests resources in developing proficient employees. Post training, identifying employees planning to change job is crucial, as this can adversely impact the company.

- **Client -** Human Resources team of the company. The HR team serves as the conduit facilitating communication and collaboration between the organization and its employee**.**

- **Solution –** Predicting whether an employee will change their data science job after completing training.

- **Goal –** Developing a predictive model to assists the Human Recourses team in anticipating whether a trainee is likely to seek alternative employment post-training, thereby aiding them in making informed decisions.

- **Dataset Selected** – We have selected the following public data set.

HR Analytics: Job Change of Data Scientists

Scope of work

This project structured into 5 main layers, which include.

I. User interface layer
II. Data wrangling and data cleansing layer
III. Data mining layer.
IV. Model building and analysis layer.
V. Data visualizing layer.

Here is the brief explanation of the above layers.

1. **User interface layer**
   This is also known as the frontend layer, where users can interact with the system. In here users can select data or input relevant data which is important for the analytics section. We are mainly focusing about user- friendliness of the system. End of the project our goal is to implement this interface with a simple questionary. This interface plays a important role in the system because it interacts with the end-users

2. **Data wrangling and data cleansing layer**
   This layer performs an important part which is data cleaning and preprocessing, which helps to identify corrupted and inaccurate records. After detecting them, it will be replaced, modified, or deleted by using appropriate preprocessing techniques. This process will help the model to provide more accurate results.

3. **Data mining layer**
   This layer primarily revolves around the analysis and gathering of dataset using algorithms to extract and transform data into structured format, which is suitable for the further analysis.

4. **Model building and analysis layer**
   This layer is used to build the predictive model. Which is a mathematical solution developed using selected dataset and used to predict the desired outcomes from newly gathered data**.**

5. **Data visualizing layer**
   This layer facilitates the graphical representation of the data set. Which will enable users to interpret predicted result generated by the model. And enhance their comprehension of the findings derived from created model.

Activities

- **Find a real-world problem and define a solution.**
  We found a publicly available dataset which is addressing a current real-world problem. Kaggle was used to find a real-world data set which is, HR analytics found that data scientist are changing their jobs after certain time period.

  For this real world problem we were able to come up with a solution, which is to build a model that will predict whether data scientist are going to change the job after certain period of time.

- **Data preparation, model building and training**
  Given the initial state of the dataset, preprocessing become important. This involves cleaning, normalization, handling null values, dimensionally reduction and overall preparation for the solution implementation. Wide research on the problem and several models were selected to facilitate implementation of potential solutions.

- **Evaluate the model.**
  From the list of prepared models, the selection of the most suitable model is the next step. A complex evaluation process will be conducted, considering the attributes such

as accuracy and error rates, final goal is to identify the best performing model for the adoption.

- **Make predictions.**
  The chosen optimal model will be employed to make predictions, effectively addressing the business problem.

- **Front-end development and deployment**
  Building front-end application will be the last part. This will give a better user experience and reduce the technical complexity of the solution. And integrate the model into a production environment.

Approach

Our project commences with the careful selection of the dataset. After going through the dataset, we determine the necessary steps for data cleaning to ensure its suitability for model building. We intend to construct two distinct model employing different techniques for binary classification. After that we will evaluate and compare the accuracy of these models, enabling us to identify the best performer. Finally, we planned to develop User Interface (UI) where user can interact with

Dataset - HR Analytics: Job Change of Data Scientists

Data preprocessing

- Remove the column which have less weight on predicting power.
- Remove the rows with null value. (Null value handling)
- Discretize the column with continuous value.
- Perform the data normalization, reduction, and integration operation on the dataset.
- Divide the data set into training and testing set**.**

Building the model

- Two models will be developed using training dataset.
- Following will be used for model building.

- Algorithm – Decision tree
- Language – Python

Analyzing and verifying the model

- Using the test dataset, model will be evaluated, and the best model will be chosen based on the accuracy of the model.

Building the interface and server

- React js will be used for the frontend.
- Backend will be developed using streamlit.

Deliverables

The primary goal of this system is to offer insights to the Human Recourses team regarding whether data scientist will switch jobs after receiving the training offered by company. Sequentially, it will assist them in making informed decision to reduce costs and time while also enhancing the quality of the training program.

Assumptions

- **Data quality assumption**
  Data is assumed to be accurate, complete, and representative of the problem.

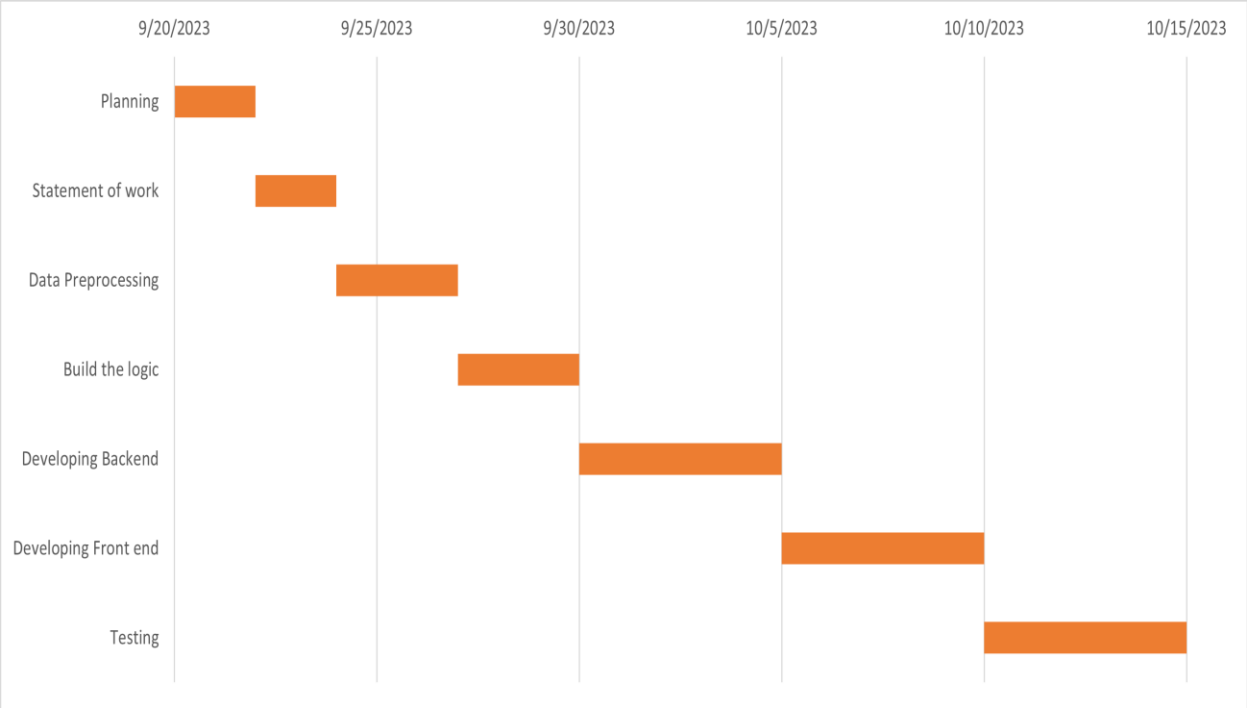
- **Independence assumption**
  Assume that value of one data point does not depend on or influence the values of other data points. This assumption simplifies the modelling process and crucial for the validity of various statistical methods.

- **Noisy data assumption**
  Dataset may contain noisy data, which are random variants or errors in the data. Data mining algorithms assume useful information is stronger than noisy data.

- **No hidden variables assumption**
  Assume that all relevant information is present in the data set and there are no hidden or unobserved variable that could impact the result.

- **Feature relevance assumption**
  Assumes that the features or variables used in analysis are relevant to the problem.

Project plan and Timeline



Gantt Chart

Project Team, Roles, and Responsibilities

	Member IT Number	Member Name	Member Role	Member Responsibility
1	IT21313370	Dissanayake D.J.R	Team Leader Solution Developer Business Analyst Solution Tester	Implement model Handle documentation Test alternate model Data analysis and process UI development
2	IT21224348	Sulakkana H.D.S.R	Solution Developer Business Analyst Solution Tester	Implement model Handle documentation Test alternate model Data analysis and process UI development
3	IT21302244	Kuhananth C	Solution Developer Business Analyst Solution Tester	Implement model Handle documentation Test alternate model Data analysis and process UI development
4	IT21224652	Manathunga M.A.O.S	Solution Developer Business Analyst Solution Tester	Implement model Handle documentation Test alternate model Data analysis and process UI development