Introduction to Data Science - Spring 2023
Anne Cuzeau - Andrew Nguyen - Eric Vandament

IBM Attrition dataset

What makes IBM employees quit?

...is it just about pay?

Table of content

The Team:

Anne Cuzeau
Andrew Nguyen
Eric Vandament

About our data

- Hypothetical dataset created by IBM data scientists (Kaggle). Very few NaN

  BUT: once the employeeID removed: duplicates!

- This dataset has **23,436 rows** and **37 columns** describing different employees profiles
- Both numeric and categorical variables
- **Main goal**: find a model to help HR predict attrition and understand how pay rate influences employee retention

Variables

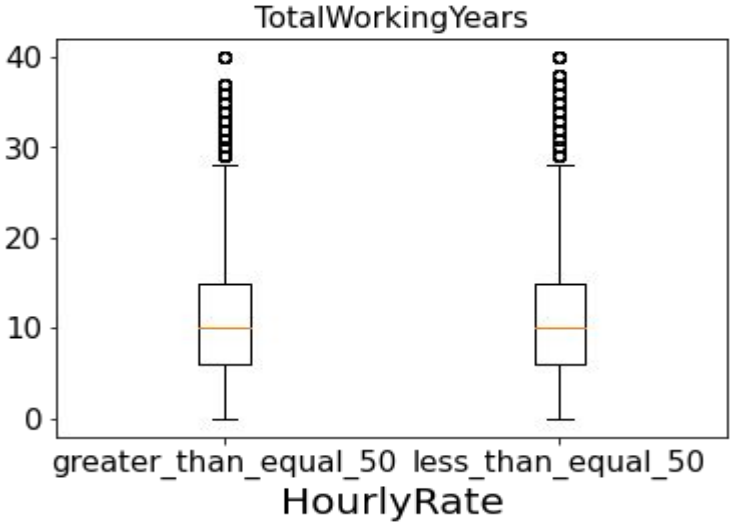
- **Job related**: Job role, Department, job level, travel frequency
- **Demographics**: Age, Gender, Distance from home, relationship status, educational background
- **Career Descriptions**: Number of companies worked for, years at company, years in current role, years since last promotion
- **Pay**: Daily Rate, Hourly Rate, Monthly Income, Monthly Rate

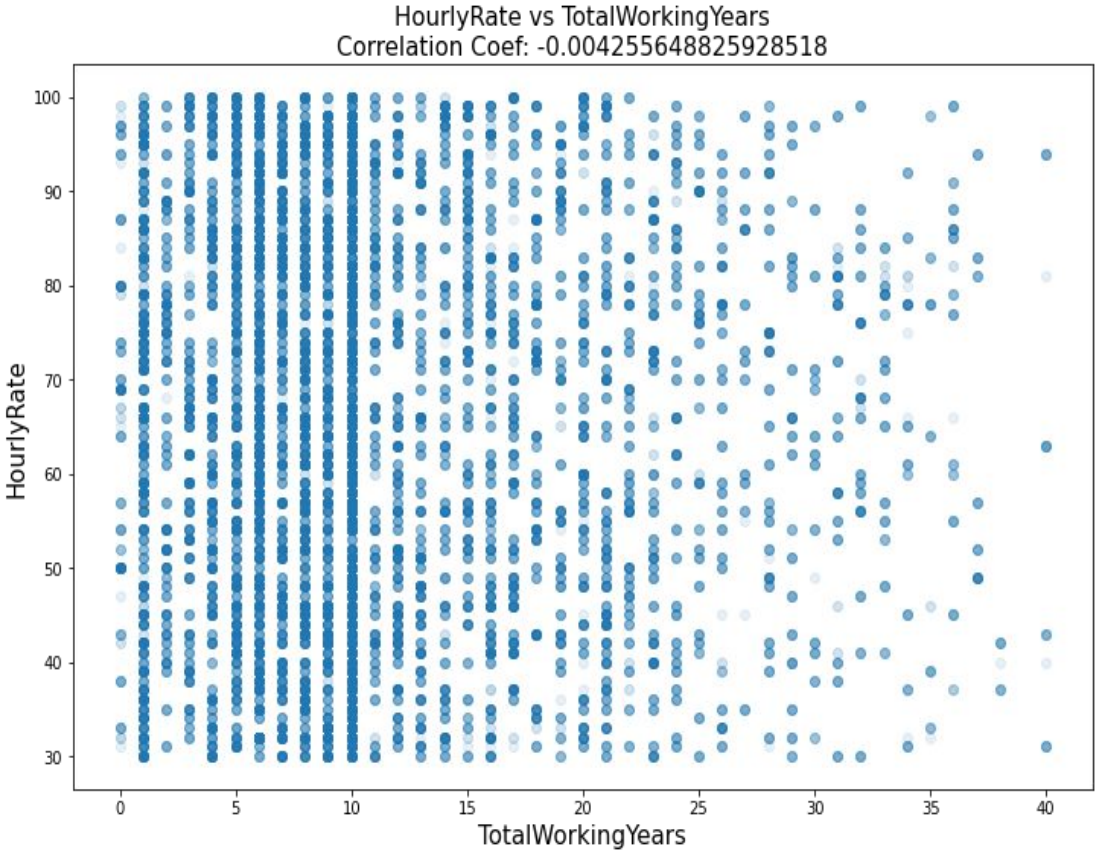
Does pay rate influence attrition?

... or is it influenced by other factors?

What factors influence pay rate?

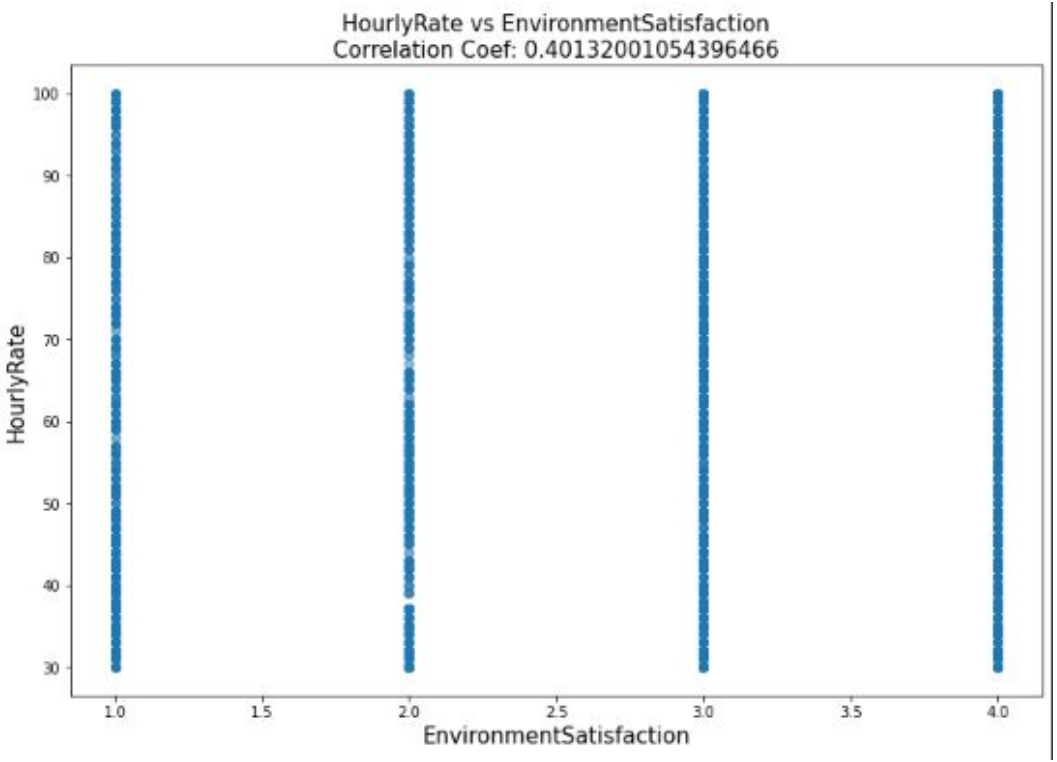
- Comparing pay rate to other factors, most of the correlation Coefficients are below .1





What factors influence pay rate?

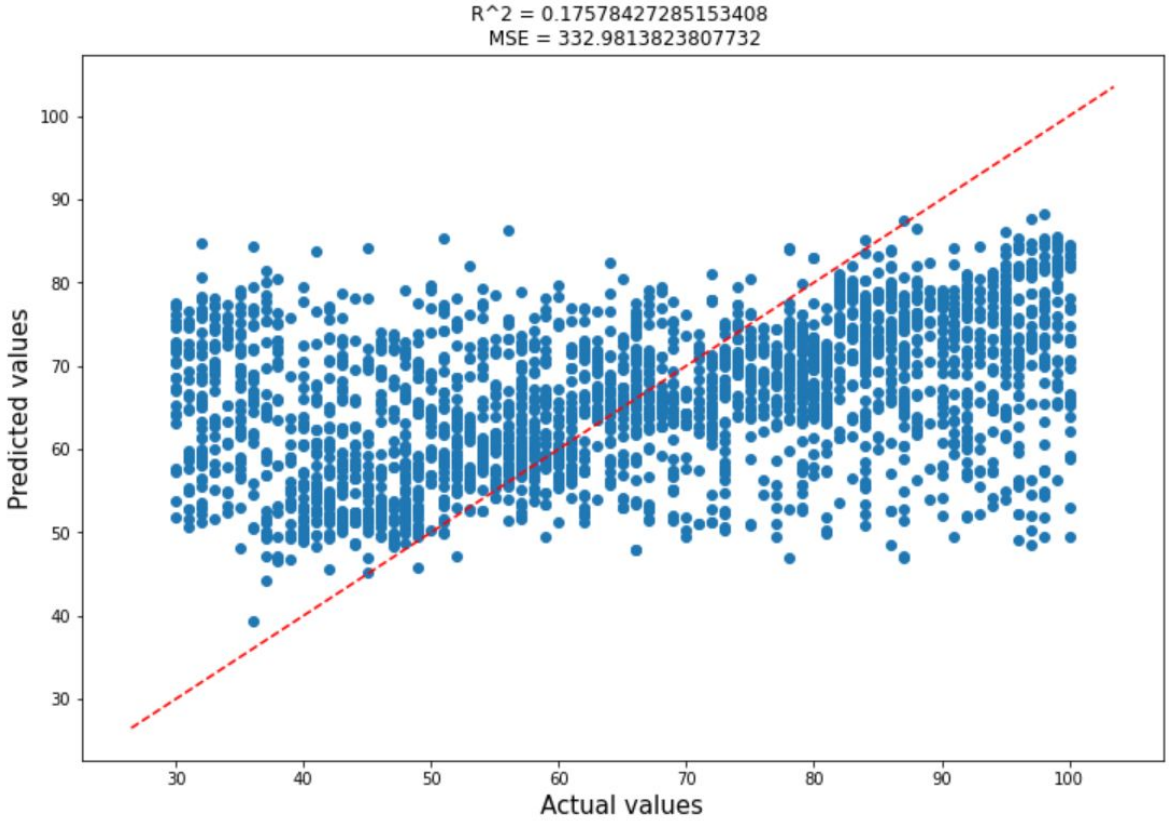
- With the highest being .4 with Environment Satisfaction



What factors influence pay rate?

Linear Regression Model

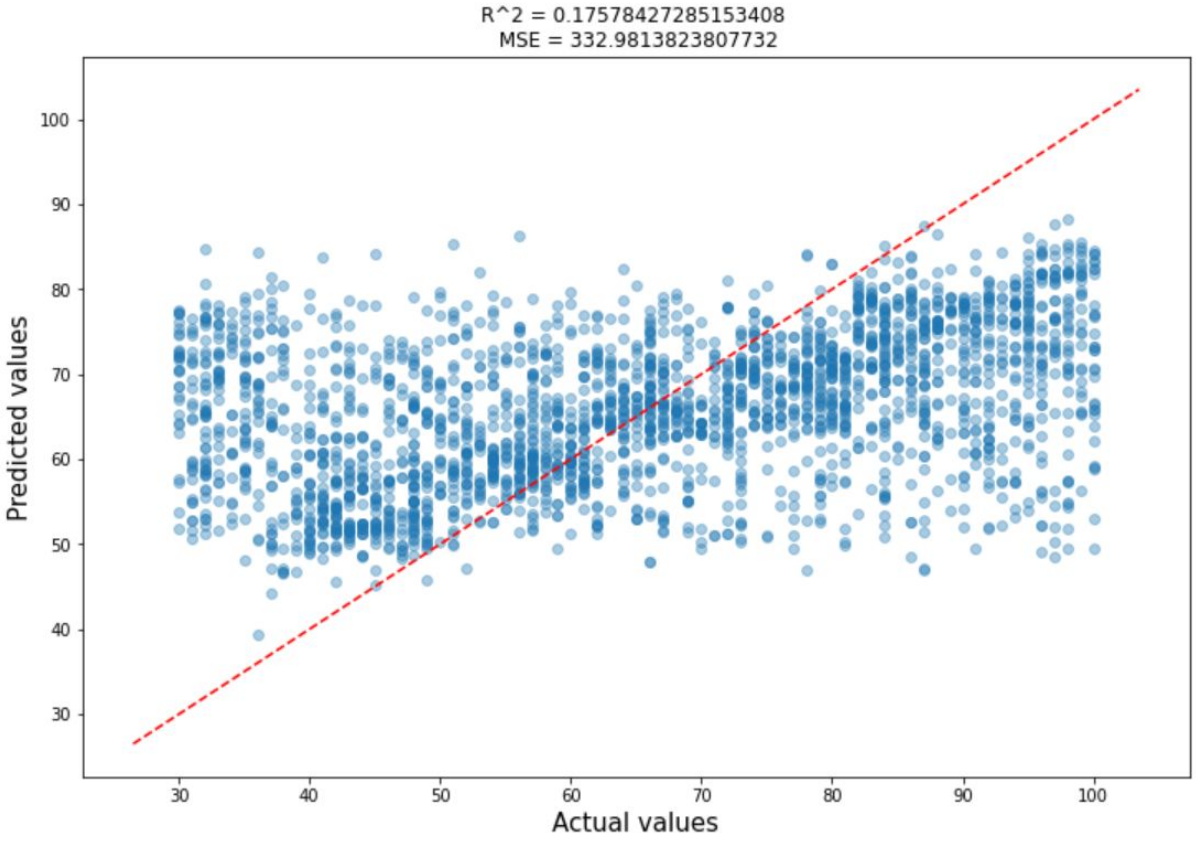
- Fitting Hourly Rate shows us no real correlation



What factors influence pay rate?

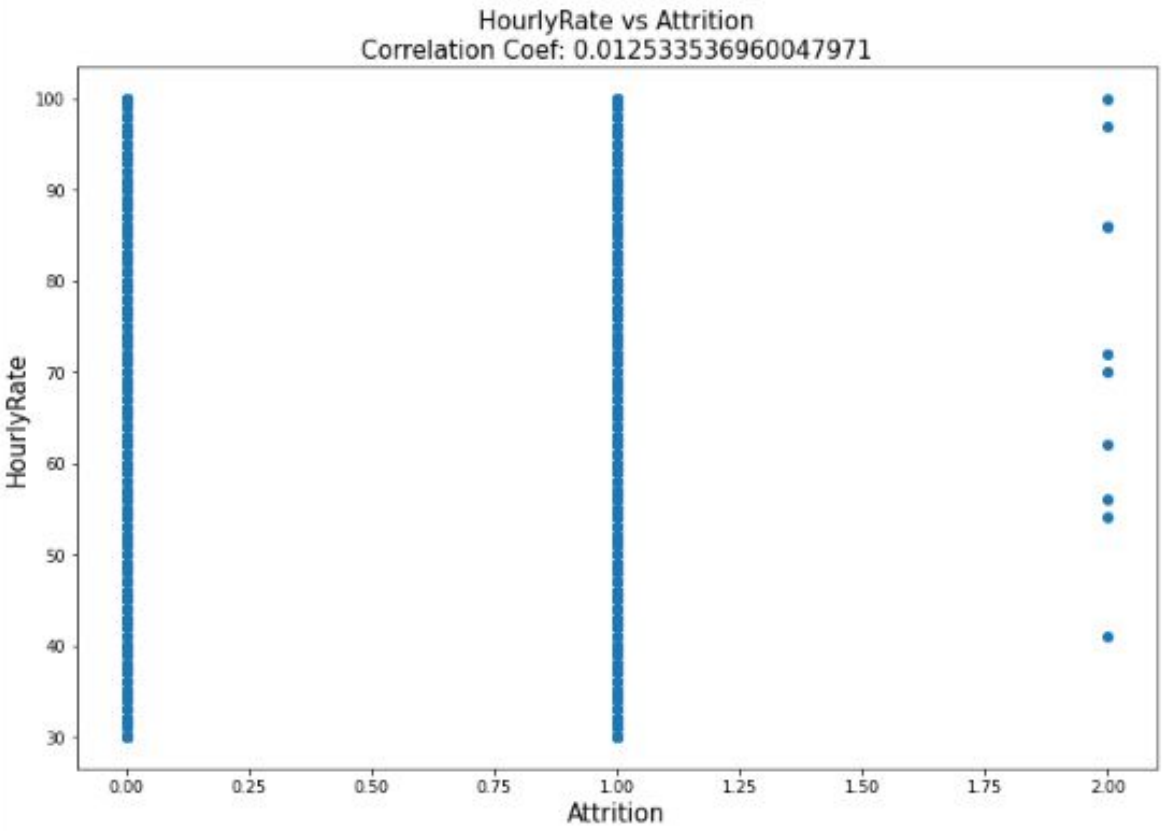
Linear Regression Model

- Fitting Hourly Rate and using an alpha of 0.375 shows a slight upwards trend but not highly noticeable.



Hourly Rate Vs Attrition

- LR model for **attrition** relies heavily on **Monthly and daily rate**



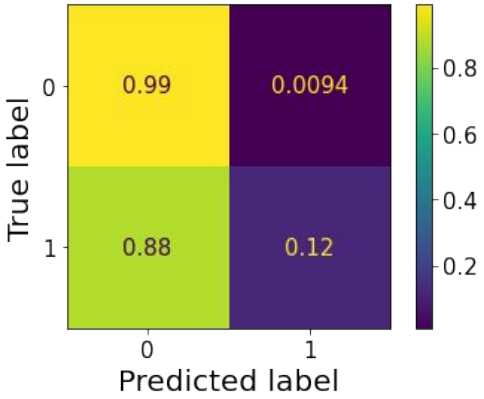
Predicting Attrition

What makes an employee likely to quit their job that isn’t pay rate?

What makes an employee likely to quit their job?

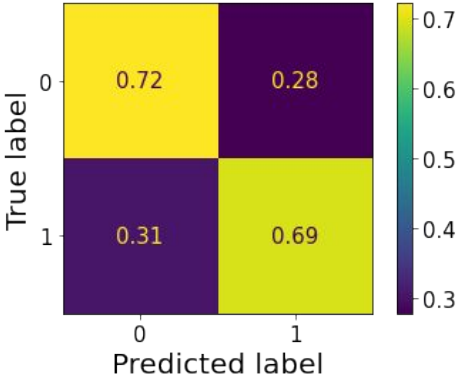
Initial findings, LR baseline

- Really good at predicting staying even if they left



LR Confusion Matrix after Balancing Data

- Worse at predicting people who will stay
- 57% increase at predicting people who will leave

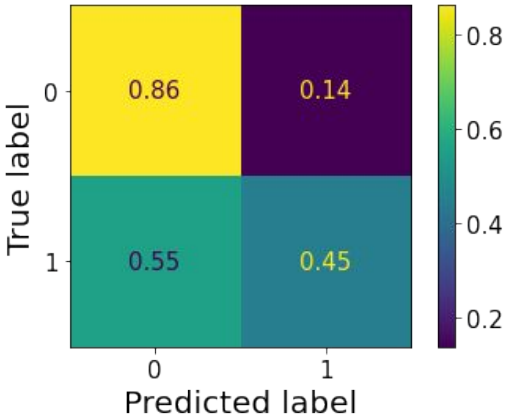


What makes an employee likely to quit their job?

Initial Decision tree and its Confusion Matrix

- It is decent at predicting if someone will stay
- Terrible at predicting those who left

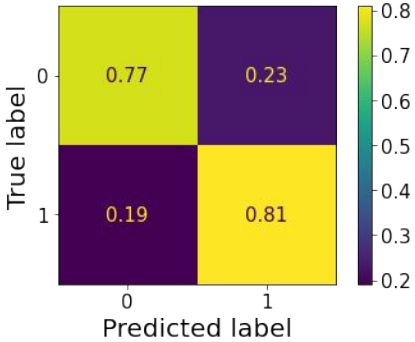
Depth = 5 Split = 10 Leaf = 5



Decision tree and its Confusion Matrix after Tuning

- Decrease of 9% for predicting employees that stayed
- Increase of 36% for predicting employees that left

Depth = 30 Split = 5 Leaf = 2

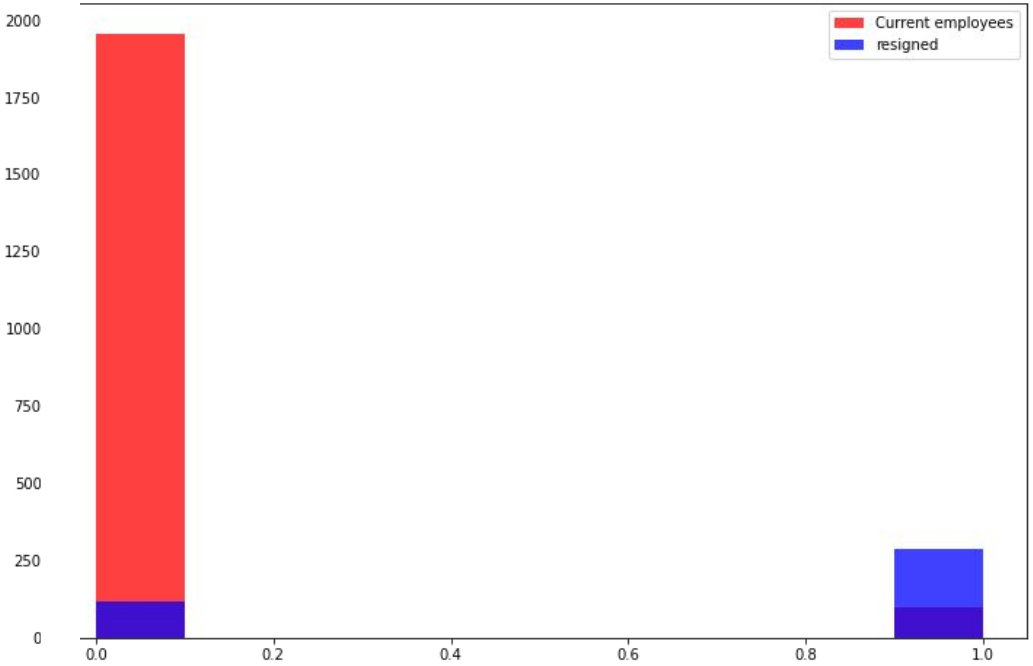


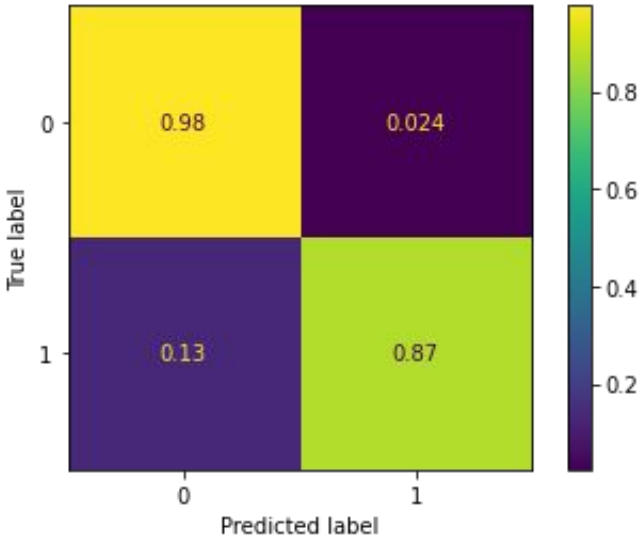
Nearest Neighbor

- Used gridSearchCV to find the best parameters
- Combined **high attrition demographics**
- Combined **high attrition roles & levels**
- KNN with winning metric: Hamming
- Ran with stratified data and balanced data
- **Nearest neighbor: 1**

Strong results

Hamming, 1 neighbour, stratified data: 89% overall accuracy





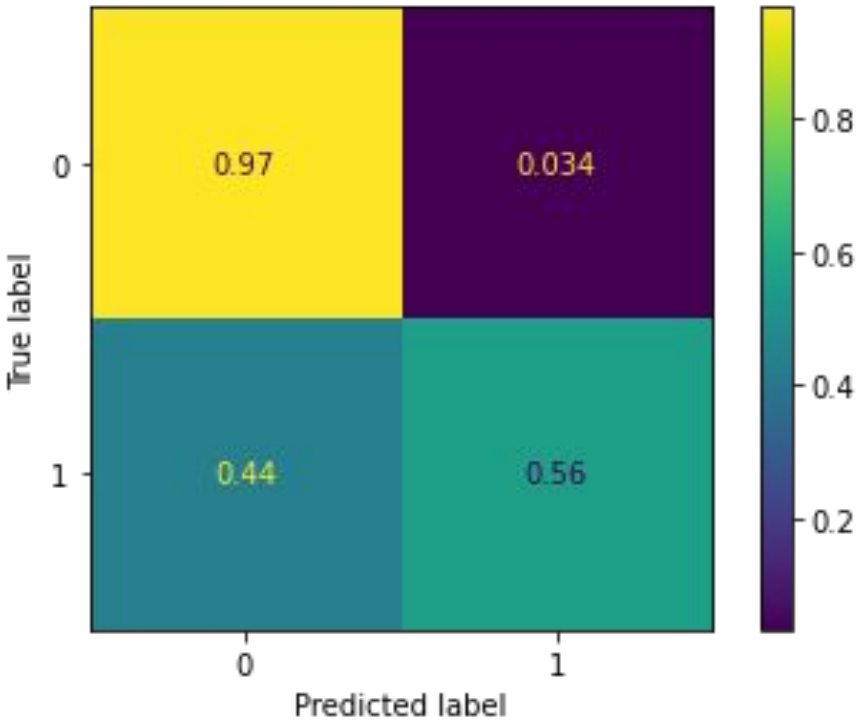
Strong results

- Looking at ‘false stayed’:
  - Very confused about R&D (61% of wrong predictions): diverse job levels, diverse pay. Somewhat higher attrition across job levels
  - Sales: wrongly predicted that executives would quit
  - Strong confidence
  - Even with new columns (broke down R&D by job levels + pay): no improvements

XGBoost

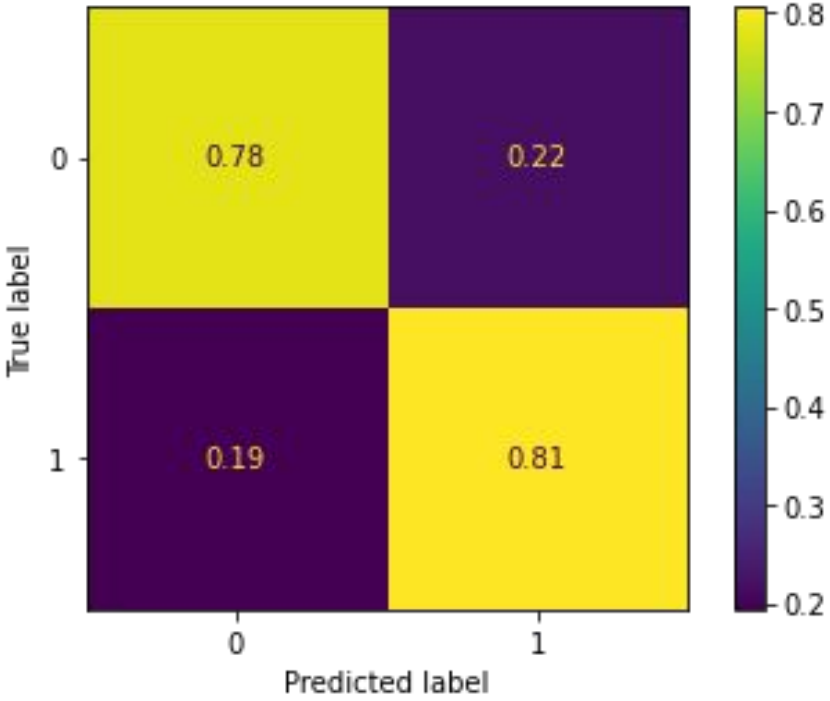
- Ran two DTs:
  - One with stratified data
  - One with balanced data (number of employees who left = number of employees who stayed)

XGBoost: Stratified data



- Great at predicting if an employee will stay
- Only 56% for predicting if they will quit
- Still confused about R&D (56% of the wrong predictions)
- Not predicting that everyone will stay: 20% of people predicted to quit

XGBoost: Balanced data



- Better at predicting if people will quit
- Very small dataset: only training on 1,875 rows
- Still confused about R&D: job that were wrongly predicted were mostly
  - in life science (R&D)
  - high job levels, except for lab techs.

Main Takeaways

What makes an employee likely to quit their job?

Our Findings:

- **Entry level** employees: Of all job roles the majority of employees leaving are level 1 or 2
- Best models: **decision trees** compared to any other model
- People **younger** in age
- Some **high attrition roles & levels**
  For instance:
  - Sales level 1 = 96% quit
  - Sales Representatives and Laboratory Technicians (46.33% of people who quit)

Thank you!

Any questions?

Extra Slides

Main takeaways

All models look at the same factors:

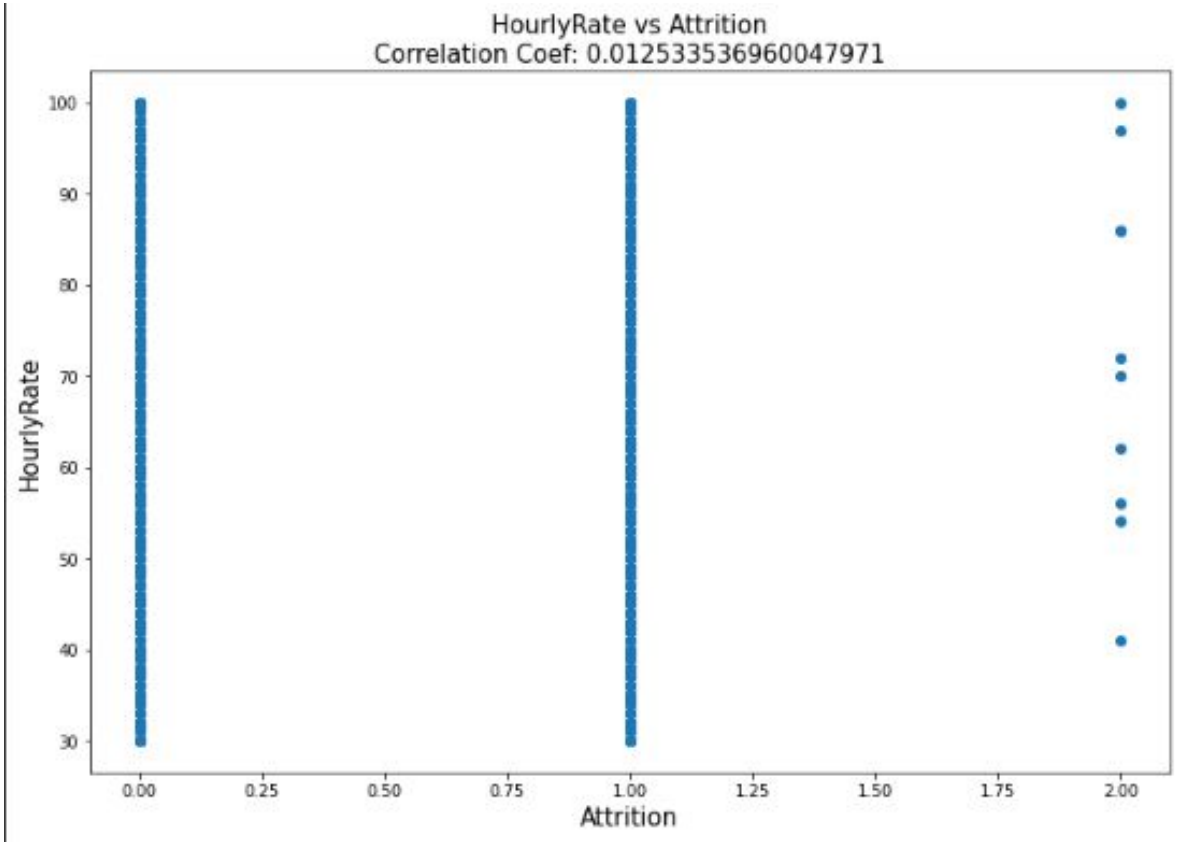
- Pay rate and attrition are related!
- But job role, job levels: huge influences

Are pay rate and an employee's attrition related?

Tying it all together

Hourly Rate Vs Attrition

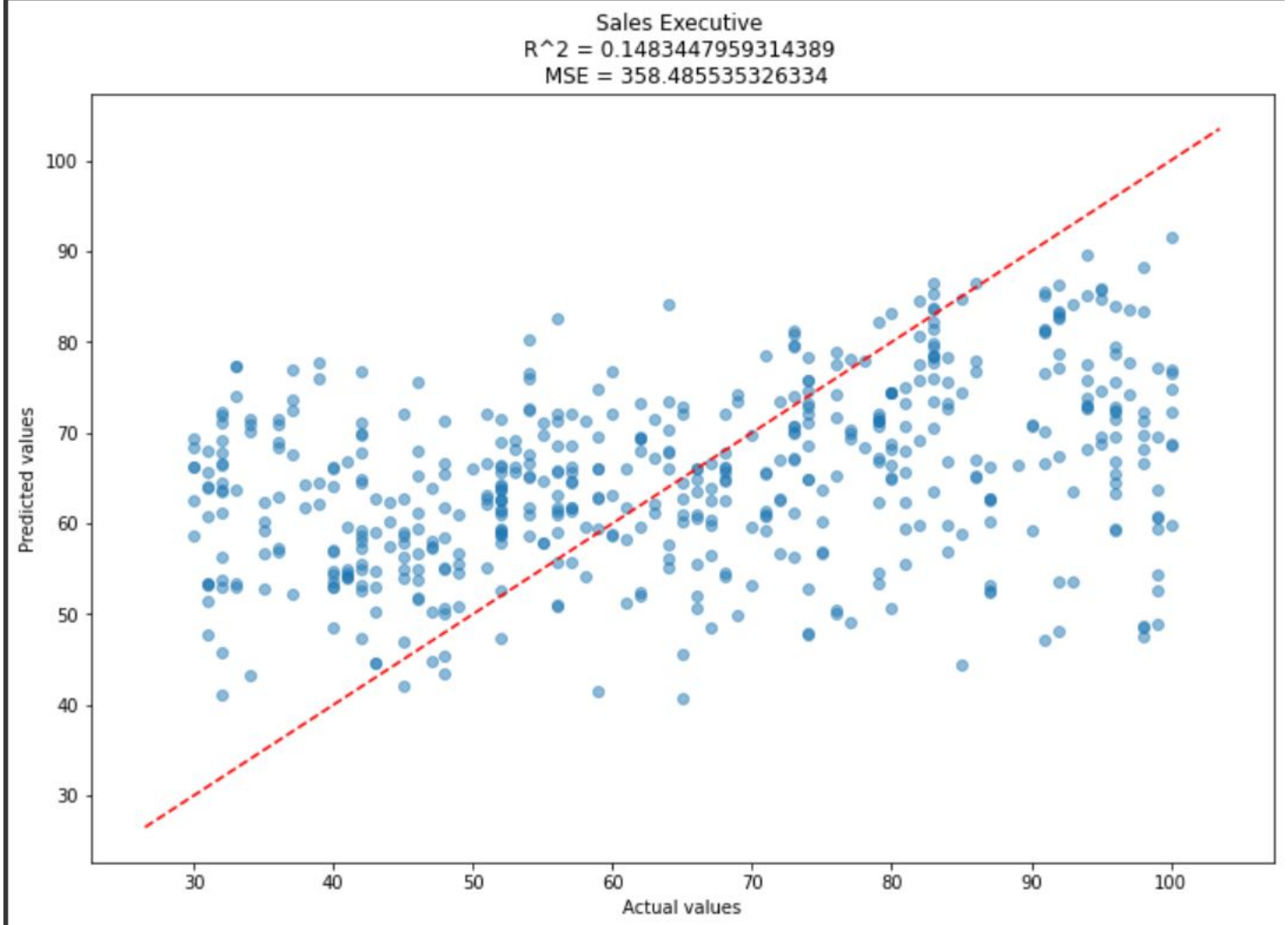
- LR model for **attrition** relies heavily on **Monthly and daily rate**

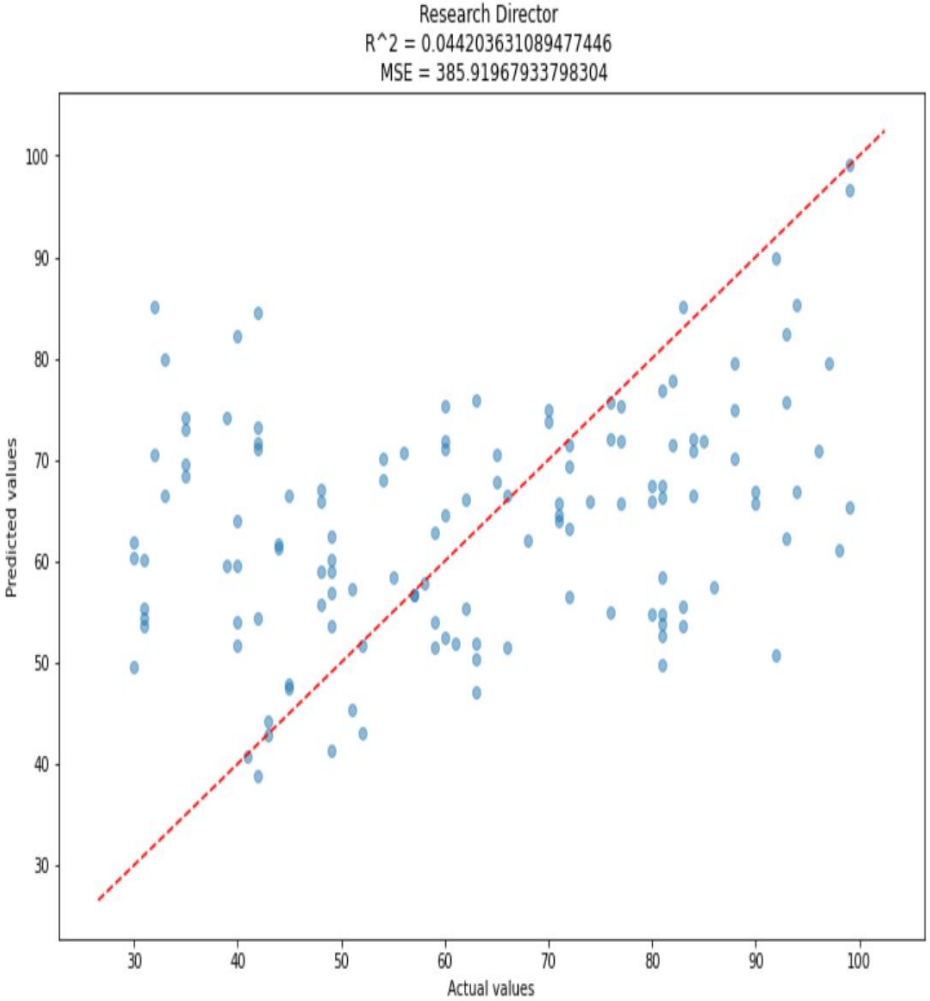


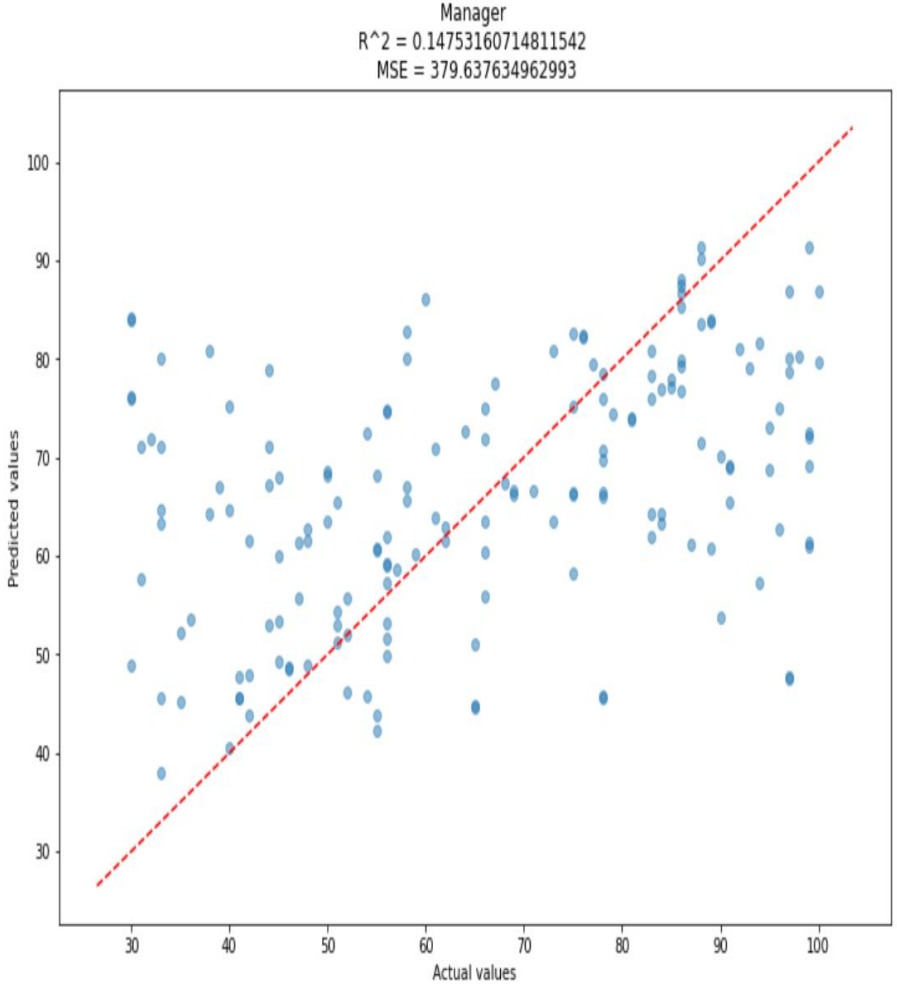
Are pay rate and an employee's attrition related?

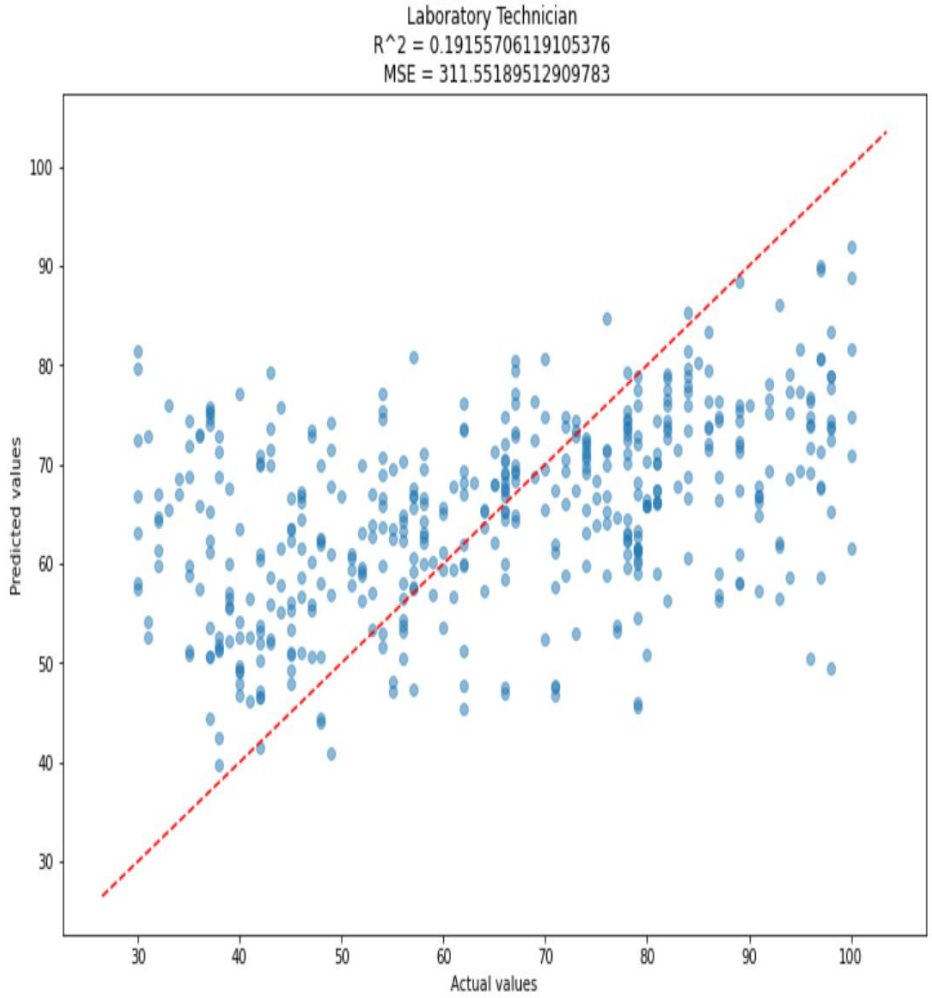
- So far: no clear linear relationship
- However: our linear regression, Knn and decision tree for attrition rely heavily on:
  - Pay rate
  - Job level (related to pay rate)

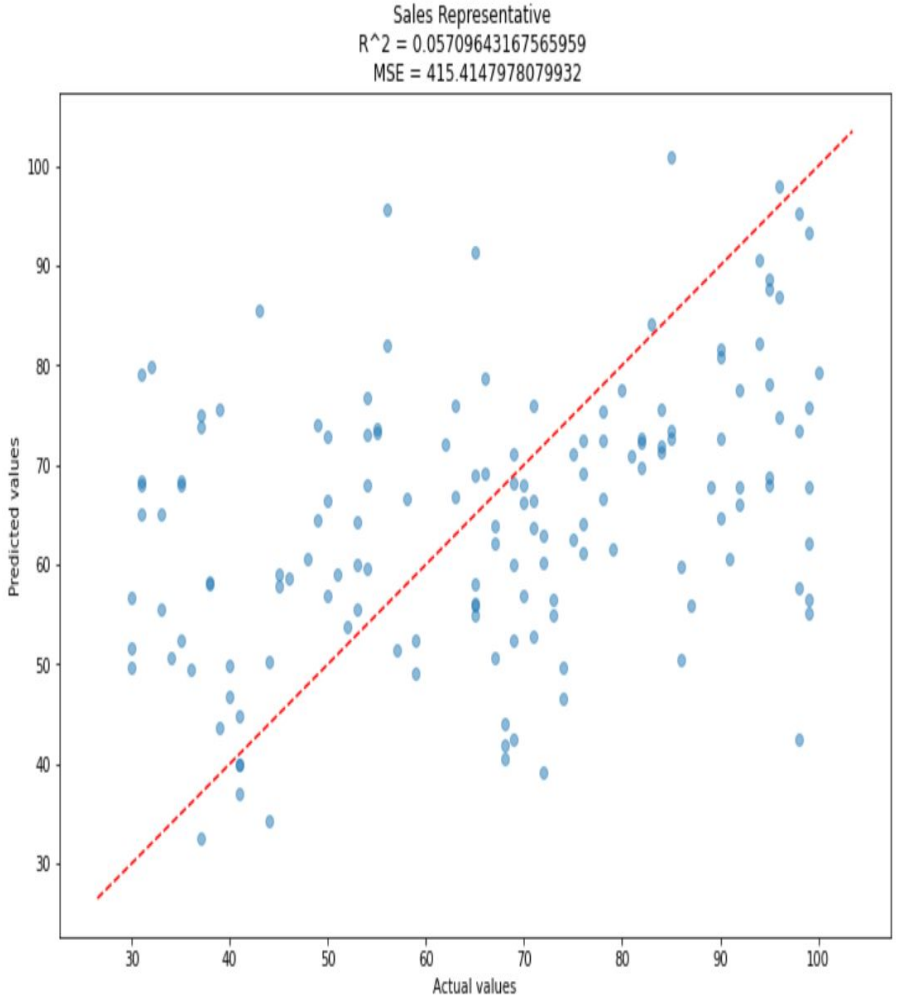

Job Role Preds Scatter Plot
Predicted values
Actual values
100
80
60
40
20
30
40
50
60
70
80
90
100
Sales Executive
Manager
Research Director
Sales Representative
Laboratory Technician
Research Scientist
Manufacturing Director
Healthcare Representative
Human Resources

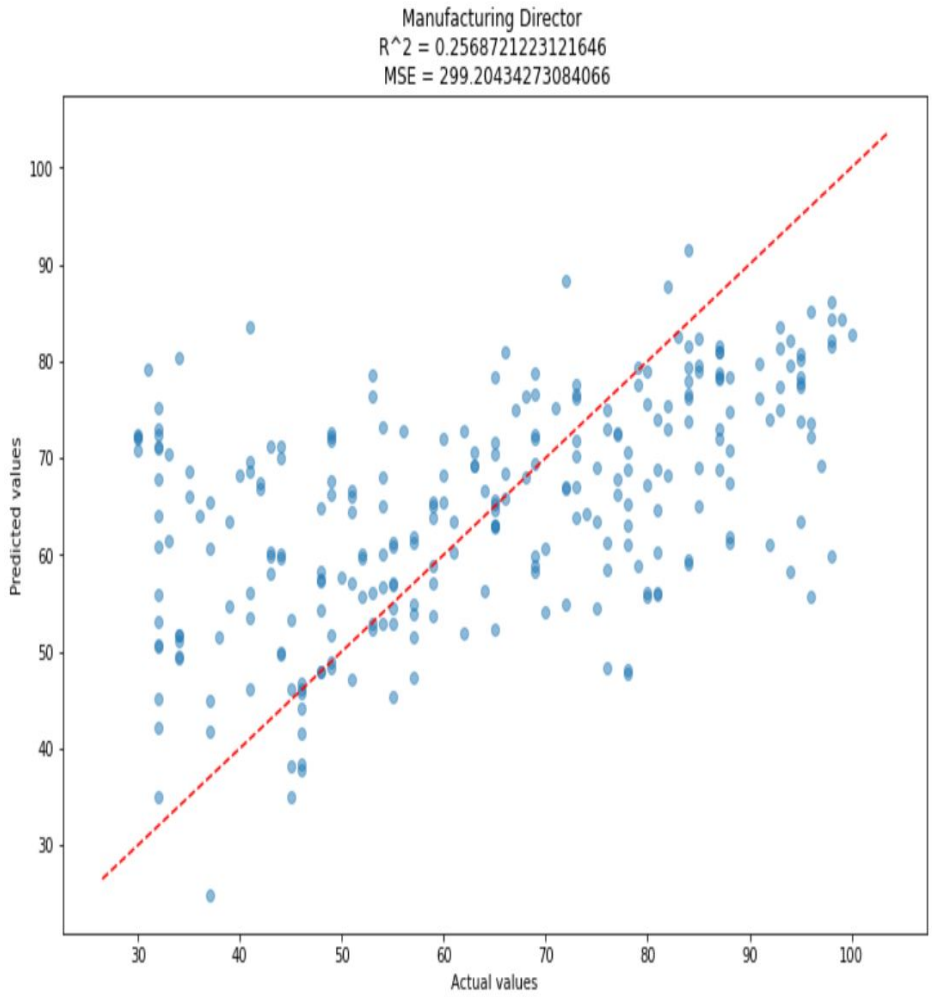

Sales Executive
R^2 = 0.1483447959314389
MSE = 358.485535326334
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100

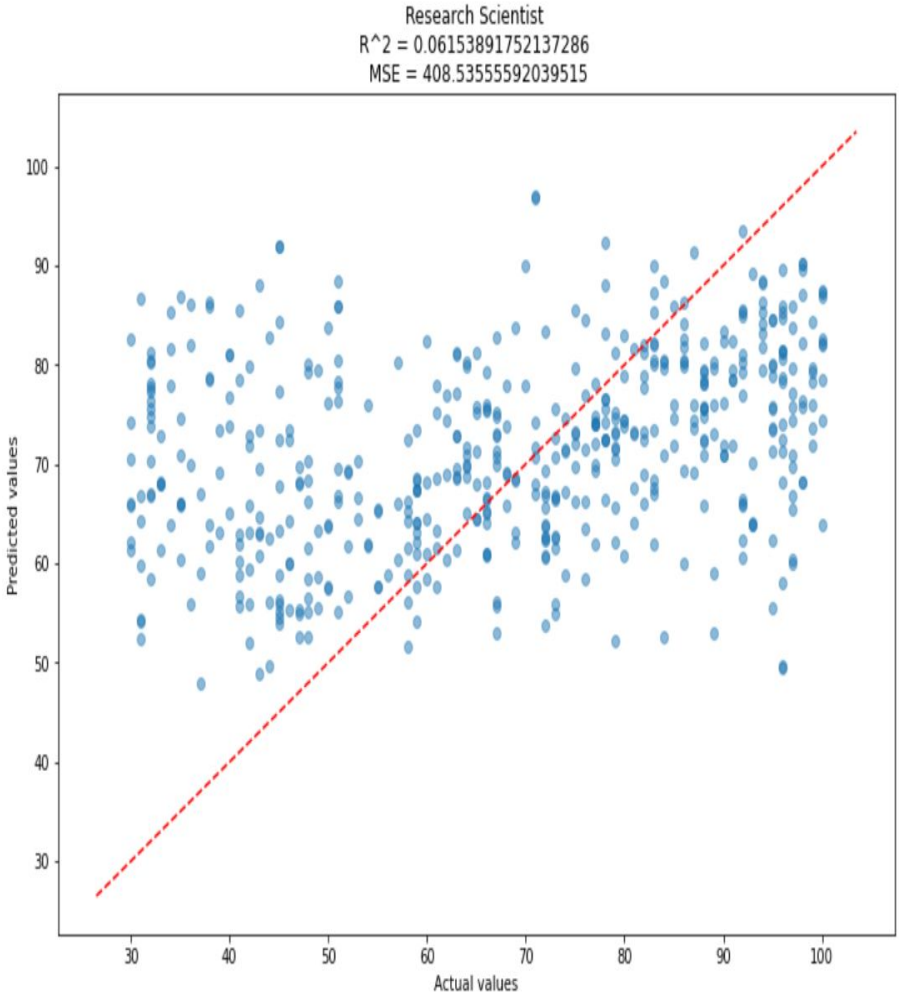

Research Director
R^2 = 0.044203631089477446
MSE = 385.91967933798304
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100

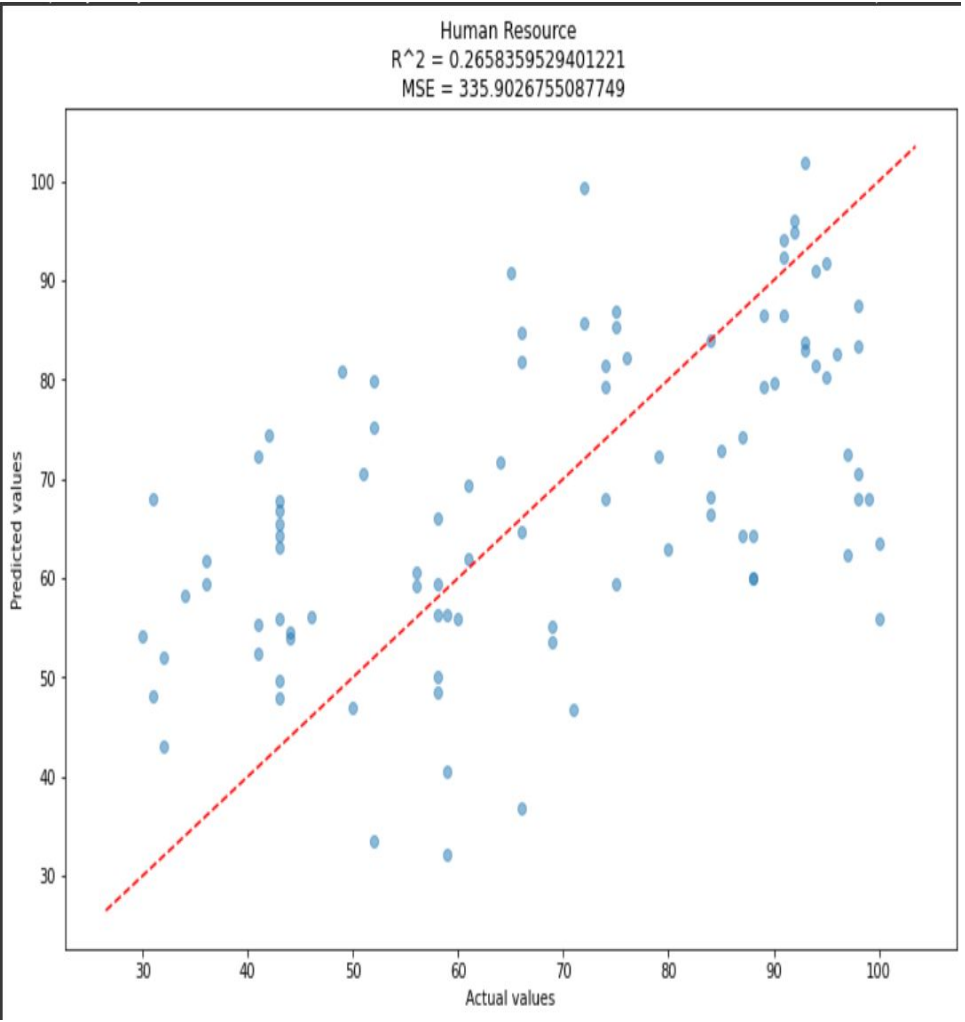

Manager
R^2 = 0.14753160714811542
MSE = 379.637634962993
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100

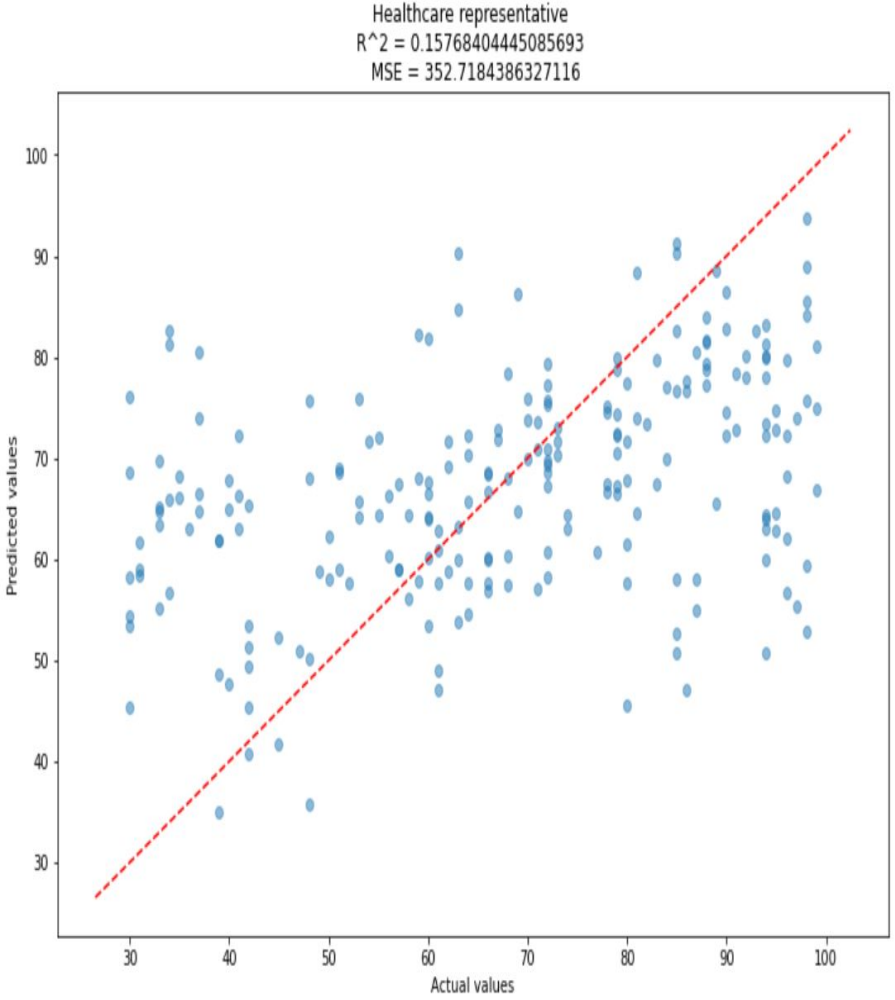

Laboratory Technician
R^2 = 0.19155706119105376
MSE = 311.55189512909783
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100


Sales Representative
R^2 = 0.05709643167565959
MSE = 415.4147978079932
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100


Manufacturing Director
R^2 = 0.2568721223121646
MSE = 299.20434273084066
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100


Research Scientist
R^2 = 0.06153891752137286
MSE = 408.53555592039515
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100


Human Resource
R^2 = 0.2658359529401221
MSE = 335.9026755087749
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100


Healthcare representative
R^2 = 0.15768404445085693
MSE = 352.7184386327116
Predicted values
Actual values
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100