Predicting Employees Attrition

Shereef Bankole

Background Statement

- Attrition is a common problem in all businesses
- HR and top executive are very much concerned about keeping their employees
- It is highly imperative to unravel the factors that leads to employees' turnover and put in place a cost effect retentive plan
- Here predictive model was performed and evaluated on employee attrition using openly available IBM HR Data

Procedures

- Data loading
- Data inspection and cleaning
- Brief statistical analysis
- Exploration data analysis –generating several plots
- Data pre-processing-
  - Checking for sample imbalance
  - •Employing SMOTE and ADASYN
  - Turning categorical variables to numeric variables
- Model Development and Selection
- Model Evaluation

Data loading

Load Data

```
raw_data=pd.read_csv('IBM_HR_Employee_Attrition.csv')
raw_data.head()
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...

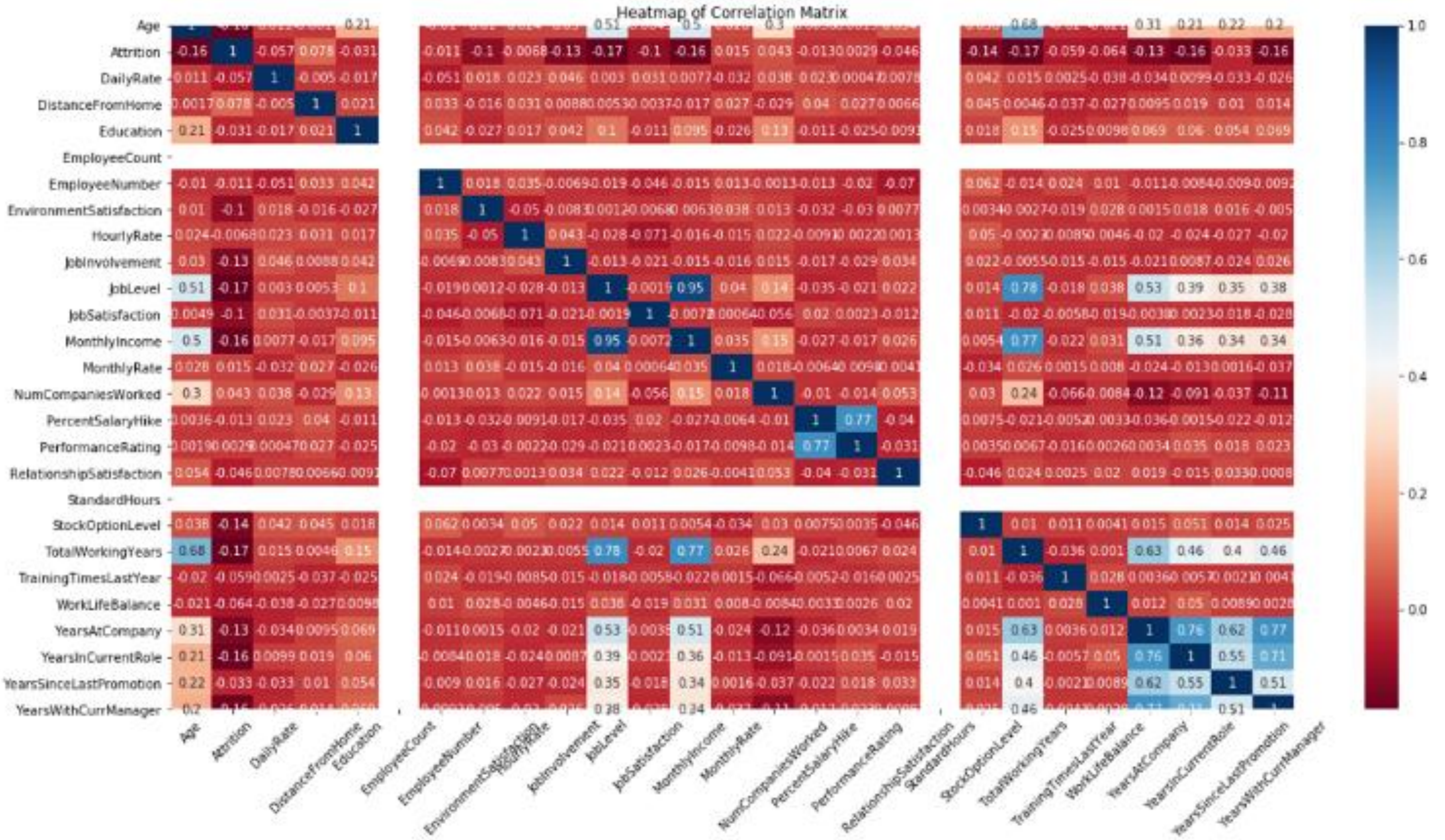
5 rows × 35 columns

Data inspection and cleaning

```
raw_data.info()
```

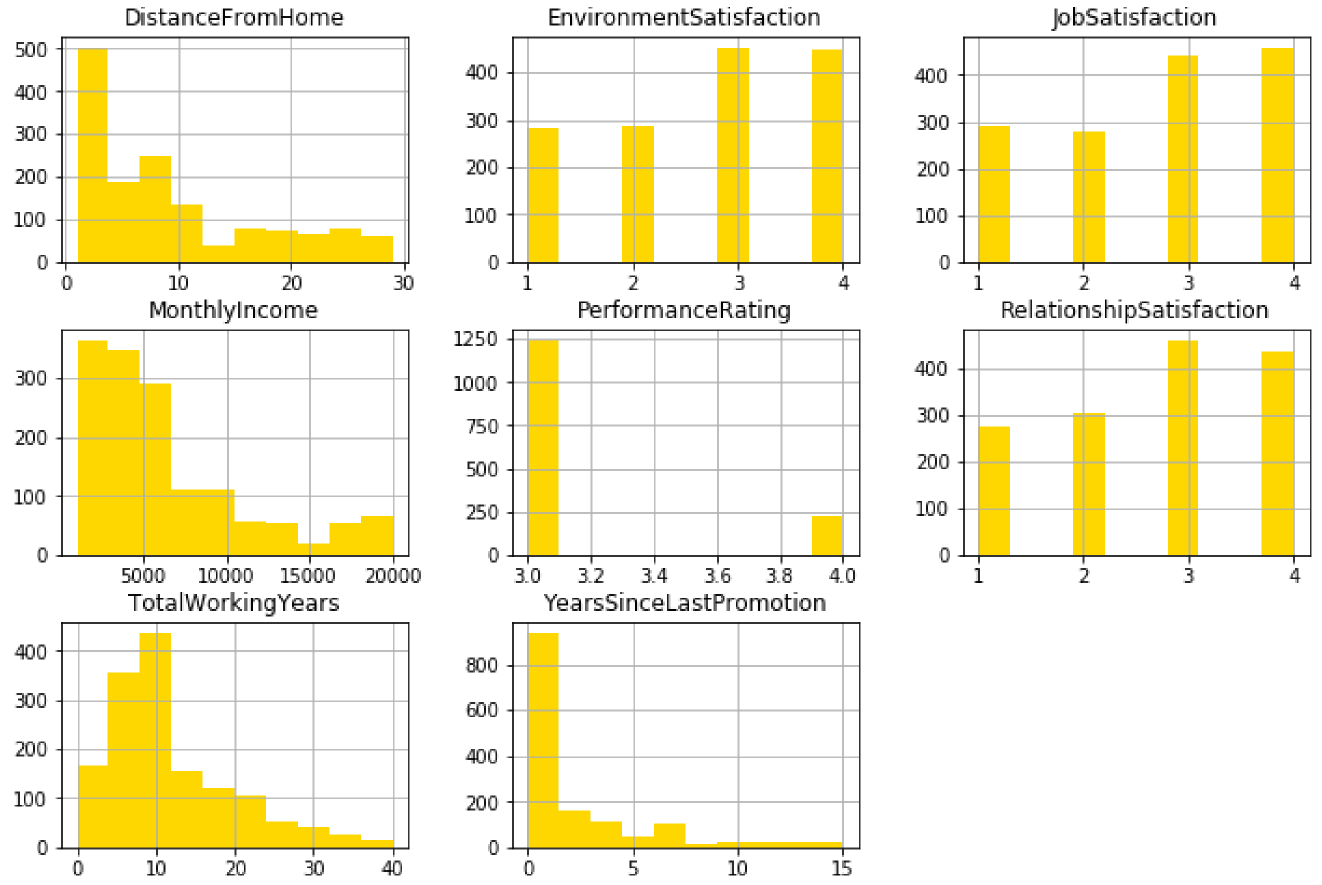
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
Age                           1470 non-null int64
Attrition                     1470 non-null object
BusinessTravel                1470 non-null object
DailyRate                     1470 non-null int64
Department                    1470 non-null object
DistanceFromHome              1470 non-null int64
Education                     1470 non-null int64
EducationField                1470 non-null object
EmployeeCount                 1470 non-null int64
EmployeeNumber                1470 non-null int64
EnvironmentSatisfaction       1470 non-null int64
Gender                        1470 non-null object
HourlyRate                    1470 non-null int64
JobInvolvement                1470 non-null int64
JobLevel                      1470 non-null int64
JobRole                       1470 non-null object
JobSatisfaction               1470 non-null int64
MaritalStatus                 1470 non-null object
MonthlyIncome                 1470 non-null int64
MonthlyRate                   1470 non-null int64
NumCompaniesWorked            1470 non-null int64
```

Brief statistical analysis



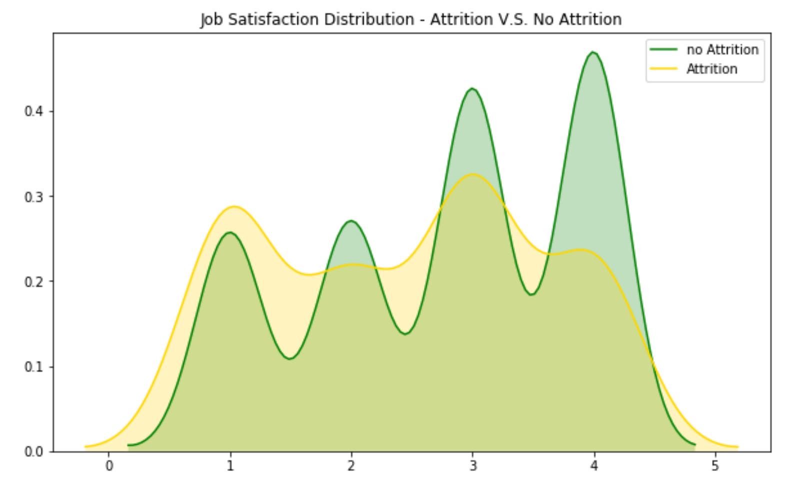
Exploration data analysis

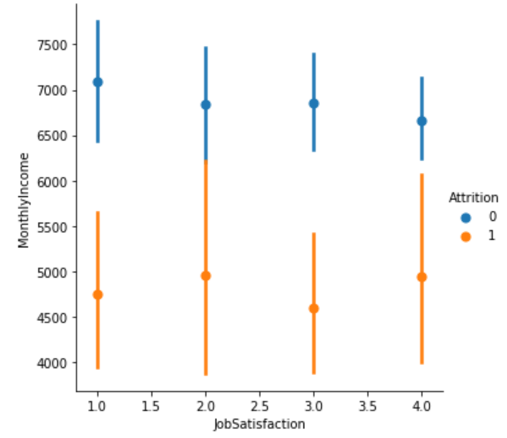
- EDA is key to performing modelling.
- More than half of the employees live 10 miles away from the office
- Two groups are identified based on environment satisfaction
- Job satisfaction and relationship follow similar trend as environment satisfaction.
- More than half of the employees earn less than $10000 monthly
- All the employee received good performance rating
- More than half have worked in the company for less than 10 years
- Many were promoted within the last two years

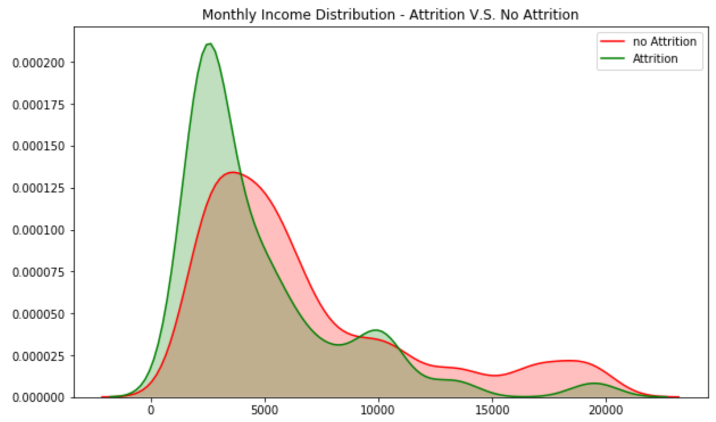


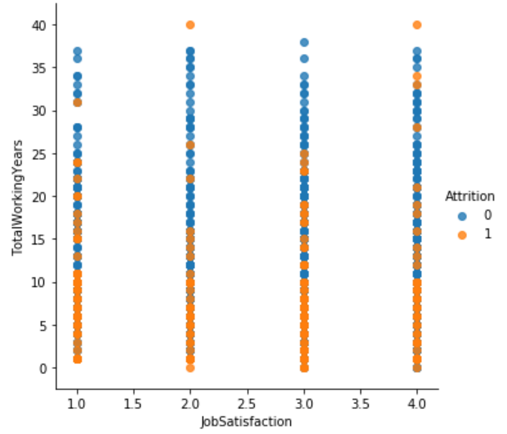
Exploration data analysis

- Employees attrition is not related to job satisfaction.
- Employee who have worked in the company for less than15 years are more likely to leave.
- Lower income earners are more susceptible to attrition



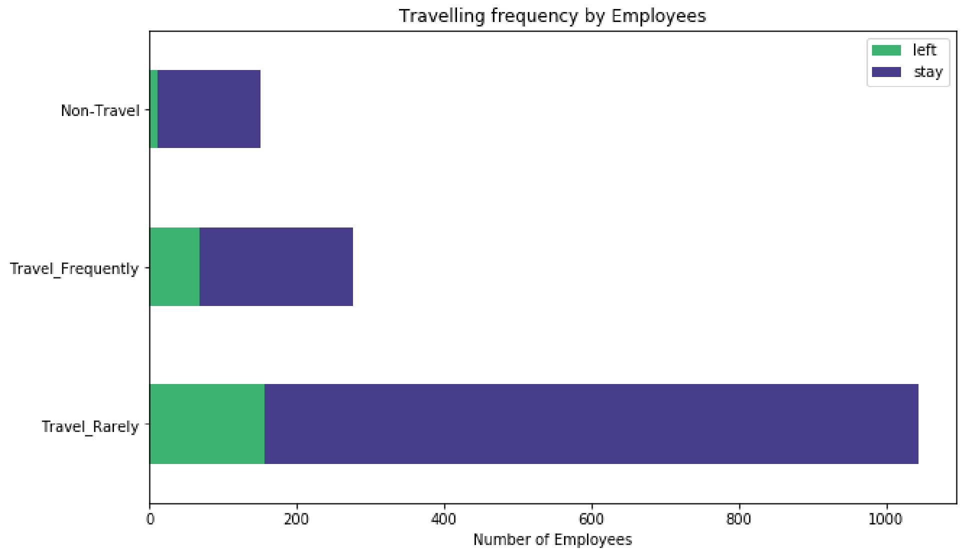


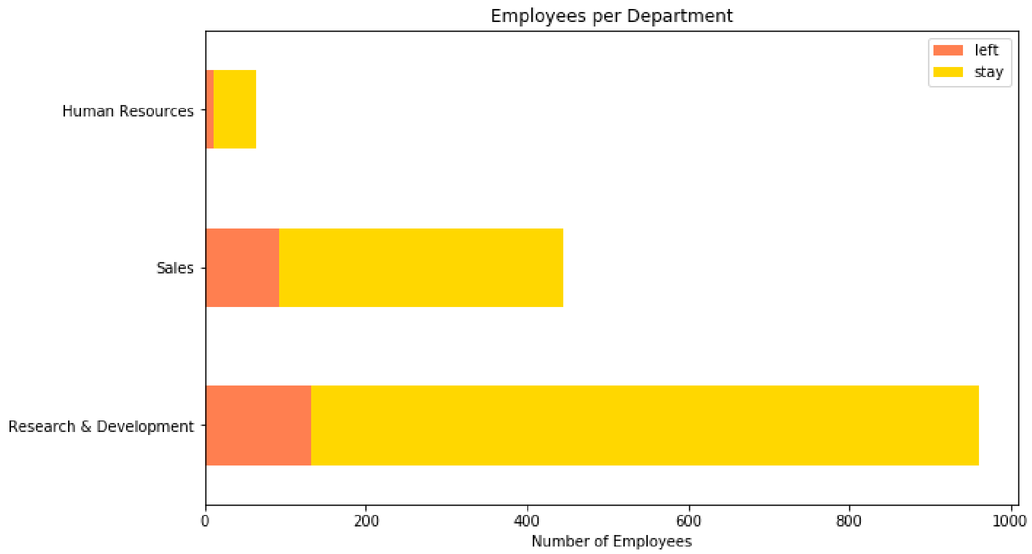




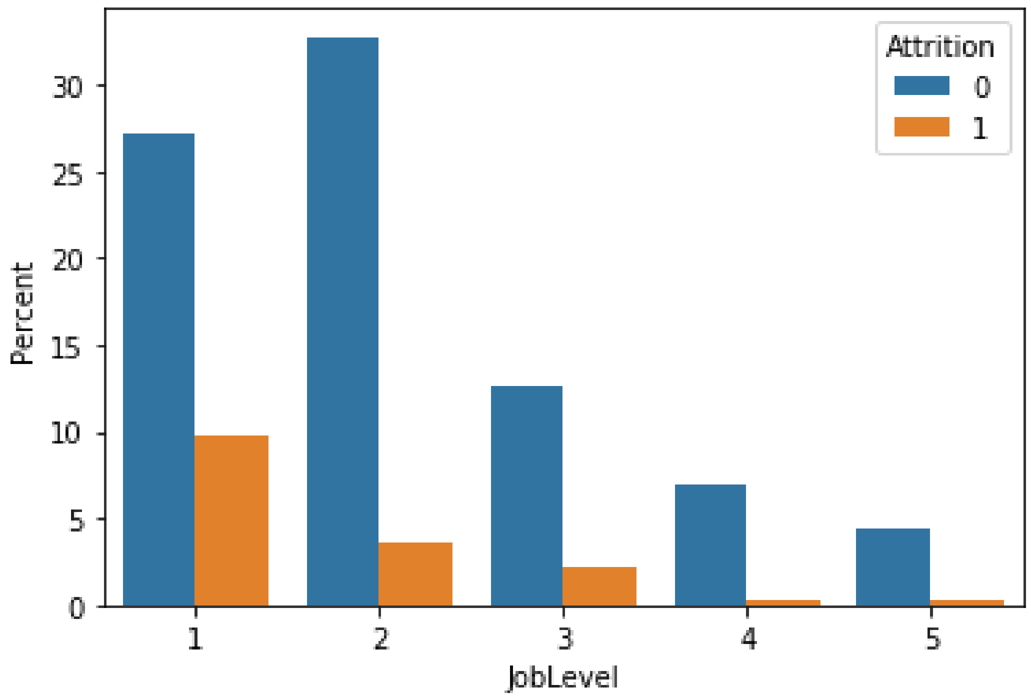
Exploration data analysis

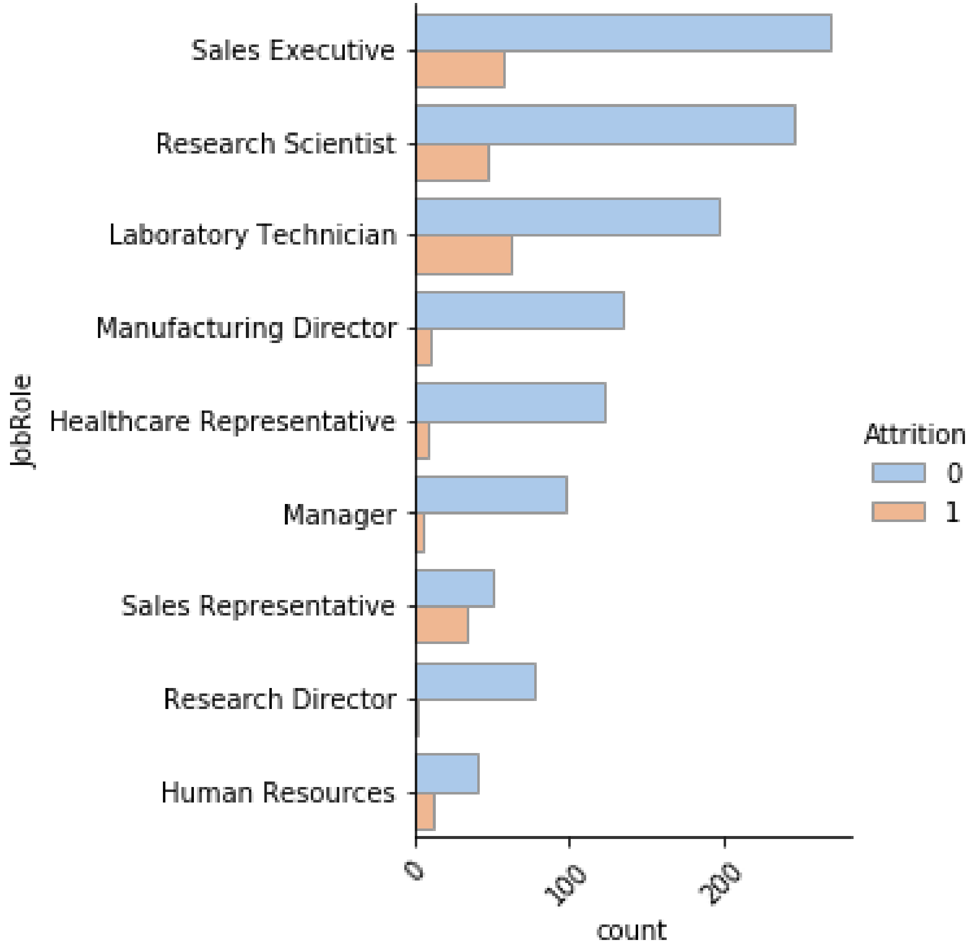
- Employee attrition by travelling and department





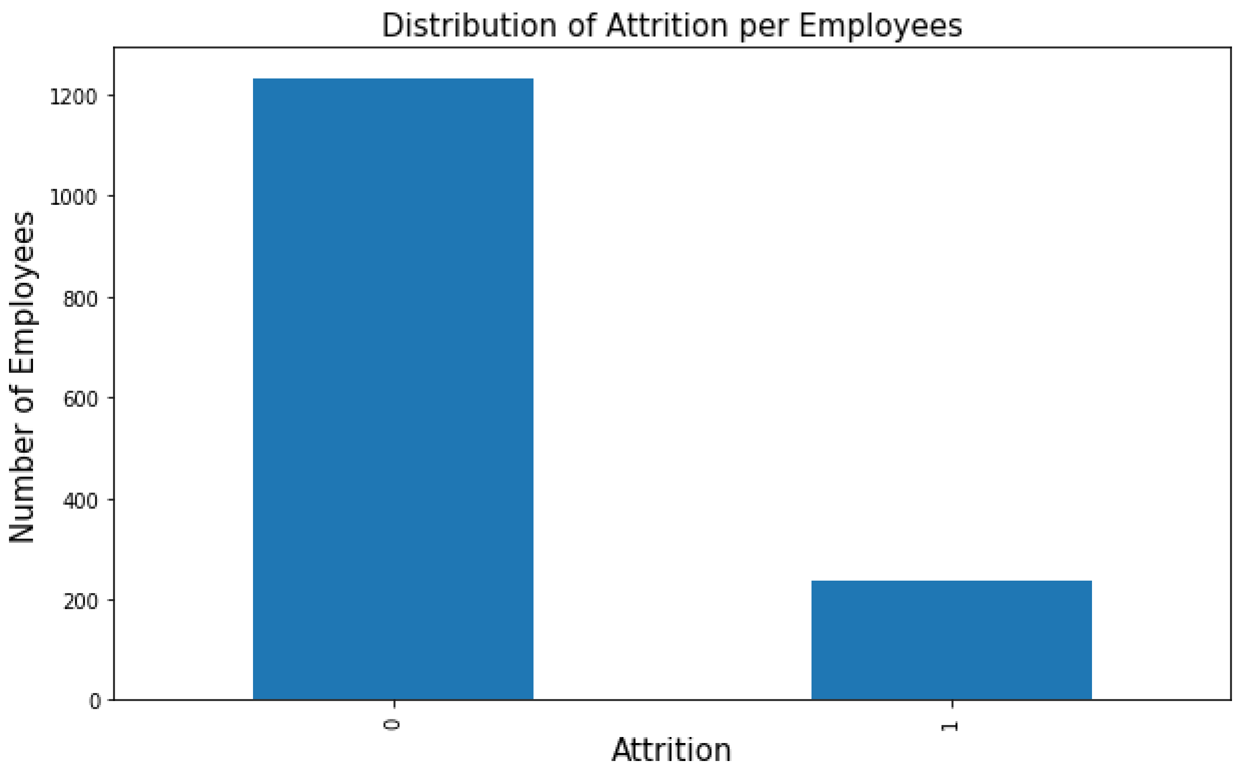
Exploration data analysis





- Employee attrition by job level and job role
- Employee in more junior role are more prone to leave for another company

Data Pre-processing



	Attrition	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely	Department_Research & Development	Department_Sales	Ec
0	1	0	1	0	1	
1	0	1	0	1	0	
2	1	0	1	1	0	
3	0	1	0	1	0	
4	0	0	1	1	0	

- Left: checking for data imbalance
- Right: turning categorical variables to numeric variables.

Model selection and development

Train Test Split

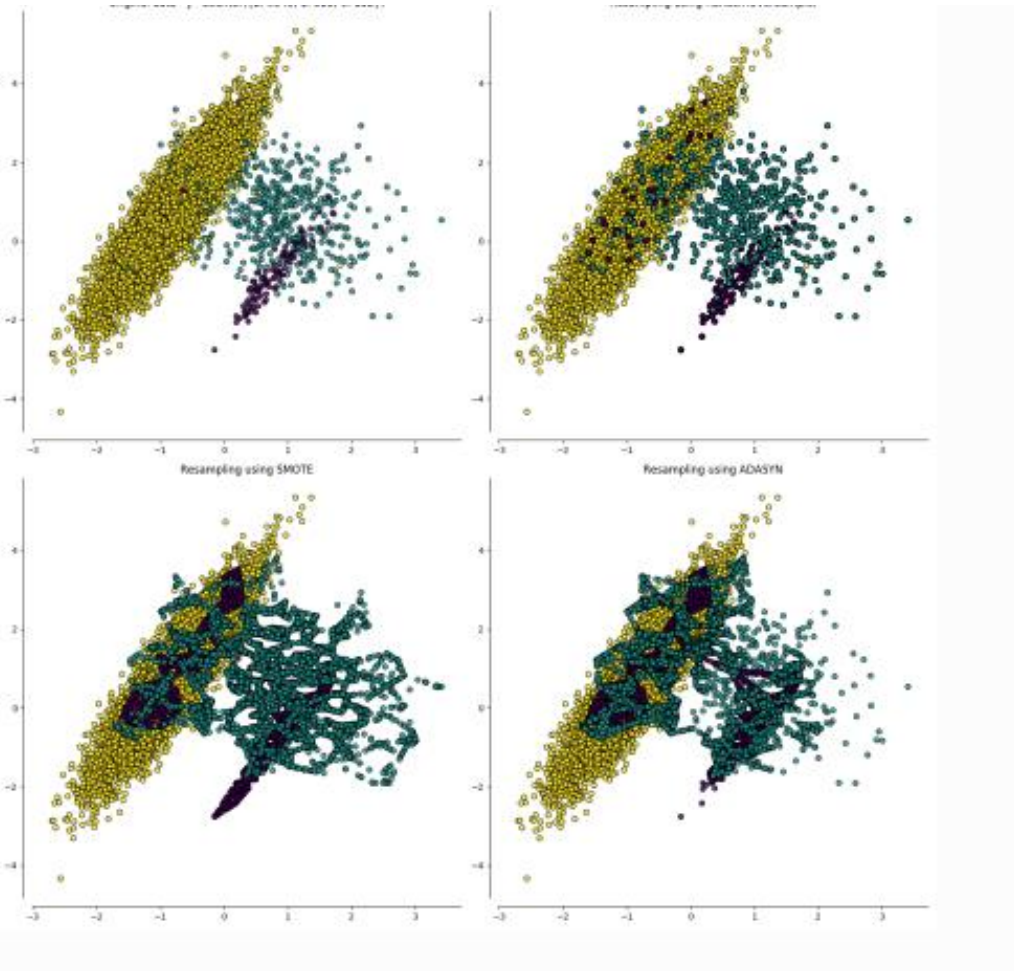
Oversampling using SMOTE and ADASYN

Train 3 Models

1. Logistic Regression
2. Random Forest
3. Gradientboosting

- SMOTE method of adjusting data imbalance performed better than ADASYN with logistic model.

- Adjusted data with SMOTE were trained further with other models



SMOTE (Synthetic Minority Over-Sampling Technique)
ADASYN (Adaptive Synthetic Sampling method)

Model Evaluation

---Gradient Boosting Model---

Gradient Boosting AUC = 0.64

	precision	recall	f1-score	support
0	0.88	0.98	0.93	370
1	0.75	0.30	0.42	71
accuracy			0.87	441
macro avg	0.81	0.64	0.68	441
weighted avg	0.86	0.87	0.85	441

---Random Forest Model---

Random Forest AUC = 0.66

	precision	recall	f1-score	support
0	0.89	0.99	0.93	370
1	0.83	0.34	0.48	71
accuracy			0.88	441
macro avg	0.86	0.66	0.71	441
weighted avg	0.88	0.88	0.86	441

---Logistic Regression Model---

Logistic Regression AUC = 0.76

	precision	recall	f1-score	support
0	0.94	0.78	0.85	370
1	0.39	0.75	0.51	71
accuracy			0.77	441
macro avg	0.67	0.76	0.68	441
weighted avg	0.85	0.77	0.80	441

- Top left table: Gradient boosting evaluation metrics
- Top left table: Random Forest metrics
- Bottom left: Logistic regression

Model Evaluation

- Top table: ROC for the 3 models
- Bottom table: features importances