SC1015 Mini Project

DSF3
HUNG KUO-CHEN
JODI TAY SEOW XUAN
YANG XIAOYUE

Table of Contents

01

Motivation

Problem definition and dataset

02

EDA

Insights and tools used

03

ML

Why and how the techniques used help achieve the objective(s)

04

Outcome

Interesting discoveries and recommendations

Motivation

- Employee attrition: the turnover rate in various job roles
- High attrition rate brings about problems
  - Hard-to-replace employees leave → lower productivity and profits
  - High costs incurred in training and hiring new employees
- Aim: uncover reasons for an employee’s resignation and recommend improvements made within IBM to retain its employees

IBM HR Attrition Dataset

- Used data from 1470 employees in IBM
    - 16.1% left IBM, 83.9% stayed in IBM
- Includes 34 independent variables based on an employee’s profile
    - Each contributed to whether an employee decided to leave or stay

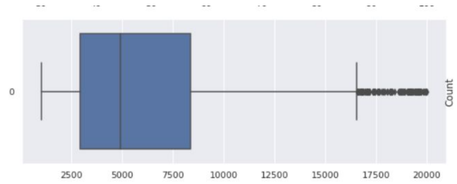
Cleaning the Data

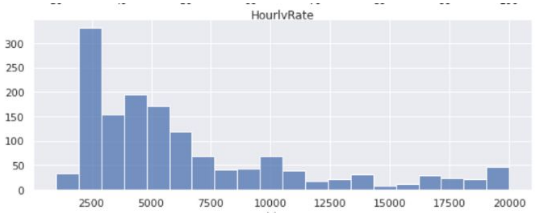
- Dropped insignificant columns
  - EmployeeCount, Over18, StandardHours, EmployeeNumber
- Checked for missing values and duplicates
  - None found

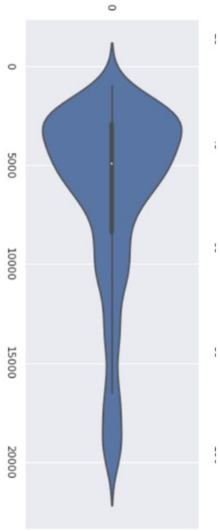
Breakdown of Variables

Numeric

- 14 regular numeric variables
- 9 factor numeric variables
- A variety of data visualization methods: box plot, histogram plot, and violin plot

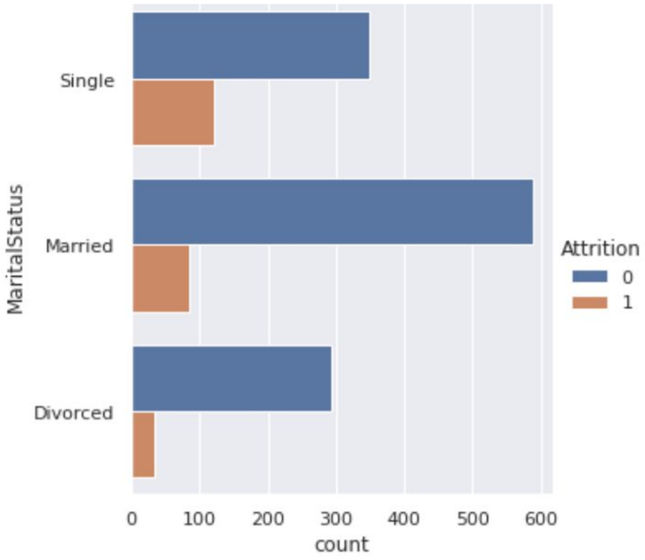






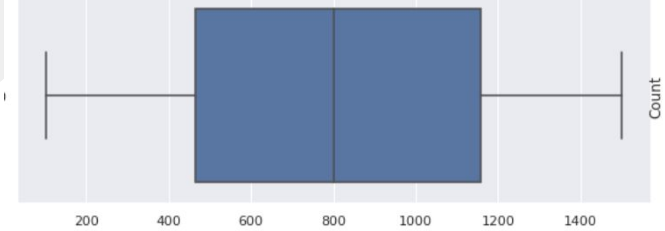
Categorical

- 6 categorical variables
- Categorical bar plot of each variable against attrition using GroupBy

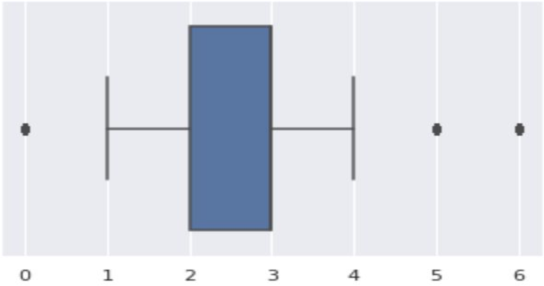


Outliers (Box Plot)

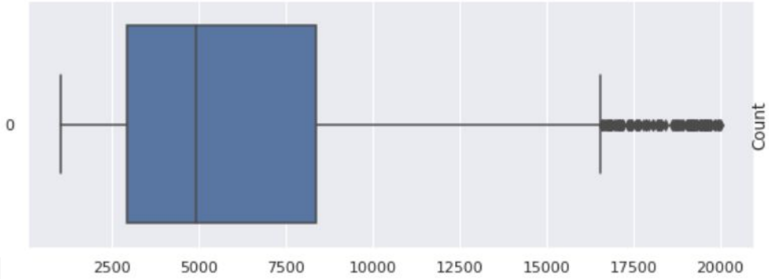
Different variables have different degrees of outliers (close to none, moderate, and large)



Box plot with close to no outliers



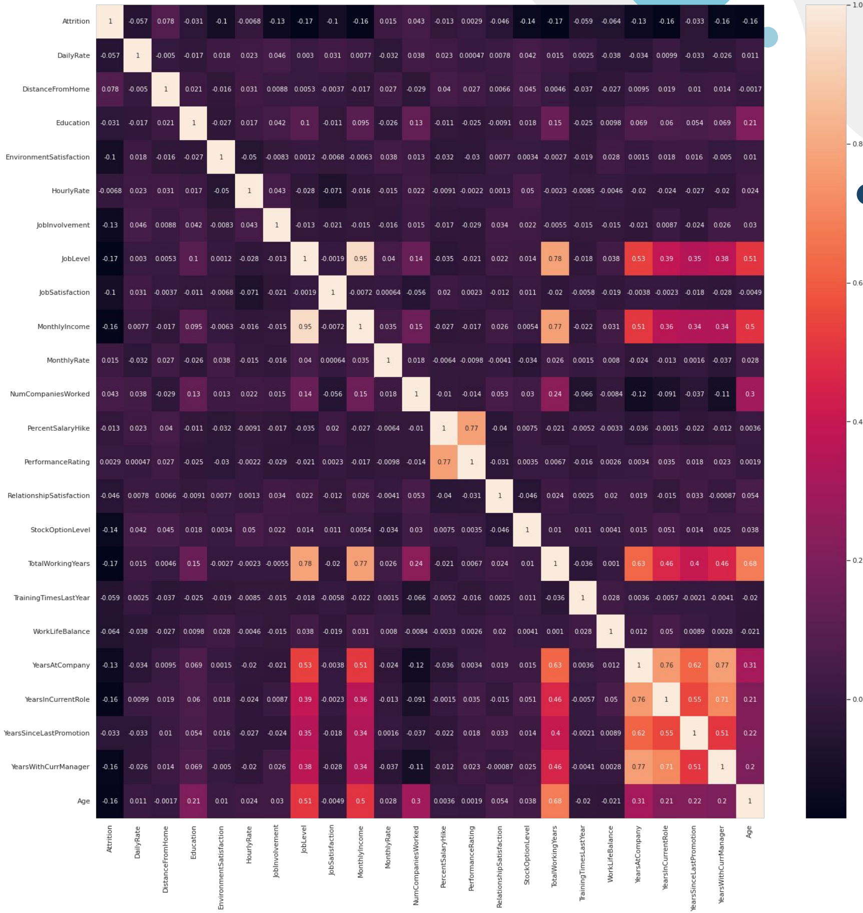
Box plot with a moderate number of outliers



● Box plot with a large number of outliers

Correlation Matrix

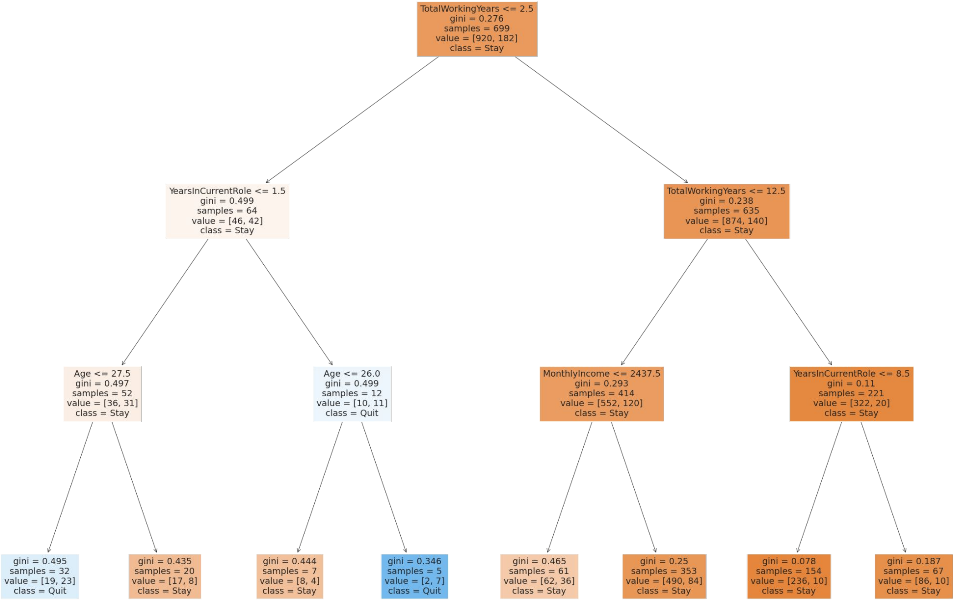
- Reclassify attrition as a numeric variable
- Plot the correlation matrix for attrition against all other numeric variables
- Interesting Findings:
  - MonthlyIncome and JobLevel (0.95)
  - TotalWorkingYears and JobLevel (0.78)
  - TotalWorkingYears and MonthlyIncome (0.77)



	Attrition	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate	NumCompaniesWorked	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	Age
Attrition	1	-0.057	0.078	-0.031	-0.1	-0.0068	-0.13	-0.17	-0.1	-0.16	0.015	0.043	-0.013	0.0029	-0.046	-0.14	-0.17	-0.059	-0.064	-0.13	-0.16	-0.033	-0.16	-0.16
DailyRate	-0.057	1	-0.005	-0.017	0.018	0.023	0.046	0.003	0.031	0.0077	-0.032	0.038	0.023	0.00047	0.0078	0.042	0.015	0.0025	-0.038	-0.034	0.0099	-0.033	-0.026	0.011
DistanceFromHome	0.078	-0.005	1	0.021	-0.016	0.031	0.0088	0.0053	-0.0037	-0.017	0.027	-0.029	0.04	0.027	0.0066	0.045	0.0046	-0.037	-0.027	0.0095	0.019	0.01	0.014	-0.0017
Education	-0.031	-0.017	0.021	1	-0.027	0.017	0.042	0.1	-0.011	0.095	-0.026	0.13	-0.011	-0.025	-0.0091	0.018	0.15	-0.025	0.0098	0.069	0.06	0.054	0.069	0.21
EnvironmentSatisfaction	-0.1	0.018	-0.016	-0.027	1	-0.05	-0.0083	0.0012	-0.0068	-0.0063	0.038	0.013	-0.032	-0.03	0.0077	0.0034	-0.0027	-0.019	0.028	0.0015	0.018	0.016	-0.005	0.01
HourlyRate	-0.0068	0.023	0.031	0.017	-0.05	1	0.043	-0.028	-0.071	-0.016	-0.015	0.022	-0.0091	-0.0022	0.0013	0.05	-0.0023	-0.0085	-0.0046	-0.02	-0.024	-0.027	-0.02	0.024
JobInvolvement	-0.13	0.046	0.0088	0.042	-0.0083	0.043	1	-0.013	-0.021	-0.015	-0.016	0.015	-0.017	-0.029	0.034	0.022	-0.0055	-0.015	-0.015	-0.021	0.0087	-0.024	0.026	0.03
JobLevel	-0.17	0.003	0.0053	0.1	0.0012	-0.028	-0.013	1	-0.0019	0.95	0.04	0.14	-0.035	-0.021	0.022	0.014	0.78	-0.018	0.038	0.53	0.39	0.35	0.38	0.51
JobSatisfaction	-0.1	0.031	-0.0037	-0.011	-0.0068	-0.071	-0.021	-0.0019	1	-0.0072	0.00064	-0.056	0.02	0.0023	-0.012	0.011	-0.02	-0.0058	-0.019	-0.0038	-0.0023	-0.018	-0.028	-0.0049
MonthlyIncome	-0.16	0.0077	-0.017	0.095	-0.0063	-0.016	-0.015	0.95	-0.0072	1	0.035	0.15	-0.027	-0.017	0.026	0.0054	0.77	-0.022	0.031	0.51	0.36	0.34	0.34	0.5
MonthlyRate	0.015	-0.032	0.027	-0.026	0.038	-0.015	-0.016	0.04	0.00064	0.035	1	0.018	-0.0064	-0.0098	-0.0041	-0.034	0.026	0.0015	0.008	-0.024	-0.013	0.0016	-0.037	0.028
NumCompaniesWorked	0.043	0.038	-0.029	0.13	0.013	0.022	0.015	0.14	-0.056	0.15	0.018	1	-0.01	-0.014	0.053	0.03	0.24	-0.066	-0.0084	-0.12	-0.091	-0.037	-0.11	0.3
PercentSalaryHike	-0.013	0.023	0.04	-0.011	-0.032	-0.0091	-0.017	-0.035	0.02	-0.027	-0.0064	-0.01	1	0.77	-0.04	0.0075	-0.021	-0.0052	-0.0033	-0.036	-0.0015	-0.022	-0.012	0.0036
PerformanceRating	0.0029	0.00047	0.027	-0.025	-0.03	-0.0022	-0.029	-0.021	0.0023	-0.017	-0.0098	-0.014	0.77	1	-0.031	0.0035	0.0067	-0.016	0.0026	0.0034	0.035	0.018	0.023	0.0019
RelationshipSatisfaction	-0.046	0.0078	0.0066	-0.0091	0.0077	0.0013	0.034	0.022	-0.012	0.026	-0.0041	0.053	-0.04	-0.031	1	-0.046	0.024	0.0025	0.02	0.019	-0.015	0.033	-0.00087	0.054
StockOptionLevel	-0.14	0.042	0.045	0.018	0.0034	0.05	0.022	0.014	0.011	0.0054	-0.034	0.03	0.0075	0.0035	-0.046	1	0.01	0.011	0.0041	0.015	0.051	0.014	0.025	0.038
TotalWorkingYears	-0.17	0.015	0.0046	0.15	-0.0027	-0.0023	-0.0055	0.78	-0.02	0.77	0.026	0.24	-0.021	0.0067	0.024	0.01	1	-0.036	0.001	0.63	0.46	0.4	0.46	0.68
TrainingTimesLastYear	-0.059	0.0025	-0.037	-0.025	-0.019	-0.0085	-0.015	-0.018	-0.0058	-0.022	0.0015	-0.066	-0.0052	-0.016	0.0025	0.011	-0.036	1	0.028	0.0036	-0.0057	-0.0021	-0.0041	-0.02
WorkLifeBalance	-0.064	-0.038	-0.027	0.0098	0.028	-0.0046	-0.015	0.038	-0.019	0.031	0.008	-0.0084	-0.0033	0.0026	0.02	0.0041	0.001	0.028	1	0.012	0.05	0.0089	0.0028	-0.021
YearsAtCompany	-0.13	-0.034	0.0095	0.069	0.0015	-0.02	-0.021	0.53	-0.0038	0.51	-0.024	-0.12	-0.036	0.0034	0.019	0.015	0.63	0.0036	0.012	1	0.76	0.62	0.77	0.31
YearsInCurrentRole	-0.16	0.0099	0.019	0.06	0.018	-0.024	0.0087	0.39	-0.0023	0.36	-0.013	-0.091	-0.0015	0.035	-0.015	0.051	0.46	-0.0057	0.05	0.76	1	0.55	0.71	0.21
YearsSinceLastPromotion	-0.033	-0.033	0.01	0.054	0.016	-0.027	-0.024	0.35	-0.018	0.34	0.0016	-0.037	-0.022	0.018	0.033	0.014	0.4	-0.0021	0.0089	0.62	0.55	1	0.51	0.22
YearsWithCurrManager	-0.16	-0.026	0.014	0.069	-0.005	-0.02	0.026	0.38	-0.028	0.34	-0.037	-0.11	-0.012	0.023	-0.00087	0.025	0.46	-0.0041	0.0028	0.77	0.71	0.51	1	0.2
Age	-0.16	0.011	-0.0017	0.21	0.01	0.024	0.03	0.51	-0.0049	0.5	0.028	0.3	0.0036	0.0019	0.054	0.038	0.68	-0.02	-0.021	0.31	0.21	0.22	0.2	1

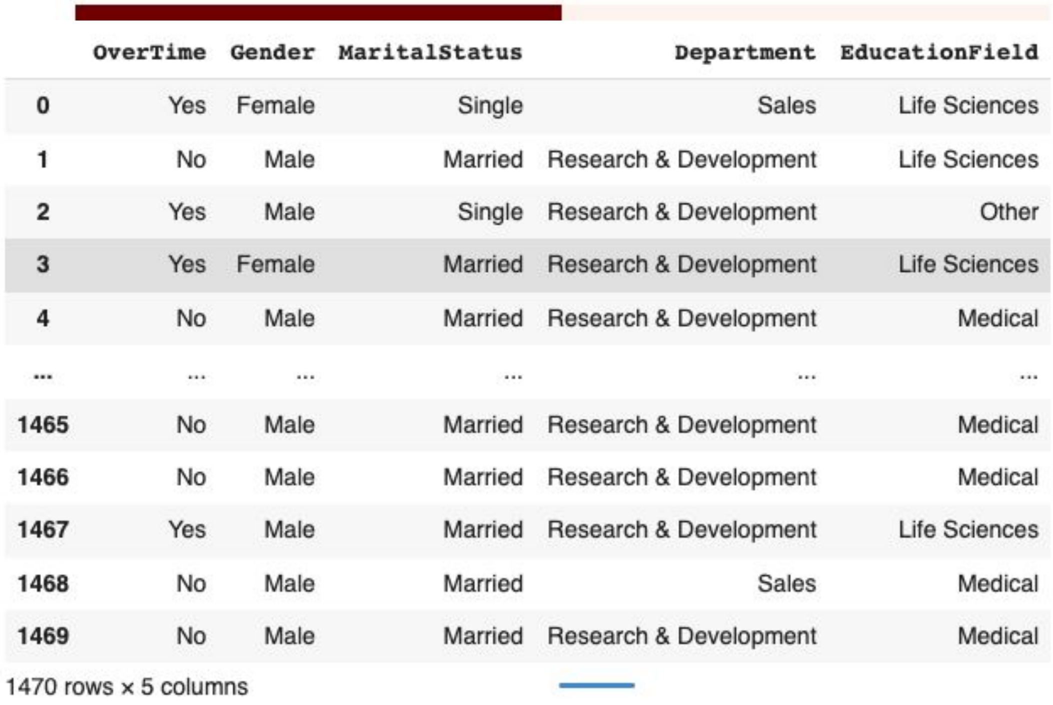
Random Forest

- Extract variables with relatively high correlation
- **Accuracy of random forest: 84.5%**
- Tune hyperparameters with random search
  - Random combinations of the hyperparameters are used to find the optimal solution for the built model



Logistic Regression

- Determine level of influence of categorical variables on attrition
- Convert each variable into numeric indicator variables with get_dummies
- Accuracy: 0.84 (train), 0.86 (test)
- Ineffective due to **extremely low precision**
  - The model only classifies 38% of employees that quit correctly



	OverTime	Gender	MaritalStatus	Department	EducationField
0	Yes	Female	Single	Sales	Life Sciences
1	No	Male	Married	Research & Development	Life Sciences
2	Yes	Male	Single	Research & Development	Other
3	Yes	Female	Married	Research & Development	Life Sciences
4	No	Male	Married	Research & Development	Medical
...	...	...	...	...	...
1465	No	Male	Married	Research & Development	Medical
1466	No	Male	Married	Research & Development	Medical
1467	Yes	Male	Married	Research & Development	Life Sciences
1468	No	Male	Married	Sales	Medical
1469	No	Male	Married	Research & Development	Medical

1470 rows × 5 columns

Neural Network

- Accuracy:

Summary of Findings

1. Random Forest

Numeric variables that have relatively high correlation with attrition can be used to predict attrition

2. Logistic Regression

Not recommended using categorical variables to predict attrition

3. Neural Network

Data Driven Insights

Reasons for Leaving

- Low salary
- Low chance for career progression
- Lack of opportunities

Improvements for IBM

- Provide more salary incentives
- Open up spots for changes in senior management

Thank You

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**