Enterprise Data Science Bootcamp

Human Resources Analysis

Predict Attrition

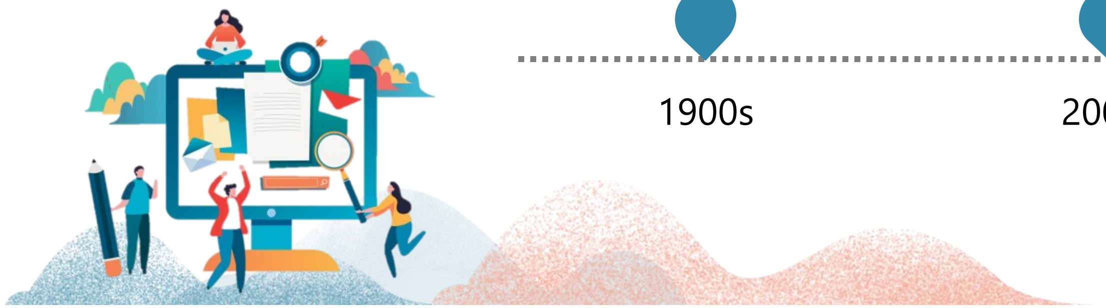
m20191415 Carolina Nascimento Alves | m20190849 Diana Ferreira

m20191417 Eliana Sobral | m20190792 Miguel Pereira

Human Resources Analytics

Without data, you're just another person with an opinion… (W. Edwards Deming)

- Human Resources analytics is a **data-driven approach** toward Human Resources Management where **Human Resources data is collected and analyzed** in order to <u>improve an organization's workforce performance</u>
- This analysis is a measured evidence of **how Human Resources initiatives** are contributing to the <u>organization's goals and strategies</u>
- Analytics helps companies **track absenteeism**, **turnover, burnout, performance** and much more
- **Human Resources decisions** are <u>no longer based on gut feeling</u>

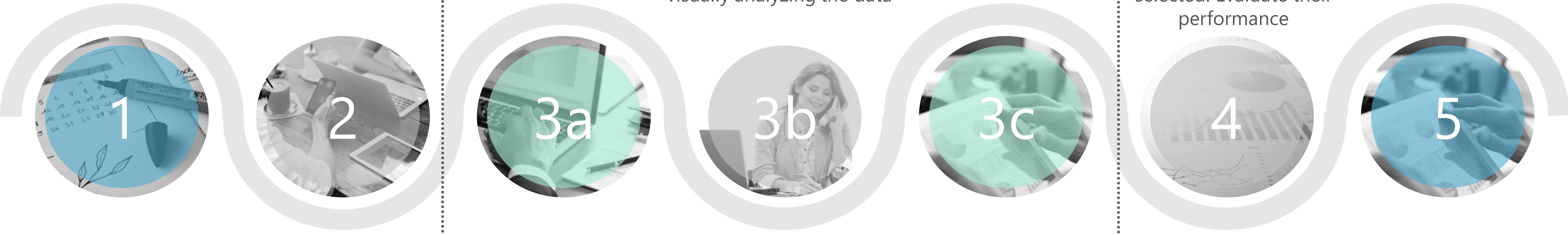


Operational Human Resources	Strategic Human Resources	**Data-driven Human Resources**
1900s	2000s	**Now**

- Make **better decisions** using data
- Create a **business case for HR interventions**
- **Test the effectiveness** of these interventions
- Move from an operational partner to a **tactical, or even strategic partner**

Methodology

- Data analysis were performed using CRISP-DM (*Cross-Industry Standard Process for Data Mining*) a structured approach to planning a data mining project such as this one



Business Understanding

Ask relevant questions and define objectives for the problem

Data Understanding

Collect all necessary data for the project

- Collect initial data
- Describe, explore data
- Verify data quality

Data Preparation

Data Cleaning

Fix the inconsistencies within the data and handle possible missing values

- Few columns without meaningful information were dropped
- Data quality leveled up to the requirements of the selected analysis techniques

Data Exploration

Form hypotheses about the defined problem by visually analyzing the data

- The dataset was firstly analyzed in Power BI to gather insights and to better understand the dataset

Feature Engineering

Select important features and construct more meaningful ones using raw data

- Categorical variables were transformed into dummies, to be used in further analysis / modelling
- New variables were created

Predictive Modelling

Train and fine tune the predictive models selected. Evaluate their performance

- Select predictive model
- Generate test design
- Build and fine tune the model
- Evaluate results

Data Visualization

Communicate the findings with key stakeholders using interactive visualizations

- Present all the main findings and conclusions to the key stakeholders
- Next steps

Business Understanding

Main Goal - <u>predict the attrition of an employee</u>, i.e. the probability of an employee with certain characteristics stay or quit his/her current job in the company

- How high is the annual employee turnover?
- How much of the company employee turnover consists of regretted loss?
- Do you know which employees will be the most likely to leave the company?
- What measures can we develop to avoid attrition?

2 Data Understanding

Data - The dataset was provided by BI4All – *Turning Data Into Insights*, with information from all employees that have been working at a specific company within a time period

Number of Records – 1470

Demographics

- *Age*
- *Gender*
- *Marital Status*
- *Education*
- *Education Field*

Income / Payment

- *Daily Rate*
- *Hourly Rate*
- *Monthly Rate*
- *Monthly Income*
- *Percent Salary Hike*
- *Stock Option Level*

Performance

- *Performance Rating*

Engagement

- *Job Involvement*
- *Job Satisfaction*
- *Relationship Satisfaction*
- *Environment Satisfaction*
- *Working Life Balance*

Job / Professional Life

- *Distance from Home*
- *Business Travel*
- *Standard Hours*
- *Department*
- *Job Level*
- *Job Role*
- *Training Times Last Year*
- *Years at Company*
- *Years In Current Role*
- *Years Since Last Promotion*
- *Years with Current Manager*
- *Total Working Years*
- *Number of Companies Worked*

Drawbacks

The dataset is lacking information that would be important to the analysis in place:

- Chronologic data >> Timestamp of the main events
- Hierarchical structure of the company
- Main drive for attrition
- Benchmarking within companies in the same industry

3 Data Preparation

Python Power BI

Pre-processing raw data into a form that can readily and accurately be analyzed and therefore gather insights and to better understand the dataset:

Data Cleaning

The columns dropped:

- *Over18* (only with 'Yes')
- *EmployeeCount* (only with '1')
- StandardHours (only with '80')

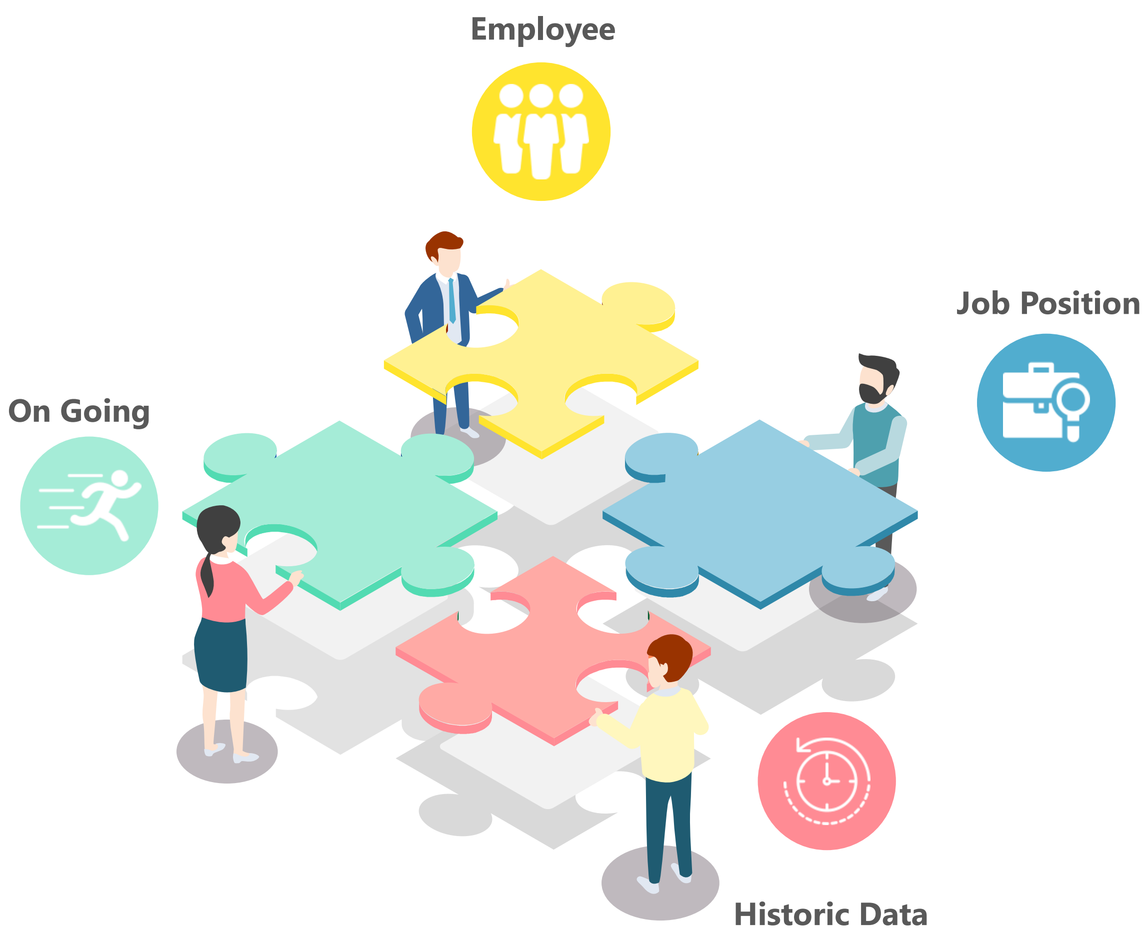
Data Exploration

Feature Engineering

- *Age_Entry* *(into the company)*
- *IncRateRatio* *(income to daily rate ratio)*
- *IncMonthlyRate* *(income to monthly rate ratio)*
- *Flag_1sJob* *(Indicates if it is first job)*
- *Perc_lftm_company* *(% work lifetime in the company)*
- *Age_Workforce* *(start of professional life)*
- *IncYearsRatio* *(income to years at company ratio)*
- *HrDailyRate* *(hourly to daily rate ratio)*
- *Avg_prev_worktime* *(average time worked at other companies)*
- *MonthlyRate* *(monthly to hourly rate ratio)*
- *Perc_tmcompany_curr_manager* *(% of time with current manager)*

A few variables were aggregated within ranges for better comprehension and meaningful insights

Data Preparation > Exploration



Data Preparation > Exploration

Python

Power BI

Same level of attrition for **both genres**

Singles are more likely to leave the company

Highest attrition in **Marketing, HR and Technical Degree**

‘The younger we are, the more likely it is for us to leave’

>>> Higher incidence on **younger ages**
70% of who leave the company have between 20 and 40 years old

Employee

‘Oh no...not again!!’

Who **travel frequently** are more likely to leave the company

Job Position

Manager and Director
Job Roles with almost NO attrition

Lower the job level, higher the tendency to leave

R&D is the largest department and with the highest turnover

HR and Sales with a higher predisposition to leave

Data Preparation > Exploration

Python

Power BI

Bad level of work life balance (satisfaction/engagement), have more tendency to leave

Greater attrition for **who did not have training** or were **not promoted** in the last years

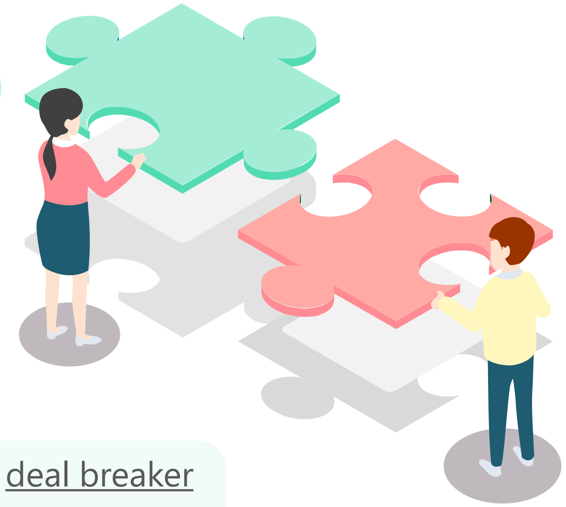
More than 50% of departures work **overtime**

All have an above evaluation
>>> No significant distinction

It may indicate a **lack of evaluation policy**. Nonmeritocratic system can cause discontent and does not encourage better performance.

>>>**Income** seems to be a deal breaker
Almost 60% of who leave has a lower income (up to 4000)

On Going



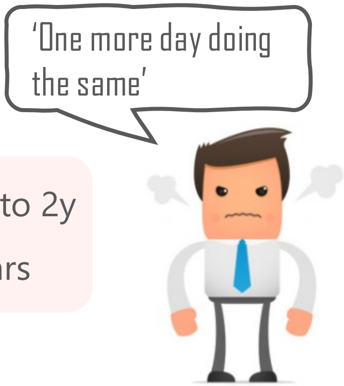
Historic Data

3/5 started in the company and **1/5 of them leave**

3/4 leave the company with **up to 10yrs of experience**

Half of the employees worked **only in the company**

4/5 spend no more than **2yrs on previous jobs**



3/5 with the **same manager** for up to 2y

3/4 in the **same role** for up to 3 years

Less likely to leave >> probably because it's happy and / or resigned

Modelling > Clustering

In order to complement our descriptive analysis, we decided to segment all employees into well defined groups by performing a **cluster analysis**

Variables All original non-categorical and transformed variables were used

Segments 4 perspectives: **Employee, Job Position, Historic and On Going**

Performance 3 Datasets to compare performance: **Original, Z-score Standardization, Min-Max Normalization**

Models Models used: **Correlation analysis, Hierarchical Clustering, K-means Clustering**

Fine Tuning An improved approach was always performed by removing some uninformative or unclear features in order to form better segments

Results **12 segments** (3 for each perspective)

Modelling > Clustering

Employee

Python

Power BI

Dataset Original non-scaled

Key Variables *Age, DistanceFromHome*

Cluster 0

Old

426 employees (29%)

Oldest group

Live close to the company

Cluster 1

Middle Age

331 employees (23%)

Medium Age group

Live far away from the company

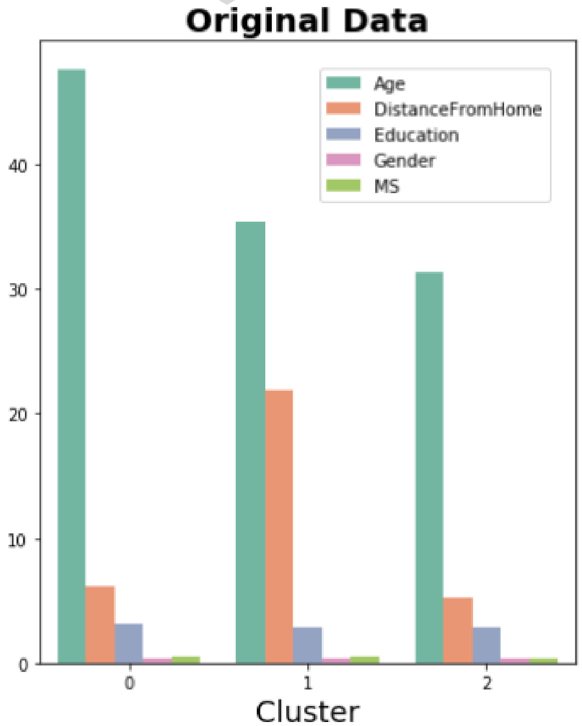
Cluster 2

Young

713 employees (49%)

Youngest group

Live close to the company



4 Modelling > Clustering

Python | Power BI

Job Position

Dataset MinMax Normalization

Key Variables *FLG_1stJob, Dep. Sales, Dep. HR, Job Level*

Cluster 0

Experienced Researcher

701 employees (48%)

Mainly **R&D Department**

Medium / High Job Level

Not 1st Job

Cluster 1

Rookie Researcher

323 employees (22%)

Mainly **R&D Department**

Lowest Job Level

1st Job

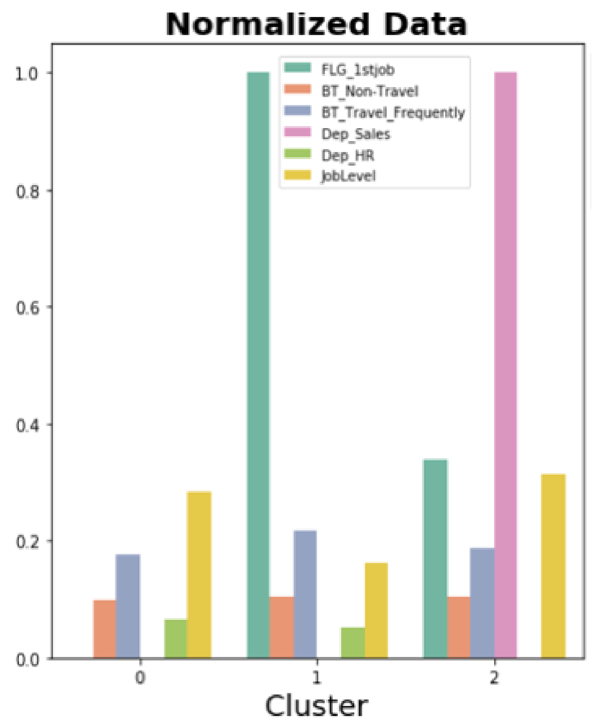
Cluster 2

Salesperson

446 employees (30%)

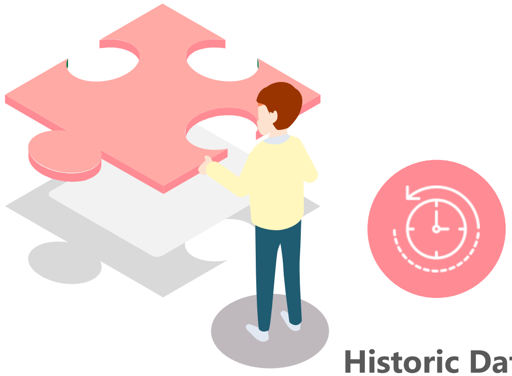
Sales Department

Highest Job Level



4 Modelling > Clustering

Python
Power BI



Historic Data

Dataset

MinMax Normalization

Key Variables

Age_Entry; Age_Workforce; Avg_prev_worktime; NumCompaniesWorked; TotalWorkingYears

Cluster 0

Senior Loyal

456 employees (31%)

Entry at **younger age**

Highest working time at the company

Medium / low number of **companies worked**

Cluster 1

Frequent Changers

475 employees (32%)

Entry at **older age**

Lowest working time at the company

Highest number of **companies worked**

Cluster 2

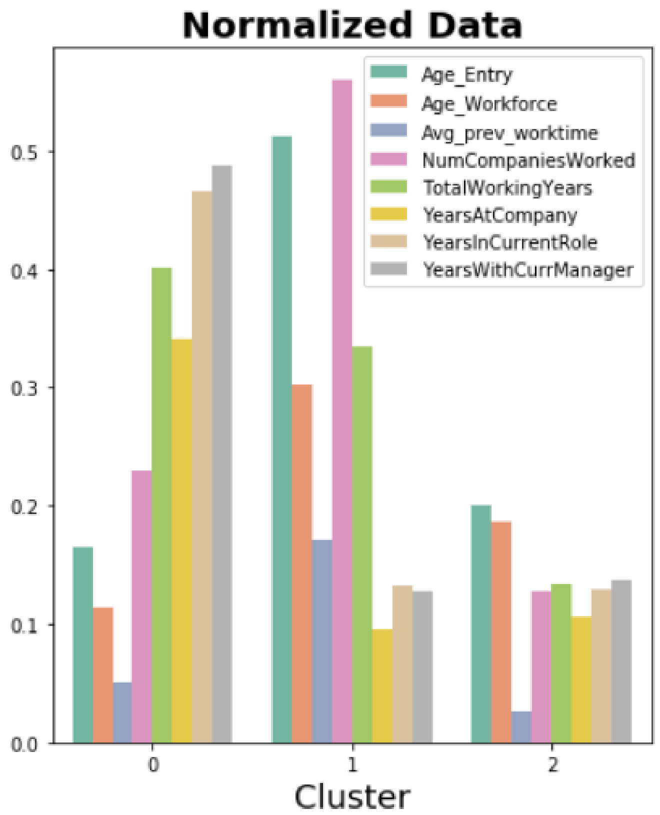
Work Beginners

539 employees (37%)

Entry at **younger age**

Low working years

Few number of **companies worked**



4 Modelling > Clustering

Python Power BI

On Going

Dataset Z-Score Standardization

Key Variables *MonthlyIncome, RelationshipSatisfaction, YearsSinceLastPromotion*

Cluster 0

Wealthy

259 employees (18%)

Highest Income

Medium relationship with peers

Promoted a long ago

Cluster 1

Friendly

714 employees (49%)

Lowest Income

Very Good relationship with peers

Recent Promoted

Cluster 2

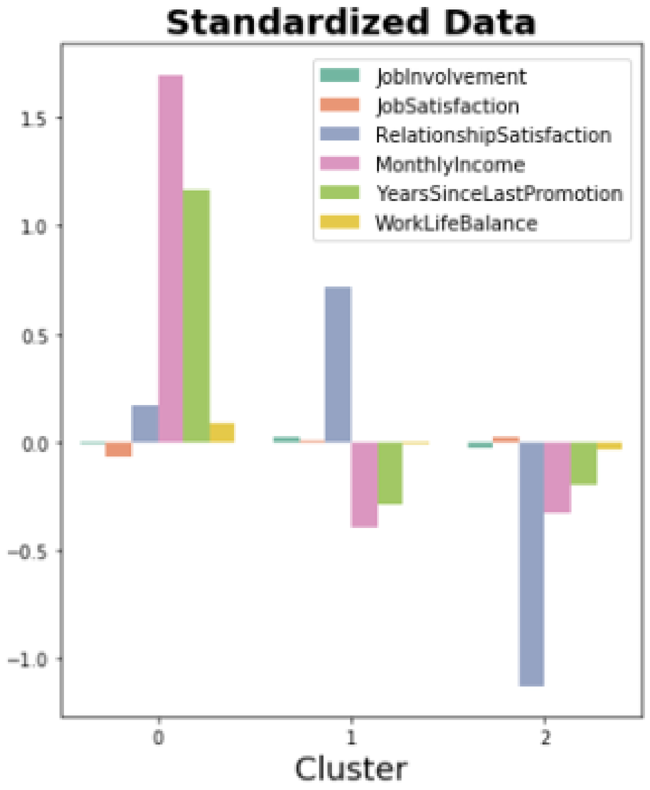
Unpleasant

497 employees (34%)

Low Income

Worst relationship with peers

Promoted in the past **few years**



Modelling > Clustering - Attrition

Base attrition: 16%

From all groups presented, we can mention the following 4 as **“most likely” to leave the company**

Beginners – Attrition:19%

- Youngest workers
- First job experience
- Low monthly income
- Good relationship with colleagues

Rotative Workers – Attrition:28%

- R&D and Sales Department
- Frequent job changers
- Older and experienced workers
- Medium monthly income
- Low years in current role and with current manager

Unfriendlies – Attrition:21%

- Bad job relationship with peers
- Younger group
- Live close to the company
- High turnover

Unknown Group

- Remaining employees that couldn’t relate with any cluster and left the company as well

4 Modelling > Correlation

Python

Power BI

Pearson and Spearman analysis

- *Job Level* and *Monthly Income* (> 0.9)
- *IncHourlyRate* and *Monthly Income* (> 0.8)
- *YearsAtCompany* and *YearsInCurrentRole/YearsWithCurrManager* (> 0.8)

Dropped variables

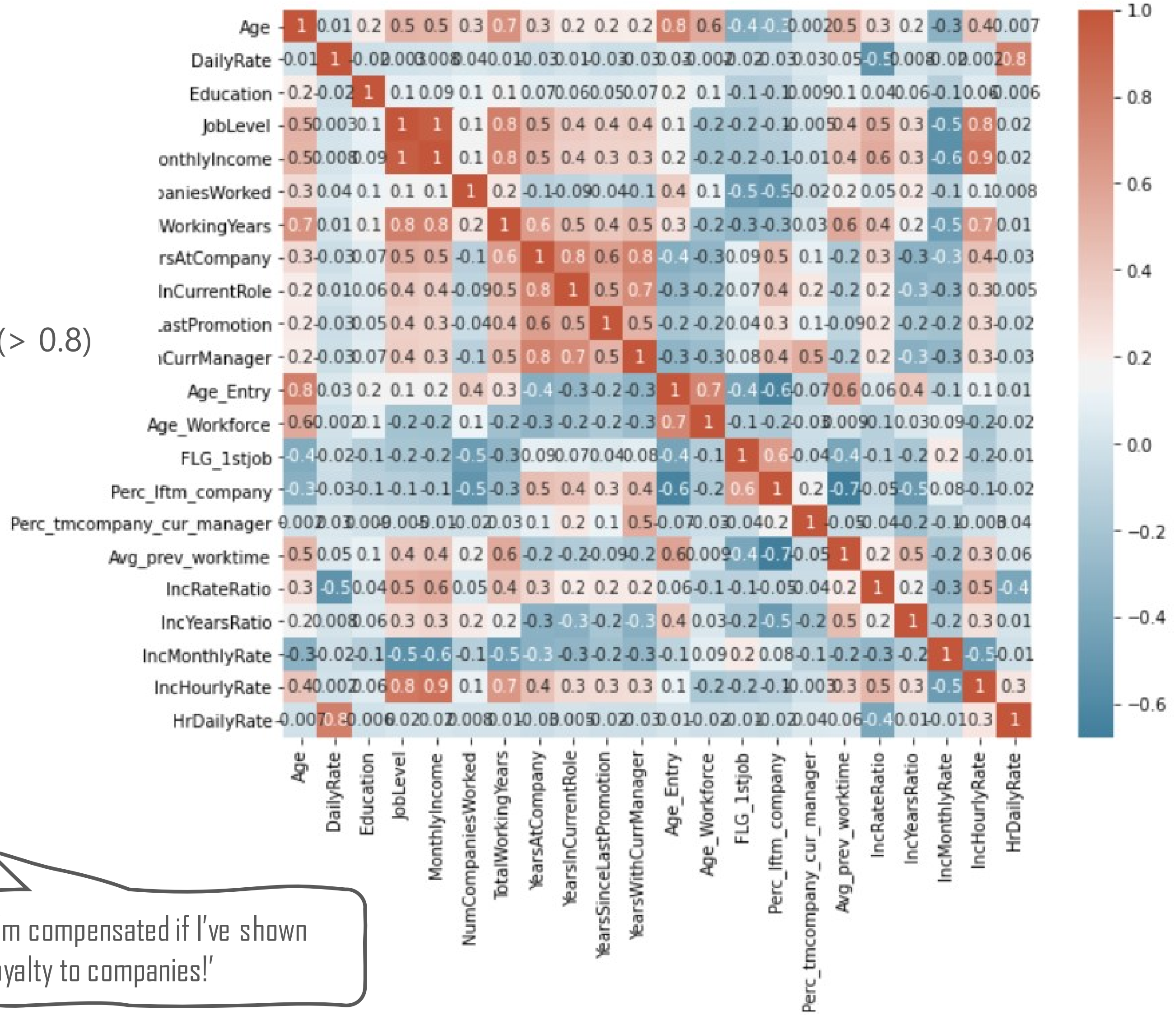
- ✕ Job Level
- ✕ IncHourlyRate
- ✕ YearsAtCompany

Income to years at company ratio has a **positive correlation** with *average time in previous companies*

Daily Rate has **no correlation** to *Monthly Income.*

'Not fair...The company charges more for my services, and I receive the same'

'I'm compensated if I've shown loyalty to companies!'



4 Modelling > Model Evaluation and Selection

an iterative process…

Selected Features

- Recursive Feature Elimination
- Decision Tree Feature Importance

Most Important Variables:

- **Monthly income** ⭐
- Daily and Hourly rate
- Overtime
- Age
- Distance from home
- Income to daily rate ratio
- Monthly rate to income ratio
- Total working years

Final Datasets

3 types of datasets:

- Data **as is**
- Data normalized with **Robust Scaler**
- Data normalized with **MinMax Scaler**

Variable nº of features of each dataset

Model Selection

- GridSearchCV find best parameters with each dataset for Decision Tree and Neural Networks
- Random Forest
- Gradient Boosting
- Logistic Regression
- Naive Bayes
- More to come…

Fine Tuning

Modelling > Models Analysis and Results

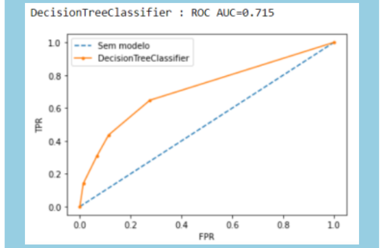
Python

Power BI

Decision Tree

Worst performer (as expected)

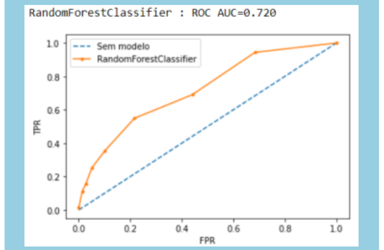
AUC = 0.715



Random Forest

Barely better than Decision Tree algorithm

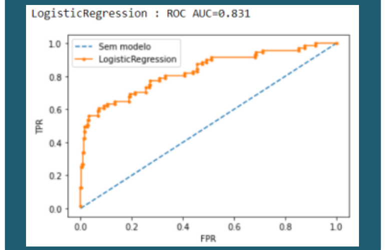
AUC = 0.720



Logistic Regression

Best performer along with Neural Networks

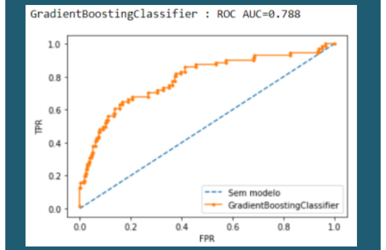
AUC = 0.831



Gradient Boosting

One of the top 3 needs parameter tunning

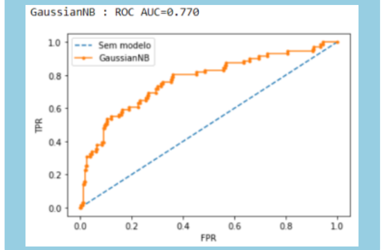
AUC = 0.788



Naïve Bayes Classifier

Best with dataset normalized with MinMax Scaler

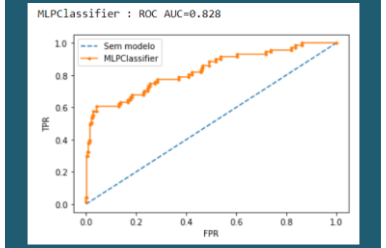
AUC = 0.770



Neural Networks

Top performer with any type of dataset

AUC = 0.828



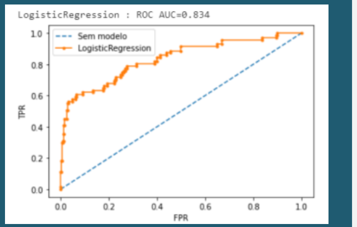
Modelling > Fine Tuning Best Models

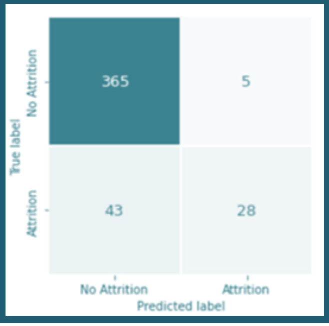
All models can be really accurate when predicting no attrition, but still have some difficulties when predicting attrition (as seen by the F1 Score)

Logistic Regression

AUC = 0.831

Marginally Improved as parameters barely change



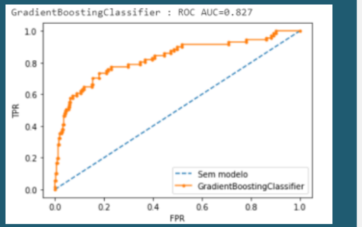


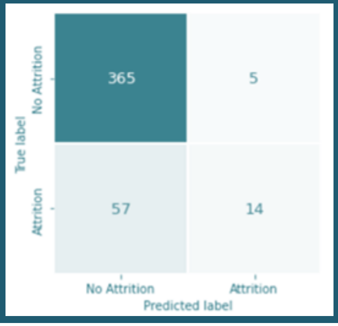
	precision	recall	f1-score	support
No Attrition	0.89	0.99	0.94	370
Attrition	0.85	0.39	0.54	71
accuracy			0.89	441
macro avg	0.87	0.69	0.74	441
weighted avg	0.89	0.89	0.87	441

Gradient Boosting

AUC = 0.827

Biggest Improvement but it has the least precision to forecast attrition



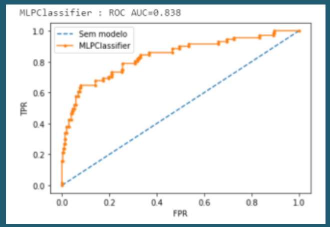


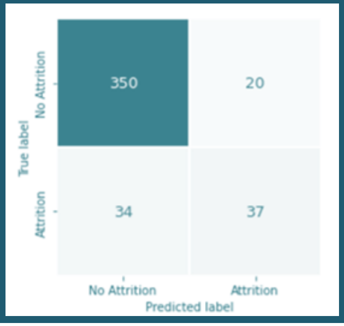
	precision	recall	f1-score	support
No Attrition	0.86	0.99	0.92	370
Attrition	0.74	0.20	0.31	71
accuracy			0.86	441
macro avg	0.80	0.59	0.62	441
weighted avg	0.84	0.86	0.82	441

Neural Networks

AUC = 0.838

Top performer even to predict attrition





	precision	recall	f1-score	support
No Attrition	0.91	0.95	0.93	370
Attrition	0.65	0.52	0.58	71
accuracy			0.88	441
macro avg	0.78	0.73	0.75	441
weighted avg	0.87	0.88	0.87	441

Retain or not retain? *There is the rub!*

In case of attrition, should we retain?

The vital few and trivial many - **Pareto's Principle**

It suggests that 20% of the workforce accounts for 80% percent of the output.

Is he the one that we truly want to retain? Is he part of the 20% group?

Should we fight to keep a star?

YES.

'With a higher bonus range, it's likely that word will get out, and then everyone else will start to feel underpaid.'

Data shows that **half of employees** who accept a **counteroffer** end up leaving **within 12 months**

Beside income, what do the employees look for?
Firm culture, reputation, opportunities, team leader, inclusion initiatives, diversity, management's support, promotions, training, etc.

NO!

Is there a long-term retention plan for the key-employees?

'I get a good sense of which folks are flight risks and then take steps to retain them, whether it's giving them bigger challenges, removing obstacles, or even finding them a more suitable role on another team.'

'I'm kicking myself for ignoring all the signs.'

'This shouldn't be a crisis. I should have had a pipeline, been more proactive about succession planning, retention—all of it.'

Employee Value

- Employee value is a function of Avg. Rate and Monthly Income

with Avg. Rate being an average rate between Daily, Hourly and Monthly Rate

$$Employee\ Value = Avg.Rate - Monthly\ Income$$

$$Avg.Rate = \frac{(DailyRate\times22)+(HourlyRate\times8\times22)+Monthly\ Rate}{3}$$

(Assumption - All services' types have equal probability to be provided)

- Employees who were **recruited** last year present a **higher Employee Value than** the ones who **left the company,** which in turn **is greater than** the one verified by those **who stayed**

- Employees with a **higher Average Rate** **do not have** a **higher Income**, in fact, the **Education field** with the **greatest Average Rate** is the one where the employees **have a lower Income**

- Employees in **management positions** (Manager / Director) present a **negative Employee** Value since their **Income is much higher than their Average Rate**

Can we easily replace the ones who leave?

Modelling > One year in advance

1 YEAR

Columns **52 to 25** **New Dataset** **Rows** **1470 to 806**

- Trying to predict attrition 1 year ahead, there are a lot of variables we **cannot** use:
  - *Marital Status*
  - *Income*
  - *JobSatisfaction*
  - …
- Variables we **can** use:
  - *Age & Gender*
  - *EducationField* *
  - *Department & JobRole* **
  - *NumCompaniesWorked*
  - *TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager*
  - *Age_Entry, Age_Workforce, FLG_1stjob*
- All year related variables, we subtract 1 year

- Every employee that had *YearsAtCompany* equal to 0 had to be removed from the dataset – they weren't there in the prior year
- Every employee with *YearsInCurrentRole*, *YearsSinceLastPromotion* or *YearsWithCurrManager* equal to 0 also had to be excluded – we don't know the reality the year before

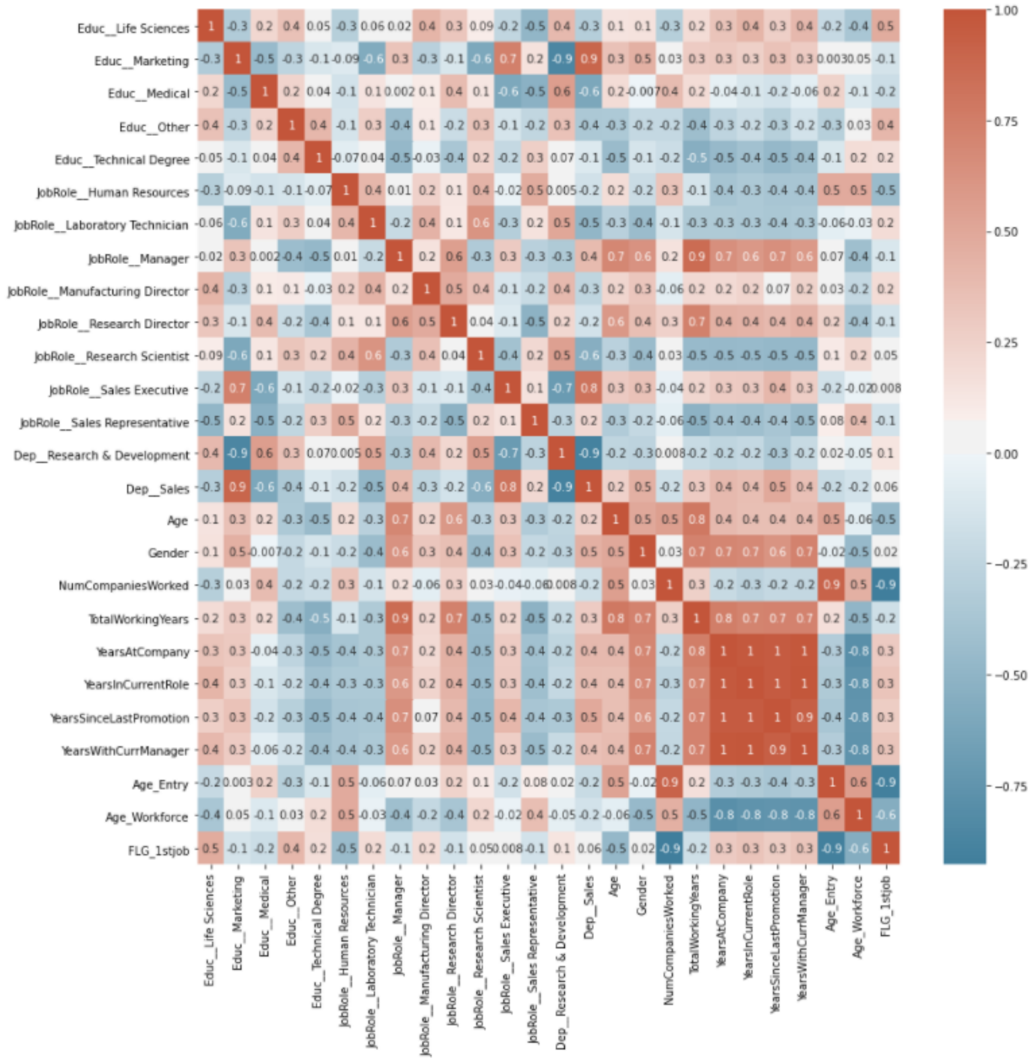
* *EducationField* is the only dubious variable: a person can change their primary education field over the years, by attending different course/post graduate programs throughout the years, but that's unlikely to happen within 1 yr.

** Since the employee have to have the same manager and current role for at least one year, we can assume, the *Department* and *JobRole* stays the same, so we're including these 2 variables

Modelling > One year in advance

Feature Selection

- *Department* variables were redundant with some *JobRole* variables (i.e. – *JobRole* of *Sales Executive* was always part of the department "*Sales*")
- Spearman correlation analysis showed that a lot of the "year" variables were correlated.
- Dropped variables:
  - *Age_Entry*
  - *Dep_Research Development*
  - *Dep_Sales*
  - *FLG_1stjob*
  - *TotalWorkingYears*
  - *YearsInCurrentRole*
  - *YearsSinceLastPromotion*
  - *YearsWithCurrManager*
- Now working with only **17 variables**



Modelling > One year in advance

Model Fine Tuning and Selection

1 YEAR

Datasets

- 2 **RobustScaled** (13 and 3 variables)
- 2 **MinMaxScaled** (12 and 6 variables)

Models

- Logistic regression
- Decision Tree (2 different ones)
- Neural Networks (2 different ones)
- Gaussian Naïve Bayes
- Random Forest
- Gradient Boosting

Chosen model: **Gaussian Naïve Bayes**, with a 13 variable dataset

Criteria**:**

Highest recall >> To make sure all the employees with intentions of leaving are identified

Precision isn't as important >> we rather have a lot of false positives and improve overall employee moral with our efforts, than have a lot of false negatives and miss the opportunity to retain talent.

	Predictions		
Reality	**0**	**1**	**Total**
0	159	52	211
1	17	14	31
Total	176	66	242

- Accuracy: 0.7149
- Precision: 0.2121
- **Recall**: 0.4516
- F1-Score: 0.2887

```
Full report:
              precision    recall  f1-score   support

           0       0.90      0.75      0.82       211
           1       0.21      0.45      0.29        31
```

'Total predicted employees who will leave after 1 year is 203, with a **21% precision**, that means that the model will **correctly predict 42 employees** that are leaving the company in one year's time.
For a total of 806 employees, we will **make effort on 25%** of them to guarantee that those **5% won't leave**.'

'With a **recall of 45%**, there are still **51 employees that the model will miss**.
More variables (like the Income a year prior) would contribute to a higher performance model.'

Overcoming Challenges

- Understanding the **meaning, range and context** of the dataset variables
  - Specifically, *Hourly, Daily and Monthly Rate*.
    - Monthly rate doesn't equate to 22 days or less of daily rate, and daily rate doesn't equate to 8 hours or less of hourly rate, being that the company might change the rate (by a lot) depending on the time period it's selling the employees' services for.
- Parameters tuning and selecting the correct number of variables for the dataset to train the models
- Iteration produces better results, but it's time consuming
- Difficulties in having high values for **Recall** in any model
- **Few** relevant variables useful for predicting attrition in 1 year

Next Steps

- Better and more variables:
    - What is the main drive for attrition?
    - Timestamp of the main events

- Implementing the model:
    - Define the frequency of the model execution
    - Define a Control Group where no measures will be implemented (BaU) to be able to ascertain if the model is useful or not
    - Periodically check the **model's performance** because it might deteriorate over time