Assignment – 2

The objective of the assignment is to predict attritions on the basis of features like age, education and many other. As training dataset is labelled, it is a classification problem of supervised learning. Now, the key to this is the proper study and research of input data (do some data pre-processing) followed by fitting some model that works well both on the training data set and test dataset sufficiently well. The key thing is to note that the model should not over-fit. Various models can be used in this regard. What we have done is that we used some very common classification techniques like logistic regression, Random forest classifier and Gradient Boosting classifier to predict attrition and tried to achieve maximum accuracy.

A) Exploratory Data Analysis:

1) It can be seen that frequency of 1 in attrition is 172 and frequency of 0 in attrition is 856. So, there are chances that given dataset is imbalanced.

Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error. But for imbalanced dataset, accuracy might not be the best option for measuring performance.

There are so many methods to deal with imbalanced dataset. Here I have listed a few methods.

a) Change the performance metric

b) Change the algorithm

c) Oversampling minority class

d) Under sampling majority class

e) Generate synthetic samples

2) Features given in the data set are of two types. There are 7 categorical features and 27 numerical features.

3) There are no null values in dataset.

4) To gain insights of data mean, min, max, standard deviation of each numerical feature can be checked.

5) To label categorical features, number of unique values in in each feature can be checked.

6) To gain more insights of data correlations between features can be observed.

7) To gain more insights histograms, boxplots, count plots can be plotted.

B) Pre-processing:

1) There are no null values in dataset. So, no need to do imputations.

2) As 7 features are object datatypes, they need to be labelled for mathematical calculations. So, these features are labelled by using label encoder.

For e.g. gender has two object values as male and female. So, they will be encoded as male= 0 and female=1.

3) As there are 53 unique values in age, these values can be put in classes.

For e.g. employees whose age is in between 20 to 30 can be put in one class.

4) By observing irrelevant features can be dropped.

For e.g. Employee number, employee count, ID are irrelevant features. So, these features can be dropped.

5) Some features have good correlation with each other, their combinations are used.

6) Feature scaling is done by using StandardScalar tool from sklearn library.

C) List of approaches:

1) Logistic regression:

First, I split the training dataset in two groups as training and validation. 20 percent of the data was separated. Then I applied logistic regression model on 80 percent of training dataset. Here I tried to maximize area under ROC curve by changing parameters of logistic regression. For e.g. By using grid search we can find best parameters and check accuracy for those parameters.

After that I did feature scaling.

After that I did some feature extractions based on the correlations.

Accuracy on training data is 0.88383 and on validation data is 0.871.

Eventually, I got maximum accuracy 0.864 around on test data.

2) Random forest classifier:

First, I split the training dataset in two groups as training and validation in the ration as 80:20. Then I applied Random forest classifier model on dataset. Here I tried to maximize area under ROC curve by changing parameters of random forest classifier. For e.g. By using grid search we can find best parameters and check accuracy for those parameters.

After that I did feature scaling.

After that I did some feature extractions based on the correlations.

Accuracy on training data is 0.9878 and on validation data is 0.8834.

Ultimately, I got maximum accuracy around 0.89393.

Final hyperparameters after using RandomSearchCV:

N_estimators = 800,

Min_samples_split = 10,

Min_samples_leaf = 2,

Max_features = ‘sqrt’,

Max_depth = 30

3) Gradient boosting classifier:

First, I split the training dataset in two groups as training and validation. 20 percent of the data was separated. Then I applied Gradient boost classifier model on 80 percent of training dataset. Here I tried to maximize area under ROC curve by changing parameters of gradient boost classifier. For e.g. By using grid search we can find best parameters and check accuracy for those parameters.

After that I did feature scaling.

After that I did some feature extractions based on the correlations.

By using RandomSearchCV, I got following hyperparameters:

Learning rate=0.48

N estimators=100

Max depth=15

Min samples split=5

Min samples leaf=4

subsample=0.58

max features=’sqrt’

Accuracy on training data is 0.9245 and on validation data is 0.8834.

By using these hyperparameters, I got maximum accuracy as 0.90909.

Results and final learnings:

Model	Accuracy (Training data)	Accuracy (Validation data)	Accuracy (Testing data)
Logistic Regression	0.8838	0.871	0.864
Random Forest Classifier	0.9878	0.8834	0.89893
Gradient Boosting Classifier	0.9245	0.8834	0.90909

- On training dataset, accuracy is like this:

Random Forest Classifier > Gradient Boosting > Logistic Regression

- On validation dataset, accuracy is like this:

Random Forest Classifier = Gradient Boosting > Logistic Regression

- On testing dataset, accuracy is like this:

Gradient Boosting Classifier > Random Forest Classifier > Logistic Regression

- From above results, it can be concluded that Gradient Boosting Classifier gave the highest accuracy on test dataset. Hence, this model is accepted for prediction purpose.

- As I mentioned earlier, dataset looks imbalanced. But it didn't matter much in this case. This dataset can be considered as balanced dataset.

- To get maximum accuracy, you will have to try many algorithms and check.

- For feature extraction, available libraries can be used otherwise you will have to use hit and trail method.