Assignment-2 Report

Name - Saurabh Parekh

Roll number - 170100016

Observations:-

1. Mean of the TotalWorkingYears attribute for the employees who stayed is more than the employees who left the company. This shows that employees who worked for more years in the company would want to stay and those who recently joined the company have higher chances of leaving the company
2. Mean of the MonthlyIncome attribute for the employees who stayed is more than the employees who left the company. This shows that employees who have higher monthly income have higher chances of staying in the company.
3. Mean of the YearsSinceLastPromotion attribute for the employees who stayed is almost equal to the employees who left the company. This shows that this attribute doesn't contribute much to deciding whether an employee will leave the company or not.
4. Mean of the YearsWithCurrManager attribute for the employees who stayed is more than the employees who left the company. This shows that employees who have worked for more years with the current manager have higher chances of staying in the company.
5. Mean of the YearsInCurrentRole attribute for the employees who stayed is more than the employees who left the company. This shows that employees who have worked in the current role for more years have higher chances of staying in the company.
6. Features like Gender, EmployeeCount, EmployeeNumber, HourlyRate, ID, MonthlyRate, DailyRate, RelationshipSatisfaction, PerformanceRating, Education and TrainingTimesLastYear contribute very less to deciding whether the employee will leave the company or not. Hence these features are removed from the training dataset as well as testing dataset.
7. Employees who travelled frequency have higher chances of leaving the company than those who travel rarely. Employees who don't travel have least chances of leaving the company.
8. Employees from the Sales department have higher chances of leaving the company than Research and Development and Human Resources departments.

9. Employees who have a technical degree have the highest chances of leaving the company where those who have their education field in Human Resources have the least chance of leaving the company.
10. Employees having a job role of Sales Representative have highest chances of leaving the company where those who are Research Director have the least chances of leaving the company. This clearly shows that JobRole feature contributes a lot to deciding whether the employee will leave the company or not.
11. Employees who are single have relatively very high chances of leaving the company than those who are married or divorced.
12. Employees who work overtime have much higher chances of leaving the company than those who don’t.
13. Employees whose Stock Option Level is 0 have very high chances of leaving the company than those whose Stock Option Level is 1, 2 or 3.
14. Employees whose Job Level is 1 have very high chances of leaving the company than those whose Job Level is 2, 3, 4 or 5.
15. TotalWorkingYears, JobLevel, Age, MonthlyIncome, StockOptionLevel, YearsInCurrentRole, JobInvolvement, YearsWithCurrManager, YearsAtCompany and EnvironmentSatisfaction are the major features which contribute to deciding whether an employee will leave the company or not.

Preprocessing Methods Used:-

1. Plotted the barplots and density plots for each of the features and analyzed which features would contribute to the attrition target column to improve accuracy of the classification model.
2. Used feature engineering to remove the unwanted features from the dataset in order to prevent overfitting of the training dataset.
3. Identified categorical features and converted them into numerical form using one-hot encoding method because the machine learning model takes only data in numerical form for training.
4. Tried out standardization to improve the numerical stability of the model and decrease the training time required. But it isn’t good in our case.

Various Approaches Used:-

1. Gradient Boosting Classification
2. k-Nearest Neighbours Classification
3. Logistic Regression

4. Random Forest Classification
5. SVM Classification using different kernels

Results and Final Learning:-

1. I learned how to use various other libraries like scikit-learn for data analysis and data mining, numpy for manipulation multi-dimensional arrays and matrices, pandas for data analysis by converting the data to pandas Dataframe, matplotlib for plotting graphs and visualising the data and seaborn which is based on matplotlib.
2. I also learned how to use different machine learning models and evaluating them using various evaluation metrics like Classification Accuracy, Logarithmic Loss, Confusion Matrix, Area Under Curve, F1 Score, Mean Absolute Error, Mean Squared Error, etc.
3. I also learned what Exploratory Data Analysis(EDA) is and how important it is to do EDA before directly proceeding to training the model.
4. I also learned a lot about feature engineering and why it is so important to filter out the useless features for classification.
5. I also got familiar with jupyter notebooks and got to know how useful they are.
6. I also learned about fine-tuning hyper-parameters and finding out the best hyper-parameters for training the model to get the highest accuracy.