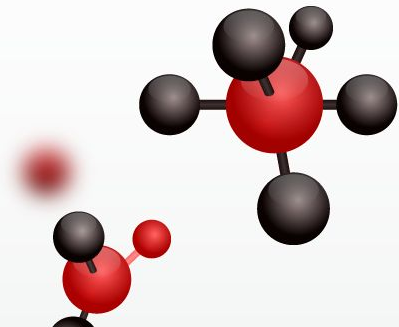




BC2406: Analytics I

Seminar Group 4 Team 3

Team Members:
AGNES NG YI RONG
ERNEST CHUA HUI SHENG
JALVIN NAI GUANG JUN
JASLYN TAN CHIEW YEE
TAN JING YU, DENIS

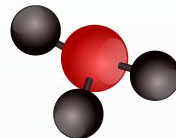
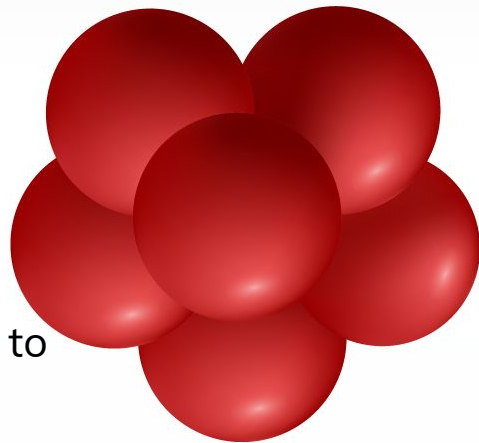


AGENDA

- Introduction
- Problem Statements
- Business Problems
- Data Cleaning
- Exploratory Data Analysis
- Model Construction and Evaluation
- Solutions
- Business Recommendations
- Limitations of the Business Recommendations
- Conclusion

INTRODUCTION

- ❖ Dataset is taken from a pharmaceutical company
- ❖ Original dataset: 23,000 rows and 37 columns
- ❖ Where we are now:
 - Manually evaluate a high volume of applications daily to shortlist potential candidates
 - Labor-intensive, time-consuming and inefficient
- ❖ Where we will be:
 - Hiring the **RIGHT** person for the **RIGHT** job
 - Reduce cost and time spent sieving out individual candidates



Breaking Down the Problem Statement



Problem Statement



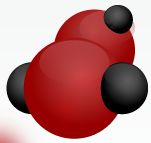
A

Accurately assess and shortlisted candidates with the relevant skill sets, experience and psycho-emotional traits



B

Match them with relevant job openings to drive operational efficiency and improve accuracy in the matching process



Business Problems

1. Education and Skills Mismatch in the Research and Development (R&D) Department

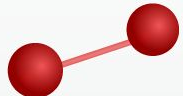
- Requires a specific skill set and education to conduct research and development on products
- Persistent mismatch of skills and education could deprive companies of the productivity and efficiency

2. Rising Burden on Human Resource Department due to Outdated Hiring Practices

- Huge amounts of time are expected to be spent screening every single applicant.
- Furthermore, mistakes in the hiring process could turn into a vicious cycle for the HR department

3. High Turnover Volatility in the Sales Department

- There is a scarce supply of graduates and the high demand for competent employees in the competitive industry.
- Even more so for the sales department





Business Problems

4. Challenge in Assessing Talent for Cultural Fit during the Hiring Process

- Likelihood that an employee will adapt to the core beliefs, attitudes and behaviors of the organization.
- Workplace culture changes from place to place and companies constantly struggle to integrate employees into the company's culture.

5. Loss of Resources Due to Turnover of Employees

- Costly to the company
- Lost its human capital investments in the form of resources spent on training, workshops and other forms of employee enhancement.

Categorizing the Variables



Variables



(1) Variables

Age + Department + DistanceFromHome +
Education + EducationField + Employee
Source + Gender + JobRole + MaritalStatus +
AverageTenure + TotalWorkingYears



(2) Variables

EnvironmentSatisfaction + JobInvolvement +
JobSatisfaction + RelationshipSatisfaction +
WorkLifeBalance + BusinessTravel +
OverTime + StockOptionLevel +
MonthlyIncome + PercentSalaryHike +
TrainingTimesLastYear



Data Cleaning



- **Change inappropriate values to NAs**
 - ◀ Eg "", "missing", "na", "Test", "Test 456", "TESTING", "?????", "TEST"



- **Remove duplicated rows**
 - ◀ 14 rows removed



Remove Redundant Columns

- EmployeeCount
 - Over18
 - StandardHours
- }
- Constant values
- ApplicationID
 - EmployeeID
- }
- Unique identifiers
- DailyRate
 - HourlyRate
 - MonthlyRate
- }
- Poor correlation, deemed unnecessary

Data Cleaning



Data Type Conversion

Factor

- Attrition
- BusinessTravel
- Department
- Education
- EducationField
- Employee Source
- EnvironmentalSatisfaction
- Gender
- JobInvolvement
- JobLevel
- JobRole
- JobSatisfaction
- MaritalStatus
- OverTime
- PerformanceRating
- RelationshipRating
- RelationshipSatisfaction
- StockOptionLevel
- WorkLifeBalance

Integer

- DistanceFromHome
- HourlyRate
- MonthlyIncome
- PercentSalaryHike



Data Cleaning



Specific Data Cleaning

- From the summary, we see that:

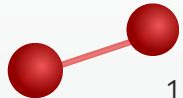
Department		Gender	
1296	: 1	1	: 0
Human Resources	: 1015	2	: 1
Research & Development	: 15343	Female	: 9393
Sales	: 7148	Male	: 14113
NA's	: 11	NA's	: 10

- Replaced with NA



Remove All Rows With NAs

- 273 NAs in 23432 rows x 30 columns
- Deemed insignificant hence removed entire row





Data Validity Check



AgeStartedWorking < 0

Derived from: $\text{TotalWorkingYears} - \text{Age}$
Illogical
Removed such entries





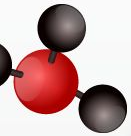
TotalWorkingYears < YearsAtCompany

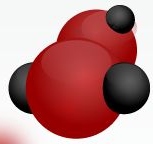
Illogical
Removed such entries

Feature Engineering



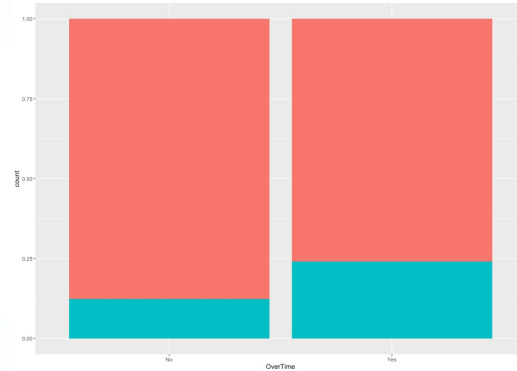
Variable	How it was derived	Rationale
 PriorYearsOfExperience	TotalWorkingYears - YearsAtCompany	<ul style="list-style-type: none">• To gauge an employee's level of experience before joining the company.
 AverageTenure	$\frac{\text{PriorYearsOfExperience}}{\text{NumCompaniesWorked}}$	<ul style="list-style-type: none">• To investigate whether employees had a "job-hopping" culture.• Suggest to the company how loyal the employees would be.



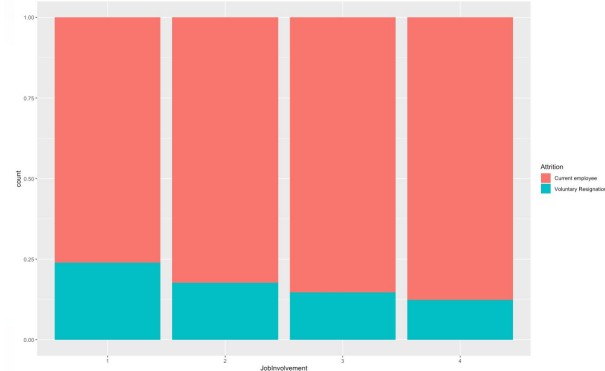


Exploratory Data Analysis

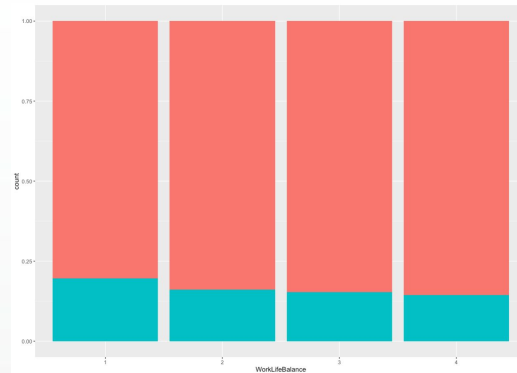
Voluntary Resignation (IBM1)



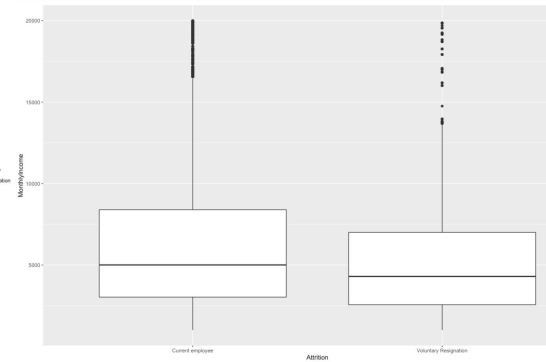
OverTime



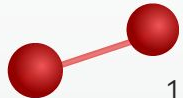
JobInvolvement



WorkLifeBalance



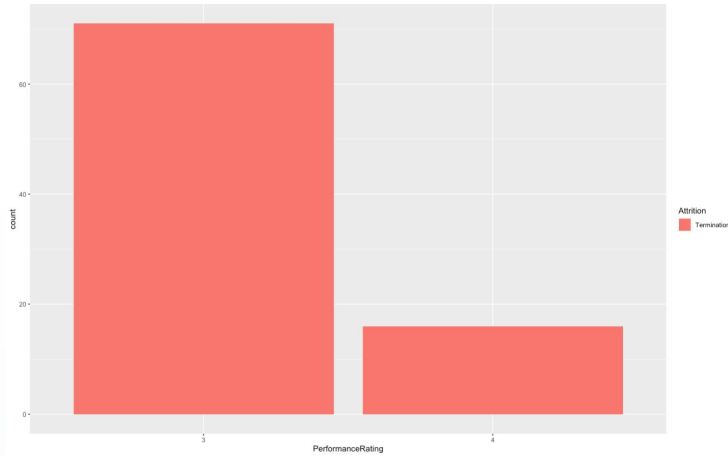
MonthlyIncome



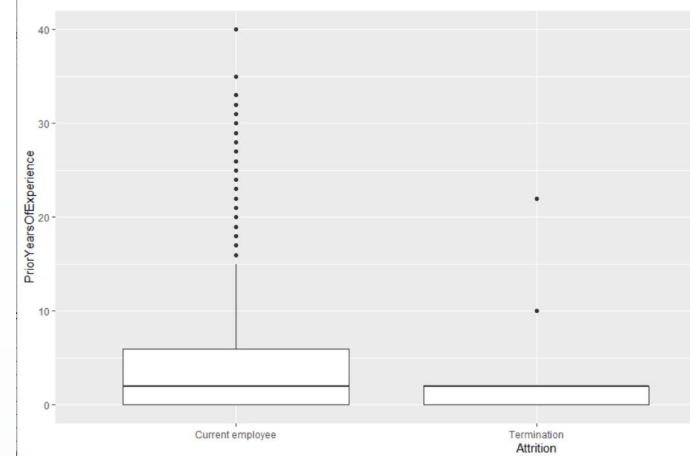


Exploratory Data Analysis

Termination (IBM2)



PerformanceRating



PriorYearsOfExperience



Model Construction and Evaluation

Model 1

Logistic Regression **Predictive**
Model

Shortlist Candidates during the
Selection Stage



Model 2

Classification and Regression
Tree (CART) **Predictive** Model

Match Shortlisted Candidates to
Relevant Job Openings



Model 3

Logistic Regression **Explanatory**
Model

Retain Current Employees





Logistic Regression Predictive Model

Purpose: Shortlist Candidates during the Selection Stage

1

Initial Feature
Selection

2

Model Refining and
Selection Process

3

Decision on
Threshold Level

4

Final Model



Logistic Regression Predictive Model

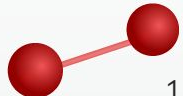
Initial Feature Selection: Domain Knowledge Approach

Group 1 Variables
Age
Department
DistanceFromHome
Education
EducationField
Employee Source
Gender
JobRole
MaritalStatus
AverageTenure
TotalWorkingYears

Must exist at selection stage ✓
Can be collected and filtered at selection stage ✓



Initial Logistic Regression
Model



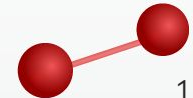


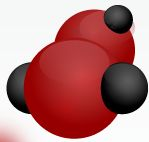
Model Refining and Selection Process

Each variable was removed one by one and the performance metrics of the model were monitored. This process is iterated until the model with the best performance was achieved. A train test split ratio of **70-30** was used for assessing performance for all models.

Statistical Opinions Considered

- ★ P-value < **0.05**
- ★ Odd Ratio Confidence Interval does not contain **1**
- ★ Generalized variance-inflation factors (GVIF) of the models were all below **2** which indicated that there were no issues of multicollinearity






Decision on Threshold Level


	Implication	Criticality
True Positive	Correctly predicted employee will leave.	High - Allows company to retain the employees who will leave via early intervention OR avoid hiring such employees in the selection stage to decrease attrition rates.
True Negative	Correctly predicted employee will not leave.	Low - Generally, not a critical issue for the company.
False Positive	Predicted employee will leave but did not leave.	Medium - Loss of potential talent
False Negative	Predicted employee will not leave but left.	High - The company would have hired and exhausted resources on investing in employees who would leave. This increases attrition rates and hurts the productivity of the company.



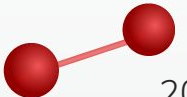
Comparison of Model Performance with Different Threshold Levels

Confusion Matrix of Logistic Regression Model with Threshold of 0.157 

```
> table(testset$Attrition, pass.hat.testm5)
      pass.hat.testm5
      Current employee Voluntary Resignation
Current employee      3557             2254
Voluntary Resignation    415             665
> mean(pass.hat.testm5==testset$Attrition) #Accuracy
[1] 0.613
> 665/(665+2254) #Precision
[1] 0.228
> 665/(665+415) #Recall/Sensitivity
[1] 0.616
>
```

Confusion Matrix of Final Logistic Regression Model with Threshold of 0.5 

```
> table(testset$Attrition, pass.hat.testm5)
      pass.hat.testm5
      Current employee Voluntary Resignation
Current employee      5793             18
Voluntary Resignation  1058             22
> mean(pass.hat.testm5==testset$Attrition) #Accuracy
[1] 0.844
> 18/(18+22) #Precision
[1] 0.45
> 12/(18+1058) #Recall/Sensitivity
[1] 0.0112
>
```





Summary of Final Model

$\log_1(\text{Attrition}) = \beta_0 + \beta_1\text{Age} + \beta_2\text{DepartmentResearch\&Development} + \beta_3\text{DepartmentSales}$
 $+ \beta_4\text{DistanceFromHome} + \beta_5\text{JobRoleHuman Resources} + \beta_6\text{JobRoleLaboratory Technician}$
 $+ \beta_7\text{JobRoleManager} + \beta_8\text{JobRoleManufacturing Director} + \beta_9\text{JobRoleResearch Director} +$
 $\beta_{10}\text{JobRoleResearch Scientist} + \beta_{11}\text{JobRoleSales Executive} + \beta_{12}\text{JobRoleSales}$
 $\text{Representative} + \beta_{13}\text{MarialStatusMarried} + \beta_{14}\text{MaritalStatusSingle} + \beta_{15}\text{AverageTenure} +$
 $\beta_{16}\text{PriorYearsOfExperience}$

Variables (7)	Threshold	Performance Metrics
Age Department DistanceFromHome JobRole MaritalStatus AverageTenure PriorYearsOfExperience	0.157	Accuracy: 0.613 Precision: 0.228 Recall: 0.616

CART Model

Purpose



1. To **match candidates** who pass the screening round to the **appropriate job openings**
2. Determine **which departments are most suited** for employees based on the **information available** in the **selection stage**.

Constructing the model



Profile the current employees against their relevant attributes to project onto shortlisted candidates

Due to the **large size of current employees (19370 entries)**, set the minimum split to 800 instead of the usual 2.

Ensures a reasonably sized optimal tree is obtained and each terminal node contains a fair percentage of the data.

CART Model

Benefits



1. **Simplify** the job of HR staff, **reducing the burden** on the HR department



2. **Uniformity** of CART, the way the employees are allocated will be **unbiased**.



3. The decision tree also has **high explainability** which makes the allocation process simple

CART Model Pruning



- We set the $CP = 0.001666$ where the test set error is at a minimum (0.7093).
- Conduct pruning at $CP = 0.001666$
- After pruning the maximal tree, the size of the optimal tree is 8 terminal nodes.

Root node error: $6402/19370 = 0.3305$

n= 19370

	CP	nsplit	rel error	xerror	xstd
1	0.247891	0	1.0000	1.0000	0.010226
2	0.007654	1	0.7521	0.7521	0.009396
3	0.005467	6	0.7129	0.7227	0.009270
4	0.001666	8	0.7020	0.7093	0.009210
5	0.000000	11	0.6970	0.7170	0.009244



Cart Model

CART Model



After checking the variable importance based on the order of splits, we realized that the top 3 most important variables are

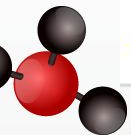
1. JobRole
2. EducationField
3. TotalWorkingYears.

The company should focus on these 3 variables when deciding which department to allocate the shortlisted candidates to.

```
> cart3.opt.ibm3$variable.importance
```

JobRole	EducationField	TotalWorkingYears	Age	Employee Source
2052.240160	151.044462	134.528068	95.414002	68.973462
AverageTenure	DistanceFromHome	Education	MaritalStatus	
64.042597	6.672478	3.032957	1.505999	

```
> |
```



Logistic Model for (2) variables

Purpose



To **identify variables** which the company could **reasonably control** and **influence** after the hiring process to **retain employees**

Constructing the model



Plotting Attrition ~ (2) variables

Refining the model



Ensure variables with **high p-value** are **removed**
Ensure variables with **Odds Ratio Confidence Intervals** do not contain 1
GVIF of the models were ensured to be below **2**

Final Logistic Model for (2) variables

$$\log_2(\text{Attrition}) = \beta_0 + \beta_1 \text{PercentSalaryHike} + \beta_2 \text{BusinessTravelTravel_Frequently} + \beta_3 \text{BusinessTravelTravel_Rarely} + \beta_4 \text{OverTimeYes} + \beta_5 \text{EnvironmentSatisfaction2} + \beta_6 \text{EnvironmentSatisfaction3} + \beta_7 \text{EnvironmentSatisfaction4} + \beta_8 \text{JobInvolvement2} + \beta_9 \text{JobInvolvement3} + \beta_{10} \text{JobInvolvement4} + \beta_{11} \text{RelationshipSatisfaction2} + \beta_{12} \text{RelationshipSatisfaction3} + \beta_{13} \text{RelationshipSatisfaction4} + \beta_{14} \text{WorkLifeBalance2} + \beta_{15} \text{WorkLifeBalance3} + \beta_{16} \text{WorkLifeBalance4}$$

Variables
PercentSalaryHike
BusinessTravel
OverTime
EnvironmentSatisfaction
JobInvolvement
RelationshipSatisfaction
WorkLifeBalance

```
Call:
glm(formula = Attrition ~ PercentSalaryHike + BusinessTravel +
     OverTime + EnvironmentSatisfaction + JobInvolvement + RelationshipSatisfaction +
     WorkLifeBalance, family = binomial, data = ibml)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.293	-0.625	-0.487	-0.399	2.603

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.04667	0.15632	-6.70	2.1e-11 ***
PercentSalaryHike	-0.02139	0.00519	-4.13	3.7e-05 ***
BusinessTravelTravel_Frequently	1.30385	0.08475	15.38	< 2e-16 ***

BusinessTravel
OverTimeYes

EnvironmentSa

EnvironmentSa

EnvironmentSa

JobInvolvement

JobInvolvement

JobInvolvement

RelationshipS

RelationshipS

RelationshipS

WorkLifeBalance

WorkLifeBalance

WorkLifeBalance

Signif. codes

(Dispersion p

(Intercept)

PercentSalaryHike

BusinessTravelTravel_Frequently

BusinessTravelTravel_Rarely

OverTimeYes

EnvironmentSatisfaction

GVIF Df GVIF^(1/(2*Df))

1.01 1

1.01 2

1.01 1

1.01 1

1.02 3

1.02 3

1.02 3

1.02 3

WorkLifeBalance4

2.5 % 97.5 %

0.258 0.476

0.969 0.989

3.127 4.360

1.617 2.212

2.126 2.470

0.720 0.926

0.782

0.774

0.749

0.576

0.482

0.783

0.983

0.893

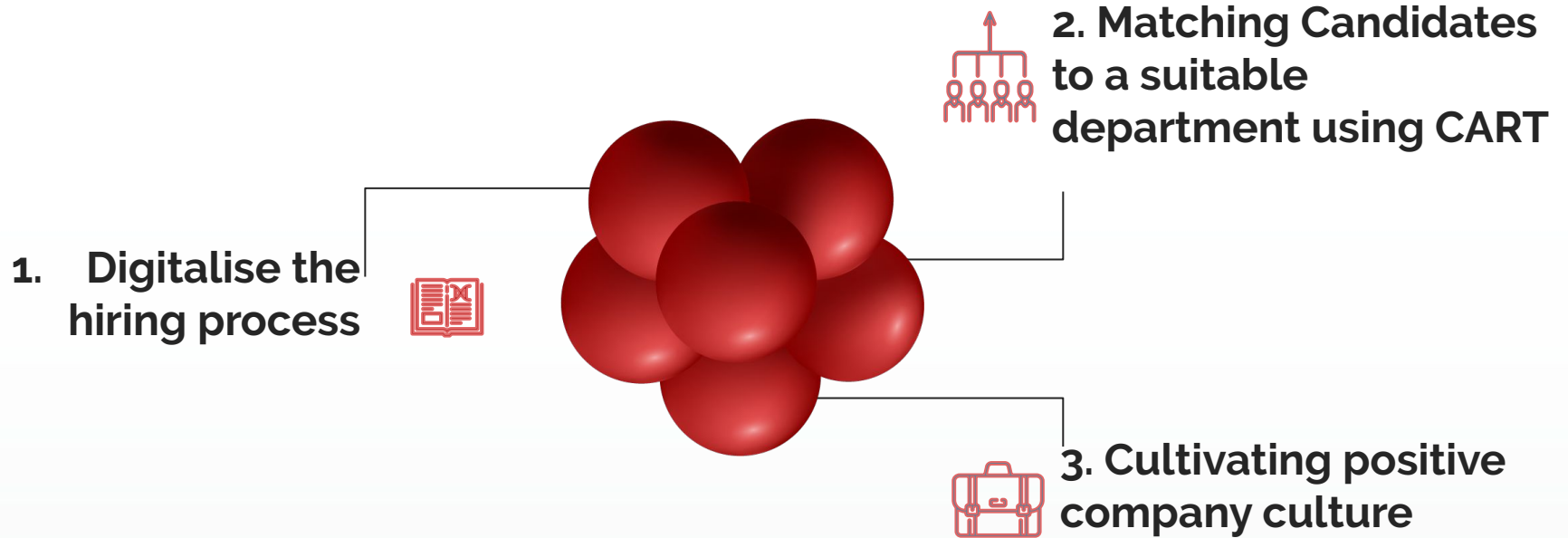
0.869

0.826

0.583 0.845

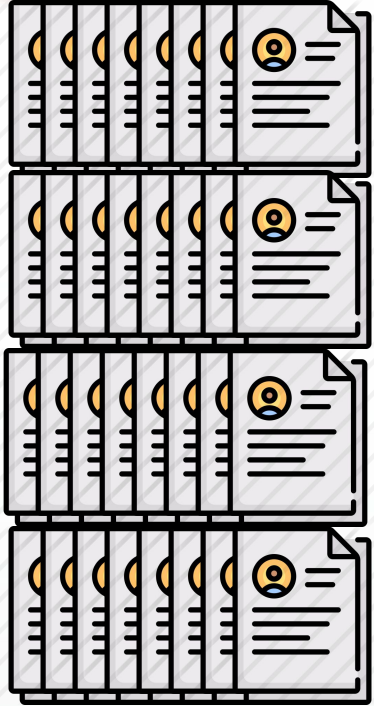
Number of Fisher Scoring iterations: 5

Business Recommendations





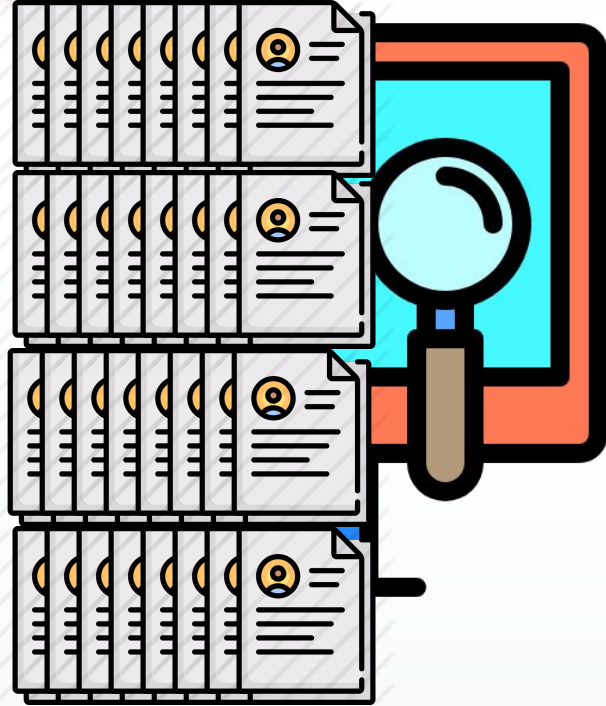
1. Digitizing the hiring process



- Extremely time consuming
- Manually sieve out candidates to move on to the next round
- Overall: Tedious and inefficient process

Furthermore, this does not guarantee a 'right' match.

1. Digitizing the hiring process



- Digitalise the hiring process with electronic versions of applicants' resume
- Stored in a database.
- Applying the logistic regression model, it would seek out applicants with the desired skill set, experience and traits.
- Shortlist only those suitable

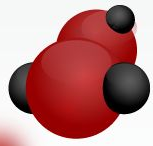
1. Digitizing the hiring process



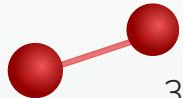
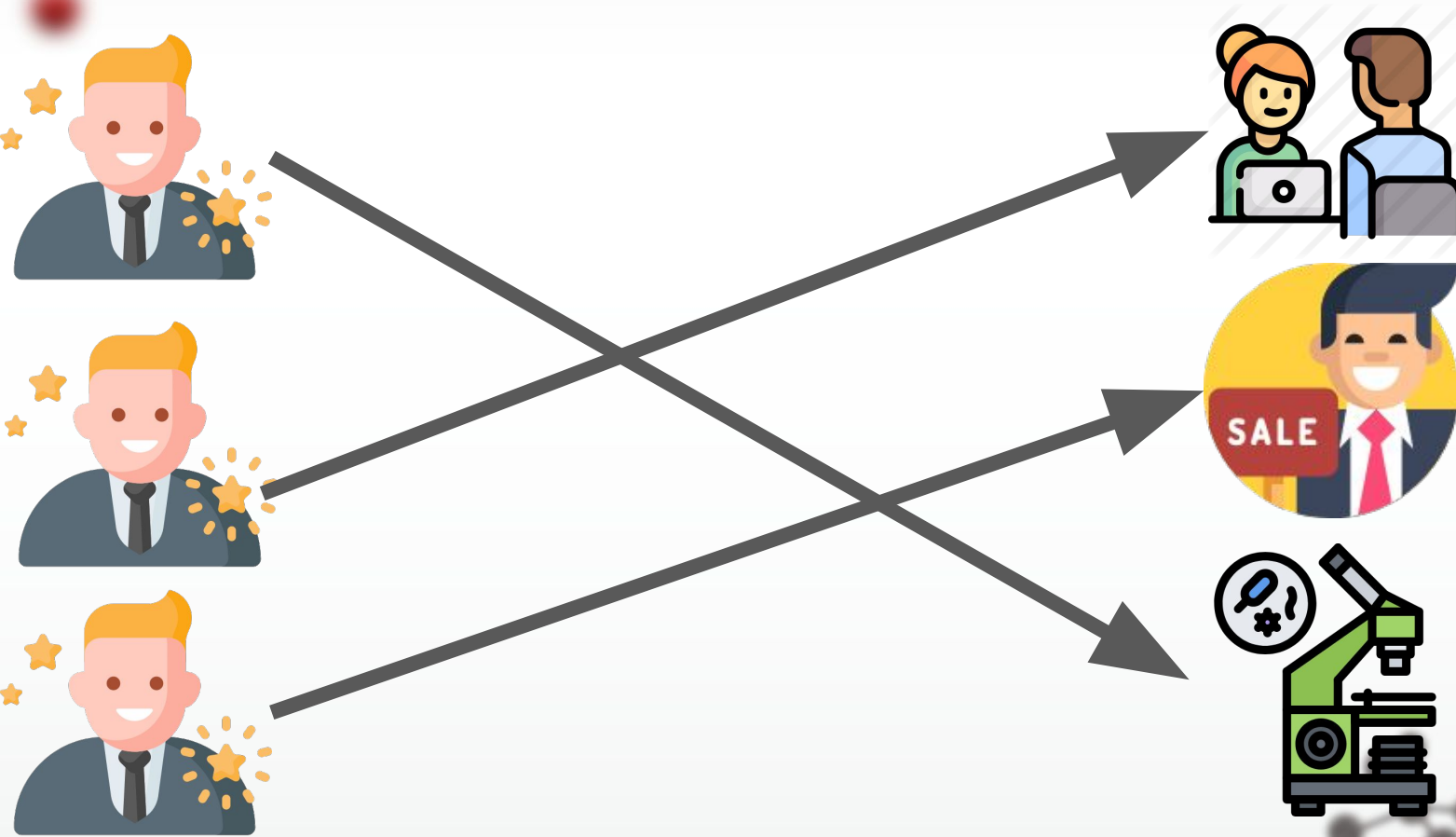
Jimmy possesses the

- Skill set
- Experience
- Traits

That the company wants in an employee



2. Matching applicants to the right department





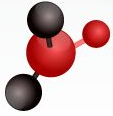
2. Matching applicants to the right department



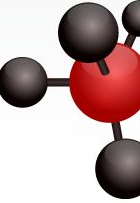
At the bootcamp, Jimmy can get a first-hand experience of life at the department or job role he applied for.

Now, both the company and Jimmy gain a better understanding - the company learns more about the candidates and Jimmy learns more about the what the job opening entails.





Significant Variables to retain employees



- | | | | | | |
|----|-------------------|----|---|----|----------------------------------|
| 1. | Overtime | 3. | Environment &
Relationship
Satisfaction | 5. | Frequency of
Business Travel |
| 2. | Work-life Balance | 4. | Job Involvement | 6. | Expected %
increase in salary |



3. Cultivating positive company culture



1. Overtime

Abolish or limit overtime.

2. Work-life Balance

Encourage work-life balance for all employees

3. Environment & Relationship Satisfaction

Improve the work space, create conducive environment.

4. Job Involvement

Allow employees to take on more responsibilities

5. Frequency of Business Travel

Regulate the frequency to minimize homesickness.

6. Expected % increase in salary



Limitations

1.Dataset

2.Models

3.Business
Recommendations



Limitations of the dataset

Fictional in Nature

Randomly generated dataset may cause analysis and insights to lack meaning pertaining to a real world context



Small proportion of termination

Extremely small proportion of the data set is employees terminated. This would result in unreliable results.



Lack of variables

To solve the business problem, we required more variables attainable at the hiring stage. However, there was a lack of such variables



Undefined variables

The variables provided by the data set were vaguely interpreted and open to interpretation which can cause issues with the modelling process.





Limitations of the model

Accuracy is only fairly high

Accuracy of the model was traded off since a lower threshold was set - this allows the company to take action on a larger pool of potential resignees



Difficulty in interpreting splitting variables

For certain variables, the splitting criteria included combinations of categories that made it difficult to draw insights from..



CART Complexity

Based on the complexity parameter table, the suggested tree would have thousands of terminal nodes. We had to exercise human judgement to attain an optimal tree which is subjective in nature.





Limitations of the business recommendations

The hiring process is susceptible to false information.

Applicants could easily lie on their resumes and the model would still shortlist them. This might end up being counterproductive.



Limited profiling

Based on the data set, the CART model could only predict a set amount of profiles which could limit the effectiveness of the job matching process..



Cost of cultivating positive company culture

There is an opportunity cost to retain employees as well. Efforts to keep employees satisfied require money could affect the company's profits.



Turnover is a natural occurrence

Despite efforts on the company's part, turnover is a natural occurrence. Other factors like office politics and career progression could still cause employees to leave.





Conclusions

Despite the restraints, we are confident that our solutions and recommendations would be able to match the right person to the right job opening and alleviate the problems faced by companies due to employee attrition.

We believe the future of HR lies in analytics.



Thank you!

