# DM6102 Project Proposal

**Group members: Renping Ge, Yuchen Ma, Zongzhu Li**

**• What problem did you select and why did you select it?**

**To sharpen our data mining skills and understand real-world data , our group decided to select the topic of job change for data scientists from a Kaggle competition: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists**

**The objective of this project is trying to predict the probability of a candidate looking for a new job after completing the training offered by the company.**

**The reason we select it is because knowing which of these candidates really wants to work for the company after training or looking for a new employment will help to reduce the cost and time.**

**• What dataset will you use? Does it need to be cleaned?**

**We would use the dataset provided by Kaggle. This dataset is collected by the company and designed to understand the factors that lead a person to work for the company(leaving their current job).**

**The dataset needs to be cleaned since it's imbalanced and contains missing values. Besides, most features are categorical, so it needs to conduct some data cleaning and preprocessing.**

**• What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?**

**Since most of the features are categorical(Nominal, Ordinal, Binary), logic regression, classification algorithm and K-mean algorithm will be used in this project.**

**• What packages will you use to implement the network? Why?**

**The following packages will be used to implement the network:**

1. **numpy to handle linear algebra**
2. **pandas to do the data processing and csv file importing**
3. **matplotlib.pyplot to generate plots**
4. **LogisticRegression and sklearn.linear_model to find the best model to present most of the data**
5. **accuracy_score, f1_score, precision_score and recall_score to measure the accuracy of the model**
6. **RandomForestClassifier to split data for training and testing**
7. **……**

**• What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?**

**In this project, EDA, preprocessing, feature selection, model building and evaluation strategy, will be applied, so all the knowledge and reference materials related to the above topic will be referred. To better understand the dataset, the description is on the following link: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists.**

**• How will you judge the performance of your results? What metrics will you use?**

**After building a model with the 80% randomly selected data from the training dataset, we will apply the model to the 20% left dataset to evaluate the model performance and fitness. The evaluation metrics will be area under the ROC curve score.**

**• Provide a rough schedule for completing the project.**

**April 4 - April 11: code for EDA, preprocessing, feature selection;**

**April 12 - April 19: code for model building and evaluation strategy;**

**April 20 - April 26: final report;**

**April 27 - April 30: presentation preparation.**