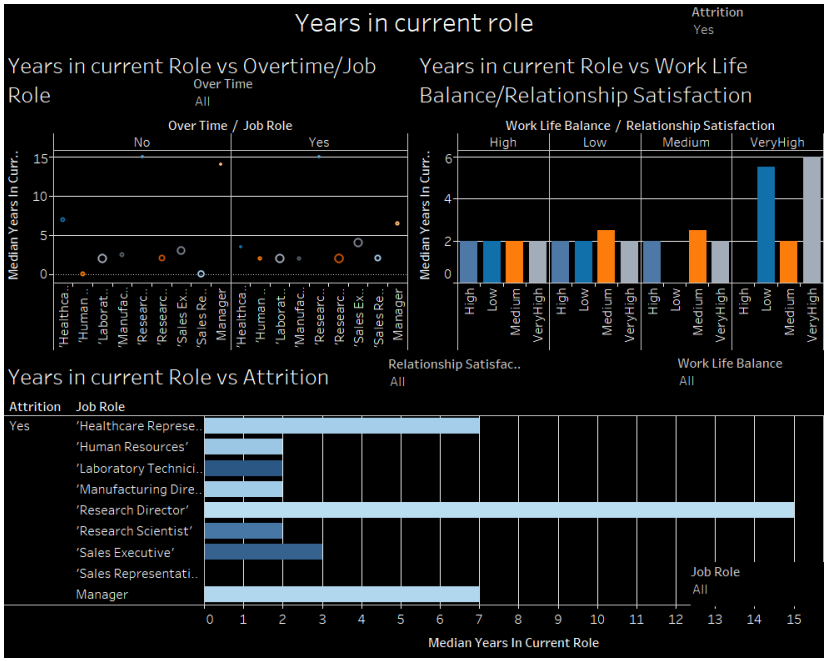
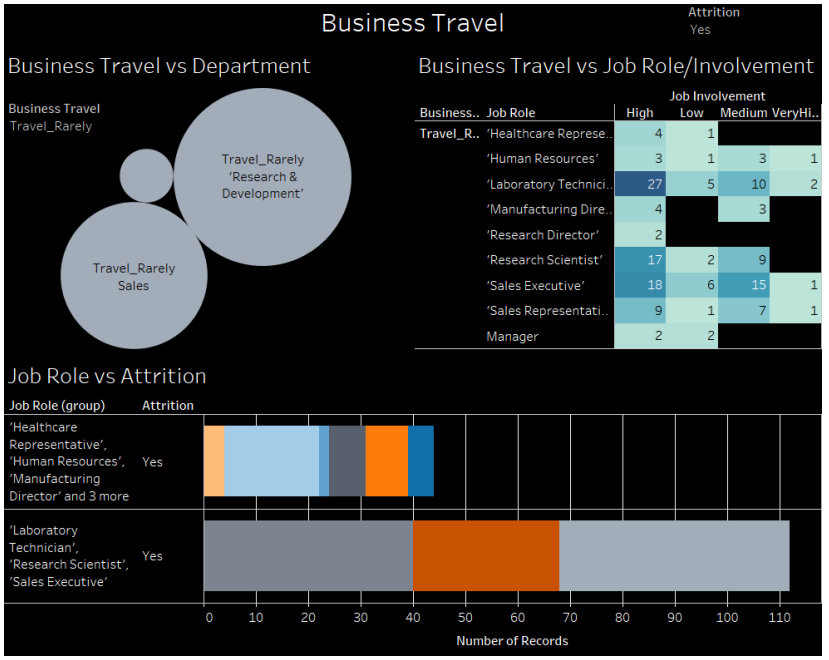
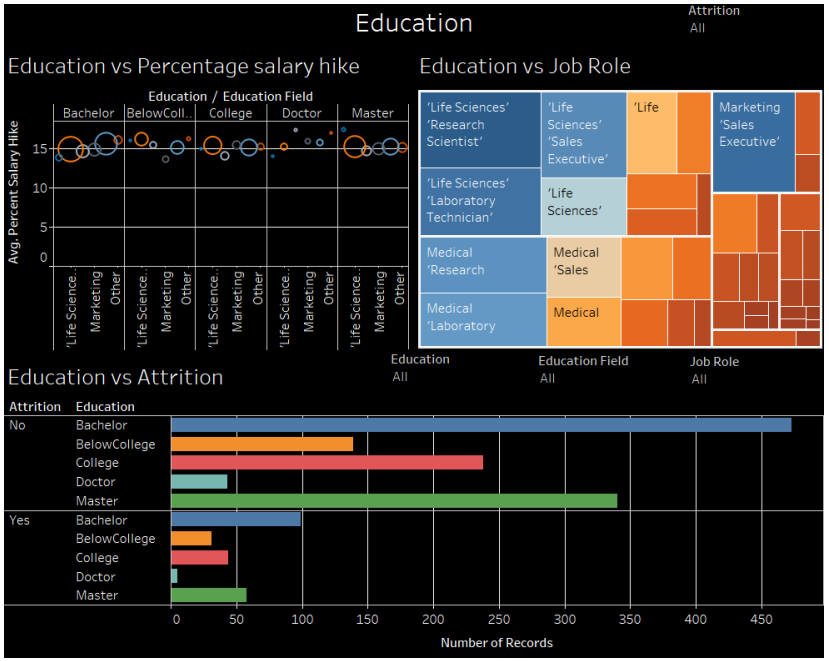
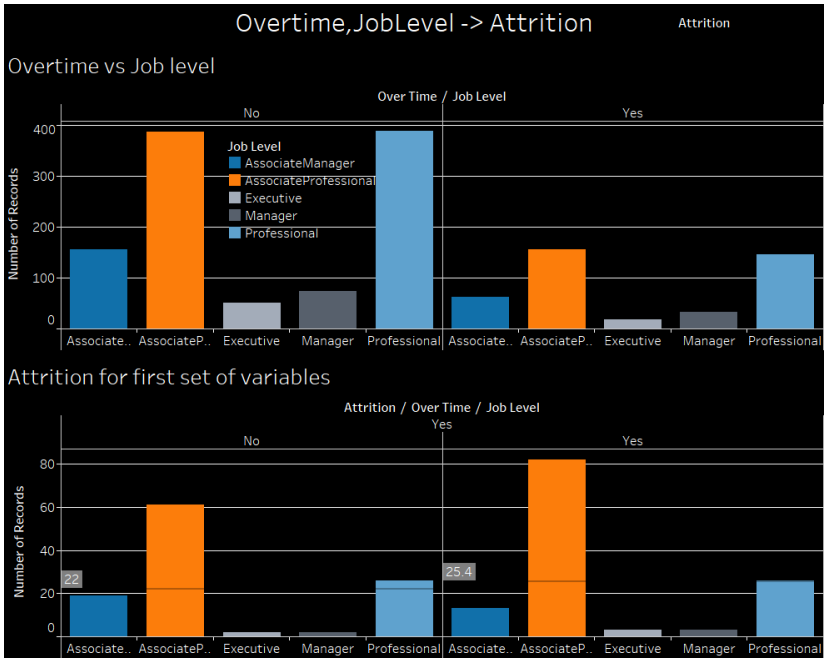
Poster – Part 1



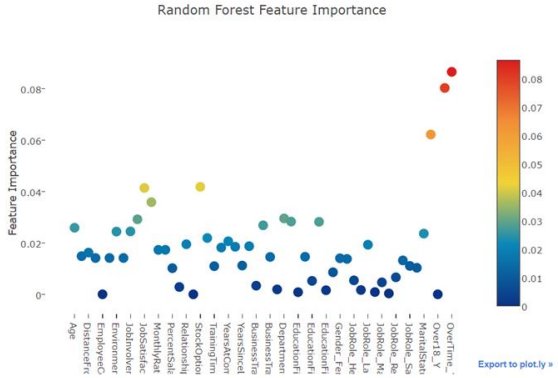




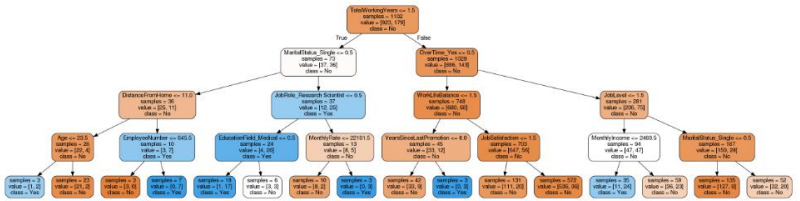
Poster – Part 2



Random Forest

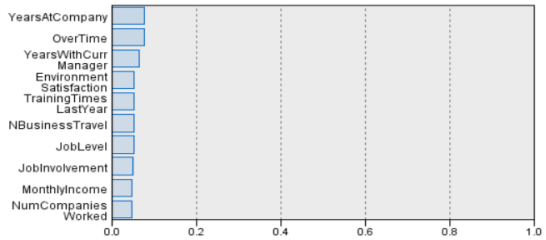


The accuracy of random forest classifier comes out to be 87.77% after feature engineering and oversampling was performed on the dataset. The outcomes of the feature importance done by random forest model states that OverTime and MaritalStatus are the two most important features to predict Attrition, which already have been selected as the set of relationship affecting Attrition in the previous dashboard.



Artificial Neural Network

Figure 2. Importance of Variables According ANN analysis



Artificail Neural Network was implemented on the dataset using Python's Keras library and TensorFlow as the backend. Without any data cleaning the important variables and confusion matrix by ANN was given. The accuracy came out to be 85.33%. To improve the accuracy, a number of steps were taken:
1. data cleaning
2. adding derived information
3. Using dummy variables and dummy traps
4. Hyperparameters
5. Training the neural network with K fold cross validation

The accuracy increased and came out out to be 97.5%

Figure 3. ANN Confusion Matrix

Comparing $N-Nattrition with Nattrition

'Partition'	1_Training		2_Testing	
Correct	922	90.39%	384	85.33%
Wrong	98	9.61%	66	14.67%
Total	1,020		450	

Coincidence Matrix for $N-Nattrition (rows show actuals)

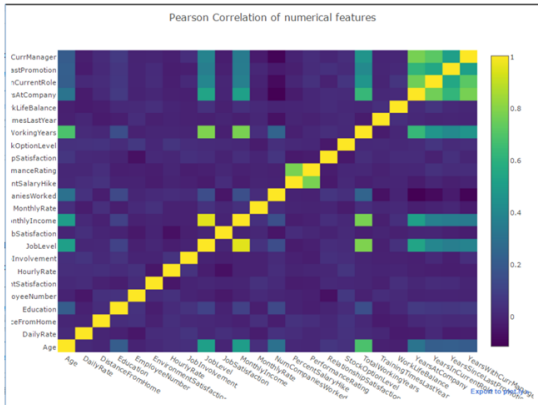
'Partition' = 1_Training	0.000000	1.000000
0.000000	852	19
1.000000	79	70
'Partition' = 2_Testing	**0.000000**	**1.000000**
0.000000	352	10
1.000000	56	32

```
print("Best accuracy: ",max)
```

```
Best accuracy:  0.974999979138
```

Statistical Modelling

	Attrition	p < 0.05
PerformanceRating	0.990075	False
Education	0.545525	False
Gender	0.290572	False
RelationshipSatisfaction	0.154972	False
TrainingTimesLastYear	0.0191477	True
EducationField	0.00677398	True
Department	0.00452561	True
WorkLifeBalance	0.00097257	True
JobSatisfaction	0.0005563	True
EnvironmentSatisfaction	5.12347e-05	True
BusinessTravel	5.60861e-06	True
JobInvolvement	2.86318e-06	True
MaritalStatus	9.45551e-11	True
StockOptionLevel	4.37939e-13	True
JobLevel	6.63468e-15	True
JobRole	2.75248e-15	True
OverTime	8.15842e-21	True



After performing EDA, the next logical step was to figure out which of the attributes have a statistically significant categorical relationship with Attrition, which is our target variable. To do this, Chi-square test was performed and the p-values of the attributes were obtained. Attributes having p-values > 0.05 had a non-significant categorical relationship and the others had a statistically significant categorical relationship

From te correlation plot, it was observed that quite a lot of columns were poorly correlated with one another. Generally, when making a predictive model, it would be preferable to train a model with features that are not too correalted with one another so that we do not have to deal with redundant features.