



EMPLOYEE TURNOVER PREDICTION WITH MACHINE LEARNING

[WENDY MARIA D'SA]

IS IT
IMPOR-
TANT TO
HAVE
GOALS?

YES!
YOU NEED
GOALS TO
SUCCEED.

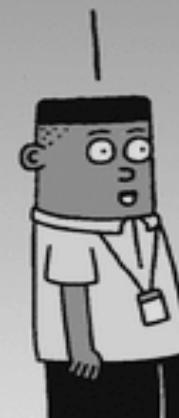


GOOD, BECAUSE MY
GOAL IS TO BECOME
AN UBER DRIVER.

I
I QUIT.

Dilbert.com @ScottAdamsSays

WHAT
IS YOUR
GOAL?



REDUCING
EMPLOYEE
TURNOVER.

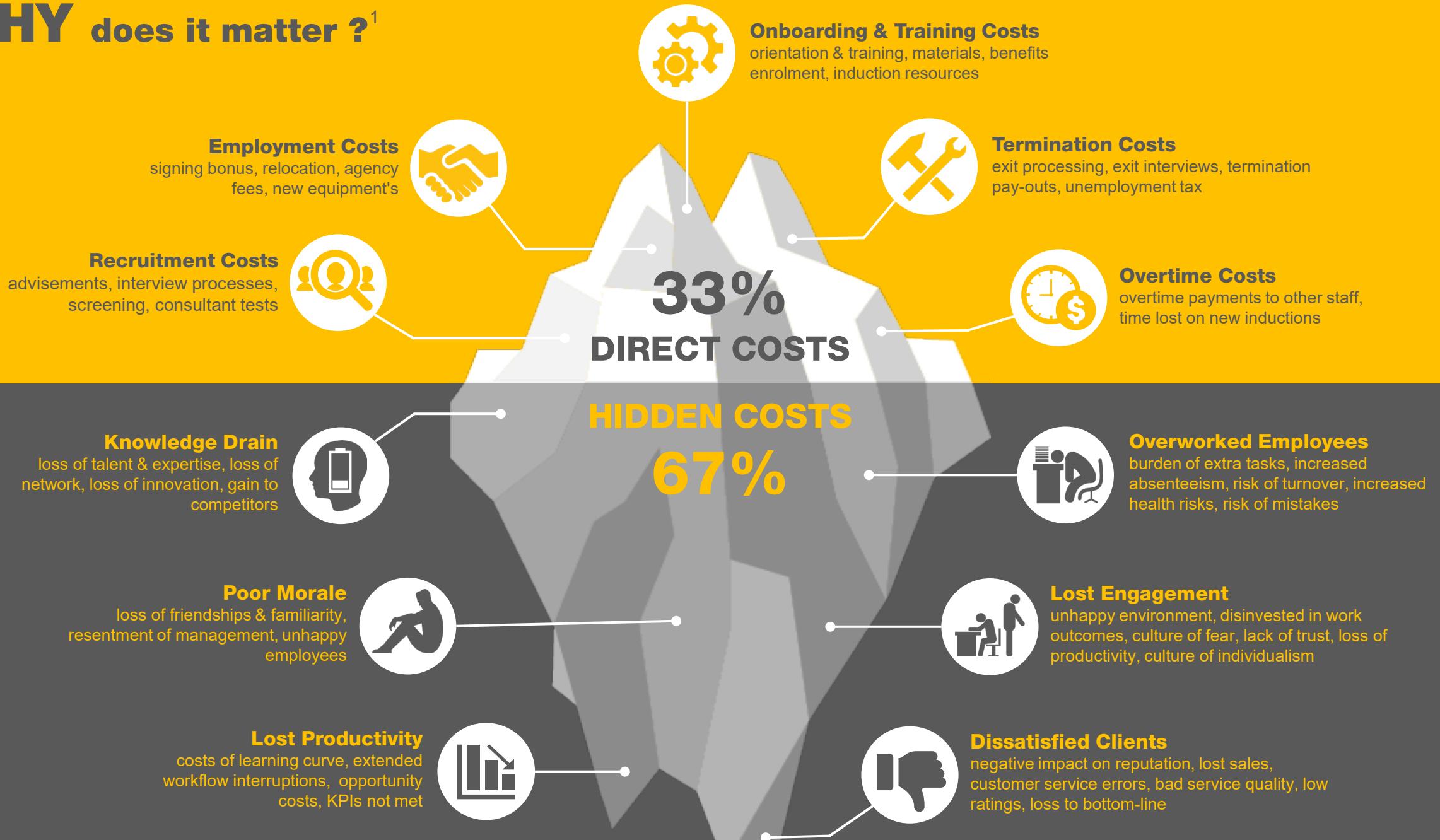


2-24-16 © 2016 Scott Adams, Inc. /Dist. by Universal Uclick

WHAT is Employee Turnover ?

- It is the **measurement of the number of employees who leave an organisation** during a specified time period, typically one year.
- It applies to an **entire organisation** but can also apply to **individual departments or demographic groups** within an organisation.

WHY does it matter ?¹



HOW much does it cost ?

UK³

In 2020, cost to employers was
~ \$100 Billion

- Retail & E-commerce 35.0%
- Hospitality 30.0%
- Administrative Services 19.3%
- Media & Publishing 17.7%
- Manufacturing 10.7%

USA²

In 2020, cost to employers was
~ \$600 Billion

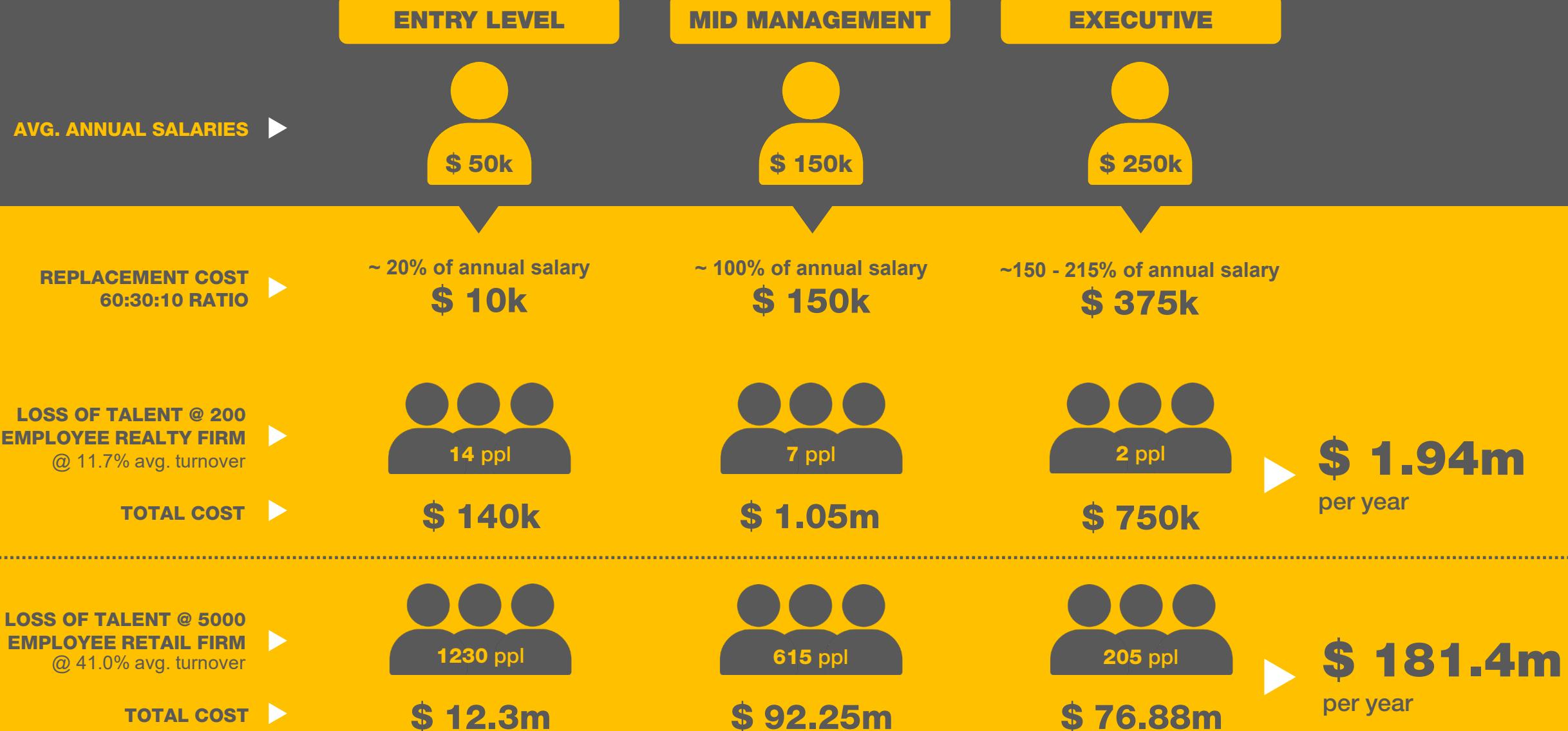
- Retail & E-commerce 30.7%
- Media & Entertainment 26.7%
- Technology 21.5%
- Life Science 20.6%
- Consulting 20.4%

AUSTRALIA⁴

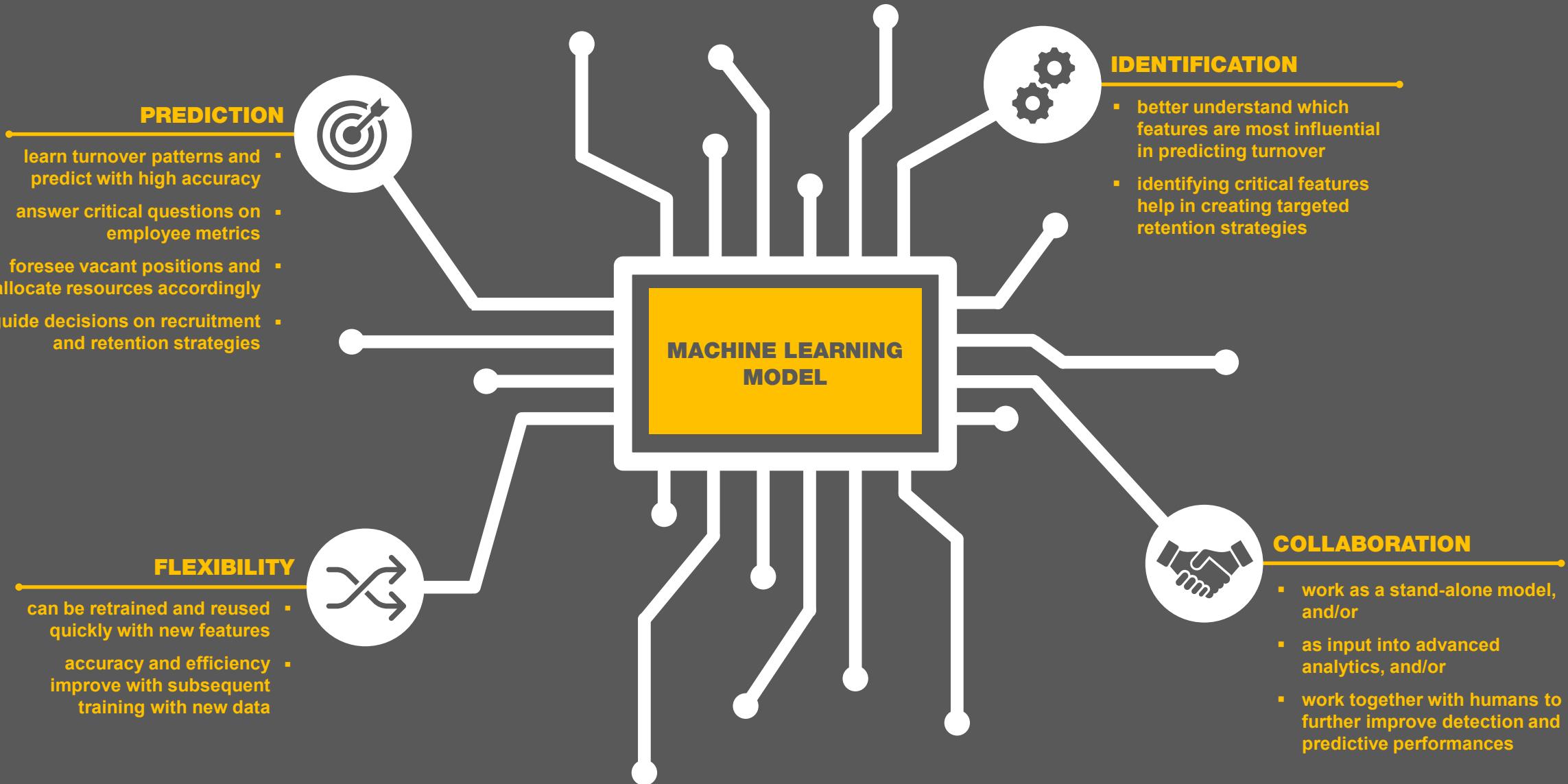
In 2019, cost to employers was
~ \$5 Billion

- Retail & E-commerce 41.0%
- Hospitality 18.0%
- Real Estate 11.7%
- Administrative Services 11.3%
- Utilities 10.7%

HOW much does it cost ?⁵



HOW can machine learning help reduce employee turnover ?



QUESTION

PROFILE



GOAL



ANSWER

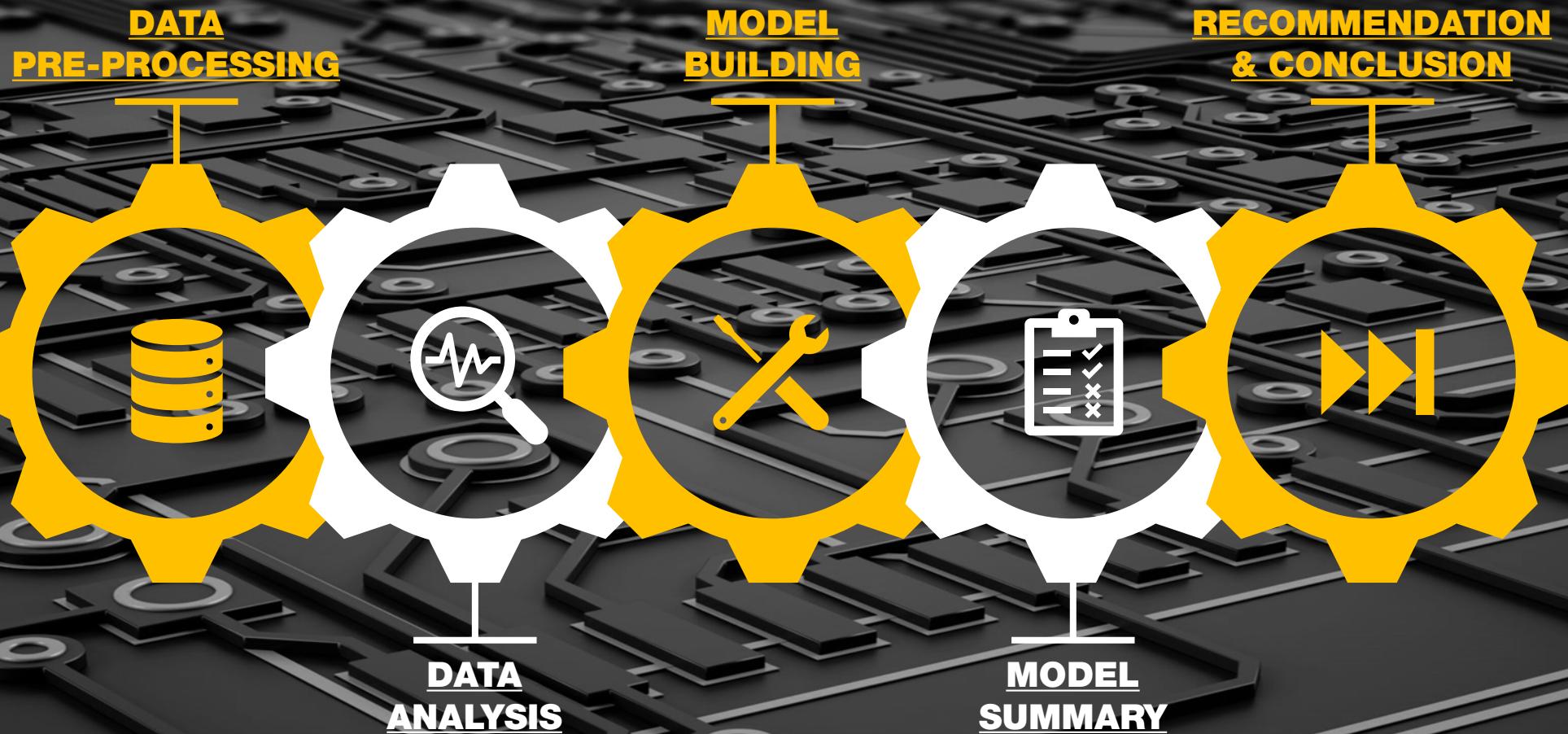


- COMPANY TYPE** ► Pharmaceutical
- COMPANY SIZE** ► 4000 - 4500 employees
- TURNOVER** ► 15.0% per year
- DIRECT COSTS** ► \$ 53 – 60 Million per year

Create a MVP
Machine Learning Prediction Model
that will assist the company reduce turnover

- WHO, WHEN, WHY of employee turnover ?
- What are the most important factors to reduce turnover ?
- What changes should be made ?

Pipeline



DATASET



DATA SOURCE

Kaggle ► <https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study>

Note ► Data consists of both numerical and categorical values



DATA PROCESSING

Observations ► 4410

About Employee ► 2 Numerical + 4 Categorical

About Job ► 10 Numerical + 6 Categorical + 3 Features engineered

About Time ► 2 Numerical time series + 3 Features engineered

About Satisfaction ► 3 Categorical

About Performance ► 2 Categorical



GENERAL DATA

- Age
- Gender
- Education level
- Education field
- Marital Status
- Distance from work
- Department
- Business travel
- Employee ID
- Job level
- Job role
- Monthly income
- No. of companies worked
- Percentage salary hike
- Standard hours
- Stock options
- Total working years
- Training time last year
- Years at company
- Last promotion
- Years with current manager
- Attrition
- Median compensation
- Compensation ratio
- Compensation level



EMPLOYEE SURVEY

- Employee ID
- Environment satisfaction
- Job satisfaction
- Work life balance



MANAGER SURVEY

- Employee ID
- Job involvement
- Performance rating

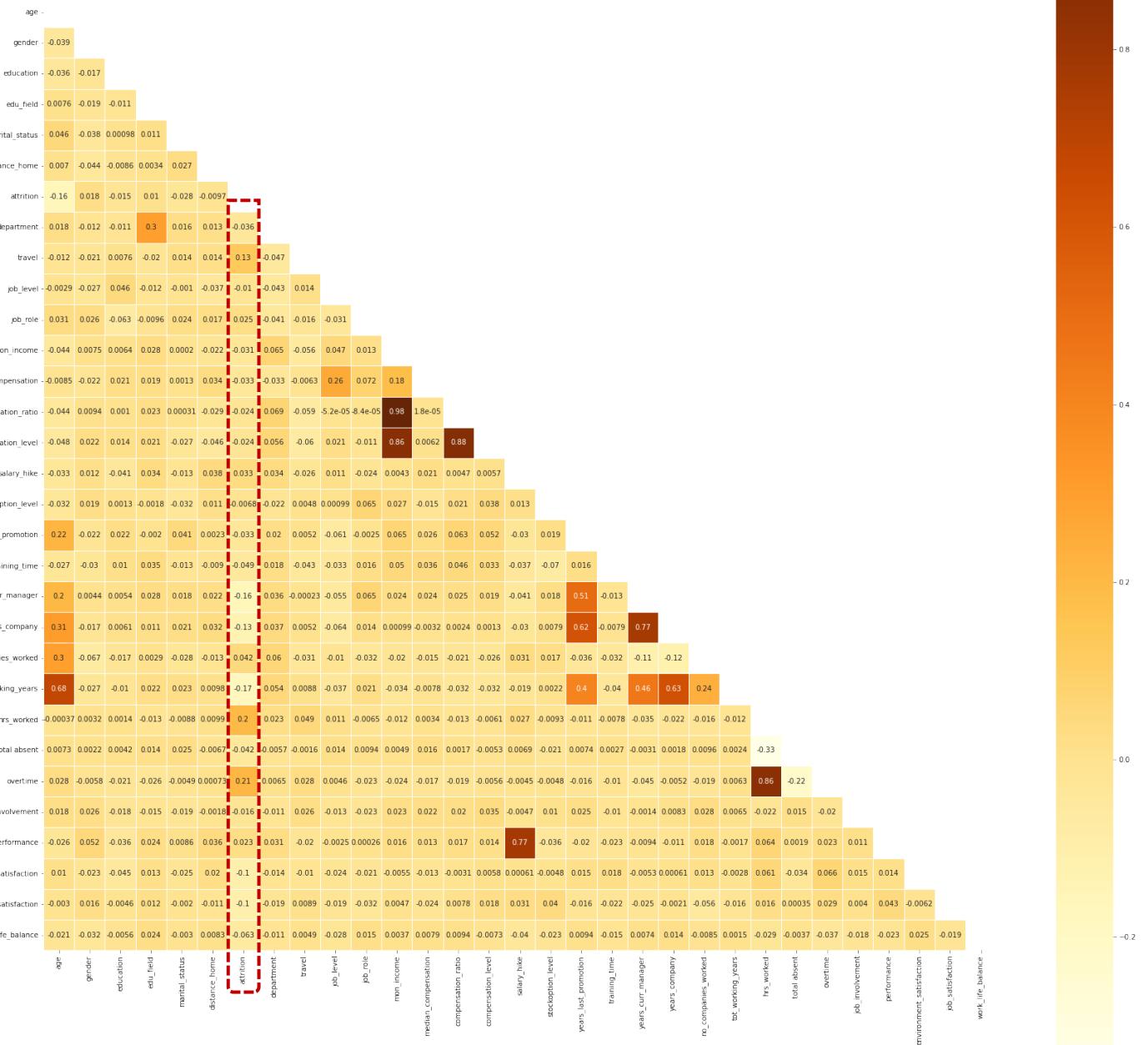


TIME IN | TIME OUT

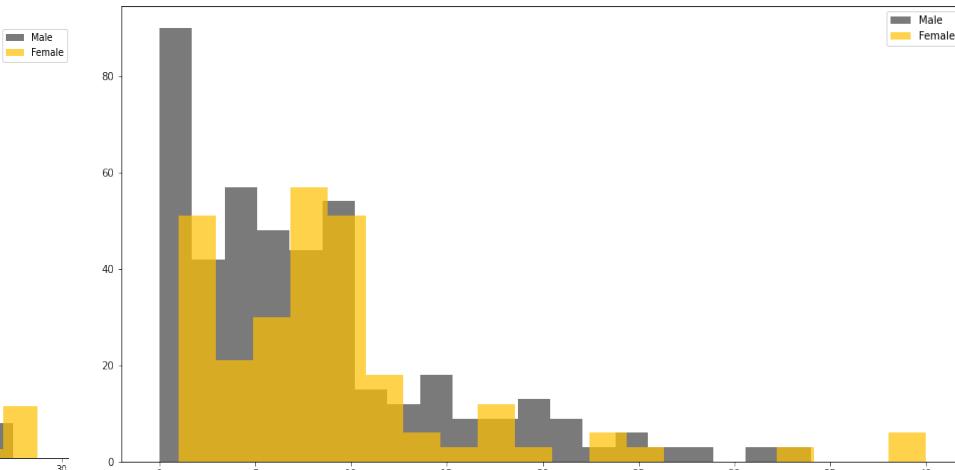
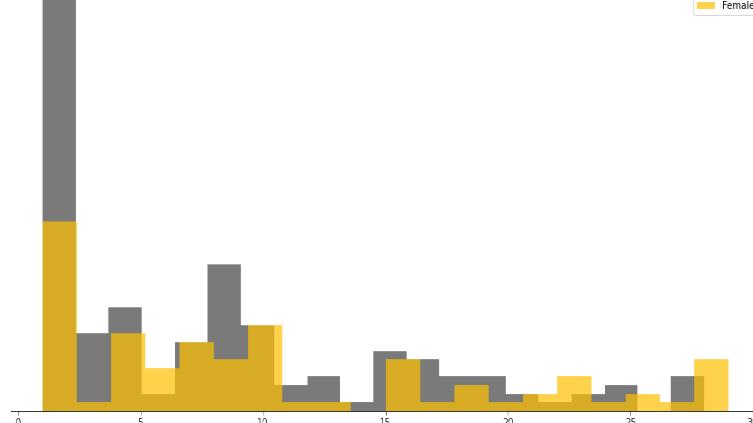
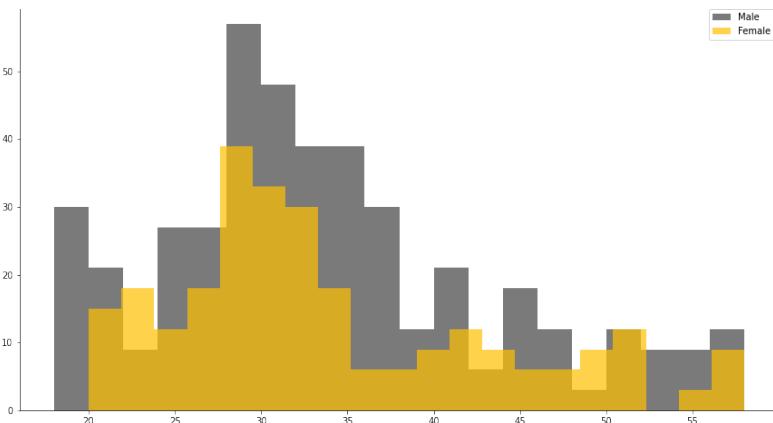
- Employee ID
- Start time – entire year
- End time – entire year
- Hours worked
- Total absent
- Overtime

OVERVIEW

- ▶ Overall, the features are **quite independent** to each other.
- ▶ There are a **few positive and negative correlations**, such as
 - **Positive:** compensation level vs. monthly income; hours worked vs. overtime, salary hike vs. performance, monthly income vs. compensation ratio...etc.
 - **Negative:** hours worked vs. total absent, total absent vs. overtime, attrition vs. total working years, etc.



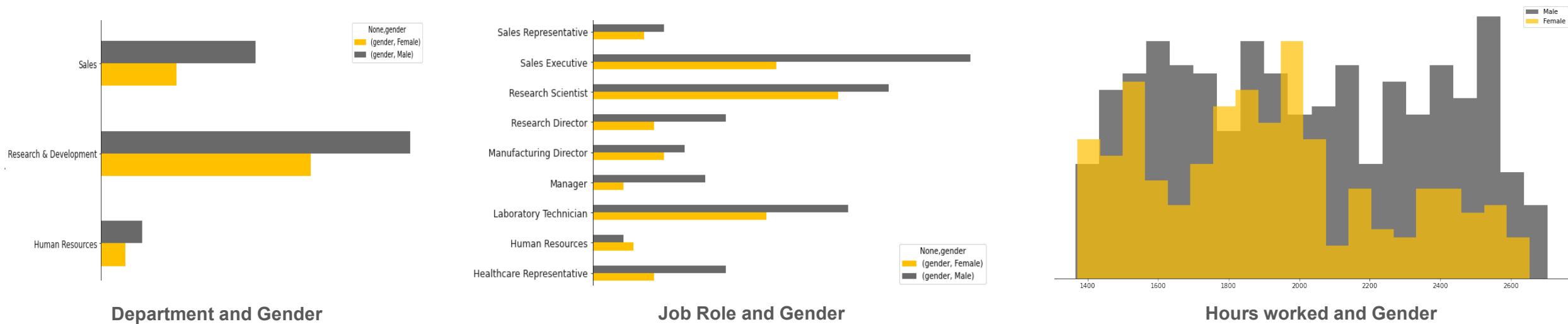
OVERVIEW



WHO

- ▶ High turnover among employees between the **ages of 25 to 35 with more male leavers than female** - could also be due to more male employees than female.
- ▶ Employees **commuting smaller distances have a higher turnover** with more female leavers travelling @ 1-10 kms.
- ▶ Turnover is higher among **employees at the start and mid career level** - which is also confirmed by the age distribution. This means that younger talent with fresher innovative ideas are leaving which could make the company less competitive.

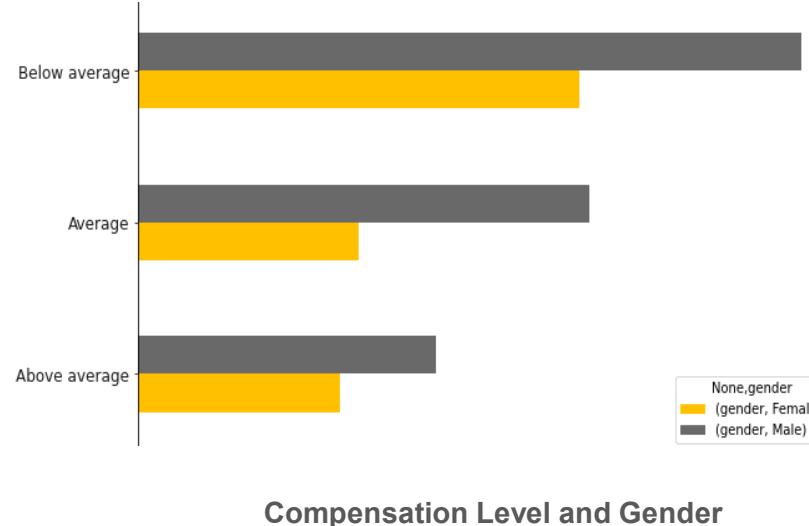
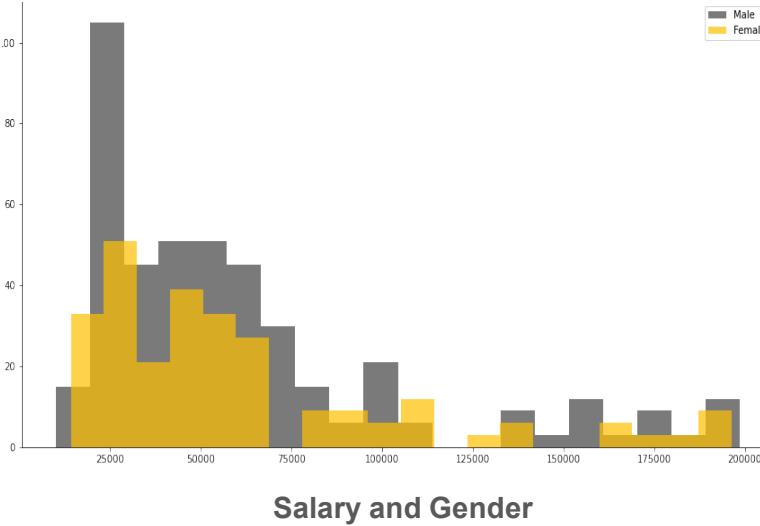
OVERVIEW



WHERE

- ▶ Turnover is high in the **Research & Development** department followed by **Sales**.
- ▶ Turnover is higher among **Laboratory Technicians and Research Scientist** - not a good sign for a pharmaceutical company as there is knowledge drain.
- ▶ The highest turnover is among male **Sales Executives** - not very good for company sales and brand value outlook for customers.
- ▶ Higher turnover among **males working longer hours than female** – it would be prudent to check against age as younger females may put in longer hours at start of career.
- ▶ Not a lot of difference in turnover and hours worked in males while **female employees working lesser hours** are quitting – this should be also checked against compensation levels.

OVERVIEW



WHY

- ▶ High turnover among **employees on the lower pay scale** which is also validated by the compensation level chart.
- ▶ Employees with **below average compensation** tend to leave more. The company may need to investigate pay disparity among employees on similar job titles and levels.
- ▶ **Male employees with lower work environment satisfaction** tend to leave, although there is not much difference with those who are more satisfied with their work environment.

PRE-PROCESSING & MODEL SETUP

1 FEATURE ENGINEERING

6 New features

2 DATASET BALANCING

Over-sampling by SMOTE | (2589, 30) (2589,
Combination sampling | (1267, 30) (887,)

3 CROSS VALIDATION

Kfold Method (10) | oversampled dataset
Leave One Out | combination dataset

4 TEST CASES

All Features | 30
Important features | Threshold Score > 0.03
15 features (oversampled dataset)
13 features (combination dataset)

LOGISTIC REGRESSION



K NEAREST NEIGHBOUR (KNN)



SUPPORT VECTOR MACHINE (SVM)



NAÏVE BAYES



RANDOM FOREST



COMPARISON

Model	Sampling Method	Cross-validation scores	
		Training Set	Score
LOGISTIC REGRESSION	OVERSAMPLED	All features (30)	0.72
	OVERSAMPLED	Important features (15)	0.69
	COMBINATION	All features (30)	0.73
	COMBINATION	Important features (13)	0.70
K NEAREST NEIGHBOUR (KNN)	OVERSAMPLED	All features (30)	0.81
	OVERSAMPLED	Important features (15)	0.88
	COMBINATION	All features (30)	0.87
	COMBINATION	Important features (13)	0.91
SUPPORT VECTOR MACHINE (SVM)	OVERSAMPLED	All features (30)	0.95
	OVERSAMPLED	Important features (15)	0.88
	COMBINATION	All features (30)	0.90
	COMBINATION	Important features (13)	0.80
NAÏVE BAYES	OVERSAMPLED	All features (30)	0.69
	OVERSAMPLED	Important features (15)	0.69
	COMBINATION	All features (30)	0.69
	COMBINATION	Important features (13)	0.68
RANDOM FOREST	OVERSAMPLED	All features (30)	0.98
	OVERSAMPLED	Important features (15)	0.98
	COMBINATION	All features (30)	0.96
	COMBINATION	Important features (13)	0.95

- ▶ All models perform well on cross validation which means that they generalize well.
- ▶ The gap between the models with all features and important features is small therefore we can be confident that most of the classification power is stored within the 13-15 important features.
- ▶ The performance of the SVM model decreases as we drop features, may need to decrease the threshold to include more features
- ▶ The best model based on the cross-validation scores is Random Forest followed by the SVM model with full features – both on the oversampled dataset

KEY METRICS

			accuracy	recall	F1	roc/auc
 LOGISTIC REGRESSION	OVERSAMPLED	All features (30)	0.74	0.70	0.47	0.79
		Important features (15)	0.71	0.67	0.42	0.75
	COMBINATION	All features (30)	0.78	0.62	0.47	0.80
		Important features (13)	0.79	0.55	0.46	0.75
 K NEAREST NEIGHBOUR (KNN)	OVERSAMPLED	All features (30)	0.76	0.94	0.56	0.94
		Important features (15)	0.87	0.94	0.69	0.97
	COMBINATION	All features (30)	0.88	0.87	0.69	0.96
		Important features (13)	0.92	0.95	0.88	0.98
 SUPPORT VECTOR MACHINE (SVM)	OVERSAMPLED	All features (30)	0.92	0.85	0.77	0.95
		Important features (15)	0.87	0.77	0.65	0.88
	COMBINATION	All features (30)	0.88	0.80	0.68	0.92
		Important features (13)	0.83	0.66	0.56	0.84
 NAÏVE BAYES	OVERSAMPLED	All features (30)	0.66	0.63	0.37	0.71
		Important features (15)	0.69	0.73	0.42	0.74
	COMBINATION	All features (30)	0.72	0.59	0.40	0.73
		Important features (13)	0.70	0.65	0.41	0.72
 RANDOM FOREST	OVERSAMPLED	All features (30)	0.97	0.85	0.91	0.98
		Important features (15)	0.98	0.89	0.94	0.99
	COMBINATION	All features (30)	0.96	0.98	0.89	0.99
		Important features (13)	0.96	0.99	0.88	0.99

- Most models perform well on accuracy with the highest belonging to Random Forest.
- KNN and Random Forest have the best recall rates on both samples, which serves our purpose as we need to minimize false negatives
- The accuracy of the combination dataset is higher and the ROC/AUC increases – while the recall rates reduce slightly (except for Random Forest), prediction performances may increase if trained on larger real information.
- Overall, the best model is Random Forest trained on the combination dataset – with or w/o full features. This indicates that the classification process for turnover is complex and non-linear.

2

1



PERFORMANCE



OVERALL BEST MODEL

on train, test, accuracy, recall, and predictive performance for turnover detection

RANDOM FOREST CLASSIFIERS

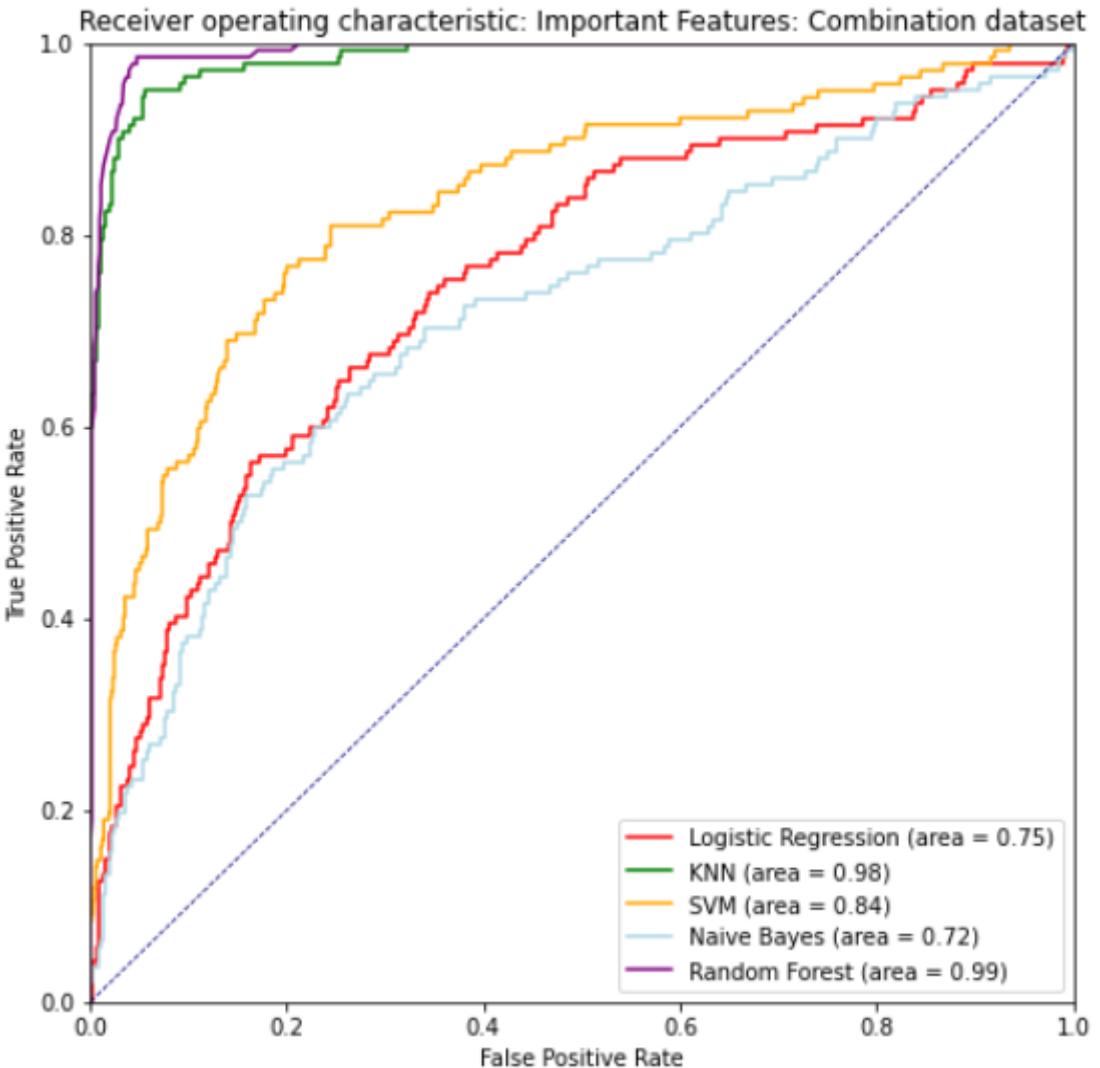


OVERALL BEST MODEL

in scenarios where the computing resource requirement is too high

K NEAREST NEIGHBOUR CLASSIFIER

- ▶ When using Random Forest, models will benefit from increased real data with an increase in minority class observations.
- ▶ Since most of the features are quite independent and contribute equally, it may work well to perform PCA feature selection to reduce number of features in order to run the models more efficiently, especially with large datasets.
- ▶ Try target encoding on categorical features to compare performances.



CONCLUSION



IMPORTANT FEATURES

- ▶ hours worked
- ▶ total career span
- ▶ salary & compensation levels
- ▶ future prospects
- ▶ work environment & satisfaction
- ▶ deep dive analytics on important features



RETAIN YOUNG WORKERS

- ▶ better training
- ▶ better work-life-balance
- ▶ better benefits



REDUCE KNOWLEDGE DRAIN

- ▶ R&D department
- ▶ work disruption, lost productivity
- ▶ dissatisfied customers
- ▶ gain by competitors



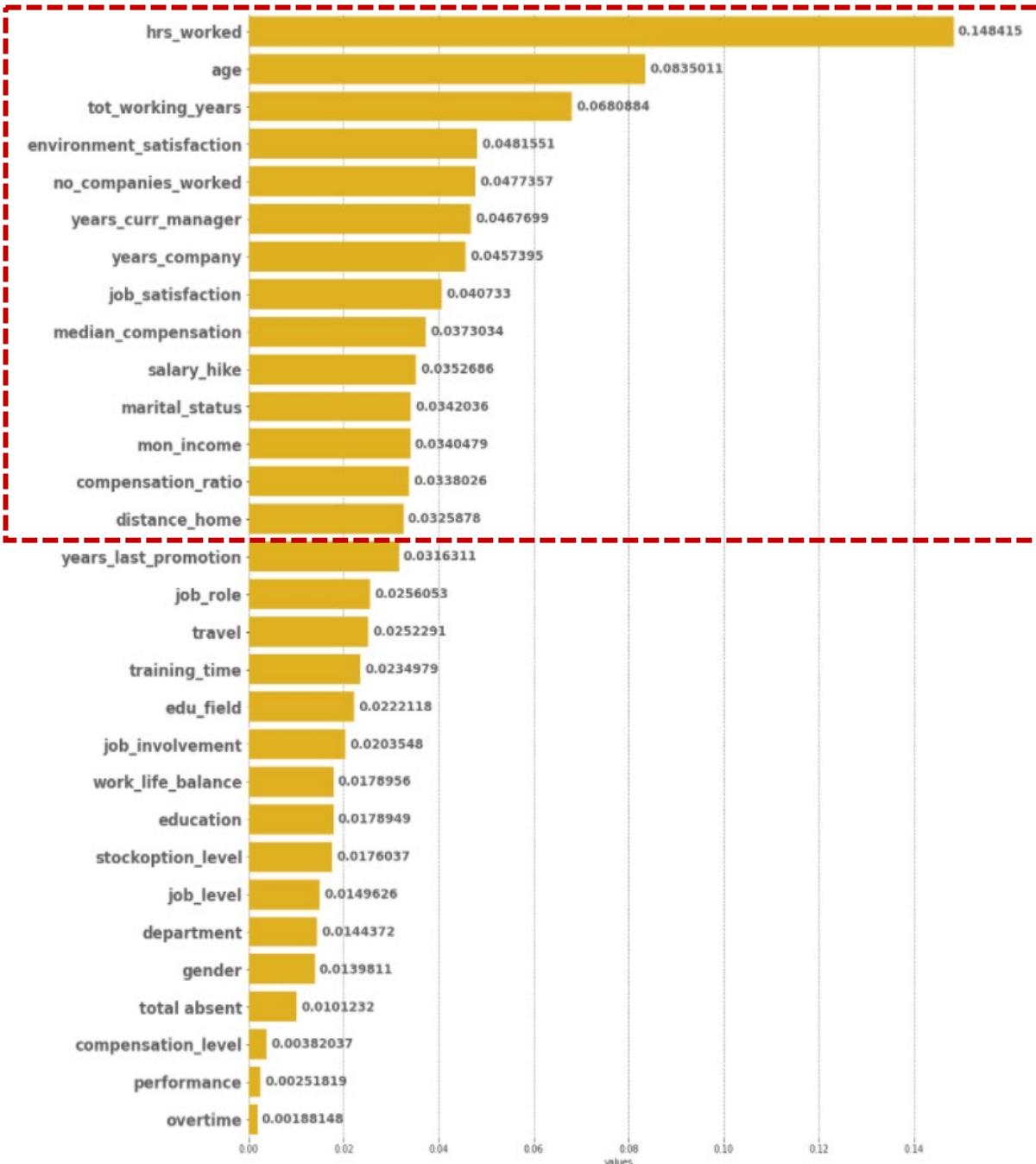
BETTER COMPENSATION LEVELS

- ▶ sales executives
- ▶ better compensation levels
- ▶ promotions
- ▶ pay disparity



PLAN HOURS WORKED

- ▶ plan back-up options based on predicted turnover
- ▶ absenteeism
- ▶ distribute workload





“

Train people well enough so they can leave, treat them well enough so they don't want to.

- Sir Richard Branson

THANK YOU

SOURCES

1.

- <https://www.shrm.org/hr-today/news/all-things-work/pages/to-have-and-to-hold.aspx>
- <https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx>
- <https://comphhealth.com/resources/infographic-the-true-cost-of-employee-turnover/>
- <https://www.forbes.com/sites/johnhall/2019/05/09/the-cost-of-turnover-can-kill-your-business-and-make-things-less-fun/?sh=71142e337943>
- <https://hhr.com.au/costs-of-employee-turnover/>
- <https://australianworkforce.com.au/blog/hidden-costs-high-staff-turnover/>
- <https://www.shrm.org/hr-today/news/all-things-work/pages/to-have-and-to-hold.aspx>

2.

- <https://financesonline.com/employee-turnover-statistics/>
- <https://www.business2community.com/infographics/the-growing-problem-and-cost-of-employee-turnover-infographic-02247963>

3.

- <https://researchbriefings.files.parliament.uk/documents/SN06152/SN06152.pdf>
- <https://thrivemap.io/reduce-attrition-and-employee-turnover/>

4.

- https://cdn.aigroup.com.au/Economic_Indicators/Fact_Sheets/2019/Labour_Turnover_in_2019.pdf
- <https://www.chandlermacleod.com/blog/2017/02/the-billion-dollar-hr-opportunity-in-australia>
- <https://sidekicker.com/au/blog/the-true-cost-of-retail-staff-turnover/>

5.

- <https://www.peoplekeep.com/blog/employee-retention-the-real-cost-of-losing-an-employee>
- <https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf>
- <https://cdn2.hubspot.net/hubfs/478187/2017%20Retention%20Report%20Campaign/Work%20Institute%202017%20-Retention%20Report.pdf>
- <https://www.hcamag.com/au/specialisation/employee-engagement/this-is-how-much-it-costs-to-hire-one-employee/192036>
- <https://www.goconqr.com/en/blog/its-not-just-the-money-the-true-cost-of-high-staff-turnover/>