SC1015 Data Science Mini-Project

Attrition Prevention

Team 11 : Najah Ismail, Apurva Mishra, Sanjay Pramod

TABLE OF CONTENTS

01

Motivation

Background

- Attrition: Employees leaving a company for various reasons.
- Naturally occurs in workforce.
- Needs to be understood and embraced by the management and HR to avoid setbacks.

Chosen Dataset

‘IBM HR Analytics Employee Attrition & Performance’

by Pavan Subhash on Kaggle

Problem Statement

Identify factors that affect Attrition, predict optimal conditions to make people stay, and devise recommendations.

02

Exploring the Data Set

1. Data Preparation

Cleaning the Data

- Drop columns irrelevant to problem
- Check to filter-out unsuitable rows
  - Retirement
  - Duplicates
  - NULL values
- Modify names - convenience and homogeneity

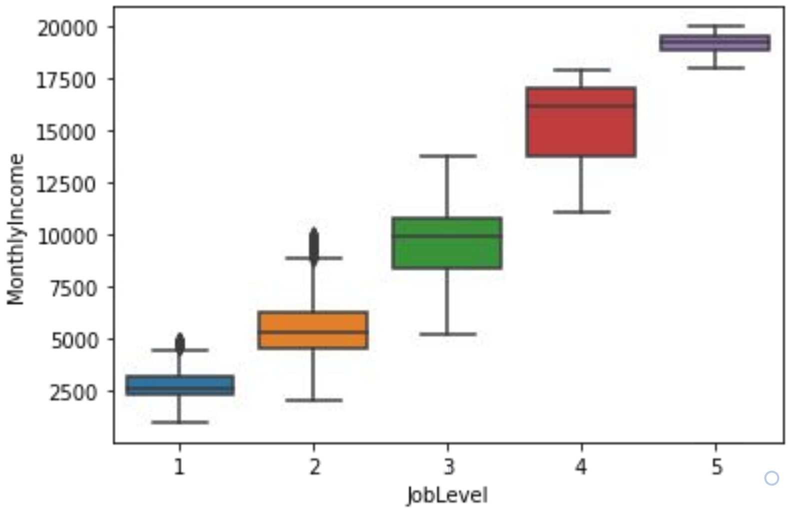
Feature Engineering

- ➢ JobSwitchRate:

  How frequently each employee switched jobs wrt years in workforce

- ➢ RelIncAtLevel:

  Standardising income across levels using median



2. EDA

EDA Overview

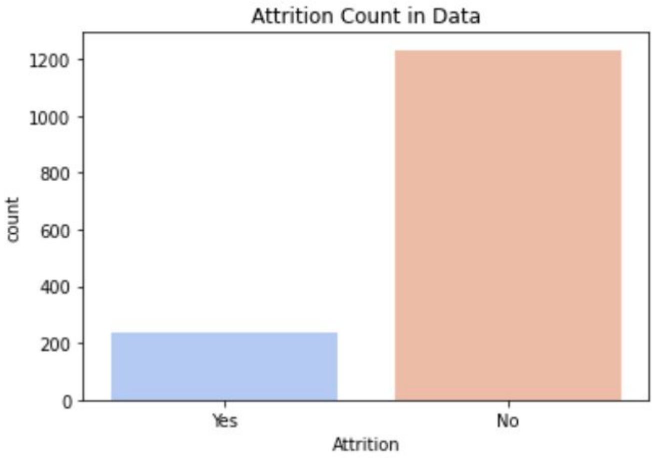
1. Attrition Value Imbalance
2. Bi-Variate Exploration : HeatMap
3. Attrition and Categorical Values
4. Attrition and Numerical Values
5. JobSatisfaction and Attrition
6. MonthlyIncome
7. Age and Gender Statistics

Mini EDA Questions

1. Employee Stagnation
2. Exploring Job Environment
3. Evaluation in a Work Space
4. Potential New Hiring

1. Class Imbalance

Exploring attrition rate in the whole data set, we observe the data is imbalanced.
Attrition=Yes has **237** data points **(16.12%)** while **Attrition=NO** has **1233** data points **(83.88%)**



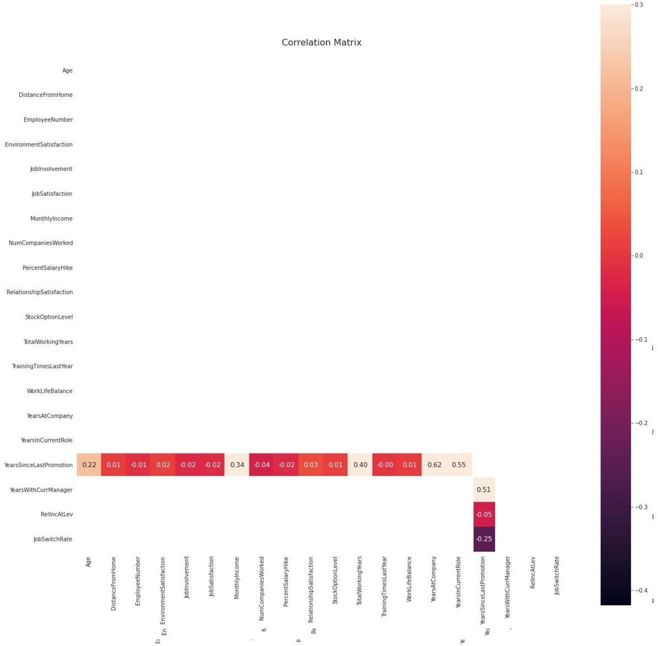
Rectifying Class imbalance

Bivariate Exploration HeatMap

TotalWorkingYears

- Age
- Joblevel
- MonthlyIncome

employees who have worked longer have higher salaries and are in higher positions.



YearsSinceLast Promotion

- Joblevel
- YearsAtCompany

employee face **stagnation** after reaching a threshold number of years.Usually those in higher job level. To be explored further in exploration question.

Categorical Variables

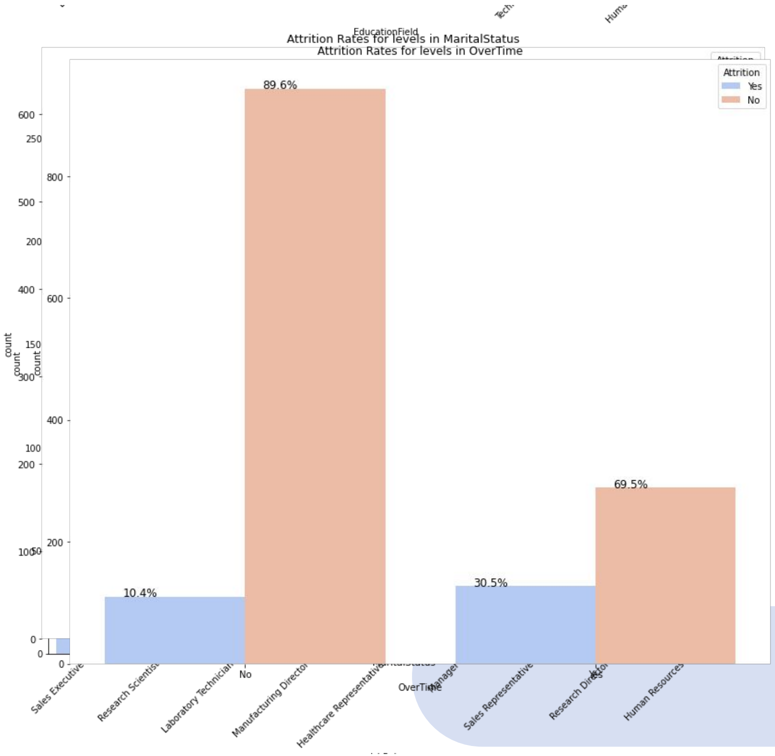
Summary:

Highest attrition Rates across categorical variables

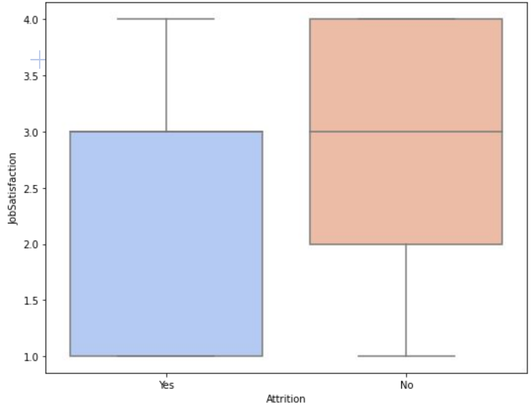
- Department :**Sales** : 21%
- Gender : **Male** : 17%
- JobRole : **LabTechniacian** : 23%
- MaritalStatus: **Single** : 19%
- OverTime : **Yes** :
- PerformanceRating : **Low and High equal** : 16%

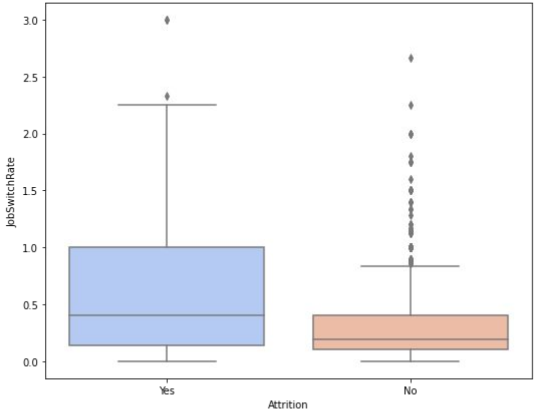
Lowest attrition Rates across categorical variables

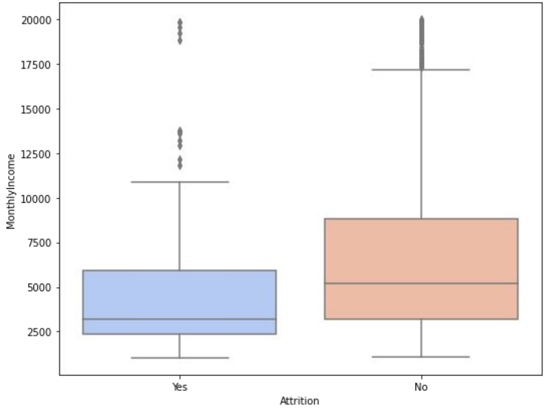
- Higher job levels & educational levels have lower attrition rate
- BusinessTravel : **None** : 8%



Numerical Variables







Attrition=YES for :

- Lower **JobSatisfaction**
- Higher **JobSwitchRate**
- Lower YearsWithCurrentManager, YearsInCompany,totalWorkingYears and Age
- Lower JobLevels
- Lower **MonthlyIncome**

Variables that seem to have little effect on attrition :

- StockOptionLevel
- employee number
- trainingTimesSinceLastYear
- RelationshipSatisfaction

Inferences

Younger employees and **newer** employees are more likely to leave. Perhaps for better job prospects.

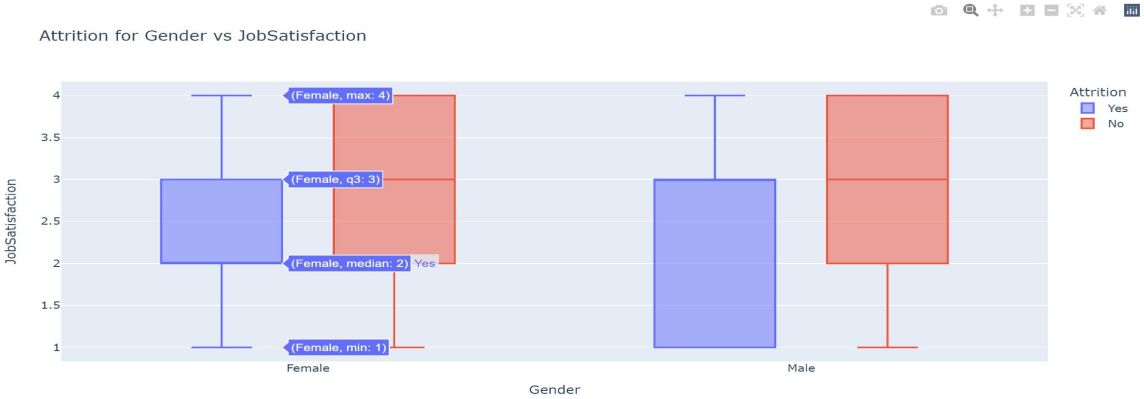
Employees at **lower JobLevels** and **lower income** are more likely to leave.

JobSatisfaction show a strong relation with Attrition. Attrited employees have **lower JobSatisfaction**

JobSatifaction and Attrition

Performance Rating

From the employers perspective the **problem** becomes clear, attrited employees with **higher PerformanceRating had lower JobSatisfaction.**



Gender

Males who left have **lower JobSatisfaction** as compared to Females.

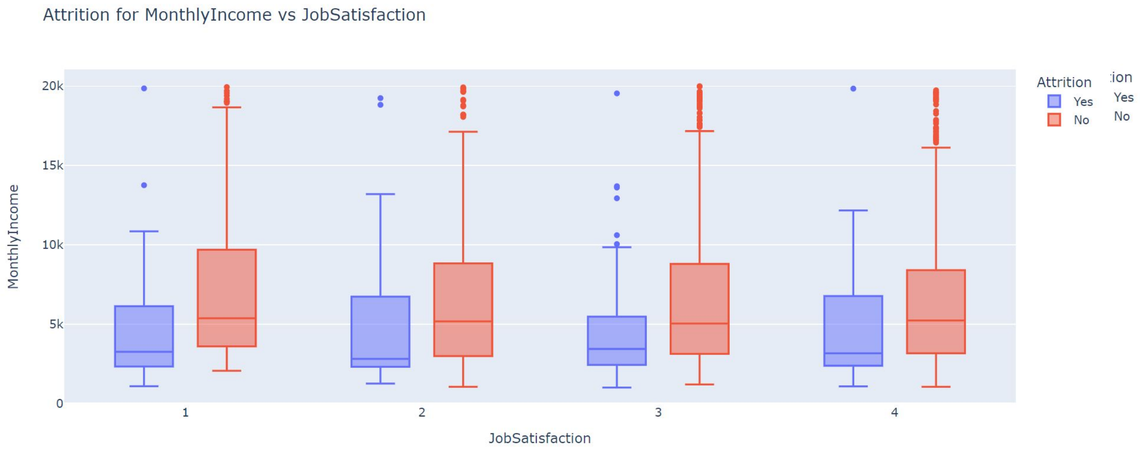
Department

Departments **HR** and **Sales** have lower JobSatisfaction for Attrited employees while it is equally high for all non-Attrited employees.

MonthlyIncome and Attrition

lower MonthlyIncome is a common trait in **Mid-Level(JobLevel = 3/4) attrited employees**. These are employees who probably have the experience and skill-set to find jobs elsewhere at higher pay.

For different levels of **JobSatisfaction**, **Lower MonthlyIncome** is a clear reason behind Attrition while median is similar for attrited and non-attrited employees

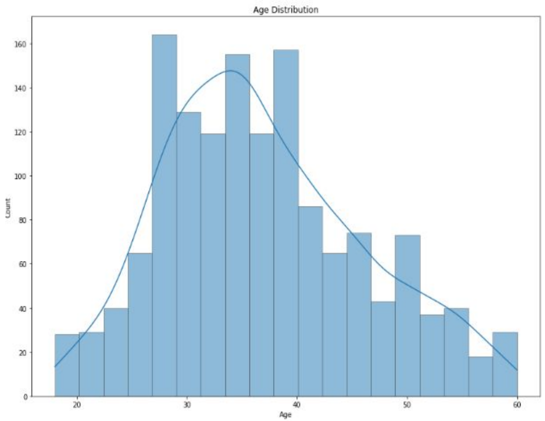


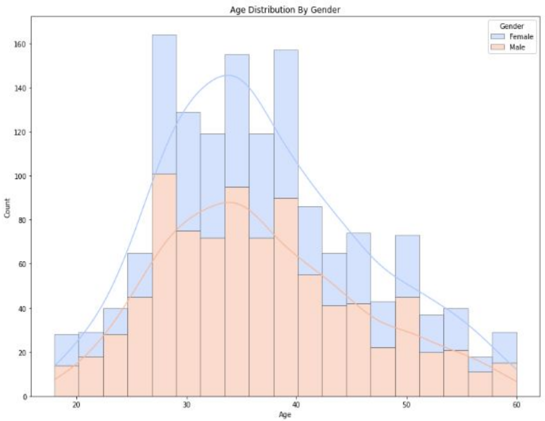
On Average:

1. Women make more money than men
2. Sales Department has highest salary
3. Research Directors and Managers make the most while Sales Representative make the least

Age

The **average age** for data set is around **37**





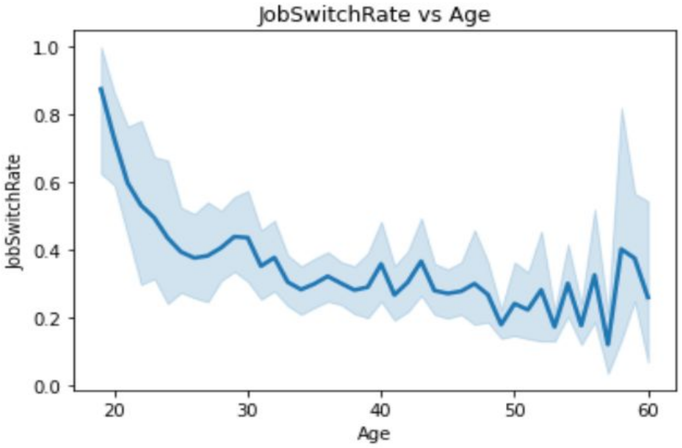

```
Overall Average Age :
 36.923809523809524
```

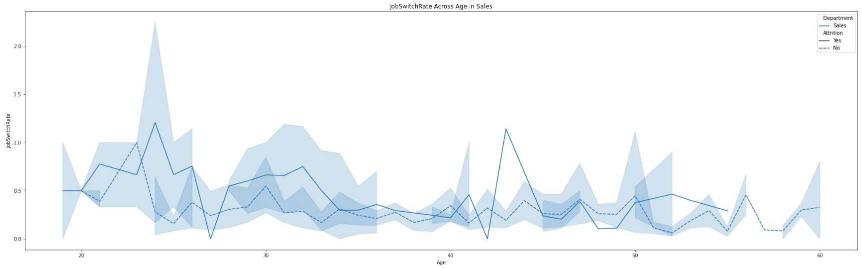
```
Average Age by :
Gender
Female    37.329932
Male      36.653061
Name: Age, dtype: float64
```

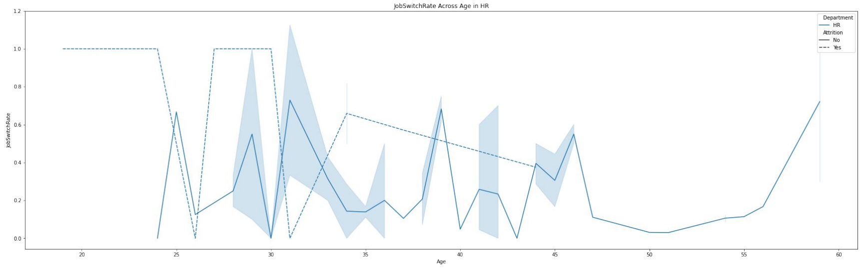
```
Average Age by :
Department
HR        37.809524
RnD       37.042664
Sales     36.542601
Name: Age, dtype: float64
```

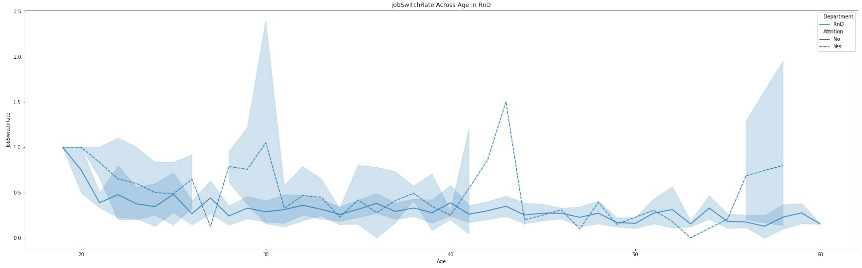
Age vs JobSwitchRate

Which Age group switches jobs most frequently? Are they likely to leave the company?









Inferences

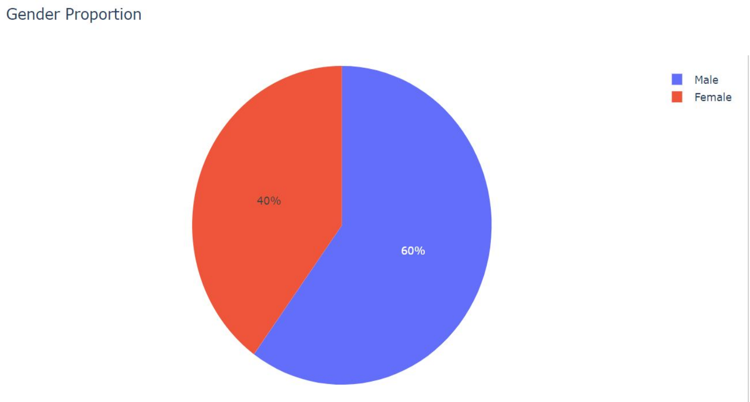
In all three department employees between the age of **25 to 35** have a general **highest switch rates** for both Attrition=Yes and Attrition=No

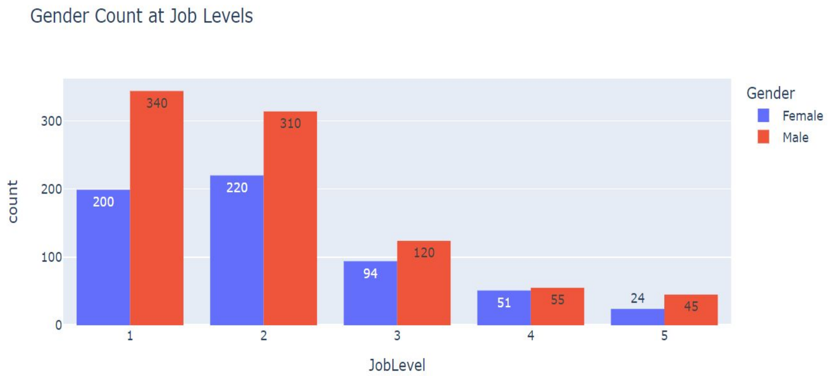
But overall there is little difference in the JobSwitchRates between the attrited and non attrited employees in each Department.

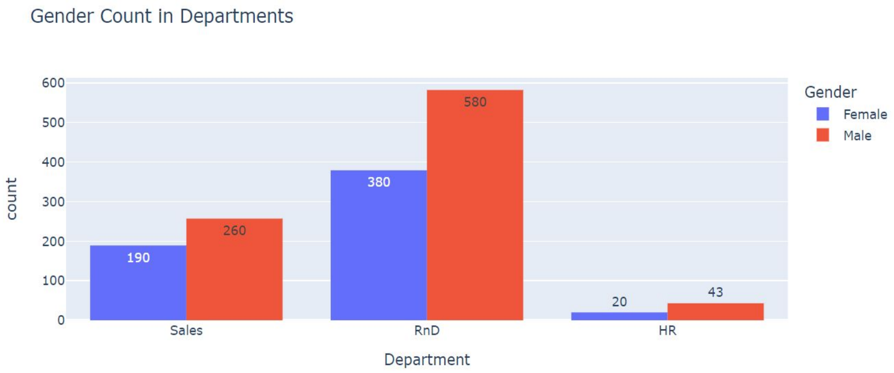
Gender

The **Female Employee count is lower** across Departments and JobLevels

Women Earn more than Men







Mini EDA Questions

1. Employee Stagnation
2. Immediate Job Environment
3. Evaluation in a Work Space
4. Potential New Hiring

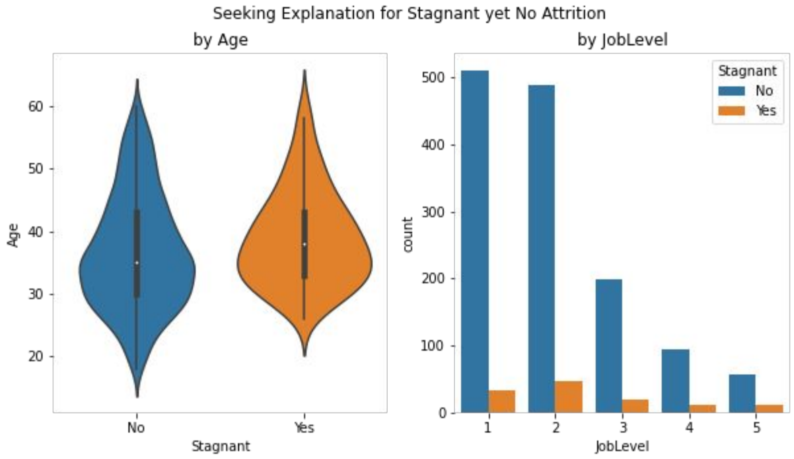
1- Stagnation - Definition

RelSalaryHike	NewManager	RecentPromotion
⇧	⇧	⇧
PercentSalaryHike	YearsWithCurrManager	YearsSinceLastPromotion

Stagnant

Low + No + No = Stagnant ; Else = Non-stagnant

1- Stagnation - Relation to Attrition



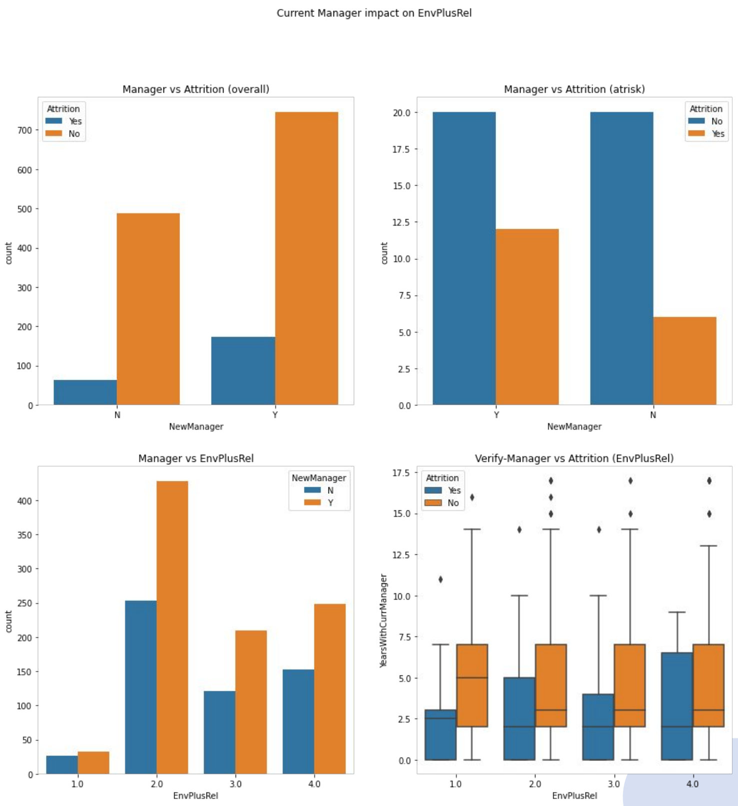
- ➢ Attrition rate is lower for ‘stagnant’ employees
- ➢ Examining relation with job level and age
- ➢ Conclusion: ‘Stagnation’ can be interpreted as settling into high level job.

2- Immediate Job Environment

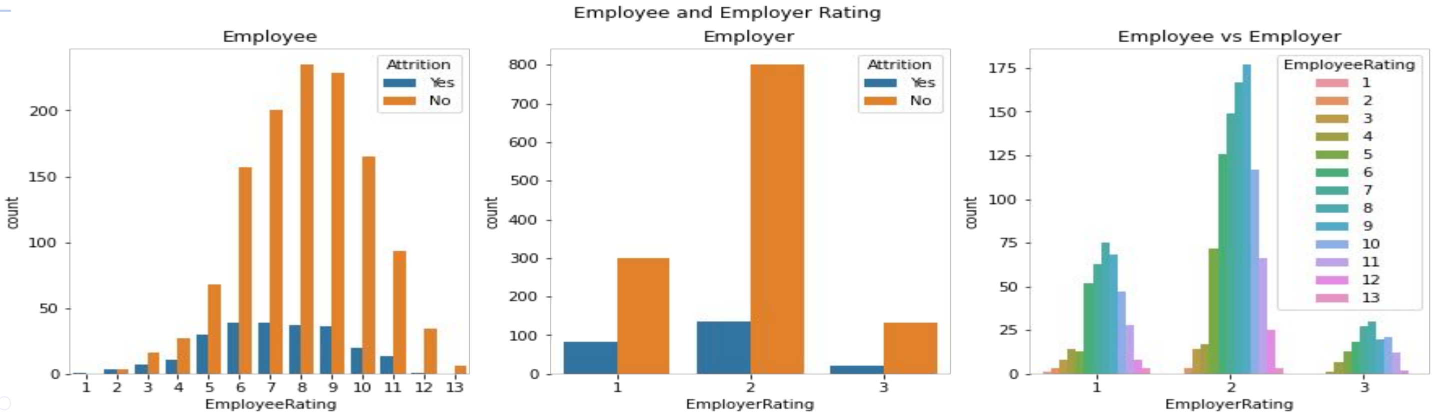
➢ Examining Environment Satisfaction, Relationship Satisfaction and New Manager
➢ EnvPlusRel <= 1 is very atrisk

➢ See if employee's discontentment (and attrition) is caused by current team

➢ No direct relation detected
  - Reshuffling futile
  - Overall issue with IBM, not team-specific



3- Evaluation in Workspace



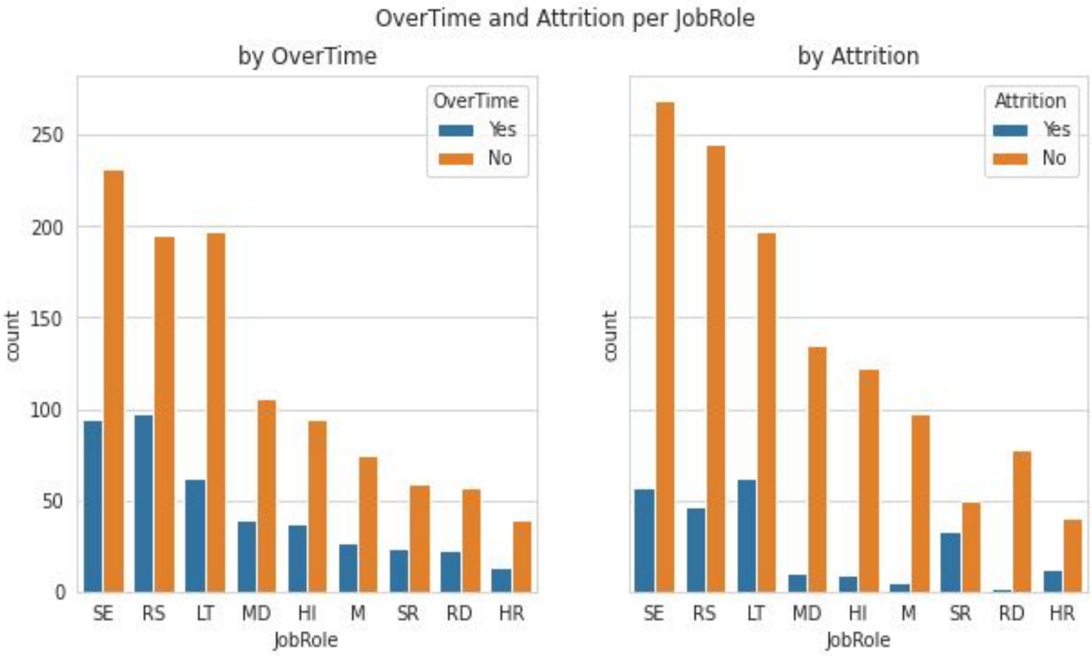
➢ Compare how employee and employer rate each other

➢ Attrition Yes employees often follow pattern - Employer rating is lower in proportion to Employee rating

➢ Well-rated employees may rate IBM low and leave

4- Potential New Hiring

- ➢ Examining Overtime and Attrition per Job Role
- ➢ Correspondence suggests some job roles are overworked
- ➢ New Hires to distribute workload
- ➢ Target highest attrition rate roles — Sales Executive, Laboratory Technician, and Human Resources



Tools and Techniques Learnt

Plotly

Used plotly for more insightful and easier to read graphs

Feature Engineering

Scaling and Altering existing variables to create new meaning.

03

Model - Predicting Attrition

Data Prep For Model

1. Dropped columns that had little effect on attrition
2. Fixed null values in new dataset and variables
3. Encoded categorical variables using ScikitLearn LabelEncoder
4. Fit to our models and tuned model parameters for best prediction

We used models that involve classification to solve it, since the problem focuses on predicting the class of attrition (yes or no) .

The attrition dataset contains 1470 datapoints , so we used simple tree based models to avoid increasing complexity.

3 model candidates

DecisionTreeClassifier

Decision tree classifier splits dataset along feature values to isolate data points belonging to each class.

Balanced Bagging Tree Classifier

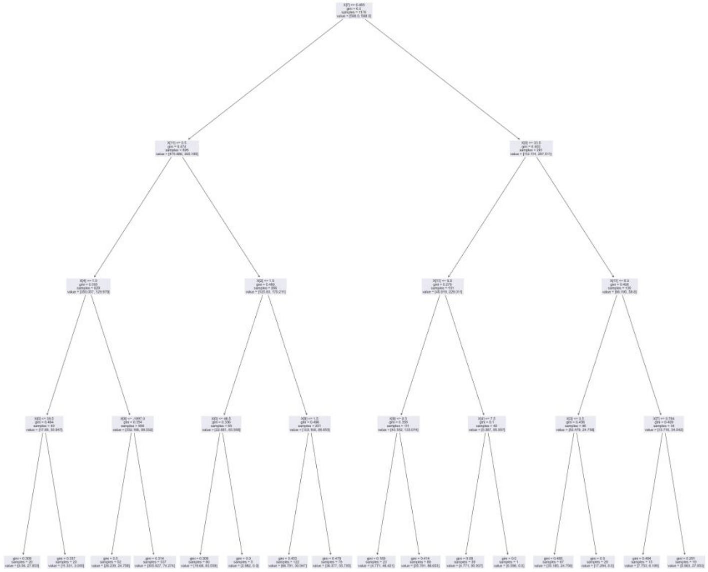
Ensemble method that fits base estimator on subsets of original data and aggregate the predictions. Also balances dataset to resolve class imbalance.

XGBoost

Ensemble method that combines several weak tree based models, and focuses on optimizing an arbitrary loss function.

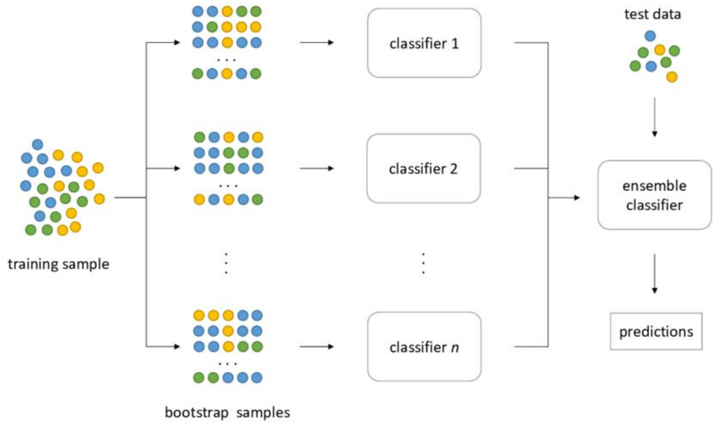
Decision Tree Classifier

- Tried different tree depths and found that depth=4 gave optimum values on the train and test data.
- Modified the class_weight attribute to “balanced” to tackle data imbalance. This ensures that the model assigns class weights inversely proportional to frequency of the class . Therefore , the model penalizes wrong predictions of the minority class more stringently.
- Train accuracy - 79.25%
- Test Accuracy - 74.15 %



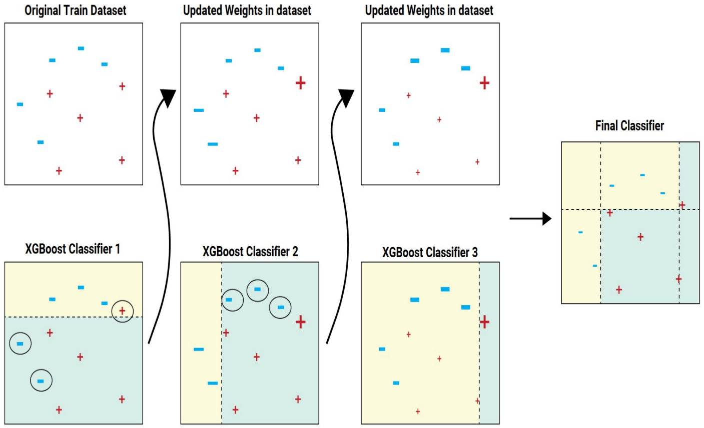
BalancedBaggingClassifier

- Used base estimator as Decision Tree Classifier to improve upon previous model. The model aggregates predictions from multiple base estimator instances.
- Classifier includes an additional step to balance the training set at fit time using a given sampler.
- Sample method - “Not majority” to resample from minority class to solve class imbalance.
- Tried different tree depths and found that depth=4 gave optimum values on the train and test data.
- Train accuracy - 86.9%
- Test Accuracy - 84.3 %



XGBoost - XGBClassifier

- We wanted to try a different ensemble method to test performance, so we chose gradient boosting.
- XGB uses ensemble of weaker Decision Trees to update the weights in the model and accurately predict classes.
- Tried different tree depths and found that depth=2 gave optimum values on the train and test data.
- Objective function of "logitraw" gave optimum accuracy with the log loss function
- Train accuracy - 88.4%
- Test Accuracy - 84.35 %



Tools/ Techniques Used

We used Decision Trees and and tweaked the max_depth attribute as learned from the lectures to get better accuracy across the board. We also changed the class_weight attribute to counter class imbalance in our dataset.

We tried out 2 ensemble methods to learn and apply something new. Both methods had parameters to help prevent overfitting and improve accuracy on imbalanced data.

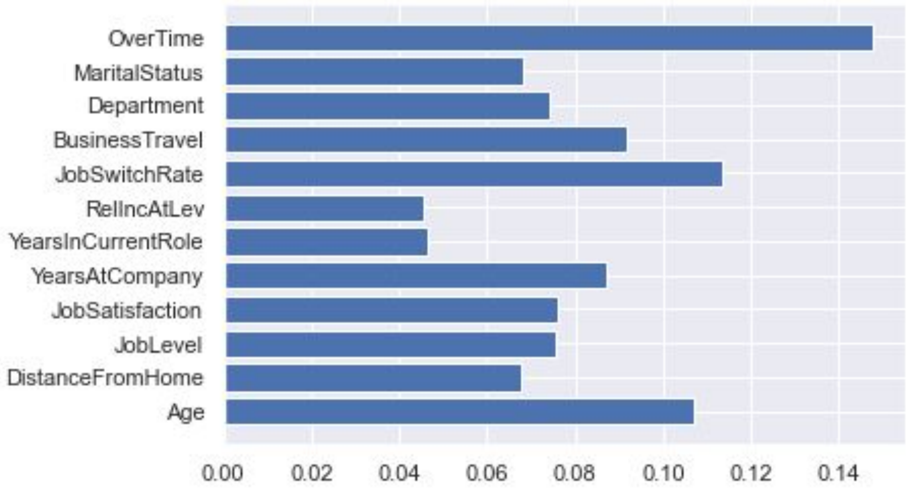
- Balanced Bagging Tree Classifier - used Oversampling.
- XGB Classifier from XGBoost

04

Verifying Recommendations

Most important feature employers can change :

1. OverTime
2. BusinessTravel
3. RelIncAtLev



Feature importance from XGBoost

4 Recommendations — 4 New Datasets

Idealised DataSet

01

Taking the ideal values for each variable, using Attrition==No values
OverTime: Mode
BusinessTravel: Mode
RelIncAtLev : Mean for each JobLevel

Ideal Inc Only

02

BusinessTravel and OverTime may be non-negotiable for some roles.

Only change RelIncAtLev to the ideal **mean value** at each JobLevel

Recommended Changes

10% Increase

04

Increase their salary by 5% and keep original OverTime and BusinessTravel values

5% Salary Increase

03

Increase their salary by 5% and keep original OverTime and BusinessTravel values

Confusion Matrix Results

DataSet1

False Negative : 235

01

02

DataSet2

False Negative : 187

Outcomes

False Negative : 169

DataSet3

04

03

False Negative : 169

DataSet4

05

Conclusion

Recommendations and Data Insights

Key Recommendations

Income

For employer, 5% increase is most preferable. >5% does not improve attrition rate as much.

The ideal value is average of Attrition No employees at that level.

OverTime and BusinessTravel

Reduce OverTime and BusinessTravel or Hire more employees for even work distribution.

Data Insight — At Risk Employees

High Performance and Low Satisfaction

Mid-level Employees
Job Level 3/4

Overworked Job Roles
Sales Executive,
Laboratory Technician,
and Human Resources

Younger Employees Age 25-30

Attrition, although inherent to the workforce, shows patterns to be navigated for both the employer and the employee to flourish.