Feature engineering

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Abhishek Trehan
People Analytics Practitioner

Feature engineering

- **Basic variables:** Set of variables available directly in a dataset
- **Derived variables:** Set of variables derived using data transformation of basic variables

Creating new features

- Age difference between an employee and their manager
- Job-hop index
- Employee tenure

Age difference

- Views
- Handling pressure
- Expectations
- Work ethics

Job-hopping

$$\text{Job-hop index} = \frac{\text{Total experience}}{\text{Number of companies worked}}$$

Employee tenure

- **Tenure**: duration of employment
- **Inactive** employees tenure

```
date_joining & last_working_date
```

- **Active** employees tenure

```
date_joining & cutoff_date
```

Deriving employee tenure

```
# Coercing date variables from dd/mm/yyyy format

library(lubridate)

org_final %>%
  mutate(date_of_joining = dmy(date_of_joining),
         cutoff_date = dmy(cutoff_date),
         last_working_date = dmy(last_working_date))
```

Calculating timespan

```
# Computing time span in years

library(lubridate)

date_1 <- ymd("2000-01-01")
date_2 <- ymd("2014-08-09")

time_length(interval(date_1, date_2), "years")
```

```
14.60274
```

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Compensation

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

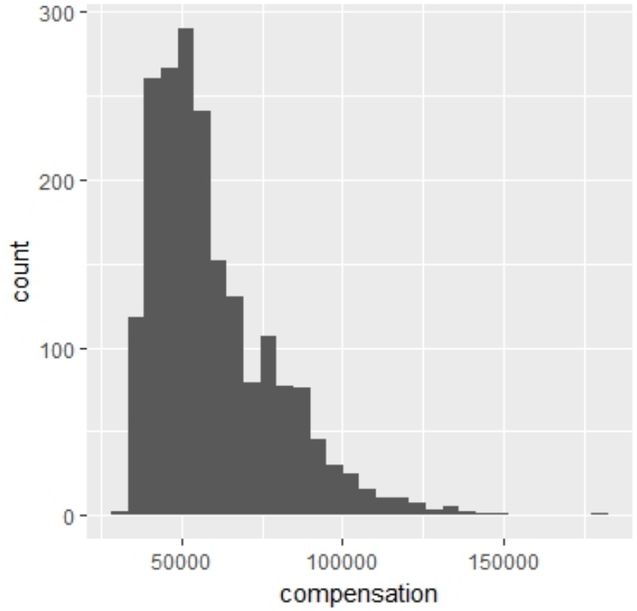
Abhishek Trehan
People Analytics Practitioner

Compensation matters

- Compensation is one the top drivers of employee turnover
- Pay matters for employee retention

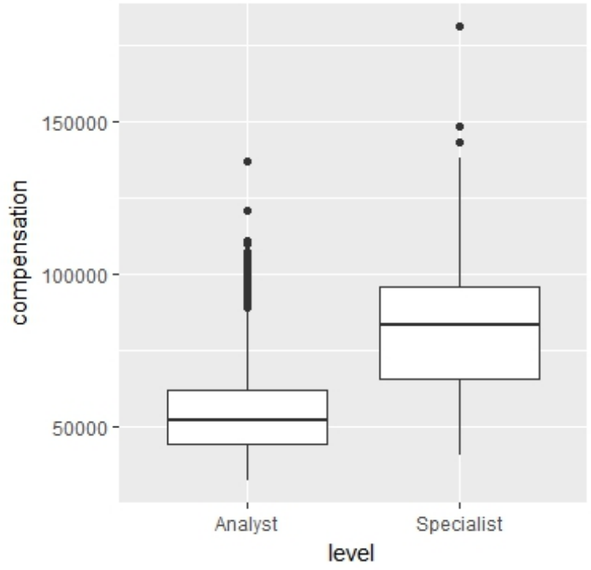
Exploring compensation variable

```
# Plot the distribution of compensation
ggplot(emp_tenure, aes(x = compensation)) +
  geom_histogram()
```



Exploring compensation variable

```
# Plot the distribution of compensation across levels
ggplot(emp_tenure,
       aes(x = level, y = compensation)) +
  geom_boxplot()
```

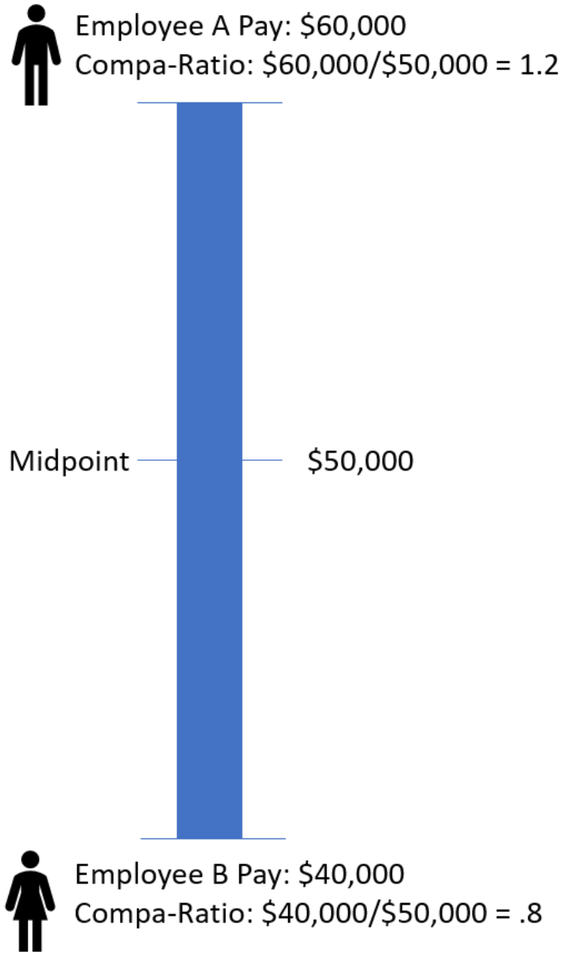


Deriving Compa-ratio

$$\text{Compa Ratio} = \frac{\text{Actual Compensation}}{\text{Median Compensation}}$$

Deriving Compa-ratio

- Compa-ratio of 1.2 or 120% means that the employee is paid 20% above the median pay
- Compa-ratio of 1 or 100% means that the employee is paid exactly the median pay
- Compa-ratio of 0.8 or 80% means that the employee is paid 20% below the median pay



Deriving median compensation & compa-ratio

```
# Derive Compa-ratio
emp_compa_ratio <- emp_tenure %>%
  group_by(level) %>%
  mutate(median_compensation = median(compensation),
         compa_ratio = (compensation / median_compensation))
```

```
# Look at the median compensation for each level
emp_compa_ratio %>%
  distinct(level, median_compensation)
```

```
# A tibble: 2 x 2
# Groups:     level[2]
  level      median_compensation
  <fct>                    <dbl>
1 Analyst                  51840
2 Specialist               83496
```

Deriving Compa-level

- Compa-ratio > 1: Above
- Otherwise: Below

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Information value

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Abhishek Trehan
People Analytics Practitioner

Understanding Information value

- Measure of the predictive power of independent variable to accurately predict the dependent variable
- Rank independent variables on the basis of their predictive power

Calculating Information value

$$IV = (\sum (\%\text{ of non-events - }\%\text{ of events})) * \log(\frac{\%\text{ of non-events}}{\%\text{ of events}})$$

Calculating Information value

```
# Load Information package
library(Information)
# Compute Information Value
IV <- create_infotables(data = emp_final, y = "turnover")
# Print Information Value
IV$Summary
```

```
                    Variable           IV
12                percent_hike 1.144784e+00
17           total_dependents 1.088645e+00
21            no_leaves_taken 9.404533e-01
31                     tenure 9.332570e-01
27          mgr_effectiveness 6.830020e-01
11               compensation 6.074885e-01
```

Information value (IV) table

Information value	Predictive power
< 0.15	Poor
Between 0.15 and 0.4	Moderate
> 0.4	Strong

- `percent_hike` : 1.14 (Strong)
- `compa_ratio` : 0.29 (Moderate)

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R