Project Report

CS-513: Knowledge Discovery and Data Mining

Final Project - Data Analysis on Employee Attrition Data

Group - Team Ducks$Indian

Aakash Rami, Rohan Ratwani, Hardik Patel, Manan Bhatt

With the employee attrition dataset, we can uncover various factors that lead to the attrition of an employee from a particular company i.e to predict the potential of employees leaving the company. The dataset contains data of around 10k employees with 27 columns. The training data set contains data for 7209 employees and the testing data set contains data for 2403 employees. As we explore the data set, first we can see that the columns EMP\_ID and JOBCODE are unique values and won't affect the model so we remove those columns. TERMINATION\_YEAR contains missing data for those employees whose status is A i.e. employees who are still working in the organization. To fill those missing values and represent them as a separate category we replaced the missing value of TERMINATION\_YEAR with 0. Apart from that REFERAL\_SOURCE and ETHNICITY also contain missing values. Rather than removing the data with missing values, we decided to impute those missing values to increase the accuracy. Since both of them were categorical columns, we replaced the missing values with the mode of the columns. It is also visible from the dataset that all the data is not under one common scale i.e. the data is very sparse. For example, ANNUAL\_RATE has values ranging from 16786 to 1250924 and other features like JOB\_SATISFACTION have values ranging from 1 to 5 so there is a need to normalize the data. We used min-max normalization to normalize the data and bring everything under a common scale.

Since the dataset contains lots of categorical values such as ETHNICITY, MARITAL\_STATUS, SEX, etc. we decided to see the feature importance through the RandomForest algorithm which will give us the importance of each column and help us decide the important features out of the entire data set. From the results, it is visible that TERMINATION\_YEAR has very high importance compared to all other features and it overpowers other features due to which the importance of other features decreases drastically. So we removed TERMINATION\_YEAR and again checked for feature importance and made note of the accuracy. We then selected the top 10 features (ANNUAL\_RATE, HRLY\_RATE, REFERAL\_SOURCE, HIRE\_MONTH, PERFORMANCE\_RATING, EDUCATION\_LEVEL, PREVYR\_1) according to the feature importance graph and again ran RandomForest with only those 10 features and noticed an increase in the accuracy.

The next step was to check and study the correlation between different columns of the data set. From the Heat Map, it is clearly visible that none of the features are highly correlated with our target variable i.e STATUS. This observation tells us that linear classification models won’t provide us any useful insights and classification and instead we need to use non-linear classification models that will take into account the effect of other non-correlated independent features. It is also visible from the heat map that ANNUAL\_RATE and HRLY\_RATE are highly correlated with each other so both of the columns will have an almost similar effect on the target variable, so we removed HRLY\_RATE. One more observation from the heat map is that PREVYR\_1 is loosely related to the PREVYR\_4 and PREVYR\_5 as compared to PREVYR\_2, PREVYR\_3. So we can include PREVYR\_4 and PREVYR\_5, but as they both are highly correlated with each other we can take anyone feature into consideration so we selected PREVYR\_4.

At this stage, the data is cleaned and prepared for running different non-linear classification algorithms. The data is splitted into 75% and 25% as training data set and testing data set respectively. We then ran different models and compared their accuracy.

1. K - Nearest Neighbor: For KNN we converted entire data into factor and ran the algorithm for different nearest neighbors and for k = 25 we got the highest accuracy of 62.08 % as compared to other values of K
2. Decision Tree: For Decision tree, we achieved an accuracy of 64.33 %
3. C50: With C50 we observed a decrease in accuracy with 61.205% compared to decision tree.
4. Naive Bayes: With the Naive Bayesian model we received an increased accuracy of 69.02% which was by far the best.
5. RandomForest: For the RandomForest model we tried by changing the hyperparameter of ntree and after comparing different values of ntree we observed the highest accuracy of 70.16% at ntree = 1000.
6. Artificial Neural Network (ANN): For ANN models we observed different accuracies by making changes in the number of hidden layers and changing other hyperparameters like threshold, repetitions, learning rate, etc. We achieved the highest accuracy of 67.862% with 2 layers, learning rate and threshold of 0.1 and repetition of 1.

After running different models we observed the highest accuracy for the RandomForest model with 70.16% accuracy.