**Project Blog: HR Analytics Project- Understanding the Attrition in HR**

Author: Kirti Arya

April 29th, 2022

Welcome to the new world of Machine Learning! We have no doubt progressed so far in the fields of data and technology. With an intent to learn more about data and curiosity to know about “What Next?”, here is my blog regarding the model that predicts the “Attrition” of the employees.

**Problem Definition:**

The attrition impacts the companies in many ways incurring additional cost to train new hire and invest time and effort affecting the productivity at work. This is the major issue and the HR try to find out the issues if the attrition in an organization goes beyond certain extent. This project contains the evaluation of the attrition data being gathered to make an insight and understand the factors boosting the attrition, their consequences and plan out strategies to overcome such situation. This is called HR analytics. This project used here works on the same principle to analyse the set of data, manipulate it and make the machine learning models out of it to understand whether the “Attrition” would be “Yes” or “No”. This model will be trained initially on data set and would be then tested on a portion of data set. The prediction made by the model would be measured in terms of the accuracy of the model. Any machine learning model with an accuracy of greater than 70% is considered as the good model performance. This project defines my analysis with the data and building model.

**Data Analysis:**

The dataset contains the structured data in the form of rows and columns with total of 1470 rows and 35 columns. There are no null values in the data. The columns labelled are - Age, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction etc. The data is mix of both numeric and string. The main intent is to predict the “Attrition” based on the factors given. The model would be trained and test over a set of data and would be run to check the accuracy of the prediction made by model. The good accuracy shows that the model can make correct predictions and can be employed for use. Since the target column has the responses in only two categories (Yes/No or 1,0) the model build can be logistic regression model or classifier model. I have made use of 5 different classifiers to predict the attrition and find the best accuracy.

The data analysis includes the exploratory data analysis to find the best possible insights from the data prior to building the model. The analysis made on this data reports certain factors. The dataset has the data in only – int64 and object data types. The columns with object type data type are converted in numeric data type using the label encoder. Since the target column has quite unbalanced responses which is balanced out using SMOTE. The data has certain outliers and skewness too. The outliers are dropped, and skewness is removed using the 'yeo-johnson' power transform. Further, the data is scaled to remove vast difference between the data values in each column. There is multicollinearity too present in the data – correlation among the input columns that is removed by using PCA. This technique merges the columns with most correlation among them and I have clubbed 2 columns pairs making the total columns in input data to be 33 instead of 35.

The best and most informative parameter is the correlation. It can be observed that how there is direct positive correlation between the Attrition column and columns such as – Over Time, Marital Status, Distance From Home, Job Role, Department, Number of Companies Worked etc. This shows a clear understanding that the employees working overtime, living far away from office, employed with max number of companies etc. have high chances to resign.

On the other hand, the Attrition column is inversely proportional to the factors such as- Years Since last promotion, relationship satisfaction, work life balance, job satisfaction, years at company, age, monthly income, total working hours etc. This indicates that employees with the least promotion, less work life balance, less job satisfaction, income, age and working hours tend to have more chance to leave the organization. This information is useful to make estimate about the employees when hiring them.

**Exploratory Data Analysis Concluding Remarks:**

The data after the exploration and tells the clear relationship between various input columns and target column called Attrition. The correlation is very helpful here. Further the skewness and outliers show that there can be certain exception too in the data. The entire data set is converted to numeric data before building model and the target column will tell the attrition as 1 or 0 (1= Yes, 0=No). The graphs shown are helpful in the clear visualization of the data result.

**Pre-Process Pipeline:**

Before we start with the model building, the data is split into two parts- consisting of input data and output data (target). The input data here is x and output data is y. The important libraries related with each different classifier are imported. The key factor to consider before starting to build model is to make sure that the entire data is numeric.

**Building Machine Learning Models:**

As mentioned, I am using 5 different types of classifiers to predict the attrition here. The classifier models used in this project are - Decision Tree, K Nearest Neighbor, Gaussian Naïve Bayes, Support Vector Classifier (SVC) and Random Forest Classifier.

**Decision Tree Classifier**

I have split the test and train data here in the 30-70 ratio. After making the certain runs, I am able to find out the best accuracy at the random state of 55 so made it final. The hyper parameter tuning is done to tune the data in order to get better accuracy. The best parameter seen here is ‘gini’ with cv= 5, so used this criterion to train the model on the train data and make the prediction on the test data ie 30% of the entire data. The accuracy of the model here reported is 83%. This can be considered as the good accuracy. Further, to validate the accuracy I have performed cross validation to display the accuracy of 5 different chunks of data. The mean accuracy is here is 77% with the highest accuracy of 83.9%. This shows that the model prediction is appropriate.

**K Nearest Neighbor Classifier**

Another model I used is the KNN model. I have split the test and train data here in the 20-80 ratio here. After making the certain runs, the best accuracy is seen at the random state of 54 in this case. In the hyper parameter tuning, the best parameter seen is - {'metric': 'manhattan', 'n\_neighbors': 3, 'weights': 'distance'}, so used this criterion to train the model on the train data and make the prediction on the test data. The accuracy of the model here reported is 86.8%, better than the previous model. This is considered as the good accuracy. Further, to validate the accuracy I have performed cross validation to display the accuracy of 5 different chunks of data and using the best parameters. The mean accuracy seen here is 84.33% with the highest accuracy of 84.58%. This shows that the model prediction is appropriate.

**Gaussian Naïve Bayes**

Next model used is the Gaussian NB model. Here, again the test and train data is split in the ratio 20-80 and the best random state selected here is 40 since accuracy delivered is maximum here. In the hyper parameter tuning, the best parameter seen is- {‘var\_smoothing': 0.0023101297000831605}, so used this criterion to train the model on the train data and make the prediction on the test data. The accuracy of the model here reported is 87.65% that is comparatively better than both the previous models. To make sure or validate the accuracy the cross validation performed displays the accuracy of 5 different chunks of data and using the best parameters. The mean accuracy seen here is 80.62% with the highest accuracy of 87%. This shows that the model prediction is appropriate however the difference between the mean and highest accuracy is maximum in this case.

**Support Vector Classifier.**

The model used next is the SVC. It has three different kernels- rbf, poly and linear. To find out the maximum efficiency here, the test and train data here, is split into 30-70 ratio and the best random state selected is 50. The hyper parameter tunning is also performed here, and it displays the best parameters as - {'C': 10, 'kernel': 'rbf'}. I have used this criterion to perform the model training and testing. In this case, the models show the accuracy of 92% which is the best accuracy so far. I have further not run the model with other two kernels since ‘rbf’ is the best reported kernel. The cross-validation step also shows that the model is performing accurately. It displays the maximum accuracy as 93.7% and mean accuracy as 87.84%.

**Random Forest Classifier**

The fifth and last model used here is the Random Forest Classifier. Here, I have kept the test and train data in the ratio 20-80 owing to the performance and accuracy of the model, hence the random state seen here is 46. Further, the hyper parameter tunning is done to improve the model performance and the best parameters seen here are- {'max\_depth': 15, 'min\_samples\_split': 2} with accuracy shown is 89%. Then, model training and testing is performed using the best parameters. It delivers the accuracy of 91% which is again a good accuracy score. The cross-validation steps here show that the mean accuracy is 87.32% and the maximum accuracy is 94.72% for 5 different chunks of data. This model too performs accurately.

**Conclusion**

Summing up the different models and their performances, the data looks like as displayed below:

* Decision Tree Classifier – accuracy= 82%, cross validation score = 77%
* K Nearest Neighbor Classifier - accuracy= 86.67%, cross validation score = 83.33%
* Gaussian Naïve Bayes Classifier- accuracy= 88%, cross validation score = 80%
* Support Vector Classifier (kernel = rbf) - accuracy= 92%, cross validation score =87%
* Random Forest Classifier - accuracy= 91%, cross validation score = 87%

Considering the data above, I have selected the Support Vector Classifier as the best model with the maximum accuracy and lest difference between cross validation score and accuracy predicted.

Further, the data set predicted is saved and model is saved using the library pickle. To perform the prediction, the model saved can be loaded from the disk. It predicts the “Attrition” with the accuracy of 98%.

Hence, the model SVC can be used for prediction.

This is my best effort to perform data analysis and build the prediction model out of it. However, it might contain certain shortfalls or discrepancies since I am still in learning phase.

Thanks for your time and patience!