Hiring Heroes

Data Driven Consulting for Human Resources

November 2nd - 2017

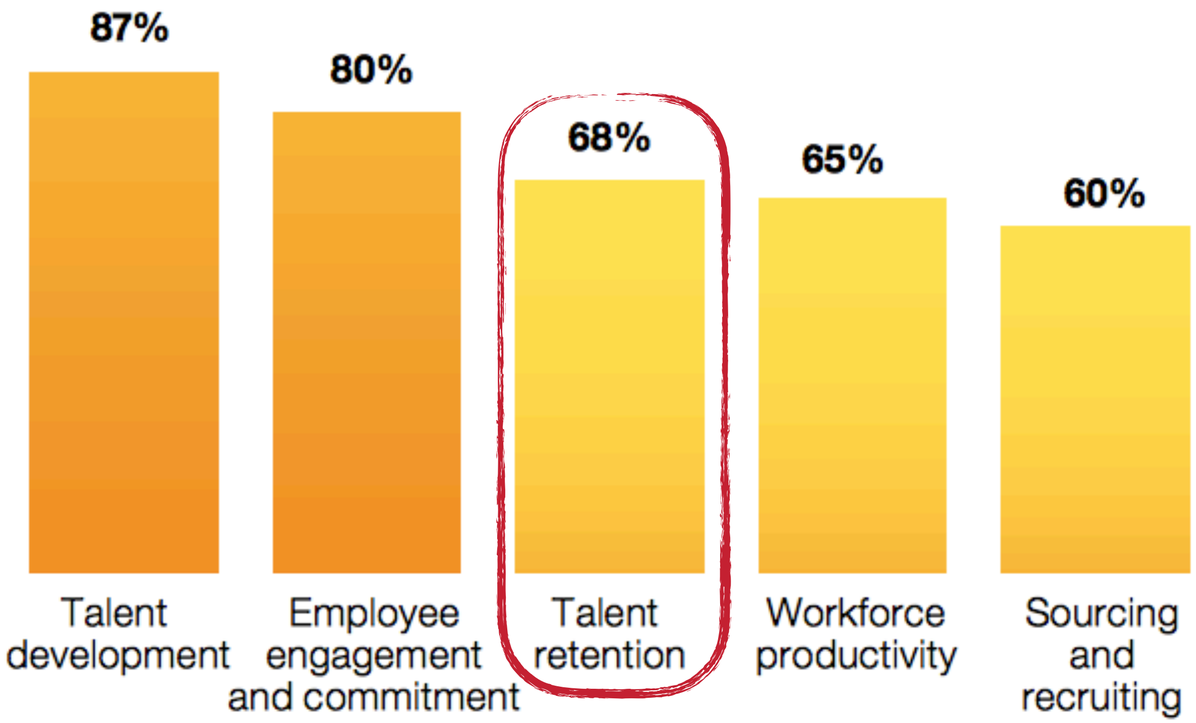
Roy Klaasse Bos, Quentin Meeus, Oscar Ulises Carreón
Jan Seipp, Eduardo Andreotti and Vincent van Mierlo

JADS Jheronimus Academy of Data Science

"Replacing a salaried employee costs 6 to 9 months' salary on average"

Society of Human Resource Managers: (2014)

Top 5 Workforce Related Challenges

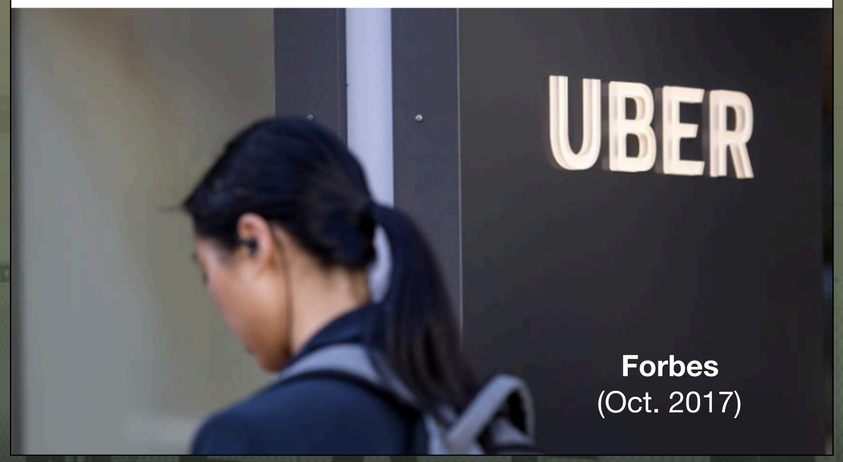


Source: **IBM (2014)**

Gender inequality

- Wage
- Job levels

Three Women Engineers Sue Uber Claiming Unequal Pay, Discrimination



Forbes
(Oct. 2017)

More than 60 women consider suing Google, claiming sexism and a pay gap

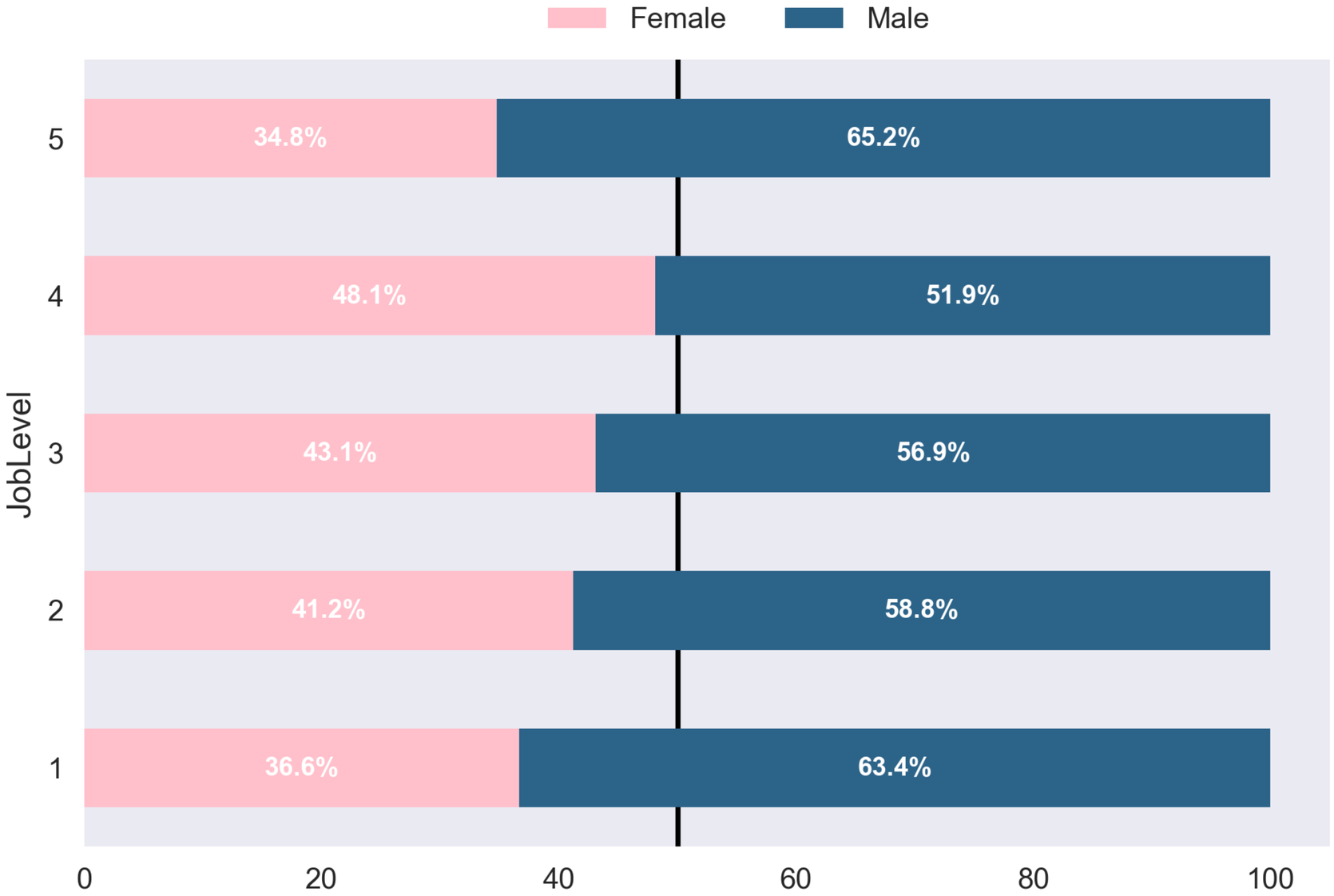
Scandal over discrimination at the company deepens as dozens of current and former staff say they have earned less than men despite equal qualifications

Sam Levin • Last modified on Wednesday 9 August 2017 23.54 BST

The Guardian (Aug. 2017)

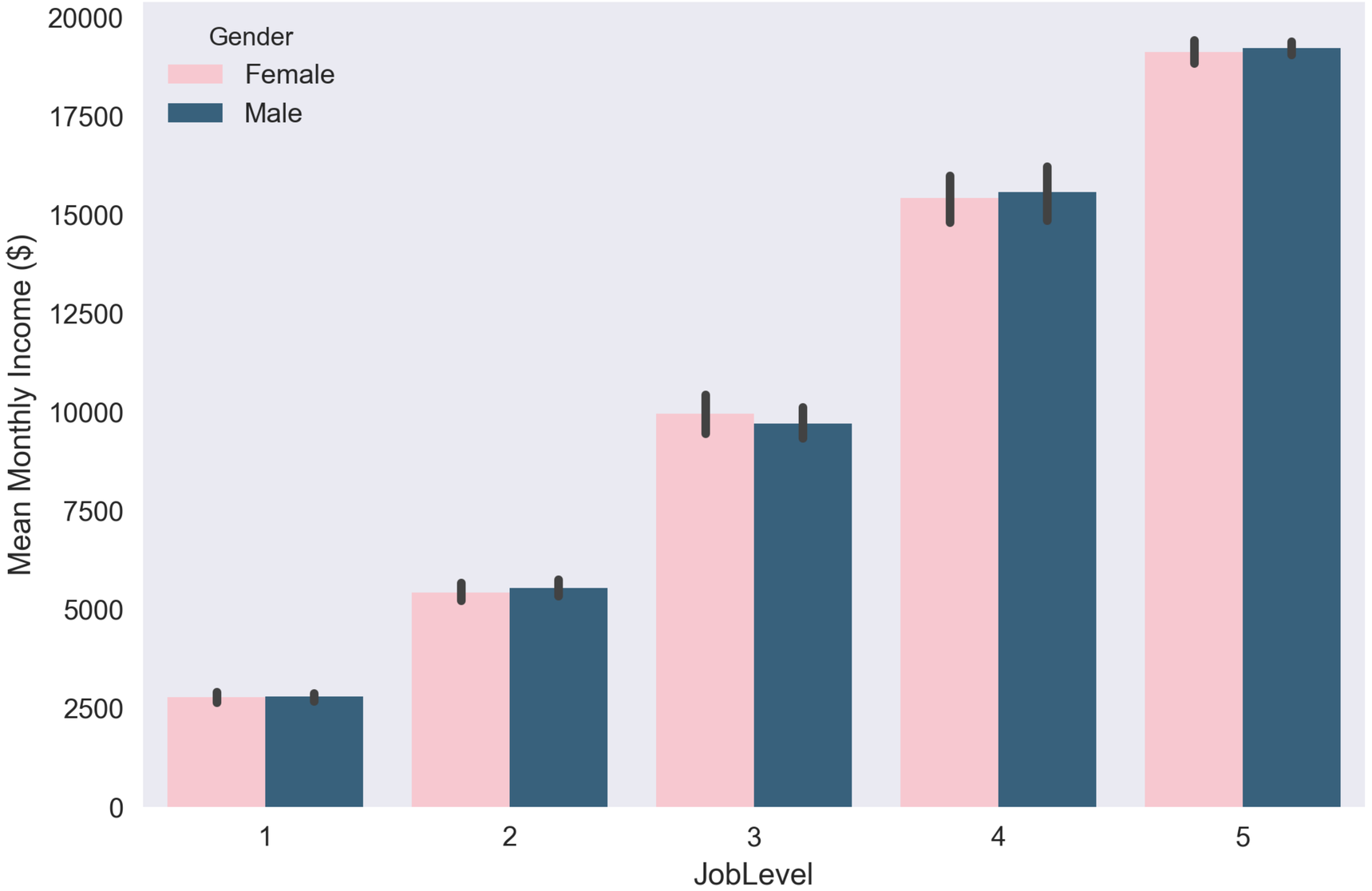
Case Description

- **Medicine Manufacturer**
- Employee data from **3 departments** (**n=1470**)
- Target variable: **Attrition**
- Input features: **Demographic Data** + **Job Characteristics**

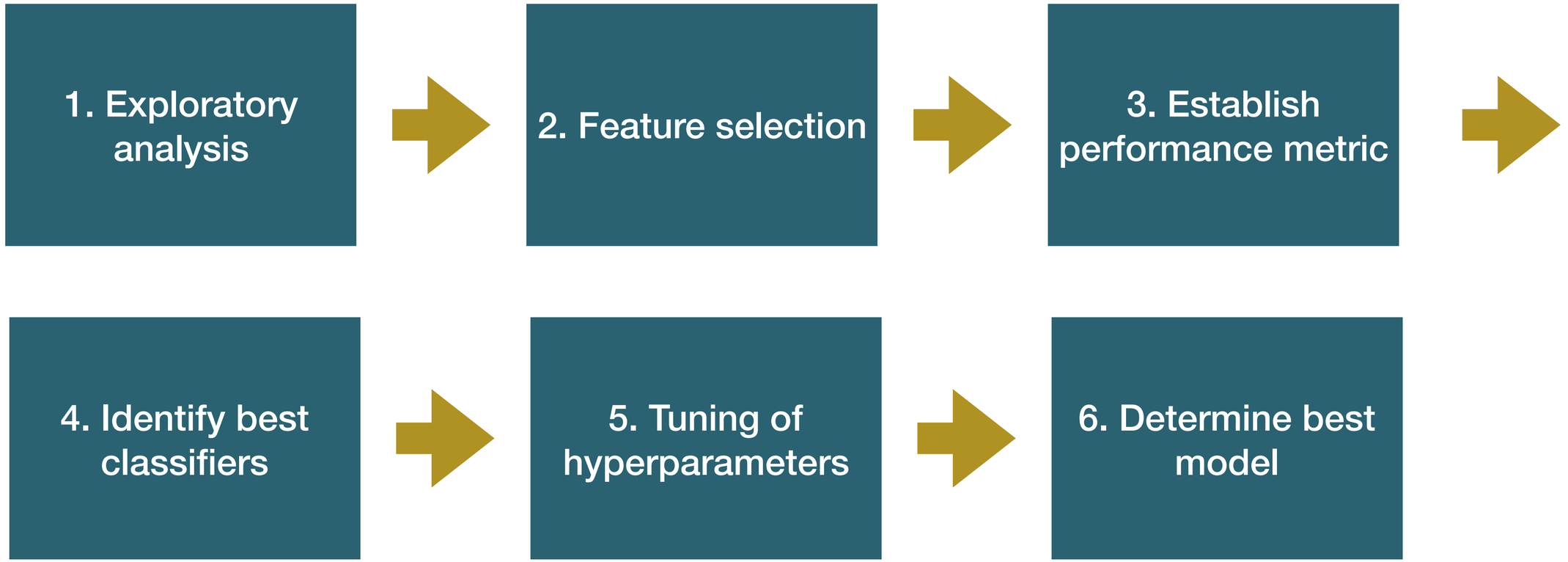

Female
Male
JobLevel
5
4
3
2
1
34.8%
65.2%
48.1%
51.9%
43.1%
56.9%
41.2%
58.8%
36.6%
63.4%
0
20
40
60
80
100

Job Level	Healthcare Representative		Human Resources		Laboratory Technician		Manager		Manufacturing Director		Research Director		Research Scientist		Sales Executive		Sales Representative	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
5							17	26			7	19						
4	3	6					23	24	8	2	10	16			7	7		
3	17	27	3	3	1	2	7	5	23	22	16	12		1	27	52		
2	31	47	3	10	17	39			41	49			27	30	98	135	3	4
1			10	23	67	133							87	147			35	41

Job Level	Manager		Research Director	
	Female	Male	Female	Male
5	40%	60%	27%	73%
4	49%	51%	38%	62%
3	58%	42%	57%	43%

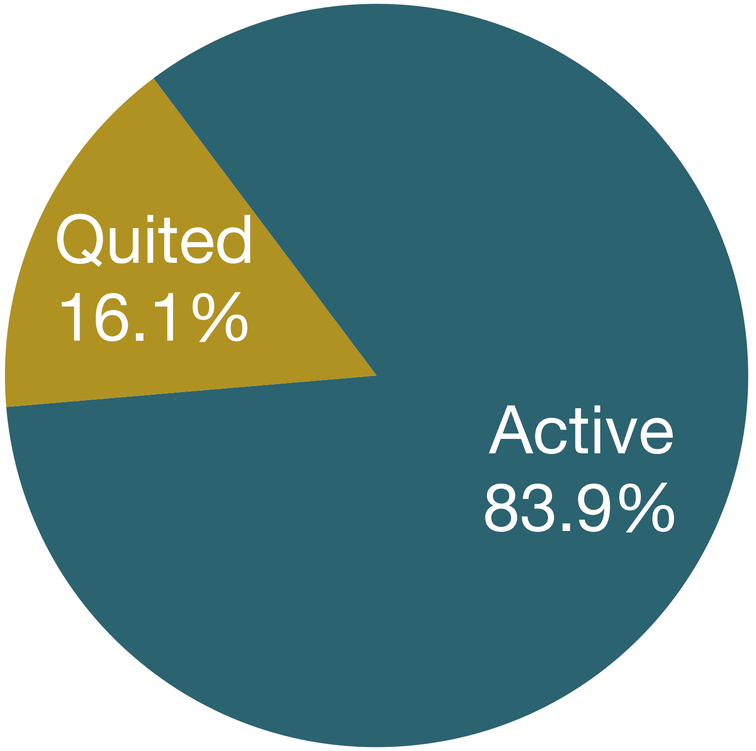

Gender
Female
Male
20000
17500
15000
12500
10000
7500
5000
2500
0
Mean Monthly Income ($)
1
2
3
4
5
JobLevel

Predictive Modelling Approach

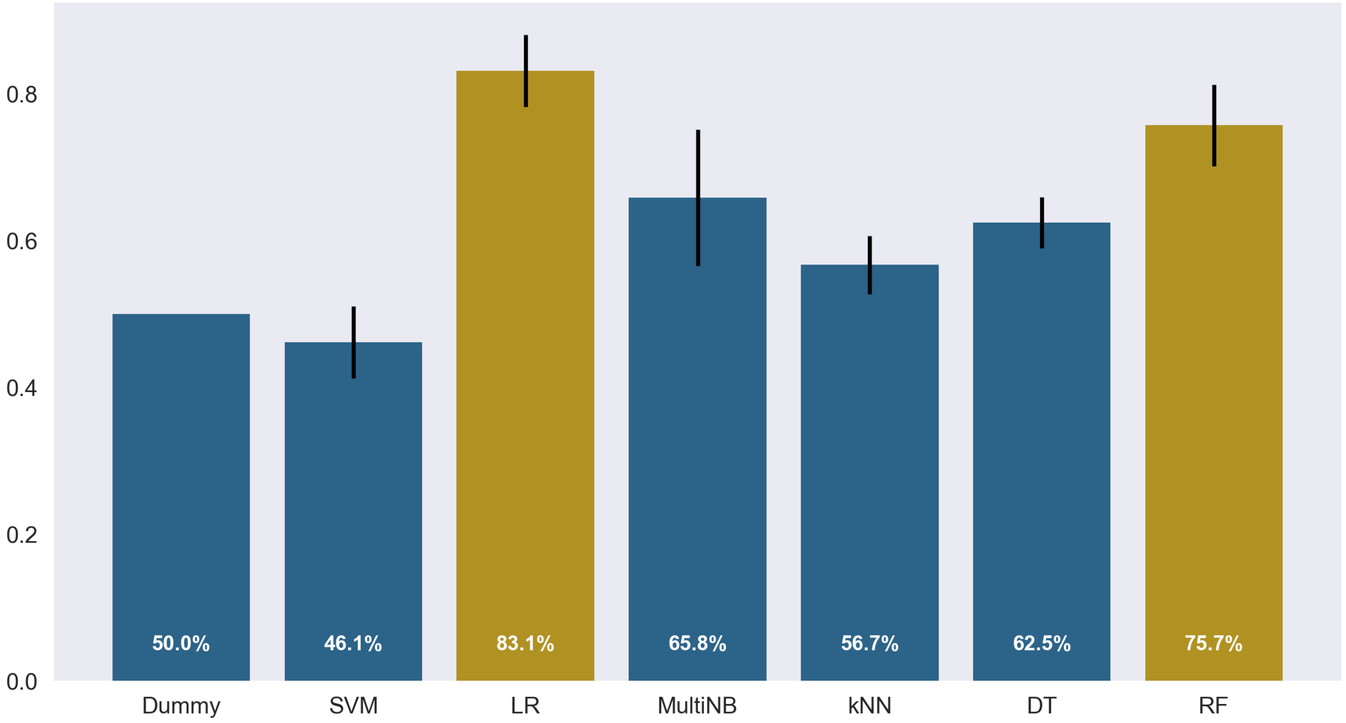


Establish Performance Metric

- **Skewed dataset**
- **Accuracy** vs **AUC score**



Identify Best Classifiers



Tuning of Hyperparameters

Logistic Regression

- Penalty parameter
- C-value
- Class-weight

Random Forest

- Maximum number of features
- Maximum depth
- Minimum number of samples per split

Determine Best Model

Classifier	AUC	Accuracy	Recall	Precision
Logistic Regression	75.9%	89.4%	63.6%	43.8%
Random Forest	52.9%	85.8%	33.3%	8.3%

#1 Moderate Working Hours

In [226]:

```
clf = best_models[0]
coef = pd.DataFrame(index=X_train.columns)
coef["Coefficients"] = clf.coef_[0]
coef.sort_values("Coefficients", ascending=False)
```

Out[226]:

	Coefficients
OverTime	2.00
JobRole_Laboratory Technician	1.08
BusinessTravel	0.86

Example

Male Sales Executive (without overtime)

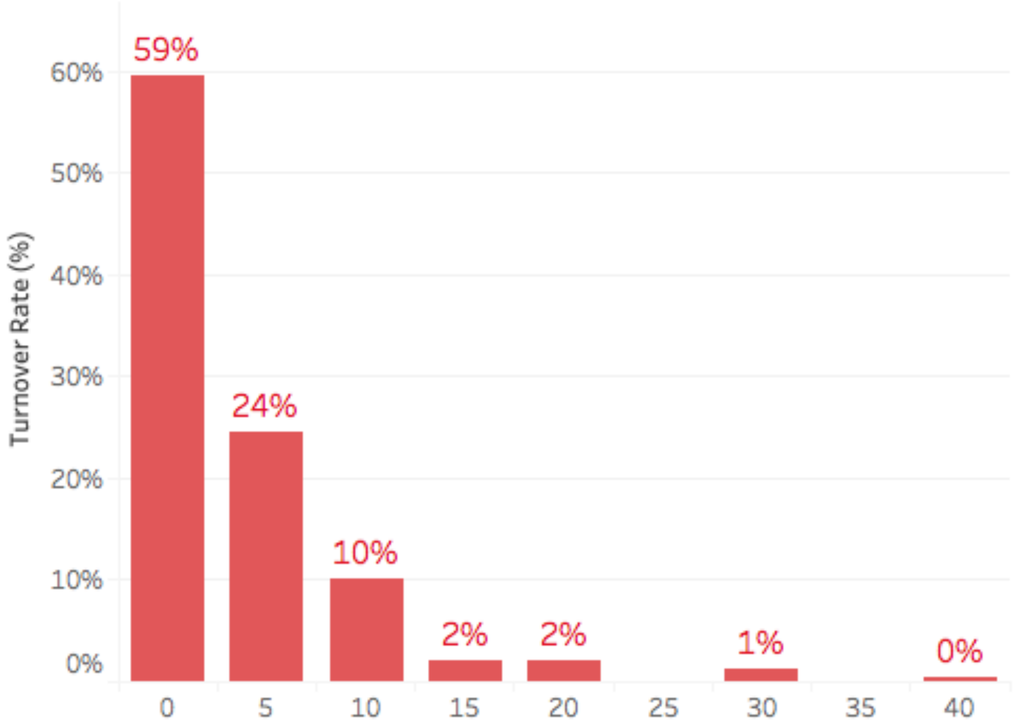
60.6% attrition probability

Male Sales Executive (with overtime)

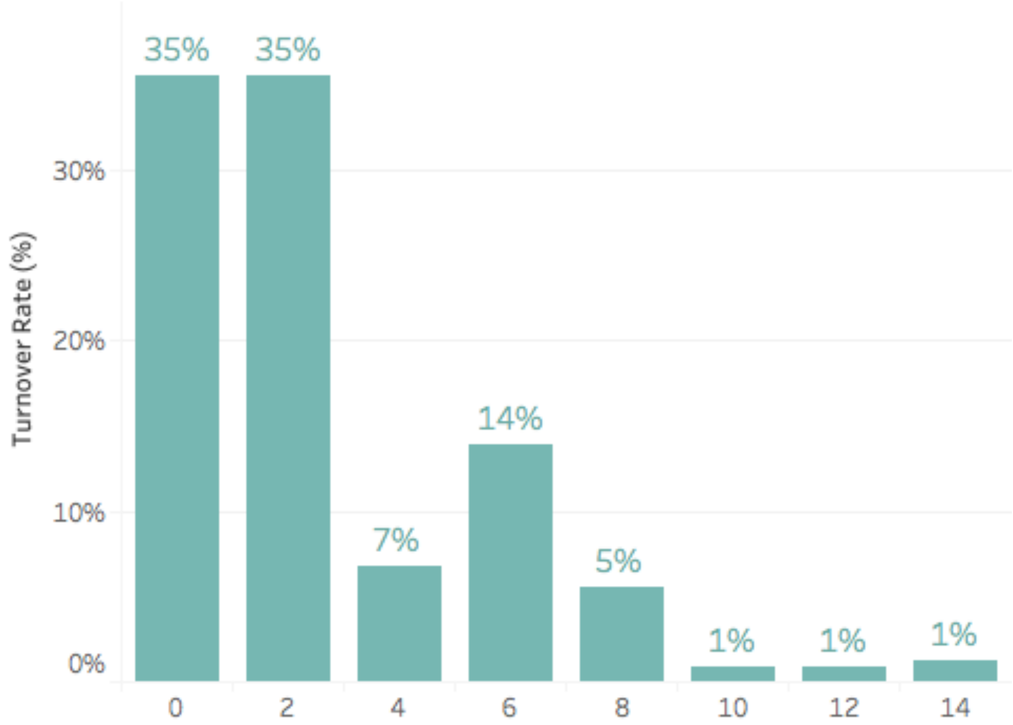
87.1% attrition probability

#2 Make New Hires Feel At Home

"Turnover Rate" vs "Years at Company"



"Turnover Rate" vs "Years in Current Role"



MONEY WON'T BUY EMPLOYEE LOYALTY

FOR THESE AND OTHER INSIGHTS:
BIT.LY/HIRING-HEROES

Back-up slides

In case of tricky questions... 😖

Assumption

“It is assumed that the **drawbacks** of any **resignation outweigh** the **potential benefits** (e.g. better performance of a new hire).

This is a reasonable assumption to make since **all employees** are evaluated either *'Excellent’* or *‘Outstanding'*.”

AUC before & after tuning

At first sight, the performance of the **Logistic Regression** and **Random Forest** model seem to deteriorate after tuning the hyperparameters.

AUC before & after tuning

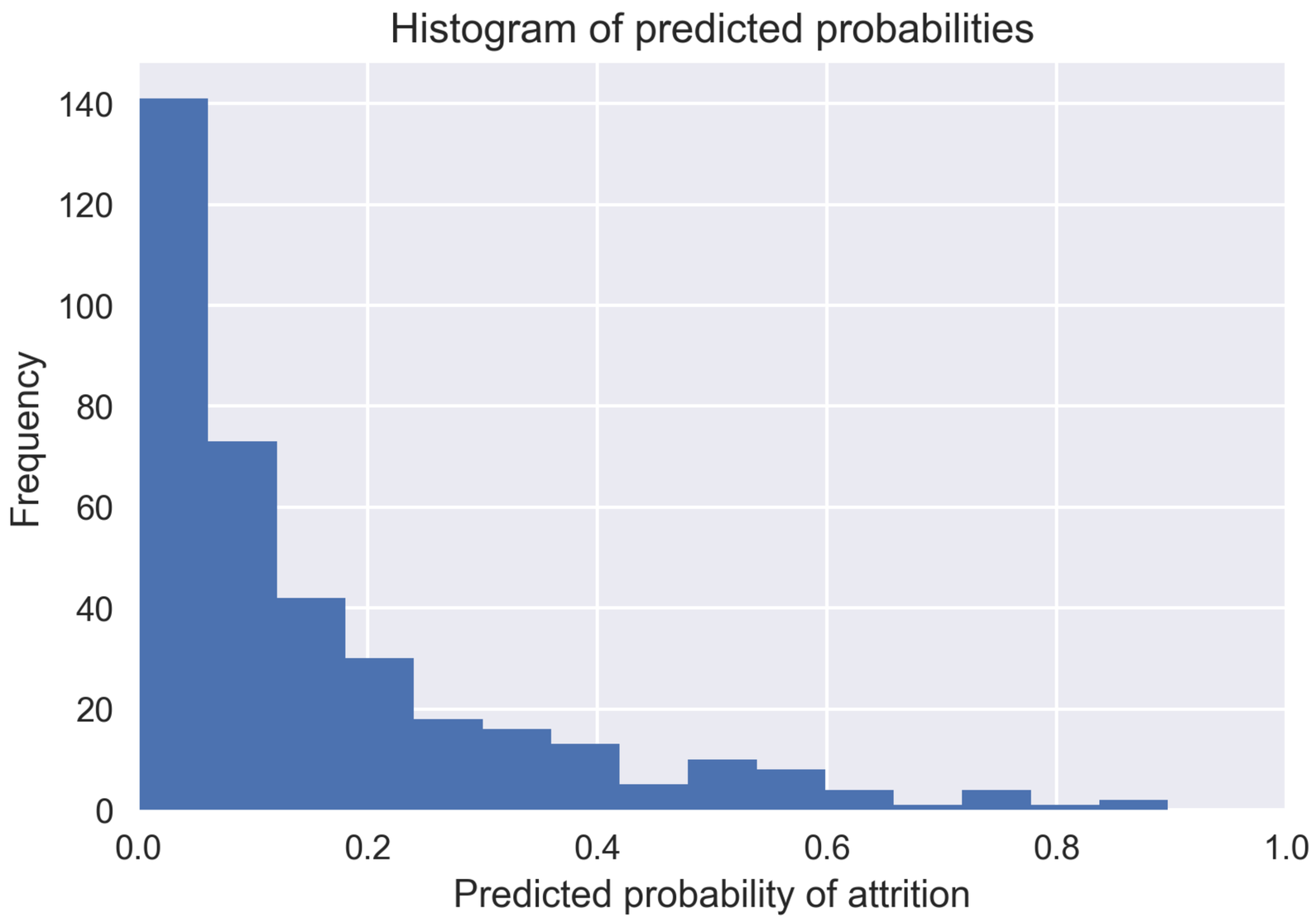
That is because of **different validation methods**:

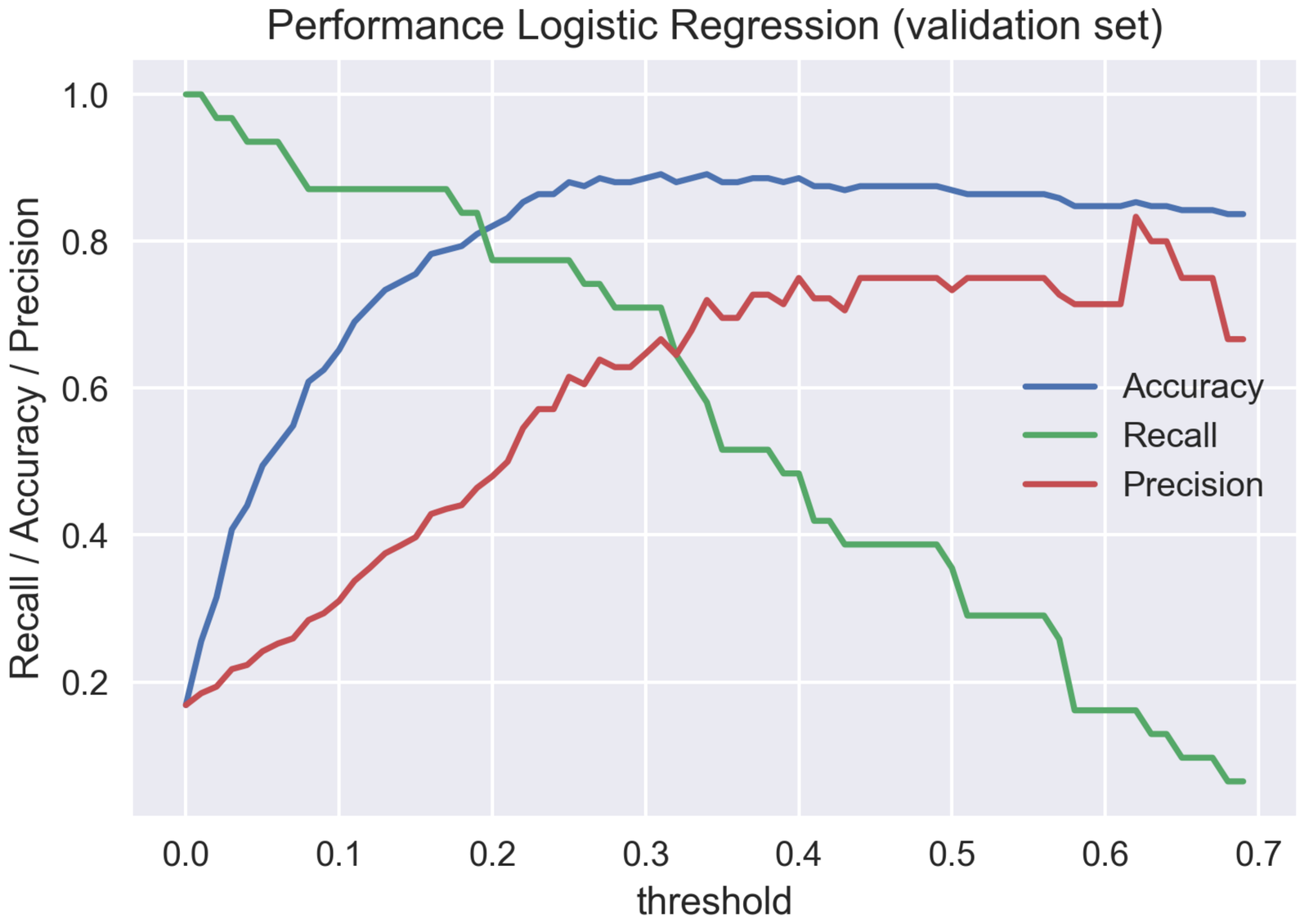
```
results = OrderedDict()
for name, model in models:
    kfold = KFold(n_splits=10)
    cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
    results[name] = cv_results
```

```
best_models = []
for (name, clf), params in zip(classifiers, [lr_params, rf_params]):
    gs = GridSearchCV(clf, params)
    gs.fit(X_train, y_train)
    best_models.append(gs.best_estimator_)
    y_pred = gs.predict(X_test)
```

AUC before & after tuning

The testing methodology that was adopted required us to first split the test in **75% training set** and **25% testing set**. Then, the cross-validation further split the training set into **10 folds**. The scores have been calculated on a sample of **7.5%** of the total volume dataset (+/- 110 rows) and **2.5%** for the testing set.


Histogram of predicted probabilities
Frequency
140
120
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted probability of attrition


Performance Logistic Regression (validation set)
Recall / Accuracy / Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
threshold
Accuracy
Recall
Precision

Explanation for imbalance job level 1

	Job Role / Gender																			
	Healthcare Representative		Human Resources		Laboratory Technician		Manager		Manufacturing Director		Research Director		Research Scientist		Sales Executive		Sales Representative	
Job Level	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
5							40%	60%			27%	73%						
4	33%	67%					49%	51%	80%	20%	38%	62%			50%	50%		
3	39%	61%	50%	50%	33%	67%	58%	42%	51%	49%	57%	43%		100%	34%	66%		
2	40%	60%	23%	77%	30%	70%			46%	54%			47%	53%	42%	58%	43%	57%
1			30%	70%	34%	67%							37%	63%			46%	54%

Logistic Regression Coefficients

	Coefficients
OverTime	2.00
JobRole_Laboratory Technician	1.08
BusinessTravel	0.86
MaritalStatus_Single	0.86
EducationField_Technical Degree	0.66
JobRole_Sales Representative	0.62
Gender	0.53
Department_Sales	0.46
PerformanceRating	0.33
EducationField_Human Resources	0.31
JobRole_Human Resources	0.26
NumCompaniesWorked	0.22
Department_Human Resources	0.18
EducationField_Marketing	0.16
YearsSinceLastPromotion	0.14
YearsAtCompany	0.13
Education	0.08
DistanceFromHome	0.04
JobRole_Manager	0.03
MonthlyIncome	-0.00
Age	-0.02
PercentSalaryHike	-0.02
MaritalStatus_Married	-0.06
TotalWorkingYears	-0.08
JobRole_Manufacturing Director	-0.09
JobLevel	-0.10
TrainingTimesLastYear	-0.12
YearsWithCurrManager	-0.12
RelationshipSatisfaction	-0.15
JobRole_Sales Executive	-0.17
YearsInCurrentRole	-0.18
StockOptionLevel	-0.20
EducationField_Other	-0.21
WorkLifeBalance	-0.23
EducationField_Medical	-0.28
Department_Research & Development	-0.28
EducationField_Life Sciences	-0.29
JobRole_Research Director	-0.32
JobSatisfaction	-0.35
EnvironmentSatisfaction	-0.36
JobInvolvement	-0.43
MaritalStatus_Divorced	-0.44
JobRole_Healthcare Representative	-0.55

Feature Selection

	Fisher	p-value
OverTime	94.6565	1.00925e-21
MaritalStatus_Single	46.607	1.2657e-11
TotalWorkingYears	44.2525	4.06188e-11
JobLevel	43.2153	6.79538e-11
YearsInCurrentRole	38.8383	6.00319e-10
MonthlyIncome	38.4888	7.14736e-10
Age	38.1759	8.35631e-10
JobRole_Sales Representative	37.2128	1.35226e-09
YearsWithCurrManager	36.7123	1.73699e-09
StockOptionLevel	28.1405	1.30101e-07
YearsAtCompany	27.0016	2.31887e-07
JobInvolvement	25.242	5.67707e-07
BusinessTravel	24.068	1.03348e-06
JobSatisfaction	15.89	7.04307e-05
EnvironmentSatisfaction	15.8552	7.17234e-05
JobRole_Laboratory Technician	14.3207	0.00016036
MaritalStatus_Married	12.2536	0.000478263
JobRole_Research Director	11.6863	0.000646838
MaritalStatus_Divorced	11.3826	0.000760663
Department_Research & Development	10.7578	0.00106275
JobRole_Manager	10.2615	0.00138749

	Fisher	p-value
JobRole_Manufacturing Director	10.1817	0.00144842
Department_Sales	9.6603	0.00191907
JobRole_Healthcare Representative	9.14808	0.00253302
DistanceFromHome	8.96828	0.00279306
EducationField_Technical Degree	7.0953	0.00781321
WorkLifeBalance	6.02612	0.0142111
TrainingTimesLastYear	5.21165	0.0225785
EducationField_Marketing	4.58191	0.0324755
EducationField_Medical	3.24981	0.0716366
RelationshipSatisfaction	3.09558	0.0787136
NumCompaniesWorked	2.78229	0.0955253
EducationField_Human Resources	1.95472	0.16229
JobRole_Human Resources	1.92786	0.165204
YearsSinceLastPromotion	1.60222	0.20579
EducationField_Life Sciences	1.5717	0.21016
Education	1.44631	0.229315
Gender	1.27459	0.259092
JobRole_Sales Executive	0.57425	0.448697
EducationField_Other	0.470382	0.49292
Department_Human Resources	0.416027	0.519027
PercentSalaryHike	0.266728	0.605613
PerformanceRating	0.0122504	0.911884
JobRole_Research Scientist	0.0001898	0.98901

Workforce composition

1 | How **loyal** are the employees towards the company?

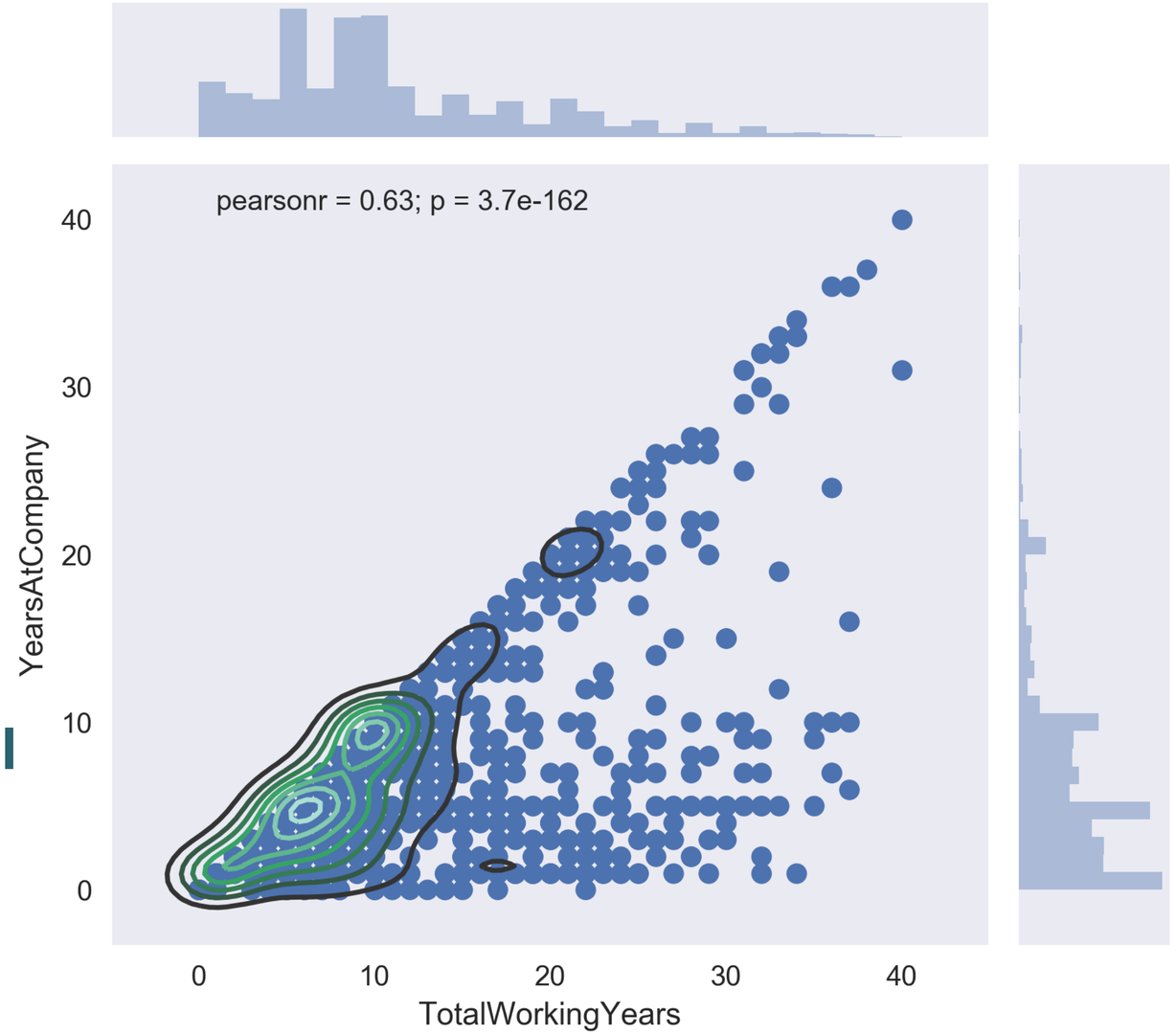
2 | How **diverse** is the workforce in terms of **education**?

3 | How does **education** relate to employee's **performance**?

Question 1:

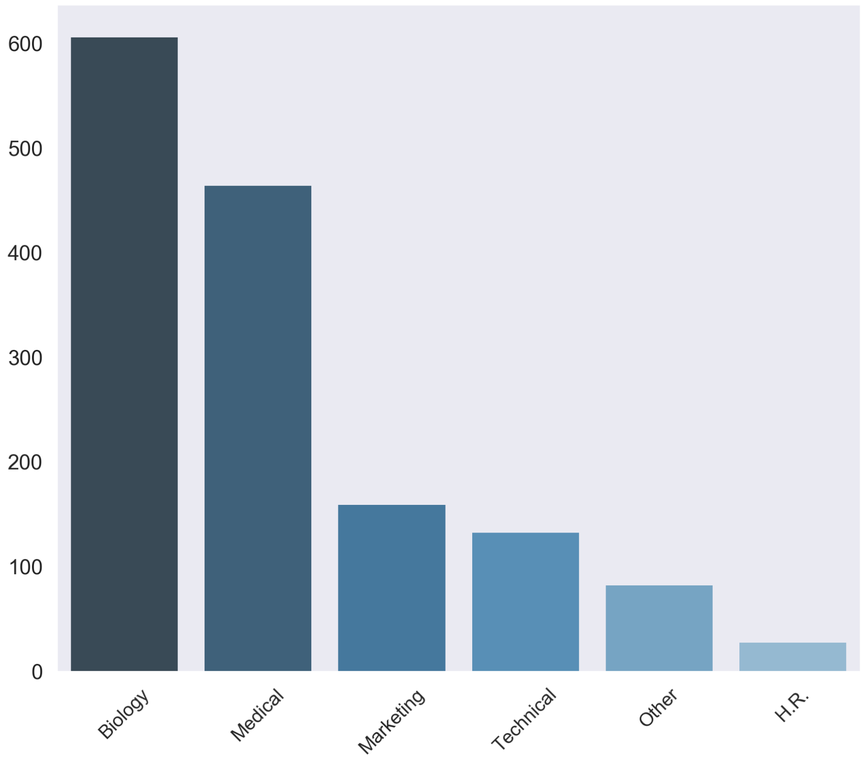
How loyal are the employees towards the company?

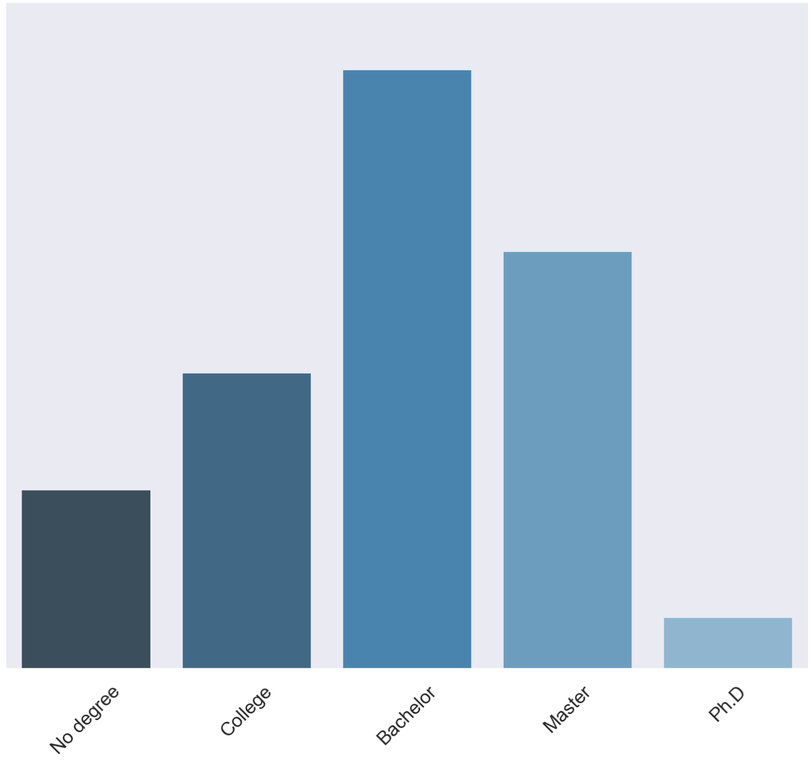
- Average years at company **7** years
- **75%** of all employees seniority > **3 year**
- Many points close to the diagonal = **loyal**



Question 2:

How **diverse** is the workforce in terms of **education**?

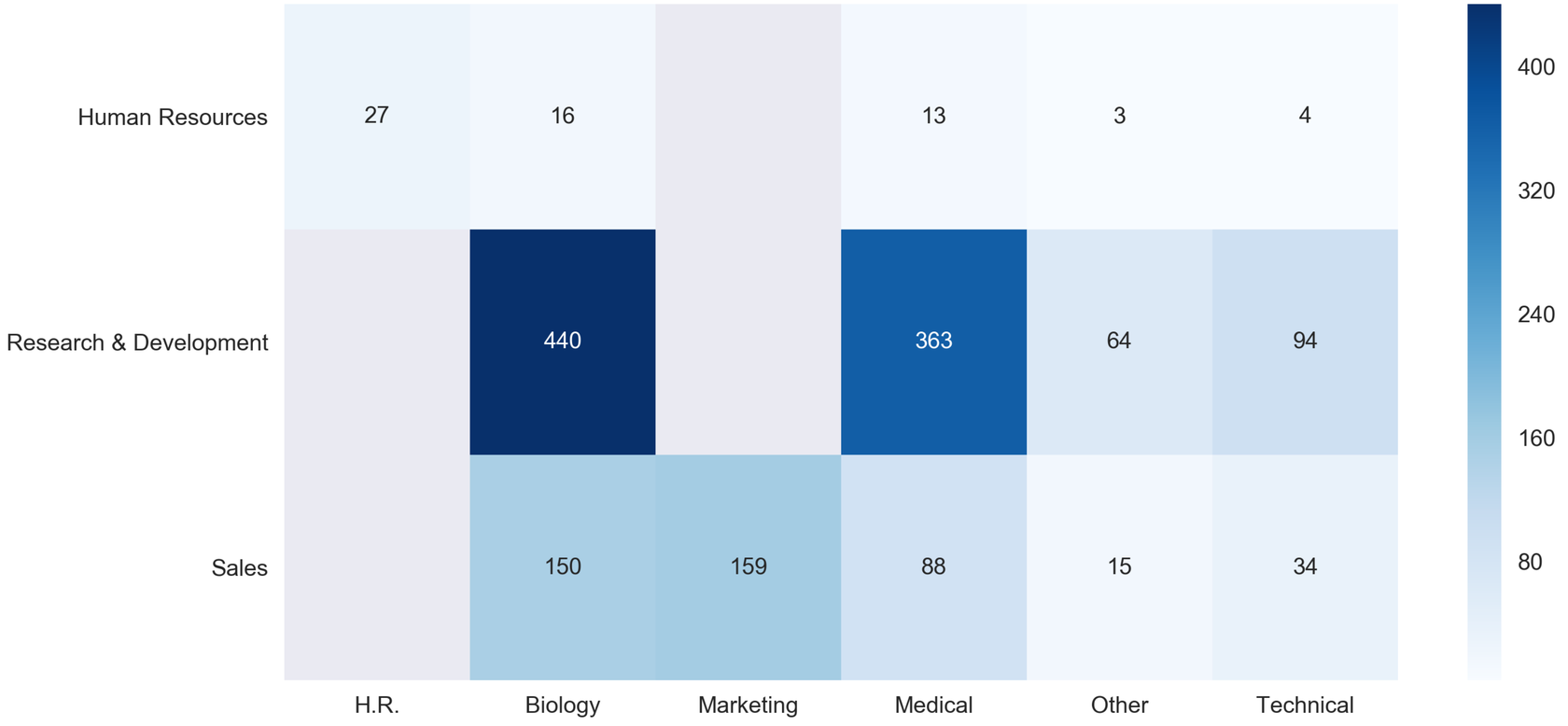




- **Biology** and **Medical** degree most popular
- **Bachelor** and **master** most common
- Few PhD graduates

Question 2:

How **diverse** is the workforce in terms of **education**?



- **Large** sales force and even more **researchers**
- **Marketing** degree only in sales department
- **HR** degree only in HR department

Question 3:

How does **education** relate to employee's **performance**?

- (Top) **performance** dependent on ***combination*** of **education** and **job role**

	PerformanceRating	PercentageTopPerformers
EducationField	JobRole	
Medical	Manager	0.38
Life Sciences	Human Resources	0.31
Other	Sales Executive	0.31
Technical Degree	Research Scientist	0.23
Other	Healthcare Representative	0.22
Marketing	Manager	0.21
Life Sciences	Manufacturing Director	0.21
Other	Manager	0.20
Life Sciences	Laboratory Technician	0.19
Other	Laboratory Technician	0.18

Gender equality

4 | How are **genders** distributed across the **departments** of the company?

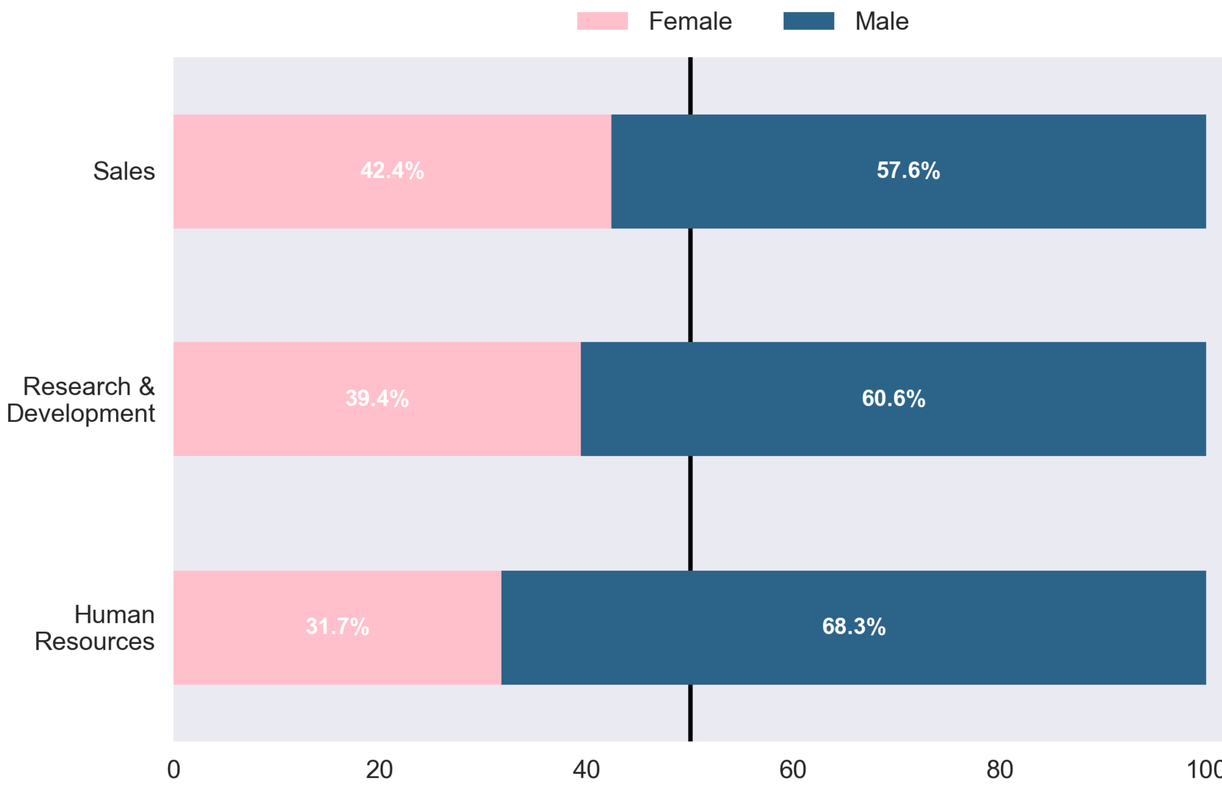
5 | How are **genders** distributed across the **hierarchy**?

6 | What **differences** exist in the **compensation** of **men** and **women** at the same position?

Question 4:

How are **genders** distributed **across** the **departments** of the company?

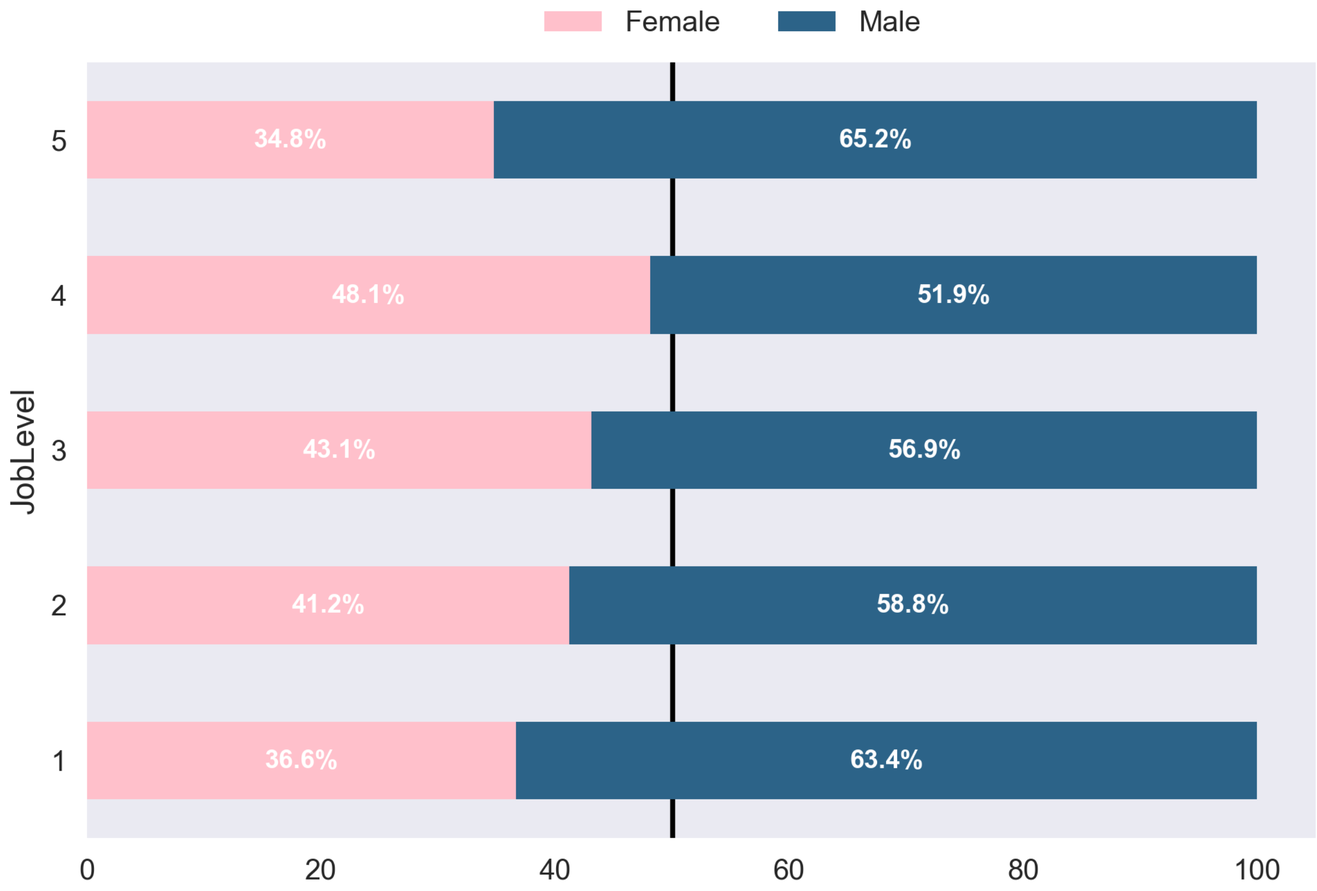
- In all departments there is a **majority** of **male** employees
- The **most significant difference** is in the **HR** department



Question 5:

How are **genders** distributed across the **hierarchy**?

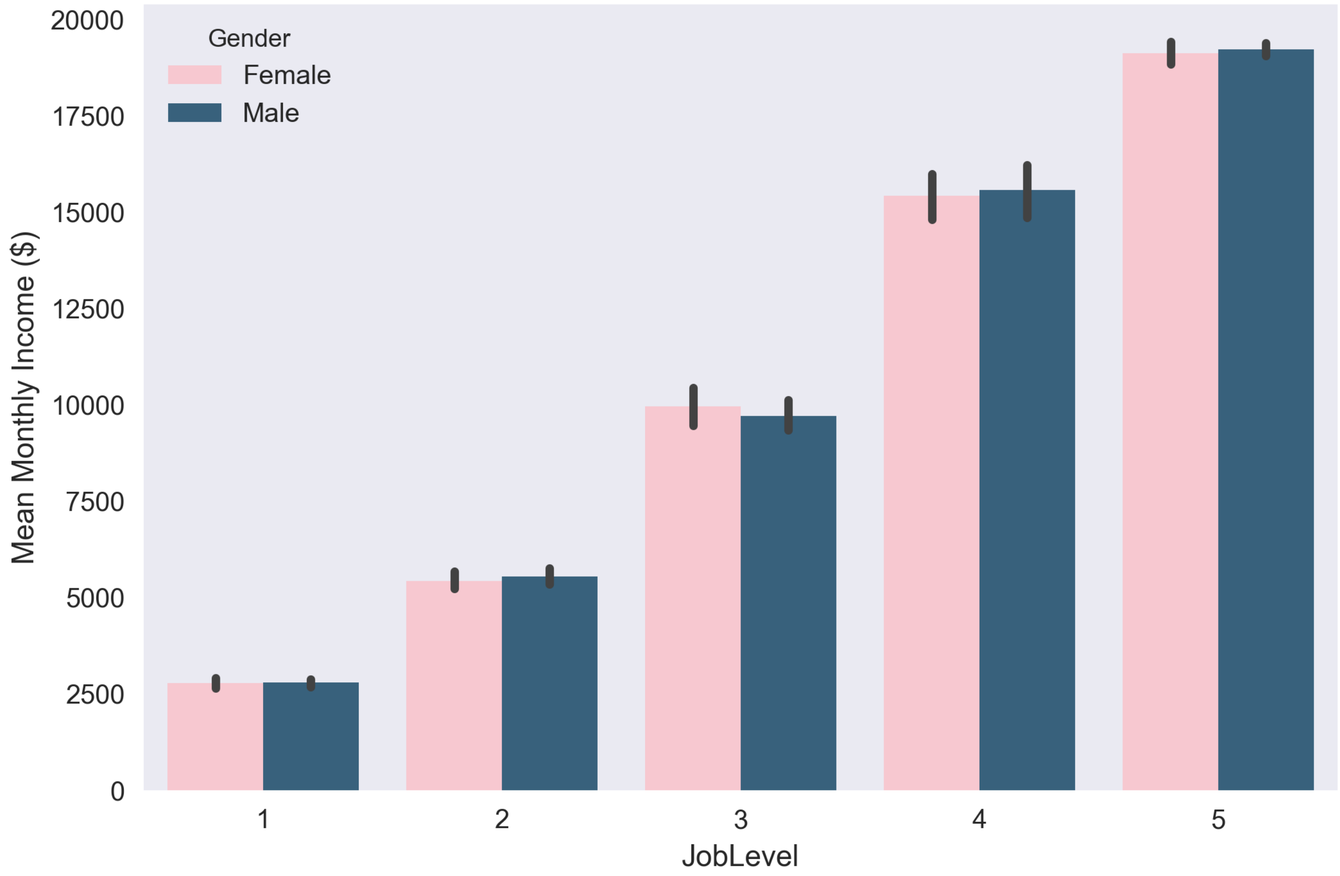
- Proportion of women **decreases** rapidly in **5th** job level
- Sign of **glass ceiling**



Question 6:

What **differences** exist in the **compensation** of **men** and **women** at the same position?

- **Ratio** between **men** and **women** wages about **1** for all job levels



Turnover

7 | What is the relation between employees' **satisfaction** and **attrition**?

8 | What impact has the **length** of the **career** on the **turnover**?

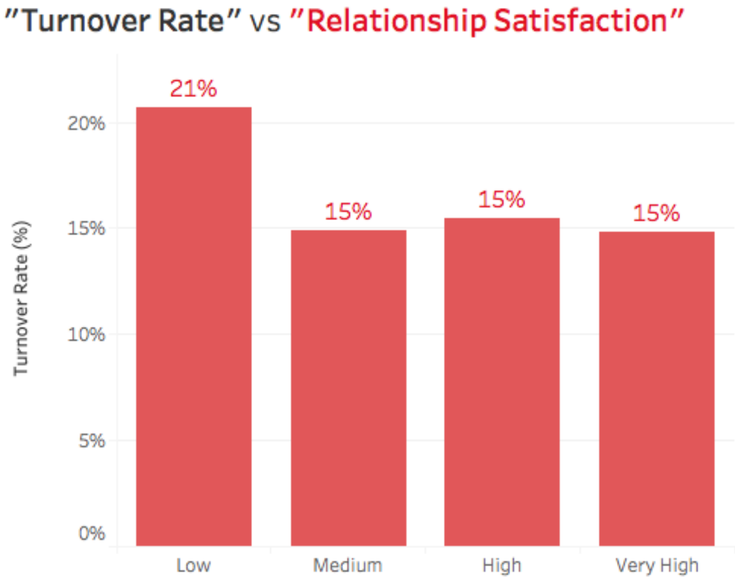
9 | How are the different **variables correlated**?

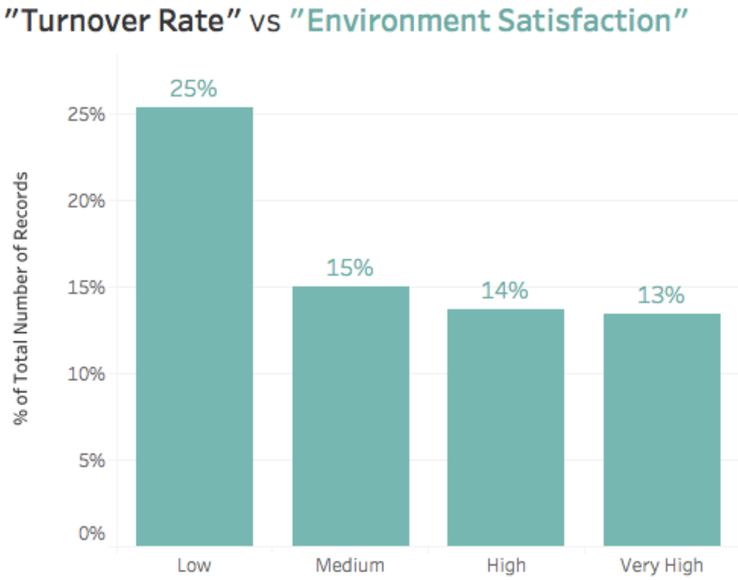
10 | How **good** are our **features** for **predicting attrition**?

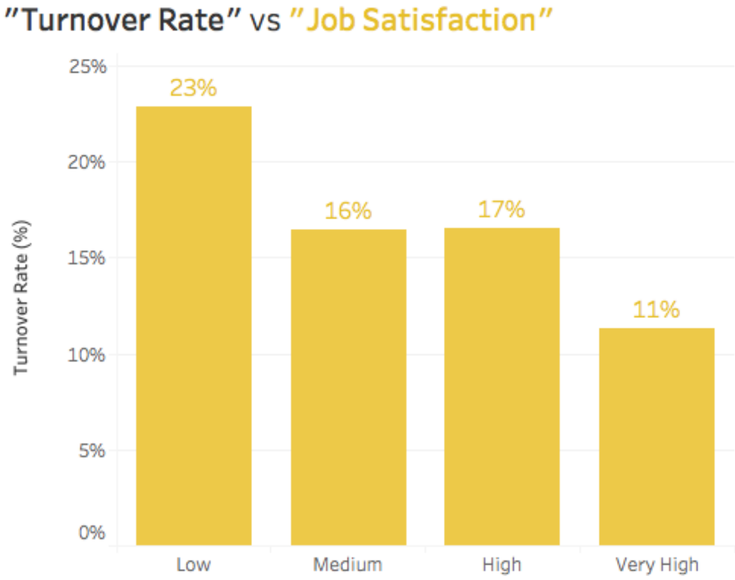
Question 7:

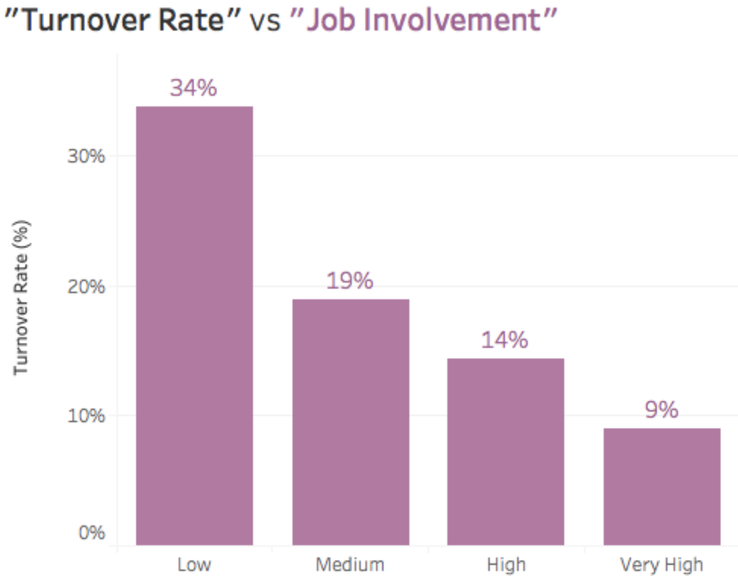
What is the relation between employees' **satisfaction** and **attrition**?

- **Low levels** of **satisfaction** are associated with **high turnover rates**





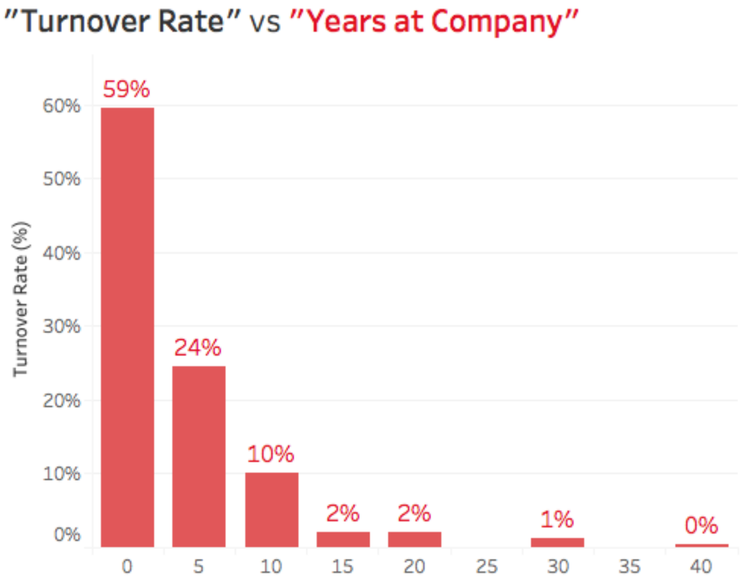


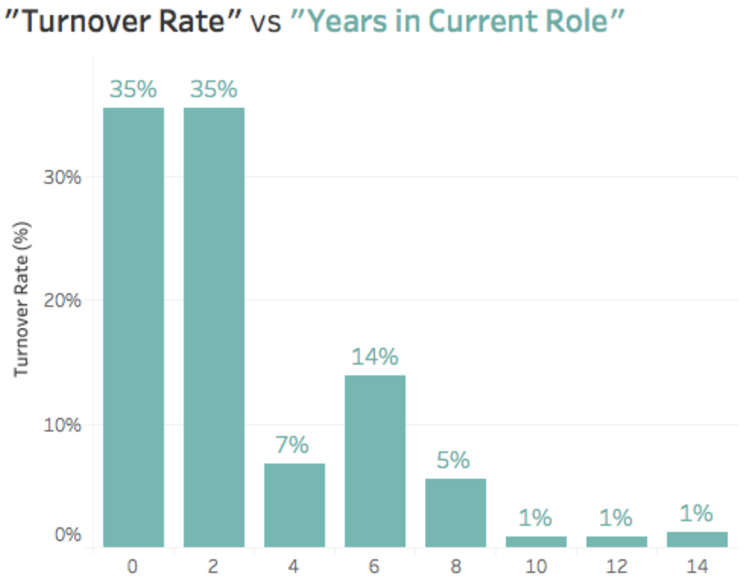


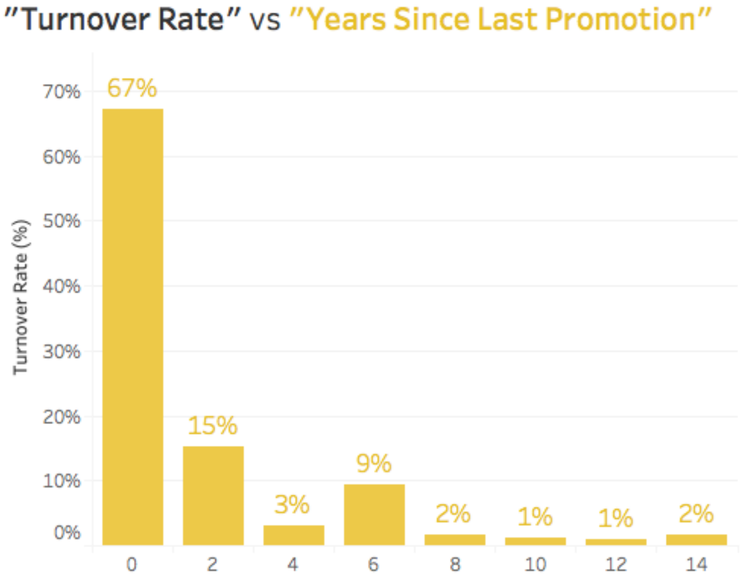
Question 8:

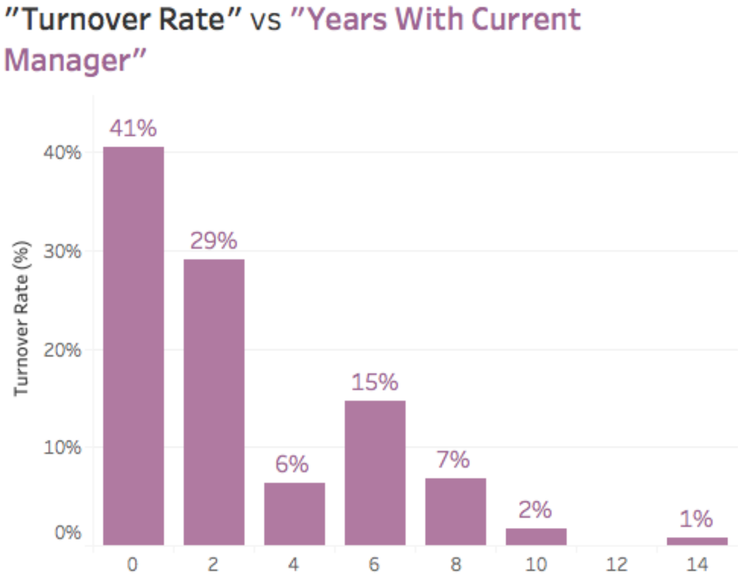
What impact has the **length** of the **career** on the **turnover**?

- **New hires** most likely to **resign**
- **Years** since last **promotion not significant**



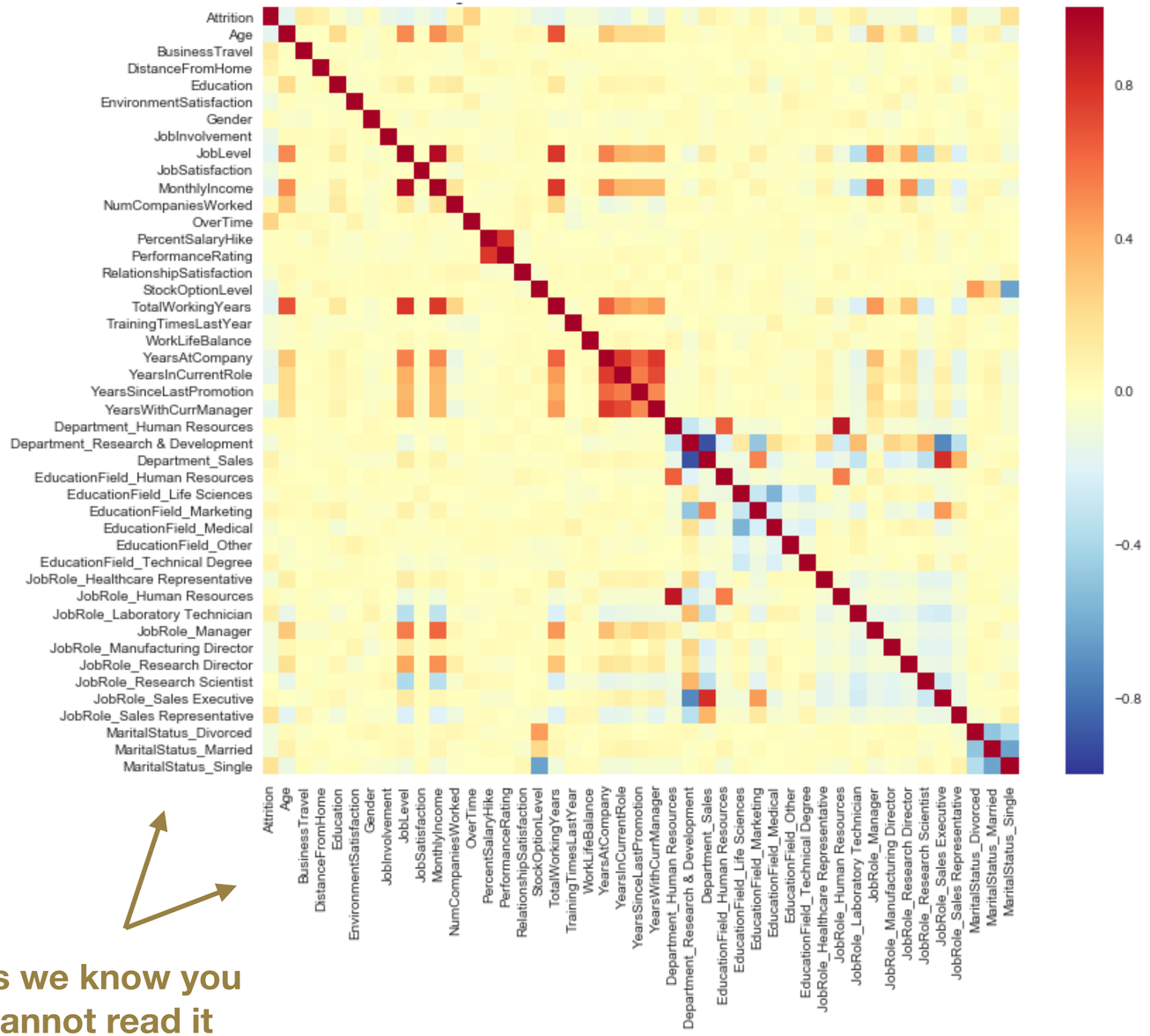






Question 9:
How are the different variables **correlated**?

- **Time** related features **highly correlated**
- **OverTime** and **MaritalStatus_Single** correlated with **Attrition**
- **Few** signs of **multicollinearity**



Yes we know you cannot read it

Question 10:

How good are our features for predicting attrition?

- For **α = 0.05:**
  26/27 features included

	Fisher	p-value
OverTime	94.66	0.00
MaritalStatus_Single	46.61	0.00
TotalWorkingYears	44.25	0.00
JobLevel	43.22	0.00
YearsInCurrentRole	38.84	0.00
MonthlyIncome	38.49	0.00
Age	38.18	0.00
JobRole_Sales Representative	37.21	0.00
YearsWithCurrManager	36.71	0.00
StockOptionLevel	28.14	0.00
YearsAtCompany	27.00	0.00
JobInvolvement	25.24	0.00
BusinessTravel	24.07	0.00
JobSatisfaction	15.89	0.00
EnvironmentSatisfaction	15.86	0.00
JobRole_Laboratory Technician	14.32	0.00
MaritalStatus_Married	12.25	0.00
JobRole_Research Director	11.69	0.00
MaritalStatus_Divorced	11.38	0.00
Department_Research & Development	10.76	0.00
JobRole_Manager	10.26	0.00
JobRole_Manufacturing Director	10.18	0.00
Department_Sales	9.66	0.00
JobRole_Healthcare Representative	9.15	0.00
DistanceFromHome	8.97	0.00

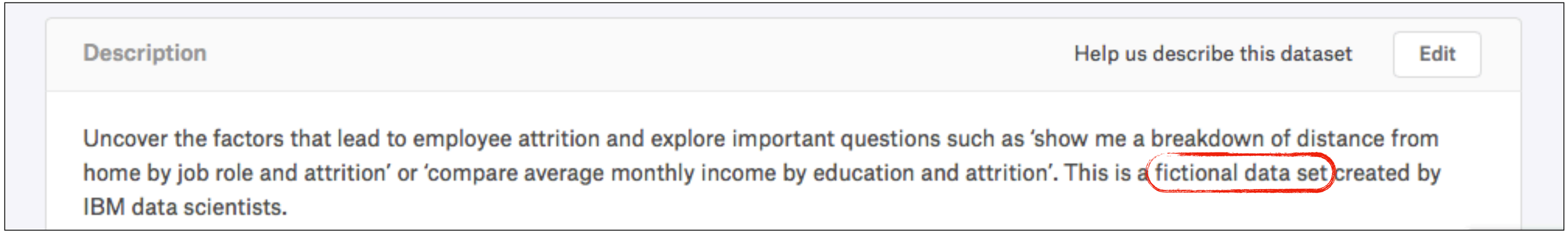
	Fisher	p-value
EducationField_Technical Degree	7.10	0.01
WorkLifeBalance	6.03	0.01
TrainingTimesLastYear	5.21	0.02
EducationField_Marketing	4.58	0.03
EducationField_Medical	3.25	0.07
RelationshipSatisfaction	3.10	0.08
NumCompaniesWorked	2.78	0.10
EducationField_Human Resources	1.95	0.16
JobRole_Human Resources	1.93	0.17
YearsSinceLastPromotion	1.60	0.21
EducationField_Life Sciences	1.57	0.21
Education	1.45	0.23
Gender	1.27	0.26
JobRole_Sales Executive	0.57	0.45
EducationField_Other	0.47	0.49
Department_Human Resources	0.42	0.52
PercentSalaryHike	0.27	0.61
PerformanceRating	0.01	0.91
JobRole_Research Scientist	0.00	0.99

Optimal Hyperparameters

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=5, max_features=17, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=5, min_weight_fraction_leaf=0.0,
            n_estimators=50, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

Representativeness findings



Source: **https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset**

Categorical Features 1/2

Feature	Categories	Definition / Comment
Attrition	{Yes, No}	Whether the employee is currently still working for the company.
BusinessTravel	{Travel_Rarely, Travel_Frequently, Non-Travel}	How frequency the employee travels for work.
Department	{R&D, Sales, HR}	The department the employee works for. An employee can only work for one department.
Education	{1=Below College, 2=College, 3=Bachelor, 4=Master, 5=Doctor}	The highest level of education obtained by the employee.
EducationField	{Life Sciences, Medical, Marketing, Technical Degree, Other}	The field of study for the most recently finished education program.
EmployeeNumber	{1, 2, .., 2068}	The employee's unique identifier.
Gender	{Male, Female}	The employee's gender (mutually exclusive).
JobInvolvement	{1=Low, 2=Medium, 3=High, 4=Very High}	To what extent the employee is involved in his/her work.
JobLevel	{1, 2, 3, 4, 5}	*Undefined* - (it is assumed a higher level implies a more senior position).
JobRole	{SalesExecutive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative}}	The current position of the employee (mutually exclusive).
JobSatisfaction	{1=Low, 2=Medium, 3=High, 4=Very High}	To what extent the employee is satisfied with his/her job.
MaritalStatus	{Married, Single, Divorced}	The current relationship status of the employee (mutually exclusive).

Categorical Features 2/2

Over18	{Yes}	Whether the employee's age is 18 or higher. Note all employees are older than 18 years old.
OverTime	{Yes, No}	*Undefined* - It has been assumed "Yes" refers to an employee whose total working hours is regularly beyond normal working hours.
PerformanceRating	{3=Excellent, 4=Outstanding}	At least to say remarkable that all evaluations are either "excellent" or "outstanding" (no "low" or "good" scores).
RelationshipSatisfaction	{1=Low, 2=Medium, 3=High, 4=Very High}	*Undefined* - (probably the quality of relationships with colleagues and managers)
StandardHours	{80}	*Undefined* - (note that all records in this column are equal to 80).
StockOptionLevel	{0, 1}	*Undefined* - (probably whether the job offer includes stock options or not.
WorkLifeBalance	{1=Bad, 2=Good, 3=Better, 4=Best}	To what extent the employee is able to combine work and personal life.

Numerical Features 1/2

Feature	Mean	Range	Definition / Comment
Age	36.9	18-60	The current age of the employee
DailyRate	802	102-1500	*Undefined*
DistanceFromHome	9.19	1-29	One-way distance from the employee's home to work. Kilometer has been assumed as the measurement unit.
HourlyRate	65.9	30-100	*Undefined*
MonthlyIncome	6.50K	1.01K - 20.0K	The monthly earnings of the employee.
MonthlyRate	1.4K	2.09K - 27K	*Undefined*
NumCompanies	2.69	0-9	The number of companies worked for previously (excluding the current company).
PercentSalaryHike	15.2	11-25	*Undefined* - (definition is unclear, especially since employees who have never been promoted (JobLevel=1) still have a salary hike of 11% or higher).

Numerical Features 2/2

TotalWorkingYears	11.3	0-40	The total number of working years at the current company plus any previous experiences.
TrainingTimesLastYear	2.8	0-6	The number of trainings the employee participated in last year.
YearsAtCompany	7.01	0-40	The total number of working years at the current company.
YearsInCurrentRole	4.23	0-18	The number of years the employee has worked in the current position.
YearsSinceLastPromotion	2.19	0-15	The number of years since the last promotion. Note that an employee can be promoted without changing postions.
YearsWithCurrManager	4.12	0-17	The number of years the employee has worked for his/her current manager.