Report for Data Contest

The prediction for the Employee Churn prediction was prepared in 2 main steps. The first step was to extract important and relevant data from the various .csv files given to us namely, *"ratings.csv"*, *"remarks.csv"* and *"remarks_supp_opp.csv"*. The second step was to train our model using a good classifier after successful data extraction / cleansing.

A feature matrix was prepared with columns signifying the necessary features extracted from the above files which would help in training our classifier. The final features were as follows:

my_id	last_rating_date	first_rating_date	weighted_rating	no_ratings

The first 5 features shown above were defined as follows:

- **my_id** : a unique ID formed by the concatenation of *emp* and *comp*.
- **last_rating_date** : the last rating date for the given employee.
- **first_rating_date** : the first rating date for the given employee.
- **weighted_rating** : ratings weighted w.r.t the date on which they were given, giving more priority to more recent ratings.
- **no_ratings** : total number of ratings given by the employee.

recd_opp_0	recd_opp_1	recd_supp_0	recd_supp_1

The above 4 features are defined as follows:

- **recd_opp_0** : number of people who opposed a given remark and had not left the company in the future.
- **recd_opp_1** : number of people who opposed a given remark and had left the company in the future.
- **recd_supp_0** : number of people who supported a given remark and had not left the company in the future.
- **recd_supp_1** : number of people who supported a given remark and had left the company in the future.

Note that as more than one remark could've been made by the same employee, and hence the final value for the features were taken as a weighted average w.r.t the length of each remark, giving longer remarks more importance.

Similarly, there were other 4 features such as *"giv_opp_0"*, *"giv_opp_1"*, *"giv_supp_0"* and *"giv_supp_1"*, which tell us the number of supports/opposes given by an employee who had in the future left/not left their company. These features were not weighted averaged.

Lastly, there were two more features called *"last_remark_date"* and *"first_remark_date"* which tell us the date at which the first/last remark by an employee was made.

All these manipulations were performed by using the pandas library on python, and finally a .csv file was exported which would then later be imported in the second step for the classification task.

(Refer to code file "data_extract.ipynb" for more information)

After the above steps, the following features were dropped:

- *last_remark_date*
- *first_remark_date*
- *last_rating_date*
- *first_rating_date*

and the following features were added:

- diff1 : the difference between the first and the last rating date
- diff2 : the difference between the first and the last remark date

These two features give an idea about the span for which the remarks and ratings have been collected from the individual employees.

After completion of the feature extraction,

- Miss_forest was used to impute the data and fill the NaN values.
- Standard Scaler was used to scale the features to the same range.
- SMOTE was used to balance the dataset if it was unbalanced by oversampling.

Then, we used the voting of the following classifiers on the data:

- Logistic regression
- XGBoost
- Random Forest
- Gaussian Naïve Bayes
- AdaBoost