The Business Analysts

HR Solutions

Team 2
Chua Hui Min (U1810936F)
Feng Jiayi (U1810441J)
Lee Xuan Yu (U1710806H)
Wong Xin Li (U1811431C)

AGENDA

1

Business Problem

Too many resumes? We are here to solve it!

2

Resume Parsing

How to select only a portion out of the millions of resumes?

3

Research

What data do we need to collect?

4

IBM Case Study: Predictive Modelling

Predicting (i) Years at Company, (ii) Job Performance, and (iii) Job Satisfaction

5

Our Recommendations

What we recommend to you.

6

Q&A

Just ask ahead!

Business Problem

Business Problem

“

Everyday, I receive 1,000 resumes and it is hard and time-consuming to read all of them.

- *HR Specialist*

Business Problem

“

Hiring and managing people is costly and time-consuming. I wish there is a way to find those who will stay longer with the company, are happy to work here, and will perform well here.

- *HR Manager*

WE HEAR YOU

1 Resume Parsing

Sift out the resumes you want

- Reduce the number of resumes to scan through
- More time to do other things

2 Model 1: Years at Company

Predict how long someone will stay with the company (company loyalty)

- Lower turnover = lesser replacement costs / training costs

3 Model 2: Job Satisfaction

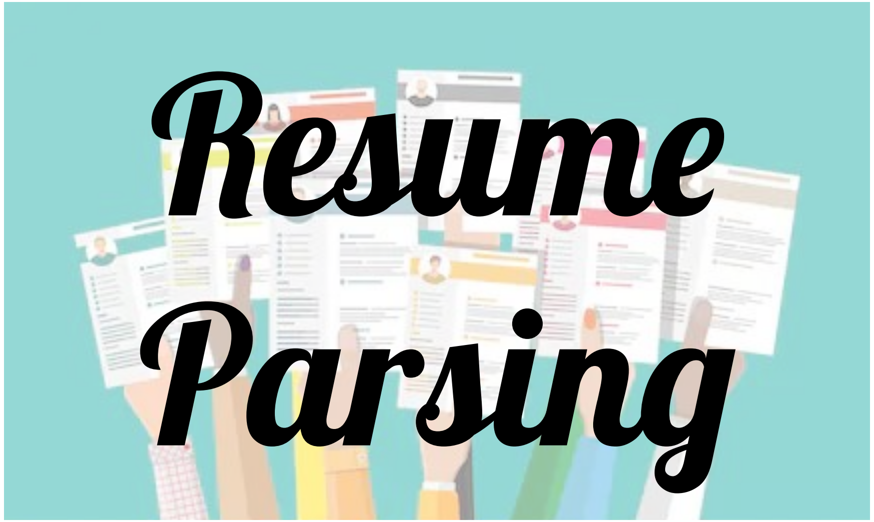
Predict how happy someone will be in the company

- Happier employees = better productivity, indirect impact on culture and surroundings
- More goal congruent

4 Model 3: Job Performance

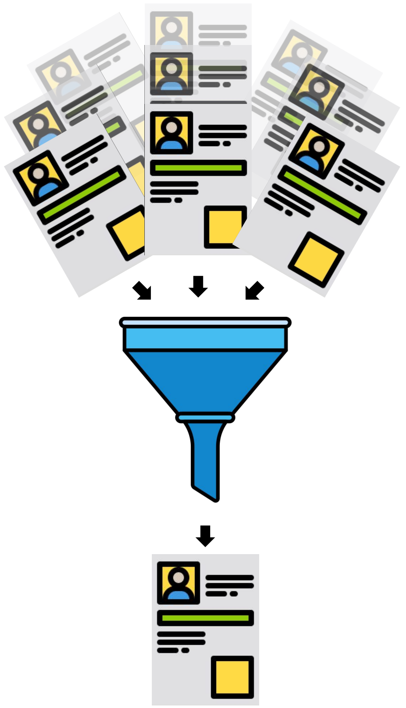
Predict how well someone will perform in the company

- Better job performance = more productive = better profits


Resume
Parsing

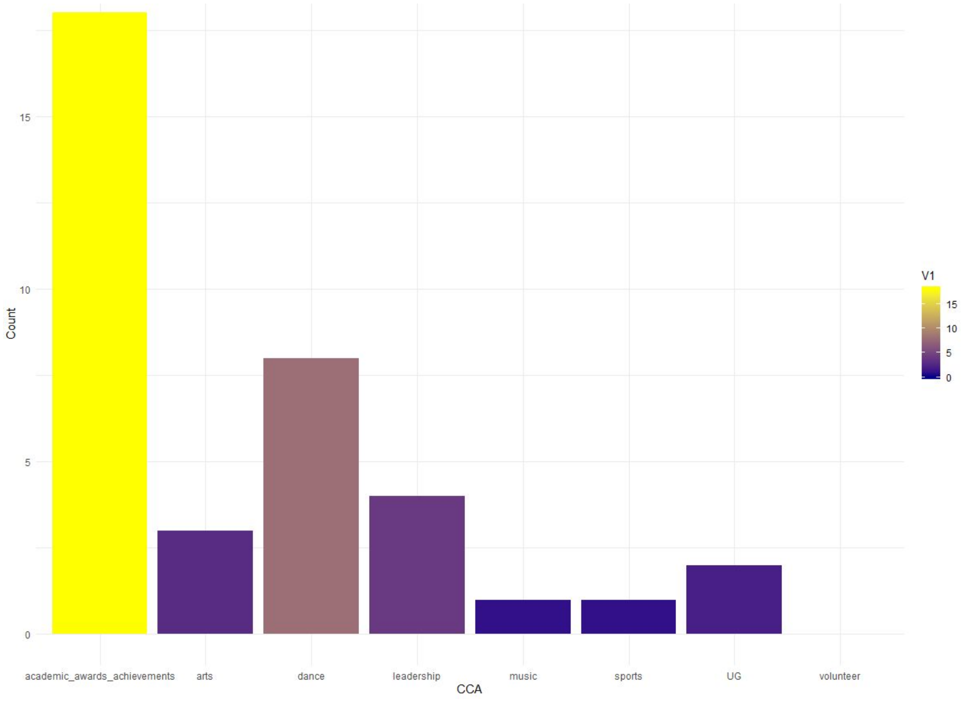
Resume Parsing

- Pass all resumes through a formula
- Calculates score based on frequency of words
- Based on CCAs / hobbies:
  - Academics / Awards / Achievements
  - Arts
  - Dance
  - Leadership
  - Music
  - Sports
  - Uniformed Groups
  - Volunteering
- Companies that value certain types of people can easily filter for them
- Can look at either (i) score in category, or (ii) total score
- Only those that pass a certain yardstick / only the top % of resumes get further processed



Resume Parsing

	document	dance	music	arts	volunteer	sports	UG	leadership	academic_awards_achievements	totalscore
1	LEE XUAN YU_Resume	8	1	3	0	1	2	4	18	37



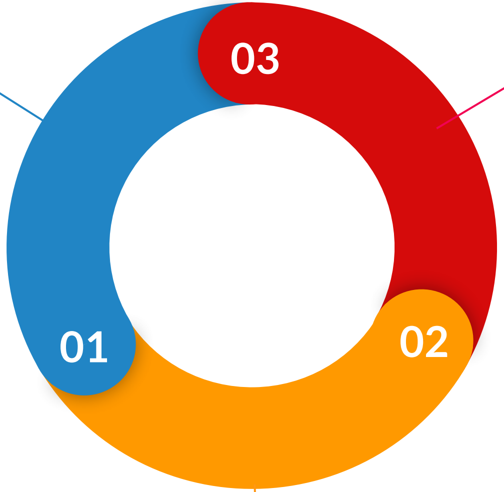
Benefits:

- Cut down on time taken to process resumes
- Auto-selection of only a portion of resumes based on stated criteria (can tailor to company's needs)
- Only target those the company wishes to target = save resources, more effective targeting

Research

Research

Number of Years Candidate Will Stay in Company
Loyalty levels;
Big Five Personality Traits (OCEAN)



Job Performance
Cognitive ability;
Conscientiousness;
Growth mindset / motivation

Job Satisfaction
Big Five Personality Traits (OCEAN);
Years of Experience on the Job

Research

1. Number of Years Candidate Will Stay in Company

Loyalty to Company

- Measured by the average number of years spent in previous jobs
- Indicator of the length of stay in a company
- Candidates with lower average length of stay in previous jobs exhibit job-hopping behaviour and will be likely to work for the company for a shorter amount of time.

Research

1. **Number of Years Candidate Will Stay in Company**

Big 5 Personality Traits (OCEAN)

Openness to experience

Individual's level of intellectual curiosity, creativity and preference for fresh ideas and variations

> Good to have high levels

Conscientiousness

Individual's level of intellectual curiosity, creativity and preference for fresh ideas and variations

> Good to have high levels

Extraversion

Individual's level of energy, positive emotions, assertiveness, sociability, talkativeness, and his/her tendency to seek stimulation in the company of others

> Good to have high levels

Research

1. **Number of Years Candidate Will Stay in Company**

Big 5 Personality Traits (OCEAN)

Agreeableness

Individual's tendency to be compassionate and cooperative towards others rather than suspicious and antagonistic

> Good to have high levels

Neuroticism

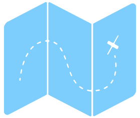
Individual's level of emotional stability and impulse control

> Good to have low levels

Research

2. Job Satisfaction

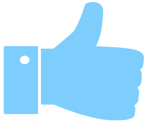
Big 5 Personality Traits (OCEAN)



Openness to experience

Conscientiousness

Extraversion



Agreeableness

Neuroticism

Years of Working Experience

Job satisfaction may increase with longer working experience because individuals would have gained a better understanding of their own wants and capabilities, thus look for jobs that better suit themselves

Research

3. Job Performance

Cognitive ability

Level of the mind's ability to learn, remember, and pay attention

Conscientiousness

Whether an individual is dutiful and thorough

Growth mindset / motivation

Level of willingness an individual is willing to learn new things

Research

4. Demonstration for Model Building

IBM HR Dataset

Turnover Dataset

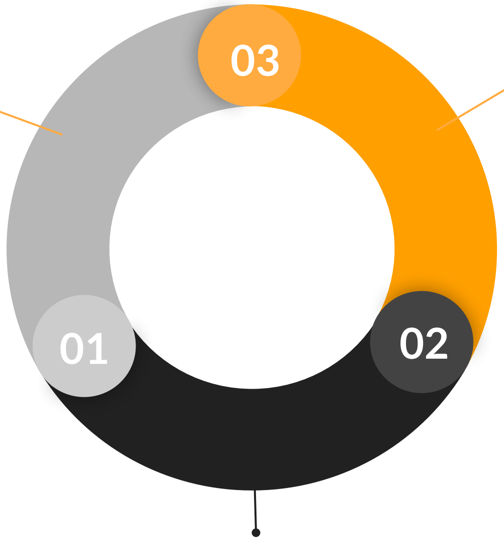
Assume that this dataset is applicable to our overall IBM case study despite being of a different origin

Predictive Modelling

Predictive Models

Length of Stay in Company

- CART model (IBM HR dataset)
- Linear Regression model (turnover dataset)



Job Performance

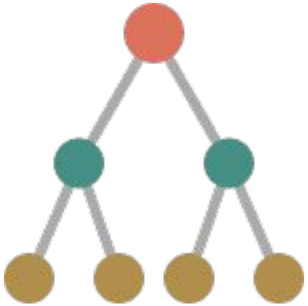
- CART model (IBM HR dataset)

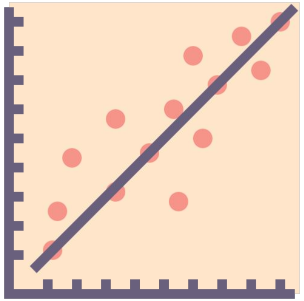
Job Satisfaction

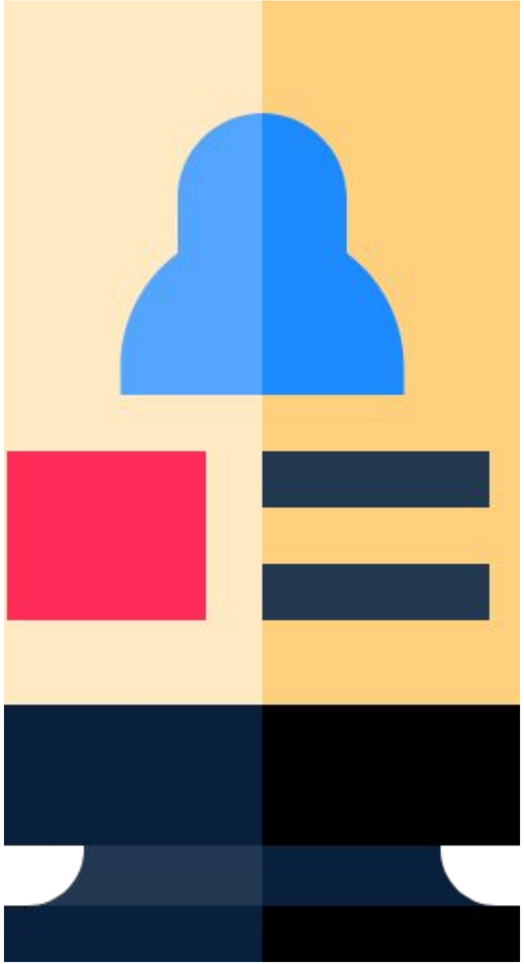
- Logistic Regression model (IBM HR dataset)

Predictive Models

1. Purpose is to make predictions about a ***candidate*** → only input variables that existed at the point of application for the job were used to build the models
2. Our optimal CART models have too many splits → impractical to analyse all decision rules → artificially pruned the tree to see the top few decision rules
3. Note that using a different dataset would yield different optimal models → crucial for our clients to provide us with accurate data of their own employees







Model 1: Predicting Length of Stay using General Characteristics and Background

Step 1: Filter the dataset to select only those employees who had voluntarily resigned

```
hr1.dt <- hr.dt[Attrition == "Voluntary Resignation"]
```

Step 2: Find the maximal tree using input variables that existed at the point of application for the job

```
cart1 <- rpart(YearsAtCompany ~
AgeJoined+DistanceFromHome+Education+EducationField+Ge
nder+NumCompaniesWorked+TotalYearsBeforeJoining+Employ
eeSource, data = hr1.dt, method = 'anova', cp = 0)
```

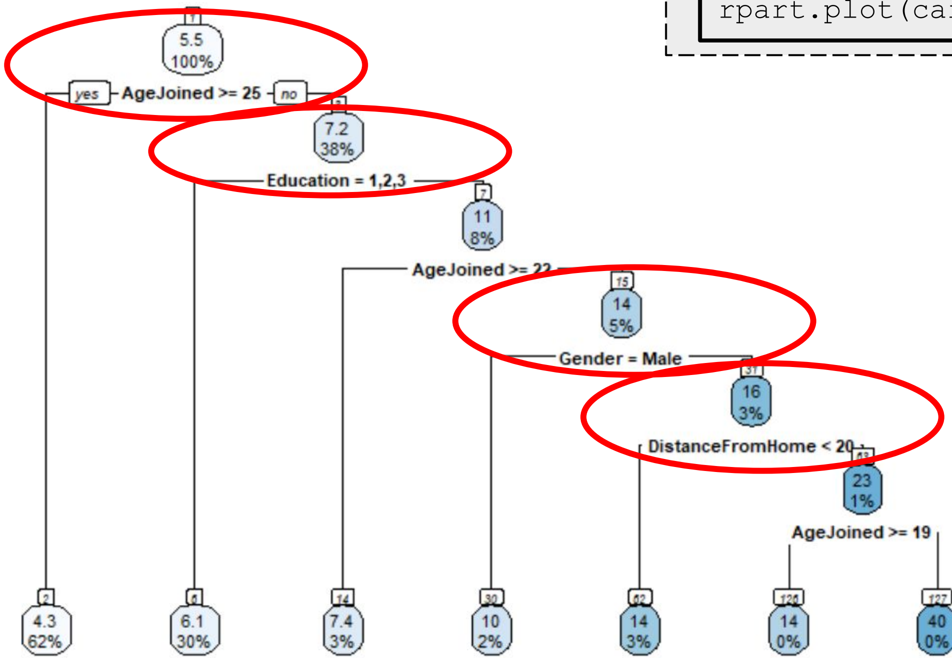
Step 3: Set cp to a value such that the pruned tree only has 6 splits, as the optimal tree is too large to analyse

```
rpart.plot(cart1, nn = T, main = "Maximal Tree")
printcp(cart1, digits = 3)
cp.opt <- 0.019
```

Model 1: Insights

Step 4: Prune the maximal tree and plot the pruned tree

```
cart2 <- prune(cart1, cp = cp.opt)
rpart.plot(cart2, nn = T, tweak = 1.3)
```



Insights:

1. Employees who are 25 or older are predicted to stay for a shorter time
2. For employees who are younger than 25, those who are more educated tend to stay for a longer time
3. For employees who are younger than 25 and are more educated, females tend to stay for a longer time
4. For these females, those who stay further away from the company surprisingly tend to stay longer

Model 2: Predicting Length of Stay using Psycho-Emotional Traits (OCEAN)

Step 1: Filter the dataset to select only those employees who are no longer in the company

```
ocean1.dt <- ocean.dt[event == 1]
```

Step 2: Run linear regression using input variables that existed at the point of application for the job

```
m1 <- lm(YearsAtCompany ~
gender+AgeJoined+Extraversion+Agreeableness+Conscientiousne
ss+Neuroticism+OpennessToExperience, data = ocean1.dt)
summary(m1)
```

Model 2: Predicting Length of Stay using Psycho-Emotional Traits (OCEAN)

```
> summary(m1)

Call:
lm(formula = YearsAtCompany ~ gender + AgeJoined + Extraversion +
    Agreeableness + Conscientiousness + Neuroticism + OpennessToExperience,
    data = ocean1.dt)

Residuals:
   Min     1Q Median     3Q    Max
-4.460 -1.790 -0.307  1.331  8.802

Coefficients:
                     Estimate Std. Error t value              Pr(>|t|)
(Intercept)            5.8447     1.4766    3.96              0.000085 ***
genderm                0.4407     0.2444    1.80                0.0719 .
AgeJoined             -0.1786     0.0120  -14.92 < 0.0000000000000002 ***
Extraversion          -0.1026     0.0775   -1.32                0.1861
Agreeableness          0.0397     0.0808    0.49                0.6231
Conscientiousness      0.1800     0.0790    2.28                0.0231 *
Neuroticism            0.0533     0.0784    0.68                0.4969
OpennessToExperience   0.1700     0.0640    2.65                0.0082 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.31 on 563 degrees of freedom
Multiple R-squared:  0.298,	Adjusted R-squared:  0.289
F-statistic: 34.2 on 7 and 563 DF,  p-value: <0.0000000000000002
```

Insight from step 2:

Only ‘AgeJoined’, ‘Conscientiousness’ and ‘OpennessToExperience’ are significant factors in predicting ‘YearsAtCompany’

Model 2: Insights

Step 3: Run linear regression again, this time using only the significant factors from step 2

```
m2 <- lm(YearsAtCompany ~
AgeJoined+Conscientiousness+OpennessToExperience, data =
ocean1.dt)
summary(m2)
```

```
> summary(m2)

Call:
lm(formula = YearsAtCompany ~ AgeJoined + Conscientiousness +
    OpennessToExperience, data = ocean1.dt)

Residuals:
   Min     1Q Median     3Q    Max
-4.301 -1.770 -0.344  1.286  8.747

Coefficients:
                     Estimate Std. Error t value
(Intercept)            5.3810     0.6984    7.70
AgeJoined             -0.1747     0.0119  -14.66
Conscientiousness      0.2361     0.0611    3.87
OpennessToExperience   0.1854     0.0622    2.98
                                Pr(>|t|)
(Intercept)            0.0000000000000059 ***
AgeJoined            < 0.0000000000000002 ***
Conscientiousness                 0.00012 ***
OpennessToExperience              0.00301 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.32 on 567 degrees of freedom
Multiple R-squared:  0.286,     Adjusted R-squared:  0.282
F-statistic: 75.5 on 3 and 567 DF,  p-value: <0.0000000000000002
```

Insights:

1. The younger the employee is when he joined the company, the longer he would stay
2. The more conscientious the employee is, the longer he would stay. This agrees with our hypothesis.
3. The more open to experience the employee is, the longer he would stay. This agrees with our hypothesis as well.

Model 3 (Logistic Model): Predicting Job Satisfaction

Step 1: Run multinominal logistic regression using input variables that existed at the point of application for the job

```
m4 <- multinom(JobSatisfaction ~
AgeJoined+DistanceFromHome+Education+EducationField+Gender
+NumCompaniesWorked+TotalYearsBeforeJoining+EmployeeSource
, data = hr.dt)
summary(m4)
```

Step 2: Use p-value test and confidence interval test to assess the statistical significance of the determining variables

```
z <- summary(m4)$coefficients/summary(m4)$standard.errors
pvalue <- (1 - pnorm(abs(z), 0, 1))*2 # 2-tailed test
p-values
pvalue

OR.CI <- exp(confint(m4))
OR.CI
```

Model 3 (Logistic Model): Predicting Job Satisfaction

Step 3: Run the new multinominal logistic regression using only those input variables that are statistically significant

```
m5 <- multinom(JobSatisfaction~ Education+TotalYearsBeforeJoining+EmployeeSource ,
data = hr.dt)
summary(m5)
```

```
> summary(m5)
Call:
multinom(formula = JobSatisfaction ~ Education + TotalYearsBeforeJoining +
    EmployeeSource, data = hr.dt)

Coefficients:
  (Intercept) Education2 Education3 Education4 Education5 TotalYearsBeforeJoining
2      -0.148    -0.0405    0.00255     -0.194     -0.486                 -0.0079
3       0.653    -0.3673   -0.26036     -0.260     -0.506                 -0.0131
4       0.642    -0.1868   -0.21439     -0.213     -0.377                 -0.0105
  EmployeeSourceJob Portal EmployeeSourceReferral
2                   0.2648                   1.48
3                   0.1026                   1.09
4                   0.0462                   1.17
```

Model 3 (Logistic Model): Predicting Job Satisfaction

$\boldsymbol{Y}$**=2:** z_2 = -0.148 + 0.0405(Education2) - 0.00255(Education3) - 0.194 (Education4) - 0.486(Education5) - 0.0079(TYBJ) + 0.2648(EmployeeSourceJob) + 1.48(EmployeeSourceReferral)

$\boldsymbol{Y}$**=3:** z_3 = 0.653 - 0.3673(Education2) - 0.26036(Education3) - 0.260(Education4) - 0.506(Education5) - 0.0131(TYBJ) + 0.1026(EmployeeSourceJob) + 1.09(EmployeeSourceReferral)

$\boldsymbol{Y}$**=4:** z_4 = 0.642 - 0.1868(Education2) - 0.21439(Education3) -0.213(Education4) - 0.377(Education5) - 0.0105(TYBJ) + 0.0462(EmployeeSourceJob) + 1.17(EmployeeSourceReferral)

Model 3 (Logistic Model): Insights

Step 4: Find the Odds Ratio of the various input variables

```
OR <- exp(coef(m5))
OR
```

```
OR <- exp(coef(m5))
OR
(Intercept) Education2 Education3 Education4 Education5 TotalYearsBeforeJoining
      0.862      0.960      1.003      0.824      0.615                   0.992
      1.922      0.693      0.771      0.771      0.603                   0.987
      1.901      0.830      0.807      0.808      0.686                   0.990
EmployeeSourceJob Portal EmployeeSourceReferral
                    1.30                   4.39
                    1.11                   2.97
                    1.05                   3.22
```

1. For the continuous input TotalYearsBeforeJoining, $OR(Y=4) = 0.990$
   → For one unit increase in TotalYearsBeforeJoining, odds of high Job Satisfaction decrease by a factor of 0.990, all other variables held constant
   → Candidate preferably have fewer working years before joining company

Model 3 (Logistic Model): Insights

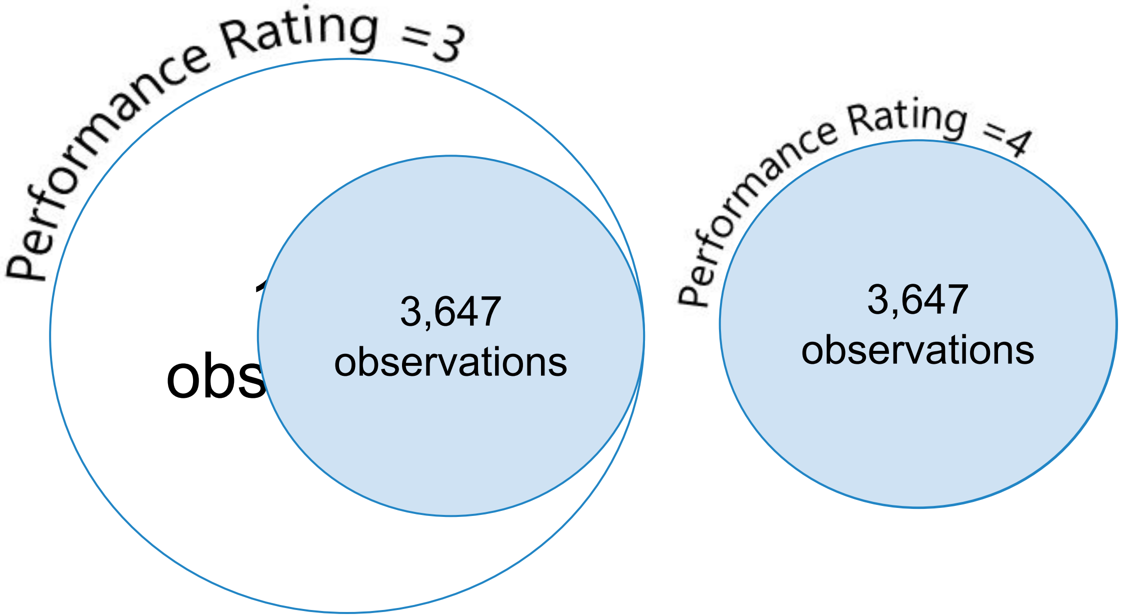
Step 4: Find the Odds Ratio of the various input variables

```
OR <- exp(coef(m5))
OR
```

```
OR <- exp(coef(m5))
OR
(Intercept) Education2 Education3 Education4 Education5 TotalYearsBeforeJoining
      0.862      0.960      1.003      0.824      0.615                   0.992
      1.922      0.693      0.771      0.771      0.603                   0.987
      1.901      0.830      0.807      0.808      0.686                   0.990
EmployeeSourceJob Portal EmployeeSourceReferral
                    1.30                   4.39
                    1.11                   2.97
                    1.05                   3.22
```

2. For categorical input EmployeeSource, $OR(Y\!=\!4)$ = 3.22
→ If the variable EmployeeSource is referral, odds of achieving high Job Satisfaction increase by a factor of 3.22, all other variables held constant
→ Candidates who are referred to the company are preferred

Model 4 (CART): Predicting Performance Rating



Step 1: Random sampling to get only 3,647 records with a PerformanceRating of “3”

```
hr.dt$ID <-
seq.int(nrow(hr.dt))

RNGlist <-
sample(hr.dt[PerformanceRatin
g == 3, ID], 3647, replace =
F)

pr3 <- hr.dt[ID %in% RNGlist]

pr4 <-
hr.dt[PerformanceRating == 4]

hr2.dt <- merge(pr3, pr4, all
= T)
```

Model 4 (CART): Predicting Performance Rating

Step 2: Find the maximal tree using input variables that existed at the point of application for the job.

```
cart4 <- rpart(PerformanceRating ~ 
AgeJoined+DistanceFromHome+Education+EducationField+Ge
nder+NumCompaniesWorked+TotalYearsBeforeJoining+Employ
eeSource, data = hr2.dt, method = 'anova', cp = 0)
```

Step 3: Set cp to get the number of splits of the optimal tree.

```
rpart.plot(cart4, nn = T, main = "Maximal Tree")
printcp(cart4, digits = 3)
cp.opt <- 0.000300369
```

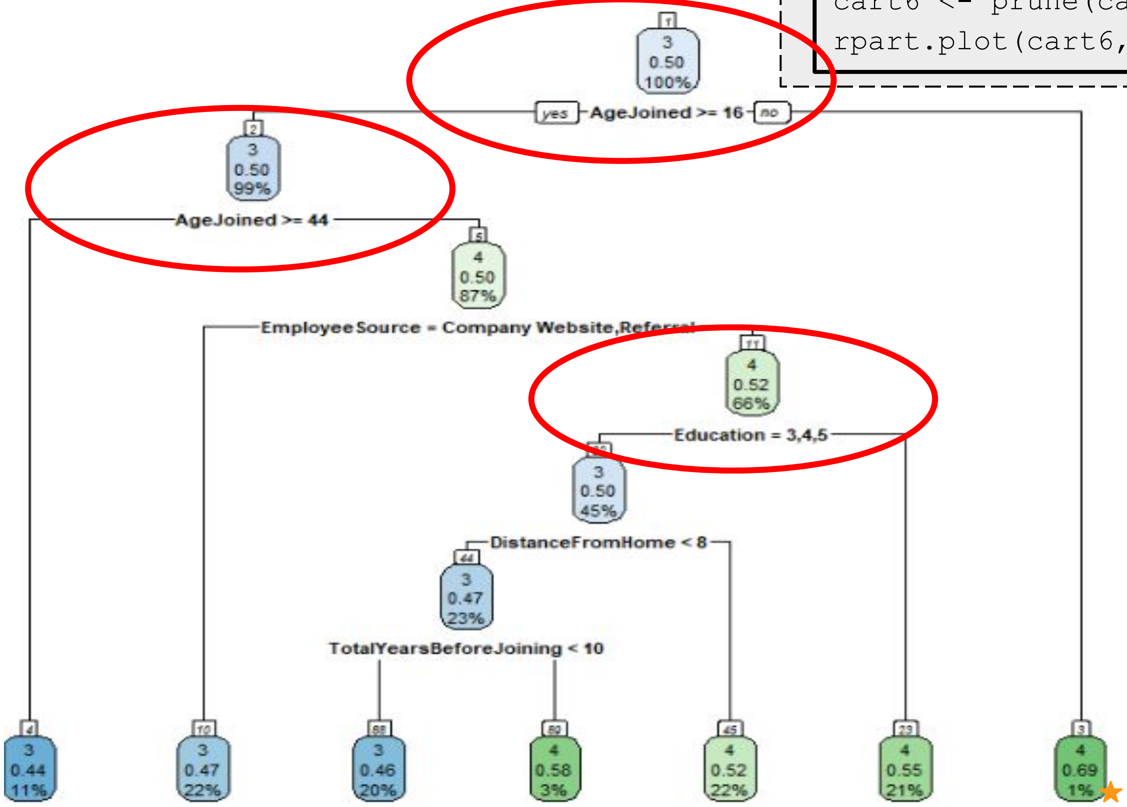
For visualisation purpose, we set the cp such that the pruned tree only has 6 splits.

```
cp.opt <- 0.0085001372
```

Model 4 (CART): Insights

Step 4: Prune the maximal tree and plot the pruned tree.

```
cart6 <- prune(cart4, cp = cp.opt)
rpart.plot(cart6, nn = T, tweak = 1.3)
```

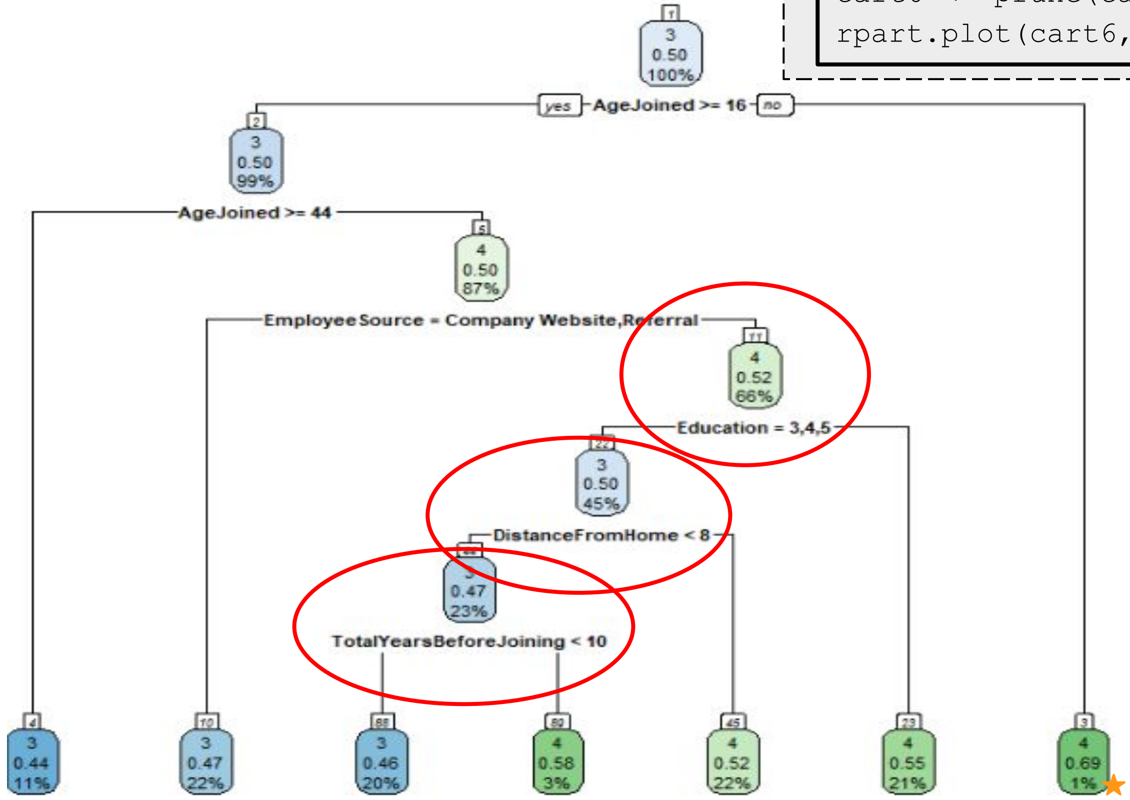


1. Candidates who join the company between ages 16 to 43 are more likely to perform better than those who join when they are above 43.
2. Candidates with a maximum of a College education are also likely to perform well.

Model 4 (CART): Insights

Step 4: Prune the maximal tree and plot the pruned tree.

```
cart6 <- prune(cart4, cp = cp.opt)
rpart.plot(cart6, nn = T, tweak = 1.3)
```



3. For those with an education level of Bachelor / Master / Doctor:
   - they tend to perform well if they live at least 8 units away from the workplace.
   - if they live less than 8 units away from the workplace, then they tend to perform worse if they only have less than 10 years of job experience.

Recommendations for IBM

Years at Company

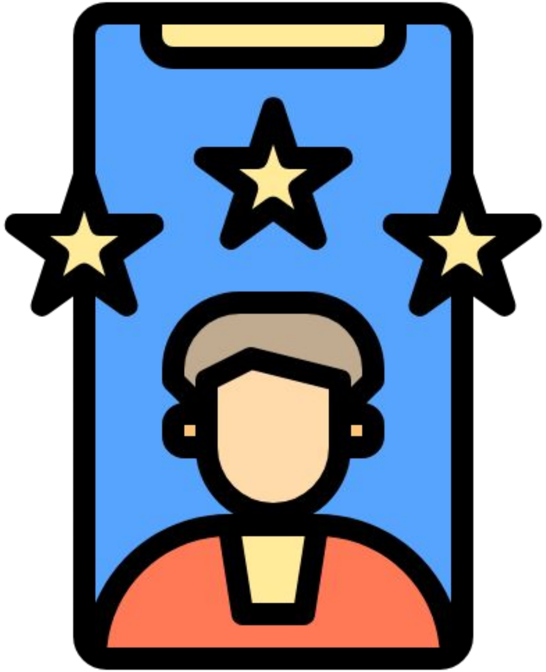
Job Satisfaction

Job Performance

Point System:

1. Give candidate a score out of 5 for each variable
    - scoring may be absolute according to a points rubrics (e.g. length of stay between 5 to 10 years is awarded 3 points)
    - or scoring may be relative across all the candidates (e.g. award 5 points to the candidate with the best result, and award the other candidates based on the ratio of their result to the best result multiplied by 5 points)

Recommendations for IBM



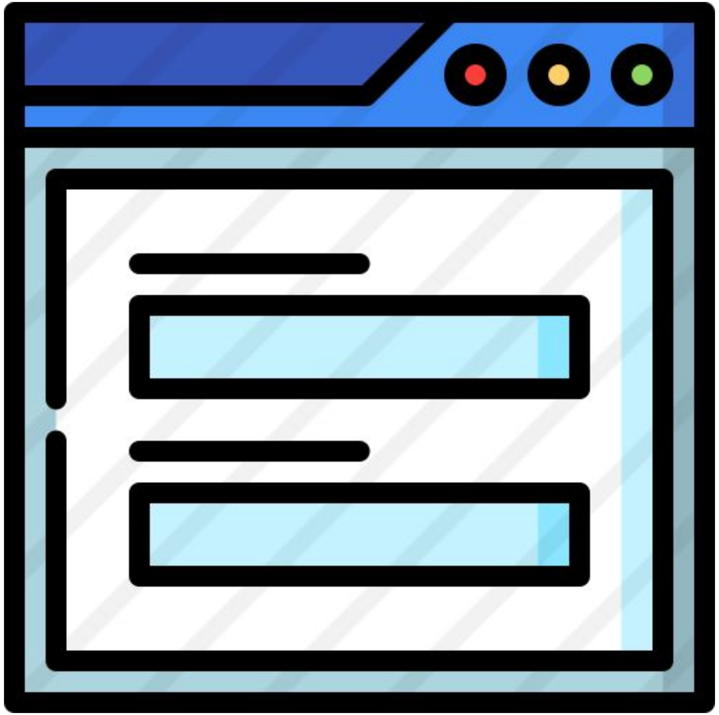
Point System:

2. IBM will total up the points to derive an overall score across all three variables

3. If IBM wishes for candidates to score at least moderately well in all the three aspects, they should first filter out those who do not have a minimum score of 3 for any variable

4. Candidates will then be ranked based on their overall scores

Recommendations for IBM



Data Collection:

1. Collect data on more possible explanatory variables from its employees (eg. cognitive ability which may help predict job performance)

2. Improve their data collection process, as there are a lot of missing data

Our Recommendations

Recommendations

1. Data collection through specially crafted forms

<table>
<tr><th>Years at Company</th><th>Job Satisfaction</th><th>Job Performance</th></tr>
<tr><td colspan="3">Age at the point of joining company, Distance from home, Education, Education field, Gender, Number of companies worked, Total working experience before joining company, Employee source</td></tr>
<tr><td colspan="2">OCEAN</td><td rowspan="2">Cognitive ability, Conscientiousness, Growth mindset / motivation</td></tr>
<tr><td>Average number of years spent in previous jobs</td><td>Years of Experience on the Job</td></tr>
</table>

Recommendations

2. Create predictive models

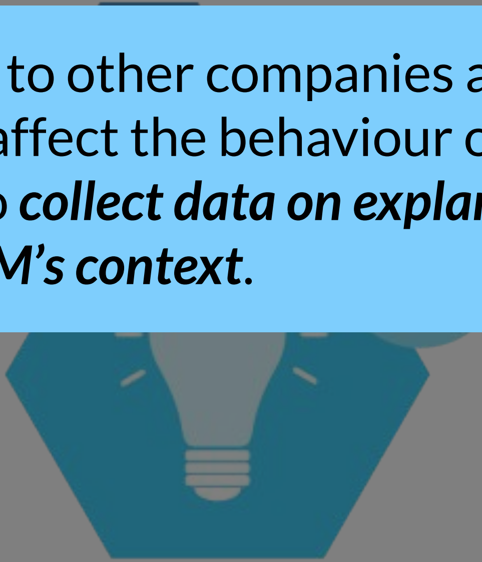
- Company as a whole
- Departments (if large dataset)
    - Take into account the **contextual differences** between departments so as to produce more accurate results

Recommendations

2. Create predictive models for each department

A company's predictive models ***may not be applicable*** to other companies as each company has their own ***unique context*** which may affect the behaviour of its employees. Hence, we still recommend our clients to ***collect data on explanatory variables which are insignificant in IBM's context***.

more accurate results



Thank You!
Questions?