Machine learning assignment

Topic: Data Pre-processing of Employee attrition data

The data has been taken from IBM Employee HR Attrition Kaggle.  
The main Business problem that is being solved here is how a system can be created to help big companies control their attrition by understanding which employee could leave so as to provide him/her some incentives to stay back.

There are around 1470 rows of data and 35 columns

The dataset has following columns

Age,Attrition,BusinessTravel,DailyRate,Department,DistanceFromHome, Education,EducationField,EmployeeCount,EmployeeNumber,EnvironmentSatisfaction,Gender,HourlyRate,JobInvolvement,JobLevel,JobRole,JobSatisfaction, MaritalStatus,MonthlyIncome,MonthlyRate,NumCompaniesWorked,Over18, OverTime,PercentSalaryHike,PerformanceRating,RelationshipSatisfaction,

StandardHours, StockOptionLevel,TotalWorkingYears,TrainingTimesLastYear,

WorkLifeBalance, YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,

YearsWithCurrManager

* Identifying the Dependent variable i.e. Attrition in our dataset and trying to drop all the

Columns which does not affect the dependent variable

* The df.info() returned me the information regarding the number of null data but this dataset doesn’t contain any null so there is no need of dealing with such data. If we had few (100-200) null data we can just drop them.
* Value\_counts()computes a histogram of the unique values in that column. It returns a Pandas series that contains the count of each unique value in the specified column, ordered by count in descending order. Here we tried on few columns like"StandardHours","Over18",”EmployeeCount" where we could infer that most of them just had one category and hence it is not effecting the Attrition variable. Therefore we dropped all these columns. Since “JobRole” can be derived from “Department” and the

“EmployeeNumber” is unique for every row of data we can drop these two columns.

* Renamed the categories of “BusinessTravel” column into numeric values i.e (0,1,2)
* conversion of categorical columns into indicator variables so we can replace these with the actual ones using the pandas get\_dummies and concat functions
* Replacing the values of Attrition Yes, No to 0,1.
* A plot of age of employees to get a age-group