**CS513 Project Presentation**

**Introduction**

IBM Company Attribution Analysis

We do some basic analysis of employee information from IBM. The data contains ANNUAL\_RATE annual rate, hrly rate, ethnicity, sex, marital status, job satisfaction, age, number of team changed, referral source, hire month, termination year, rehire. By analyzing the relationship amongst those, we can provide a model for the hiring management to better hire people and make best use of human resources.

**Data detail**

The dataset is from IBM, and the dataset contains 9612 IBM employees’ basic information; We used this to data to see if we could build a classifier that could predict whether or not this employee would stay on IBM; There are 27 columns. 21 of which are employee attributes, one column for ID name, and other columns are not considered this time.

**Research Question to handle**

The Employee information has many different dimensions, we want to extract hints from these dimensions to help management to predict if this employee would stay or leave. We sought out to answer the following questions:

Are there specific relationships between those attributes?

Which methodology is the most accuracy one among those methodology?

Can we give someone recommendation of one employee who would leave?

**Method applied**

K-Nearest-Neighbor, Cart, C50, Random Forest, Artificial Neural network, H-Clustering & K-means, Support Vector Machine

We will introduce them respectively later.

**KNN** is one of the most common data mining algorithms of classification. However, before we using KNN function, we need to initialize our data first. The first step is to import the data and replace the missing data to NA. Then what I did inn next step is discretizing the multiple data like ETHNICITY, REHIRE, ANNUAL\_RATE, HRLY\_RATE, JOBCODE, AGE. Then omit data with NA. Furthermore, we split the entire data into training and testing at a ratio of 7 to 3. Finally, we predict the status situation by applying k from 1 to 20, and we can get our rate of error is 0.3898717.

**C5.0 algorithm** (an update of the C4.5 algorithm) always compare with CART, and there are some differences with decision tree in details. In our project, we normalized variables, discretizing the multiple data like ETHNICITY, REHIRE, ANNUAL\_RATE, HRLY\_RATE, JOBCODE, AGE. Then omit data with NA. Furthermore, we split the entire data into training and testing at a ratio of 7 to 3. Apply C5.0 and then get our summary and tree plot. In the end, we can see the final error rate is only 0.3753035

**Classification and Regression Tree**

The model of decision tree works by repeatedly partitioning the data into different groups, thus the outcomes in each smallest leave groups is as same as possible. This approach called Classification and Regression Tree technically. By using CART, we can classify each attributes of data to get connections between them and our target attributes.

Slides part (page 10): This is an original approach called Solution 1 to handle raw data to create classification and regression tree by using rpart function. And STATUS is our target dataset column in the formula.

From the left-bottom graph, we can see that our CART model is not really ideal with almost 56% error rate on predicting the test dataset. The right graph is the plot of the original CART model.

Slides part (page 11): This is a simple introduction about new CART approach with examples. The two bottom graphs are for comparing changes in plot after we operated one small part of the raw datasets. Personally, I think it could describe the importance and necessary that we need to operate the raw data to be discrete or continuous for optimizing the tree structure and tree plot.

Slides part (page 12): This is an improved approach called Solution 2 to create more optimal and precise tree model by using rpart function with initialized discrete or continuous data. And from the outputs comparing we can conclude that if the data is more discrete , it can make the CART to divide leaves and create tree model more clear and better, which also could help company to manage their employees better in reality.

Slides part (page 13): This page contains different graphs of plots with dissimilar formula of rpart function. By choosing different formula of CART approach, we could get different plot of outputs from it. From the results, we could easily get the connections between values in datasets for training with the target dataset.

**Artificial neural networks (ANN)** or connectionist systems are computing systems vaguely inspired by the [biological neural networks](https://en.wikipedia.org/wiki/Biological_neural_network) that constitute animal [brains](https://en.wikipedia.org/wiki/Brain). Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules. In our project, we normalized variables, discretizing the multiple data like ETHNICITY, REHIRE, ANNUAL\_RATE, HRLY\_RATE, JOBCODE, AGE. Then omit data with NA. Furthermore, we split the entire data into training and testing at a ratio of 7 to 3. Apply ANN to predict whether or not one employee leaves IBM and the prediction error rate is just 35% since we set the threshold of possibly leaving IBM is 0.7 and it does make sense for HR to assign human resources.

**Support Vector Machine**

Under the condition of controlling classification error and separation boundary, the training data set training is separated by SVM, to obtain a proper separation interval to predict the "STATUS" target property of the test data set. From the results, some attributes (columns like "ANNUAL\_RATE", "AGE", and "EDUCATION\_LEVEL", etc.) of the raw data have an important effect on the interval between the data, and a reasonable and realistic STATUS is obtained.

Slides part (page 17): There is the core code and outputs in my SVM approach. I am using the datasets like "ANNUAL\_RATE", "JOB\_SATISFACTION", etc. to create a SVM model named “svm”, and then use this model to predict the test dataset.

From the outputs graph, we can see our error\_rate is reduced to 36.8% after raw datasets connection analyzing and transforming raw data to discrete dataset.

Slides part (page 18): About graph-left: This is my approach to illustrate the model of SVM with more specific details. Firstly, we use subset function to subset the data with three labels “Annual\_Rate”, “HRLY\_Rate” and STATUS, and we divide STATUS by its own values. Secondly, call the plot function to draw a ScatterPlot, where HRLY\_Rate is the x-axis and Annual\_Rate is the y-axis, still, our target dataset is STATUS. In the end, set the cost factor to 1 and use the subset dataset to train the svm.model.

About graph-right-top: we use points function to mark the support vector with a blue circle.

About graph-right-bottom: use abline function to add the dividing line.

**Random Forest Algorithm** is an effective approach to develop a classification model and predict the dataset. Meanwhile, it can also identify the importance of features.

The first step of analysis is to initialize the datasets.

The columns of ETHNICITY, REHIRE, ANNUAL\_RATE, HRLY\_RATE, JOBCODE, AGE should be discretized. Then the columns of EMP\_ID, REFERRAL\_SOURCE, HIRE\_MONTH, TERMINATION\_YEAR, JOB\_GROUP should be deleted because the datasets are incomplete.

After the computation by random forest, the rate of error is 0.379119. Moreover, the importance of features is showed in the picture. For example, the first three important columns are PREVYR\_1, PREVYR\_5, and JOBCODE\_A if you do it by using the “Mean Decrease Accuracy”. Plus, you will find that PERFORMANCE\_RATING, EDUCATION\_LEVEL, and NUMBER\_OF\_TEAM\_CHANGED are the first three important features when using “Mean Decrease Gini”.

**Clustering** is often performed as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream, such as neural networks.

The first step of analysis is to initialize the datasets. The features including ETHNICITY, SEX, MARITAL\_STATUS, NUMBER\_OF\_TEAM\_CHANGED, REHIRE, IS\_FIRST\_JOB, TRAVELLED\_REQUIRED, IS\_FIRST\_JOB, DISABLED\_EMP, DISABLED\_VET should be transferred numeric from the style of categorical. Then, the training datasets should be standardized.

After using **H-cluster** algorithm, we can get a cluster dendrogram and we try to cut it into 6 parts. So, we can get 6 different clusters with different features.

Next using **k-means** algorithm and when we cut the datasets into 6 clusters, the clusters 1st 3rd 6th, more employee tend to stay at company; the clusters 2nd 4th 5th trend to leave the company. After checking the plot, we can simply set the two clusters up. Finally, the first cluster have a trend to stay at company; the others have a balance between leaving and staying.