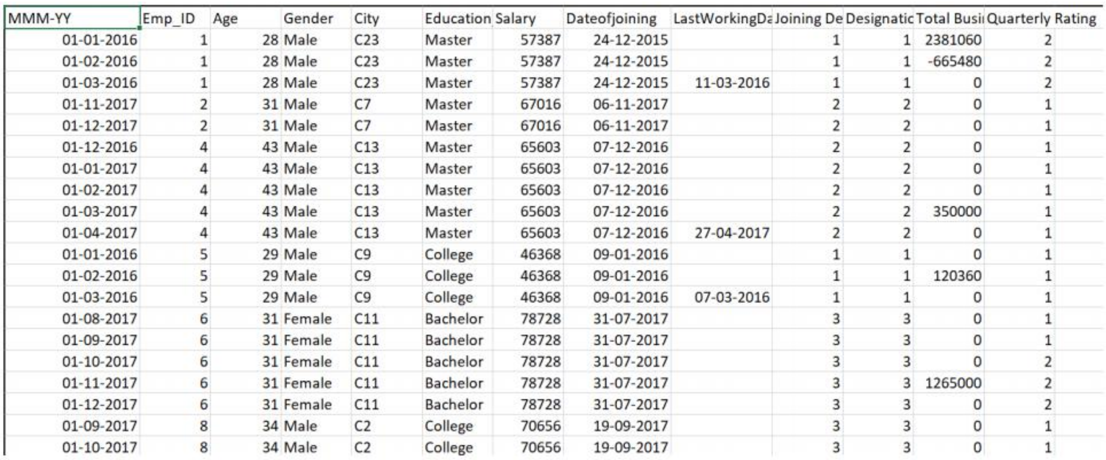
Approach

Problem Statement:

You are working as a data scientist with HR Department of a large insurance company focused on sales team attrition. Insurance sales teams help insurance companies generate new business by contacting potential customers and selling one or more types of insurance. The department generally sees high attrition and thus staffing becomes a crucial aspect.

To aid staffing, you are provided with the monthly information for a segment of employees for 2016 and 2017 and tasked to predict whether a current employee will be leaving the organization in the upcoming two quarters (01 Jan 2018 - 01 July 2018) or not, given:

1. Demographics of the employee (city, age, gender etc.)
2. Tenure information (joining date, Last Date)
3. Historical data regarding the performance of the employee (Quarterly rating, Monthly business acquired, designation, salary)



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

As the objective was to predict if an employee will leave the organization in the upcoming two quarters, the target variable was taken such that if an employee leaves the organization within 180 days of review it was taken was 1 and 0 otherwise i.e., if the last working day is 25-11-2017 and a review was conducted on 01-05-2017(208 days prior), target would be 0 and for the next review conducted on 01-06-2017(177 days prior), the target would be 1. The training data was taken only till 01-08-2017 as a full 180 days was required for prediction. The predictions had to be done at review level for each employee otherwise there would not be sufficient data and the changes in employee performance /behaviour might be difficult to catch if data was minimized to one row per employee.

Data Pre-Processing/Feature Engineering:

In the dataset, there are 13 features which are Emp_ID, Reporting Date, Age, Gender,City,Education,Salary,DateofJoining,LastWorkingDate,Joining_Designation, Designation, Total_Business_Value, Quarterly_Rating.

First step in Building a Model is to understand the Data-Set, and after understanding I came to know that, there are '2200' Duplicate values present in the 'Emp_ID' column (primary key). After that I'd Drop all the Duplicate values on the basis of last 'Reporting Date', and we get the Distinct 'Emp_ID' column.

The Next step is that the target variable is not specifically mentioned in the train data. For constructing the target variable as shown in the definition, one should first look at the 'LastWorkingDate' column. Wherever the column has null values, that means the employee is continuing his/her work at the organization at least in the next year. Wherever any date record is appearing, that means the employee has left the organization on that particular date. So as per definition, we will put 0 where 'LastWorkingDate' column is null and 1 where 'LastWorkingDate' column has a date.

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. Quarterly_Rating is the rating the employee was given the last time it was reported.

Model Building:

Now, before building the model, the categorical feature 'Gender', 'Education-Level', 'City', 'Quarterly Rating' was One-hot encoded. All the numerical features were scaled using StandardScalar. Then search for the parameter values like 'n-estimators' and 'max-depth' which gives the best f1-score using GridSearchCV.

Model Selection:

Before finalizing on Decision Tree; few classification models like LogisticRegression, 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 Decision Tree gave a good f1-score = 0.6966, this model was selected to predict the employee attrition.