COMP6237 Data Mining Group Project Brief
Mario Bros

Introduction to The Project

The project is going to choose a decent dataset, to analyse every given feature as well as implementing some technical data mining skills on the data in order to build a best predictive model and to find solution of the problem from the dataset.

The Dataset and Problem

This project would use the dataset provided from IBM Watson Analytics [1]. The data is taken from human resources department and it contains 35 features such as age, education, department, income, as well as whether they have left their jobs or not. The aim of the project is to find what are the most factors driving the employees to leave their job.

Data Analysis and Evaluation

The dataset has to be cleaned before applying the algorithm to train a model. Some columns can be omitted as they will not take any reasonable effects on the problem such as employee number, standard hours as all data contain the same number of 80 and 'over18' as they are all over 18 years old. In addition to this, the data will also be simplified a bit due to some large amount of differences within a column. For example, the monthly income column contains different amount of incomes. These will be simplified as low, medium or high based on a range of incomes. In this stage, the dataset also needs analysing by finding the correlation of the data by comparing the attributes. For example, most of the employees who have left their job also contain very high value of 'NumCompaniesWorked'.

Algorithm Approaches

The project would apply some machine learning algorithms such as decision tree, k-means clustering, random forest and SVM using Python, R and Matlab. Due to the early stage of the project, they are still under consideration. The main purpose of applying them is to build a predictive model to be able to solve the problem. Decision tree may also be used to find the balance of the data to divide the data to have further analysis such as finding what take the most effect on the problem within a specific department.

Details of Group Members

Name	Email
Zilan Huang	zh4u17@soton.ac.uk
Bang Du	bd1m17@soton.ac.uk
Zihan Wang	zw1u17@soton.ac.uk
Yang Zhou	yz12a17@soton.ac.uk
Da An	da1g17@soton.ac.uk
Mardhiyyah Rafrin	mr1n17@soton.ac.uk

References

[1] M. Stacker IV, "IBM," 2 April 2015. [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/.