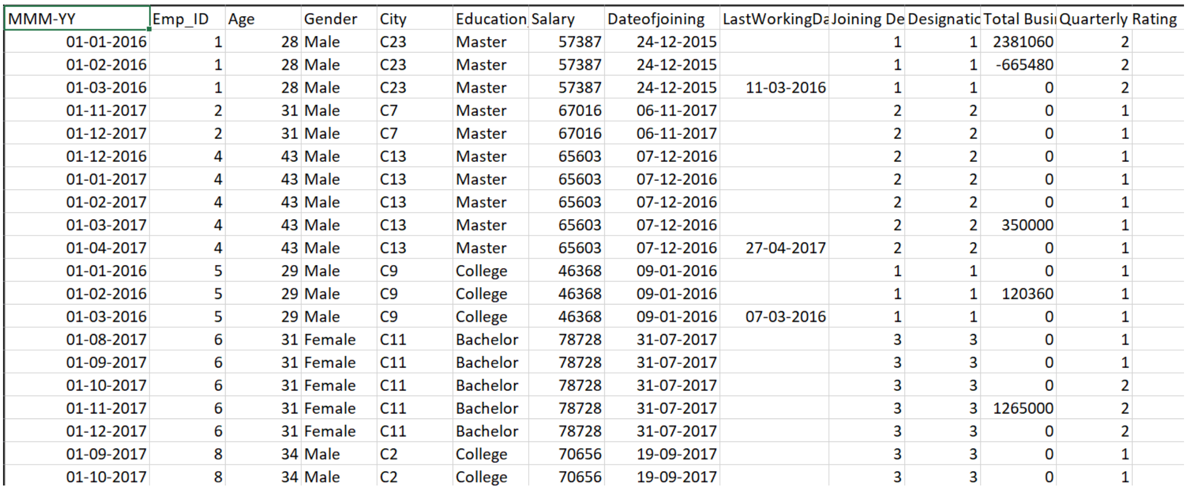
Approach

The problem statement is to predict the employee attrition of the insurance company.

The train dataset has the features while the test dataset just contains the Emp_ID. Now, since we have to predict whether the employee will leave the organization or not, it is a binary classification problem. The snapshot of the data given:



MMM-YY	Emp_ID	Age	Gender	City	Education	Salary	Dateofjoining	LastWorkingDa	Joining De	Designatic	Total Busi	Quarterly Rating
01-01-2016	1	28	Male	C23	Master	57387	24-12-2015		1	1	2381060	2
01-02-2016	1	28	Male	C23	Master	57387	24-12-2015		1	1	-665480	2
01-03-2016	1	28	Male	C23	Master	57387	24-12-2015	11-03-2016	1	1	0	2
01-11-2017	2	31	Male	C7	Master	67016	06-11-2017		2	2	0	1
01-12-2017	2	31	Male	C7	Master	67016	06-11-2017		2	2	0	1
01-12-2016	4	43	Male	C13	Master	65603	07-12-2016		2	2	0	1
01-01-2017	4	43	Male	C13	Master	65603	07-12-2016		2	2	0	1
01-02-2017	4	43	Male	C13	Master	65603	07-12-2016		2	2	0	1
01-03-2017	4	43	Male	C13	Master	65603	07-12-2016		2	2	350000	1
01-04-2017	4	43	Male	C13	Master	65603	07-12-2016	27-04-2017	2	2	0	1
01-01-2016	5	29	Male	C9	College	46368	09-01-2016		1	1	0	1
01-02-2016	5	29	Male	C9	College	46368	09-01-2016		1	1	120360	1
01-03-2016	5	29	Male	C9	College	46368	09-01-2016	07-03-2016	1	1	0	1
01-08-2017	6	31	Female	C11	Bachelor	78728	31-07-2017		3	3	0	1
01-09-2017	6	31	Female	C11	Bachelor	78728	31-07-2017		3	3	0	1
01-10-2017	6	31	Female	C11	Bachelor	78728	31-07-2017		3	3	0	2
01-11-2017	6	31	Female	C11	Bachelor	78728	31-07-2017		3	3	1265000	2
01-12-2017	6	31	Female	C11	Bachelor	78728	31-07-2017		3	3	0	2
01-09-2017	8	34	Male	C2	College	70656	19-09-2017		3	3	0	1
01-10-2017	8	34	Male	C2	College	70656	19-09-2017		3	3	0	1

Next, we want a dataset which has the feature variables and the target variable. The target variable is the variable which tells if the employee has left the organization or not. In the dataset given, there are a number of observations for each employee. The task was to have just one row for each employee.

Feature Engineering

In the new dataset, there were 13 features which are Emp_ID, Age, Gender, City, Education, Salary, Joining_Designation, Designation, Total_Business_Value, Last_Quarterly_Rating, Quarterly_Rating_Increased, Salary_Increased and finally the target.

First, we need to get the distinct Emp_ID and this would become the first column of the dataset we created. Next, we take the age of that employee the last it was reported. Gender and City were taken from the dataset given. Education and Salary were also taken the last time it was reported. Joining Designation is taken as it is from the dataset. Designation is the designation of the employee at the last time it was reported. Total Business Value is the sum of the Total Business Value acquired by the employee. Last_Quarterly_Rating is the rating the employee was given the last time it was reported. Quarterly_Rating_Increased is assigned 1 if the last quarterly rating is greater than the first rating assigned and 0 otherwise. Salary_Increased is assigned 1 if the last Monthly Salary is greater than the first Monthly Salary and otherwise 0. Target was assigned 1 if the employee had a non-null last working date and 0 otherwise.

Model Building

Now, before building the model, the categorical feature 'Gender' was label encoded and all those features which had more than 2 categories, were one hot encoded. All the numerical features were scaled using MinMaxScaler. The features which were encoded and scaled were dropped as well as the Emp_ID is dropped before building the model.Random Forest is applied on the dataset created.

We search for the parameter values like ‘n_estimators’ and ‘max_depth’ which gives the best f1_score using GridSearchCV. Random Forest with max_depth=10 and n_estimators=250 was built.

Model Selection

Before finalizing on Random Forest; few classification models like KNN, SVM, XGBoost and GradientBoost were also applied on the dataset. XGBoost led to overfitting the data. SVM, Gradient Boost and Random Forest performed well on the data. Since Random Forest gave a good f1-score, this model was selected to predict the employee attrition.

---The End----------------------------