**Slide 1: Opening**

Hello everyone, Today we will be presenting our analysis on employee attrition. Let me introduce you to my group mates Meghana, Akhil, Rajat, and Sanam.

**Slide 2: Introduction**

So what does staff attrition mean?  Staff attrition refers to the loss of employees through a natural process, such as retirement, resignation, elimination of a position, personal health, or other similar reasons. With attrition, an employer will not fill the vacancy left by the former employee. Attrition is often viewed as a way that companies can decrease labor costs. Take for instance more severe forms of attrition—like layoffs—which are often deployed as a labor reduction technique. If an employee leaves naturally, and their spot no longer needs to be filled, organizations can freeze hiring to reduce staffing costs.

**Slide 3: Agenda of the Presentation**

So first we will talk in brief about the summary of the project and our goals followed by which we will be discussing the dataset. Then we will be discussing the various methods for data selection, cleaning, and our observations and finally the conclusion of the project.

**Slide 4: Brief Summary & Goals**

The project was intended to detect potential attrition from a slightly imbalanced dataset. We Performed data analysis and developed machine learning models to detect potential attritions. This dataset was provided by Professor Khasha Dehnad and the accuracy and genuineness of the dataset were never disclosed by the organization who provided the data set to the professor and hence could not be passed on to the students. In this dataset, there were 9612 observations with 27 variables and the attrition year starting from 2004 up till 2017.

**Slide 5: Dataset Description**

In this data set there a total of 27 features like Job satisfaction, Sex, Job Group, Annual rate, hourly rate, termination year, etc. Data set has 3 columns that have unspecified data (NA) that need to be considered. The target variable is Status which has two different classes Terminated or Active. There are a total of 9166 samples. In this data set, we don’t find any noisy data i.e. where there is no employee id or no JOB\_GROUP. The annual rate includes some extra credits like overtime pay and bonuses and thus can not be directly deduced from the hourly rate. The column REFERAL\_SOURCE has both NA values and UNKNOWN values. They should not be considered the same.

**Slide 6: Dataset selection, cleaning, and observation**

For the algorithms that need the data set to be divided into test & training, we have gone with a ratio of 70% training and 30% test. In the data set provided, we found that the majority of the values that were missing were in the column “TERMINATION\_YEAR”. The missing values were for the employees who were currently active in the company so that is logically correct. To proceed with further analysis, we changed the data from number to Boolean values for some of our algorithms. if the employee has been terminated then the value becomes true and vice versa. Also, Termination year does not provide any insight into the current employee status and its attrition. Since the data set had values with a wide range, we went ahead and normalized the same. We used the following methodologies for our analysis and cleaning of the data. Correlations, Artificial Neural Network(ANN), Kohonen, Naïve Bayes, Random Forest, CART, Decision Tree, K-Means, Support Vector Machine (SVM), H-Clust, & K-nearest neighbor (KNN).

**Slide 7: Correlations**

The term "correlation" refers to a mutual relationship or association between quantities. In almost any business, it is useful to express one quantity in terms of its relationship with others. Correlation can help in predicting one quantity from another.

**Learning**, The correlation plot clearly shows the importance of variables plot against other variables. From this, we can easily figure out which columns are important and which ones are not. That is how we have removed the ID variable and Termination year in our data set before implementing algorithms.

**Slide 8: Artificial Neural Network (ANN)**

* An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain
* Hidden layer accepts the input from the previous layer and modifies it in a nonlinear fashion

**Learning**, From the above plot, we can see how the neural network relates all the variables and then comes down to its conclusion and which is pretty accurate when compared with the data. We got an accuracy of about 60%

**Slide 9: Kohonen**

It defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. These models are computed by the SOM algorithm. A data item will be mapped into the node whose model is most similar to the data item, i.e., has the smallest distance from the data item in some metric.

**Findings**, Since the table is spread across multiple columns we used Kohonen to bring down the columns to understand better. From the graph, we can see how the training process is going on for the algorithm. We used somgrid to form the grids to use the SOM function.

**Slide 10: Naïve Bayes**

* We are applying Bayes rule with strong (naive) independence assumptions as it should be.
* In this algorithm, we want to find the posterior probability in light of prior data

**Learning**, if the categorical variable has a category (in test data set), which was not observed in the training data set, then the model will assign a 0 (zero) probability and will be unable to make a prediction. As we all know Naïve Bayes is a very basic algorithm in the field of data science, the accuracy was not that good in this but still, we managed to get an accuracy of 62%.

**Slide 11: Random Forest**

* The concept is that - a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.
* This algorithm performed well on this dataset
* The key thing here is the low correlation between individual tree

**Learning**, Normalization on RF is not a good idea because it did not perform well when features are a monotonic transformation of other features

By using the Random forest, we got to know the importance of different variables in the data set and then we decided to remove the columns which did not affect the outcome. We found ANNUAL\_RATE, HRLY\_RATE & JOBCODE played a very vital role in deciding the outcome. The success rate for this algorithm is 88.38 percent.

**Slide 12: CART**

* It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value Status in our case
* And it can handle both numerical and categorical data and mirrors the human decision-making approach**.**

**Learning**, First of all, we are so sorry for not being able to show up the plot in such a small space but if we project the plot to a TV, you can see the trees and the percentages of the deciding factors. For data including categorical variables with different numbers of levels, (JOB\_Group in our case) information gain in decision trees is biased in favor of attributes with more levels. However, we tried this with removing the job group and other columns and we got an accuracy of 71%.

**Slide 13: Decision Tree**

* It uses a tree-like model of decisions and their possible consequences

**Learning**, An optimal decision tree is one that accounts for most of the data, while minimizing the number of levels (or "questions"). With the decision tree, we were able to predict the outcomes with an accuracy of 71.7% which is shown in the table. This plot represents the deciding factors and the percentages of them. If this image is projected to a TV, the plot will be clear.

**Slide 14: K-Means**

* Algorithms aim to partition *n* observations into *k* clusters (2 in our case) in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean)
* Intuitively, *k*-means implicitly assumes that the ordering of the input data set does not matter as a fact of clustering

**Learning**, Normalization is not always required, but we did it. For example K-means: K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. Using K-Means, we were able to classify the data successfully, in this case, we got a success rate of 51.24%, which is shown in the slide. You can see the tabulation as well on the slide.

**Slide 15: Support Vector Machine (SVM)**

* It is a Non-parametric algorithm
* It tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible
* Also uses Kernel functions to project and to project back the decision boundary from higher dimensions

**Findings**, powerful algorithms that perform equally well on test and train data. This algorithm has given comparatively better results, in this case, we got 69.85% as the success rate in predicting the status of an employee which is good considering the data set.

**Slide 16: H-Clust (Hierarchical clustering)**

* Algorithms put each data point in its cluster.
* It identifies the closest two clusters and combines them into one cluster.
* provide the data in the form of a distance matrix

**Findings**, We can see the cluster dendrogram of the data set. In this picture, it is not clear since we need to project this to a big screen.

**Slide 17: K-nearest neighbor (KNN)**

* K Nearest Neighbor is a nonparametric approach
* It is based on feature similarity techniques to predict by examining the majority vote from K nearest neighbor selected

**Findings**, Small value of K cause model sensitive to noise, higher value cause unnecessary computation and polling. KNN Algorithm was successfully implemented on the data set when we ran with k=3, we got an accuracy of 56% and when we ran with k = 5, the accuracy increased to 58%. We tried this for k beginning with 1 to 21 using a loop until the results converge. This method was useful in predicting the outcome taking into consideration the nearest neighbors.

**Slide 18: Conclusions**

Multiple conclusions were drawn from the analysis of the data set. First of all, we used correlation and random forest to find unnecessary columns so we can get rid of the same. By using ANN algorithm we were able to correctly predict 58.16% of the results to predict if an employee might leave the organization or not. The best prediction rate was achieved using the Support Vector Machines algorithm were able to get an accuracy of 69.85%. You can see the accuracy of the other algorithms that we have run on the data set. Employees who have changed more number of teams tend to have less job satisfaction. Using our program, the employer can easily find the potential employees who might leave the organization and the company can resolve out their issue by providing financial benefits or look into the other aspects and help them in changing their team to retain the employee. Employee's annual rate plays the most important factor in the determination an employee who might leave the organization. The people whose first job is with this organization are more likely to change their team.