IBM HR Analytics Employee Attrition & Performance

Name : Ajay Nirmal

1 Introduction

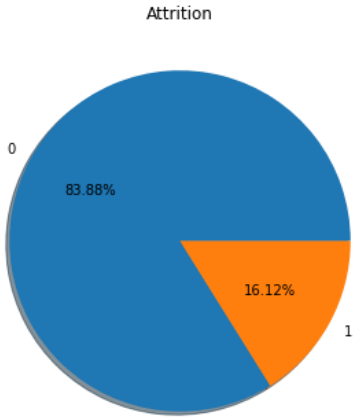
The aim of the project is to predict the attrition of an HR Analytics Employee. The project includes building and training of a model capable of predicting the attrition label using the fictional dataset provided by IBM.

2 Data Analysis

2.1 Dataset Used

The dataset contains employee attrition data created by IBM.

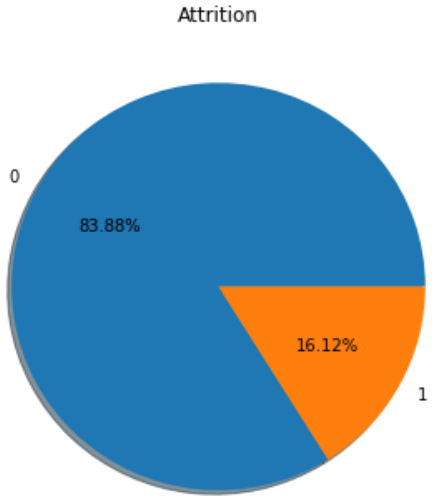
- The dataset contains 35 columns including the target label i.e Attrition
- Each column contains 1470 entries.
- The dataset includes numerical (14 features) as well as categorical (20 features) data.
- It is seen that there is data imbalance in the target label: Attrition.
    - Yes(1): 237
    - No(0): 1233



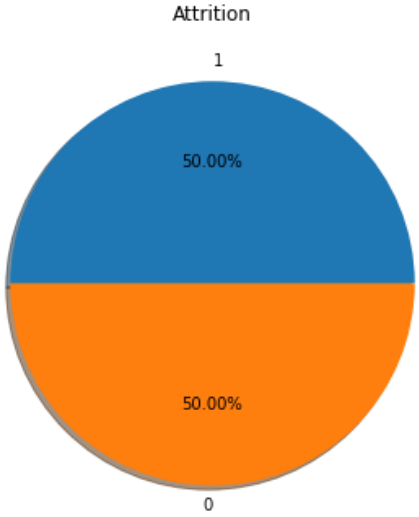
- Plots for all the numerical as well as categorical features are there in the notebook.
- Features vs Attrition graphs have also been plotted to draw some relevant inferences.

2.2 Handling data imbalance

As mentioned earlier, the data in target label: Attrition is highly imbalanced. That will affect the accuracy and evaluation metrics of the model. Thus, that data was re-sampled to remove the data imbalance. The graphs for the same can be seen below:



Before resampling



After resampling

3 Approach

3.1 Goals

The aim is to predict the attrition of an employee based upon IBM data available to us.

3.2 Tasks

- Removing data imbalance
- Splitting the dataset into test and train set
- Pre-processing the dataset for further analysis
- Trying various models and find the best one for Attrition prediction
- Training the final
- Tuning the parameters of the final model
- Prediction made on the test data

4 Experiments

4.1 Data pre-processing

Pre-processing or cleaning of data leads to a more efficient model and higher accuracy.
• Removal of unnecessary features containing only single or 2 unique values such as 'Over18', 'EmployeeCount', StandardHors' and 'EmployeeNumber'
• Categorical Encoding
• Removal of strongly correlated features. The correlation heatmap is also there in the notebook.
• Scaling the numerical features.

4.2 Finding the best model

I tried to build nine models on this dataset. The comparison can be seen in the table below on the train data:

Models	Accuracy score	Precision	Recall	F1-score
Logistic Regression	0.82	0.82	0.82	0.82
KNeighbours	0.84	0.76	0.98	0.86
Random Forest	0.91	0.92	0.89	0.91
LDA	0.82	0.82	0.82	0.82
SVC	0.82	0.83	0.81	0.82
Decision Tree	0.71	0.74	0.66	0.70
AdaBoost Classifier	0.88	0.88	0.89	0.89
LGBM Classifier	0.92	0.92	0.92	0.92
Naïve Byes	0.77	0.75	0.81	0.78

As observed from the table above, LGBM Classifier model performs the best.

5 Overall Results

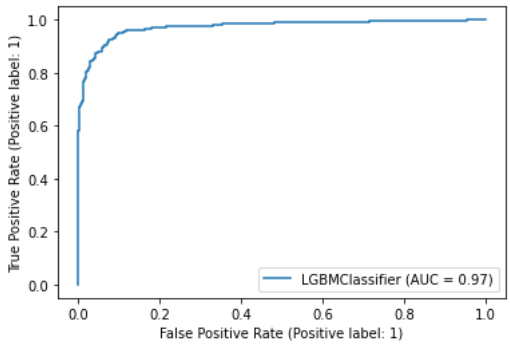
5.1 Final Model: LGBM Classifier

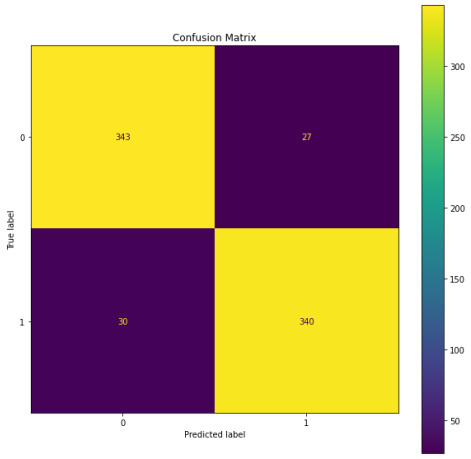
In our Attrition prediction job, LGBM Classifier outperformed all other models, attaining the greatest accuracy, precision, and F1 scores.

5.2 Final results on test data

The parameters for the LGBM Classifier are tuned further to optimize the results. The results are as follows:

```
Train Accuracy : 1.00
Test Accuracy : 0.92
```





5 Conclusion

The LGBM Classifier worked the best and gave appropriate results during prediction of the Attrition label. It outperformed all the model giving good accuracy. The imbalance in the data was removed to improve the results. At the end, the model gives 100% accuracy on the train data and 92% accuracy on test data.