ML Assignment - 2

VARSHA S
193079005

Observations made from the dataset.

1) It is seen that the ratio of men and women leaving the company is almost the same i.e. leaving the company isn't gender biased.
2) Around 35% of the employees who voted '2', 26% who voted '1', 14% who voted '3' and 13% who voted '4' as EnvironmentSatisfaction have left the company.
3) More employees who have done overtime work have left the company.
4) Around 19% of the employees who have previous work experience in one company have left the company. This is the highest percentage compared to employees with other years of previous experience.
5) The percentage of employees who travel frequently have quit the job, is greater than the other two categories in the Travel field, which is around 24%.
6) Around 20% of employees who worked in Sales have quit the job, which is greater than the other two categories in the Department field.
7) The number of employees with Marital status as Single who have quit the company is higher than the other two categories in the MaritalStatus field.

Preprocessing methods used and the reason for their usage.

1) Label encoding - This is done to convert all the string literals to numerical values, basically a machine-readable form.
2) Dropped the columns - `'EmployeeNumber'`,`'EmployeeCount'`,`'ID'`.These columns basically mention the identity of the employee and aren't the features which contribute to our classification problem.
3) Normalisation of data - Since the ranges of data values are different, normalisation is needed to bring the data to a common scale(0-1). **Though I observed that the model performed better without the normalisation. Maybe one explanation is that the difference in the ranges is not very large or that the data is proportional so normalising doesn't provide the right estimators.**

List of various approaches used from the start of the competition till final approach for best accuracy

1) Using a varied array of classifiers as we all know that all classifiers won't work well for all sorts of classification problems.
2) Tried different labels while enforcing label encoding by mapping to different numbers, using **`applymap`**(), depending on how the feature influences the classification(as observed in the data analysis).
3) Tried One-Hot Encoding to label the string values instead of a random number distribution to the features with string values i.e categorical data.
4) Tried using a **`MultiColumnLabelEncoder`** class to perform label encoding which used the class **`sklearn.preprocessing.LabelEncoder.`**
5) Performed cross-validation to find the best model to test.
6) Plotted Confusion Matrix and calculated accuracy score to find the best classifier.

Results

1) With the GradientBooster() classifier, with applymap() function, I got a score of **0.909**.(the numerical labels were given considering the observations made from the dataset, higher value to the high influencing parameter)
2) With AdaBoost() classifier, with LabelEncoder to convert categorical data to numerical data, I got a score of **0.904**.

Final learning achieved through this competition

1) I learnt to handle pandas libraries to create and handle data.
2) I got to know about confusion matrices and its usage in classification problems.
3) I had an hands-on k-fold cross validation learning experience.
4) I got to know about the various classifier functions which have varied applications.
5) I learnt briefly about data analysis from csv files.
6) I learnt about One-Hot encoding technique to convert categorical data into numeric data.