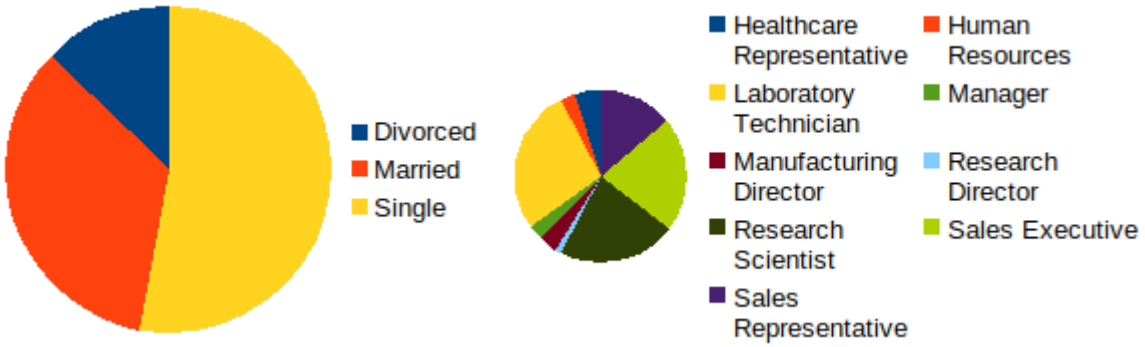
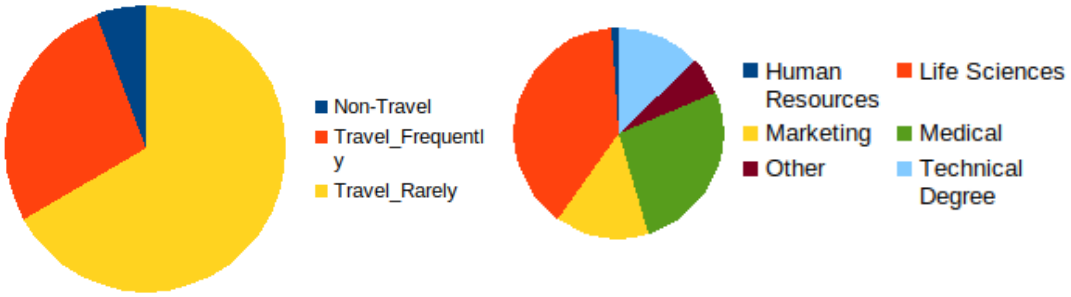
Assignment_2
Prashant Sharma
183079037

Observations you made of the dataset provided for this competition

1. A total of 33 features are provided in train.csv out of which no info is conveyed from “EmployeeCount” and “ID”, so 31 features vector.
2. 7 are string type and 24 are numeric type.
3. Total of 1028 observation is provided (Active and Ex-Employee data)
4. This is the histogram chart of numeric data
5. Image
6. In the training data set Out of 1028, 856(83.3%) are active employees and 172(16.7%) are Ex-Employee.
7. It’s an imbalance class, so while classification we need to set hyperparameters accordingly.
8. Parameters like “DistanceFromHome”, “NumCompaniesWorked” have the most positive correlation.
9. Parameters like “TotalWorkingYears”, ”JobLevel” have the most negative correlation.
10. Specific feature detail -
    a. 52% Single employees
    b. About 16.2% of leavers after 1-year of work.
    c. 29% of leavers have distance from home of 1,2,3. Quite strange
    d. 66% of leavers travel rarely
    e. 55% of leaver work overtime.
    f. 27% of leavers are Laboratory Technician
    g. 38% of leavers are from life science
    h. 41% of leavers have worked for 1 company





Above pi-charts shows portion of Ex-Employee/leavers (172).

What all preprocessing methods you used and why

1. Encoding, since Machine Learning algorithms can typically only have numerical values as their predictor variables. Hence Label Encoding becomes necessary as they encode categorical labels with numerical values. To avoid introducing feature importance for categorical features with large numbers of unique values, we will use both Lable Encoding and One-Hot Encoding
2. Feature Scaling, eature Scaling using MinMaxScaler essentially shrinks the range such that the range is now between 0 and n. Machine Learning algorithms perform better when input numerical variables fall within a similar scale. In this case, we are scaling between 0 and 5.
3. Splitting train.csv into training and testing sets, using training set we will develop model and using testing set we will improve validation error.

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

1. Logistic Regression, accuracy = 87.9%
2. Logistic Regression after fine tuned, accuracy = 88.3%
3. Random Forest, accuracy = 85.2%
4. SVM, accuracy = 85.2%

Results and Final learning you achieved through this competition.

1. Finally Logistic Regression was used.
2. Hyper parameter of logistic regression were fine tuned to achieve a score of **.90909** in kaggle competition.
3. Imbalance class need to be balanced for good accuracy.
4. AUC - ROC curve is a performance measurement for classification problem at various thresholds settings.
5. ROC is a probability curve and AUC represents degree of separability. It tells how much model is capable of distinguishing between classes.

Note – Use “183079037_LR_finetune.csv” for final score.