Utsav Chandresh Savaliya

Utsavchandresh.savaliya@sjsu.edu

Lab 1 report

CMPE 257 MACHINE LEARNING

# **PROBLEM STATEMENT AND DATASET**

This Lab works aims to find that whether a employee at a company is looking for a job change or not. The data set consists of information about the candidates. So a machine learning model is required to predict that candidate is likely to shift a job or not.

The dataset has the below information:

• index: Unique ID for candidate

• city: City code

• city\_ development \_index: Development index of the city (scaled)

• gender: Gender of candidate

• relevent\_experience: Relevant experience of candidate

• enrolled\_university: Type of University course enrolled if any

• education\_level: Education level of candidate

• major\_discipline : Education major discipline of candidate

• experience: Candidate total experience in years

• company\_size: No of employees in current employer's company

• company\_type : Type of current employer

• last*new*job: Difference in years between previous job and current job

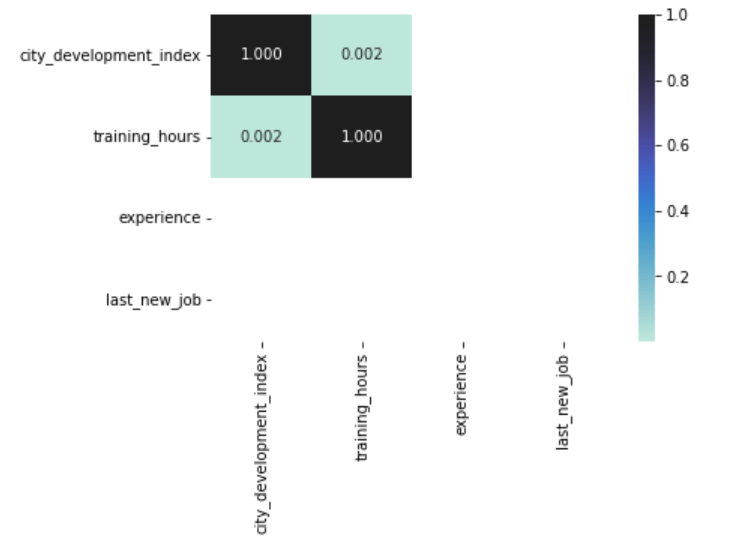
• training\_hours: training hours completed

• target: 1 – Looking for a job change, 0 – Not looking for job change

The outcome of the problem statement is the ***target*** from the above dataset. So it is classification type of problem. The dataset given is an imbalanced type of dataset. We will do required pre-processing before beginning training on the model.

# **REQUIRED PREPROCESSING:**

The following below operations are performed on each of the 13 features (columns available in the dataset):

* Handling missing values for each of the features.
* Analyzing the data based on Min, Max, Mean, Median and Mode values, this helps in understanding the deviations in the dataset if there is any.
* Removing any special symbols and apply scaling on continuous columns in which it is needed.
* Doing certain type of operation based on the type of feature. In the given data set the categorical features are city, gender, experience, university enrollment, major in the university, education level, type of the company in which the employee is working and continuous features are city development index, number of training hours, experience at job, how many last jobs are held by the employee.
* Fill the missing values with mode for all the above categorical features and with median for all the continuous features.
* All the above operations should be performed for both of the training and test data set.
* Conversion of the categorical features to binary values using one hot encoding in order to find some relation between features.
* Plotting of Correlation heatmaps in order to understand if the initial observations of data coming from scatter plots supports multi collinearity not.

**Heatmap to understand correlations between continuous features**

# **BUILDING THE MODEL**

There are five models which are applied on the given dataset for the prediction of the target variable. Following are models applied during the training and testing

1. Logistic Regression
2. SVM – poly
3. SVM – Sigmoid
4. SMV – rbf
5. Random Forest

On the evaluation metrics all models are giving results but we can easily point out that random forest gives the best results. On the evalution metrics following parameters are measured: Accuracy, Precision, Re-Call,F1-Score

We have used hyper parameter tuning to make our model best as per our requirement. There can be drastic changes in the scores if these parameters are not properly tuned.

# **TESTING ON THE MODEL**

We have test the X\_test file given with the labwork. The best results are generated are with the Random Forest model as it uses decision trees in the process. Also the time taken by Random forest model is less.