Predicting Turnover

A Data-Driven Approach to Employee Retention

Marah, Rayshawn, Cyril, Domenic

OBJECTIVES OF THE PROJECT

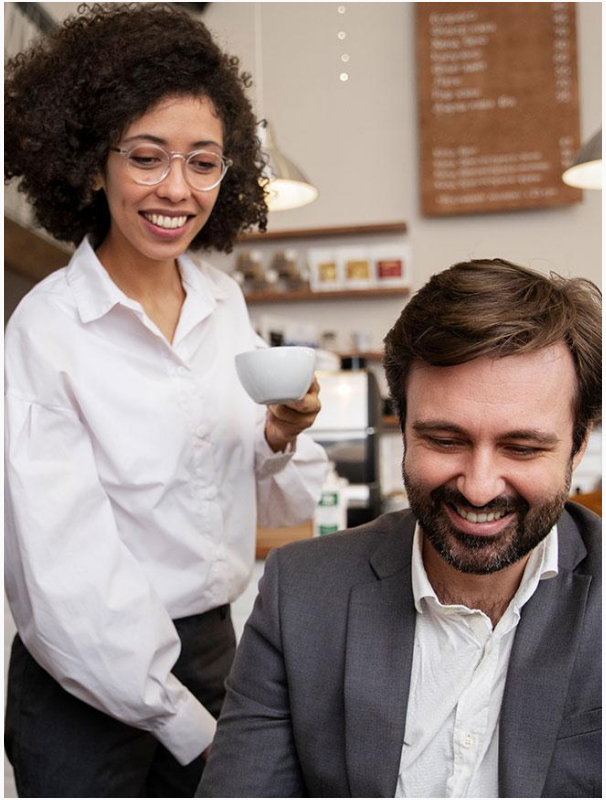
OBJECTIVES

OUR AIM

Develop a classification model to predict whether an employee is at risk of leaving a company.

THE GOAL

Identify individuals who may be at risk of leaving the company, so that the company may enact appropriate measures of outreach and support.

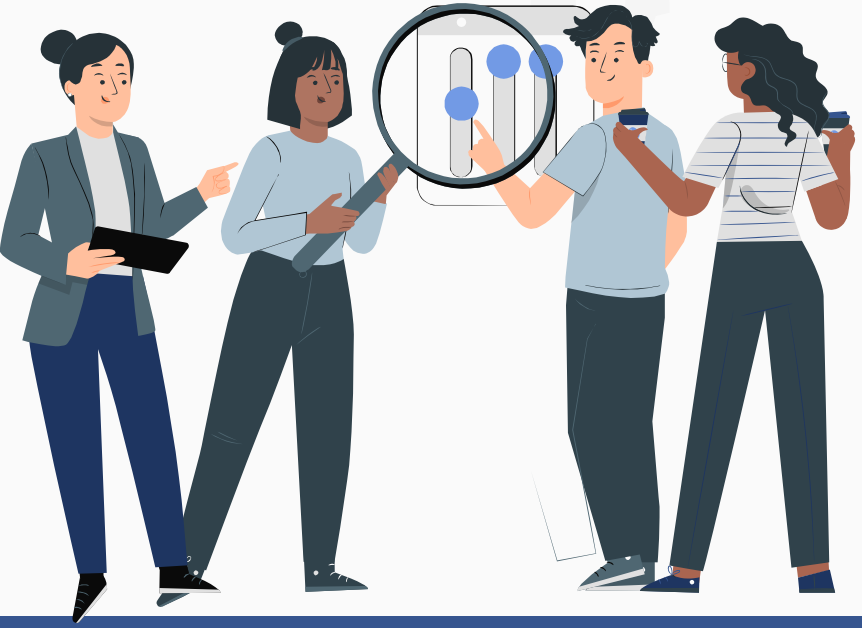


Project Overview

1. Exploring which components of employee data matter most towards determining whether or not an employee will leave their company

2. Visualizing how each component relates to the number of employees leaving each year

3. Determining a classification model that best predicts employees who are at risk of leaving



1

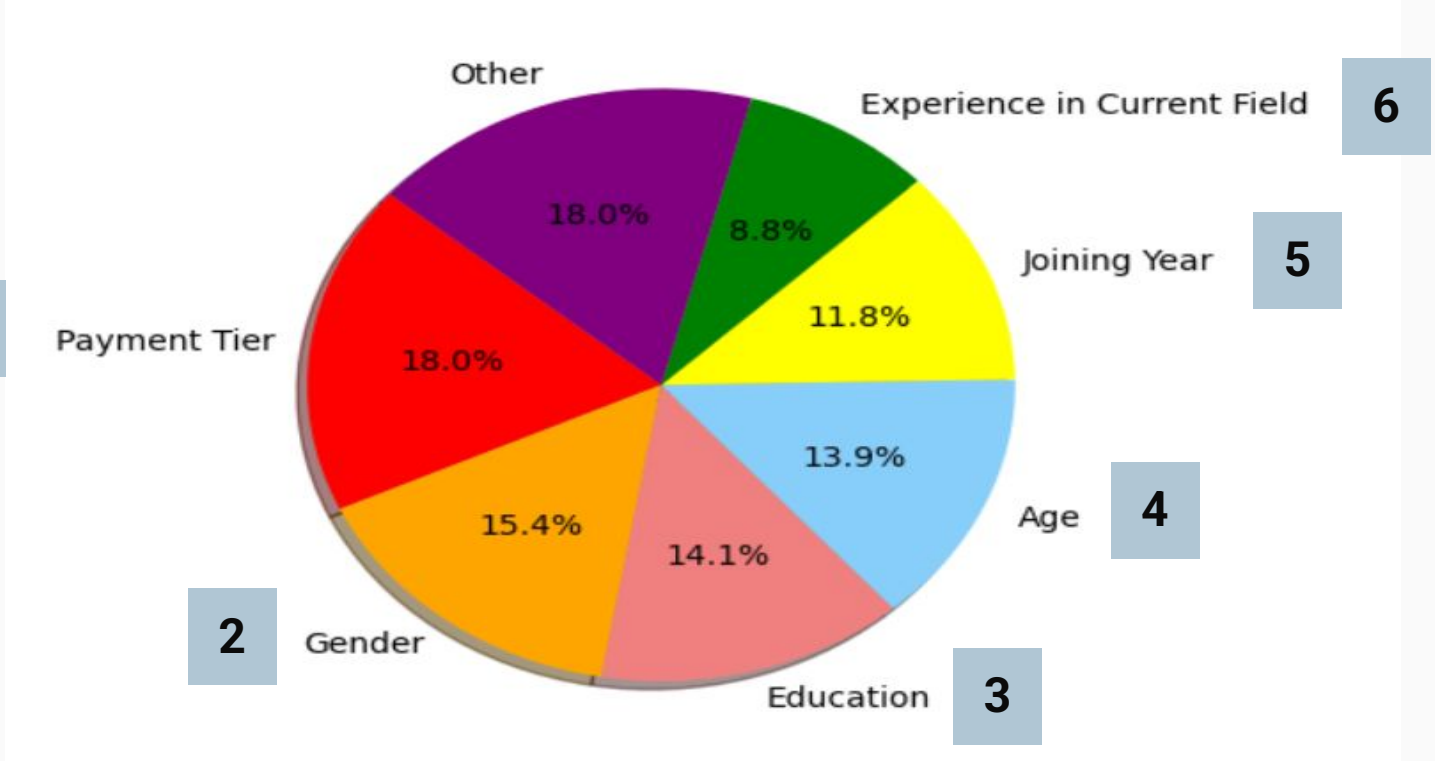
EMPLOYEE DATA

Factors Influencing Employee's Decision: Key Features for Classification

1. Payment Tier (1, 2, or 3)
2. Gender
3. Education (Bach., Mast., PhD.)
4. Age
5. Joining Year
6. Experience in Current Domain

- Did the employee leave

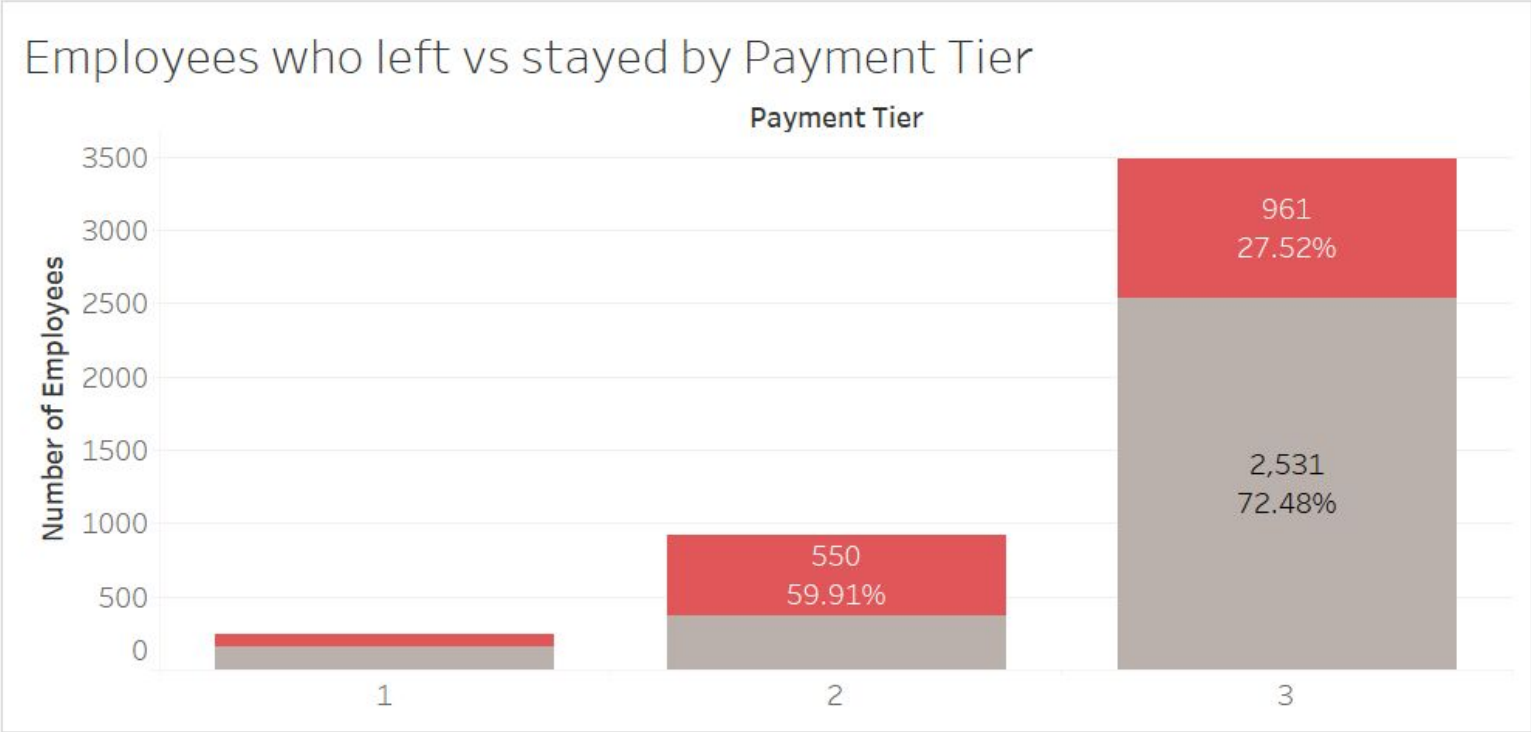
Breakdown of Feature Importance



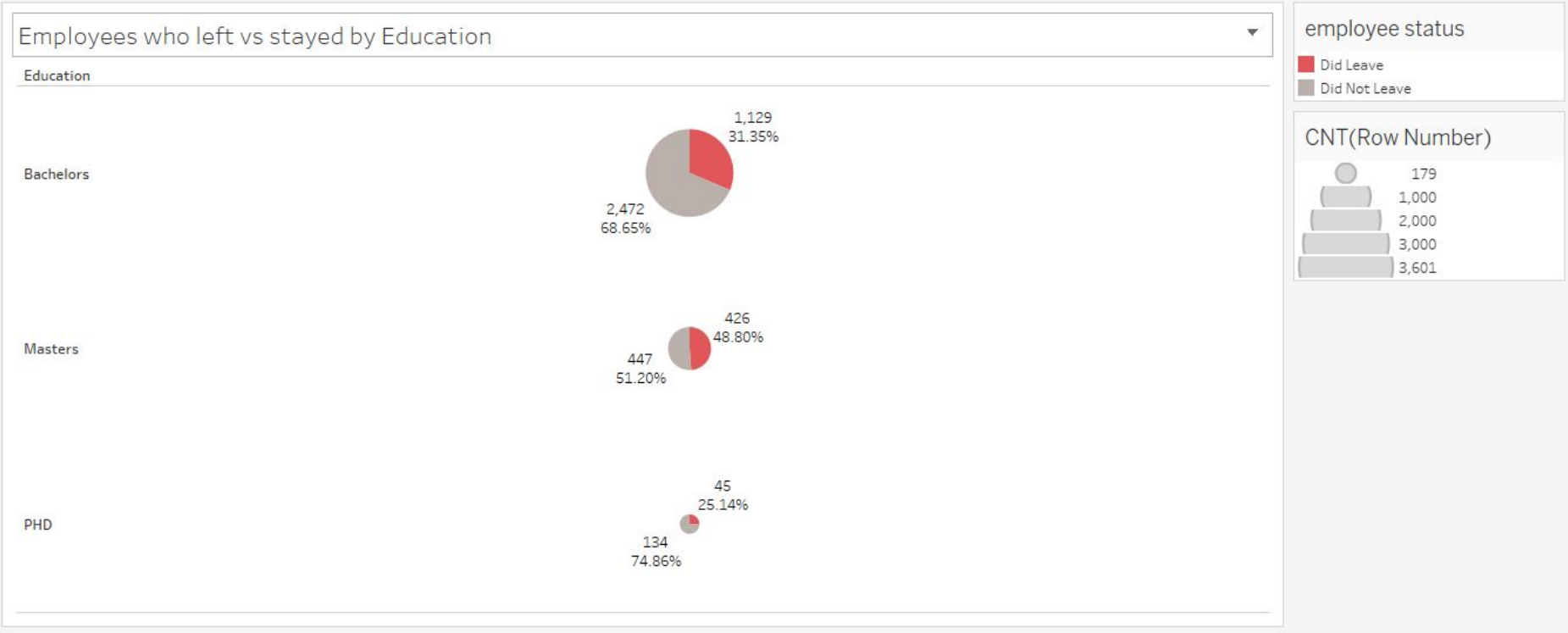
2

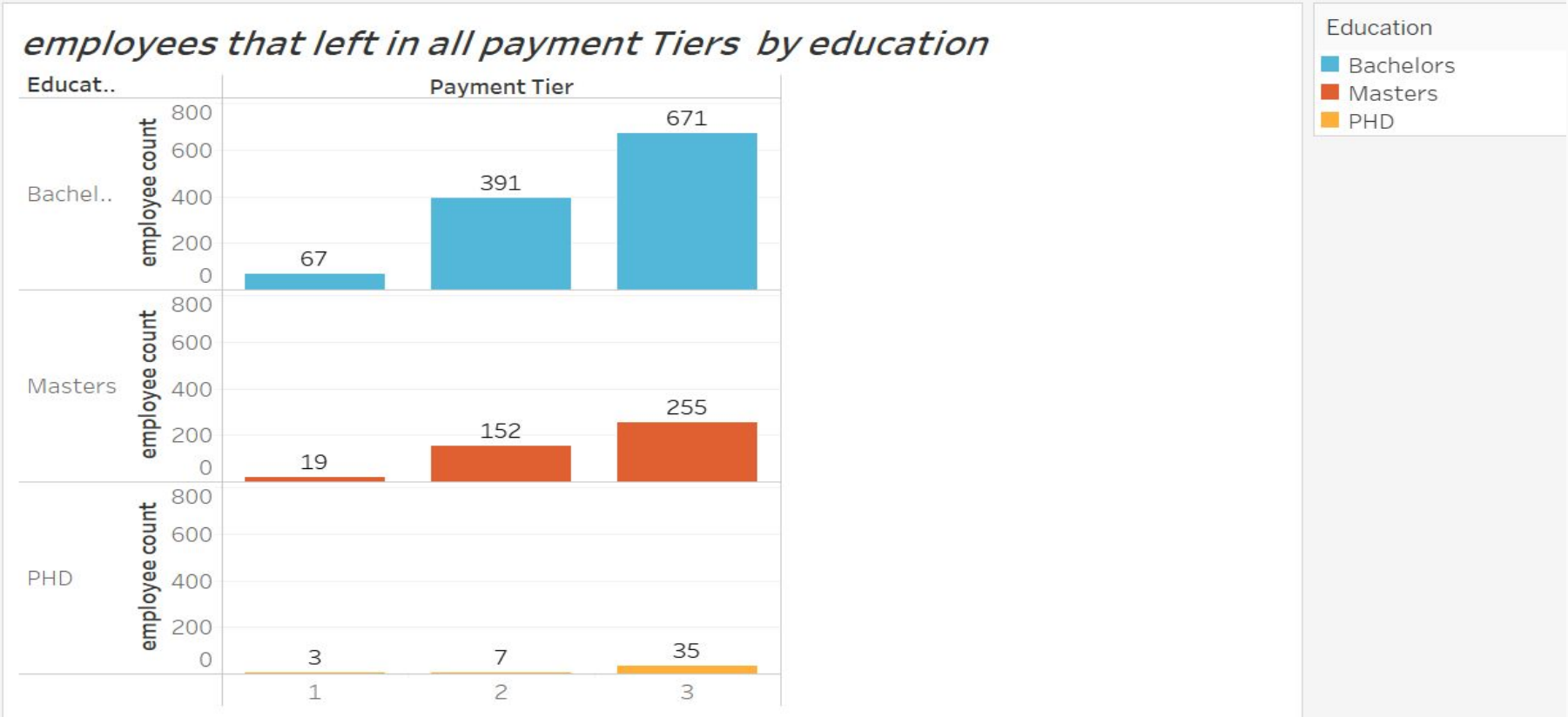
VISUALIZING KEY FEATURES

#1 – Payment Tier

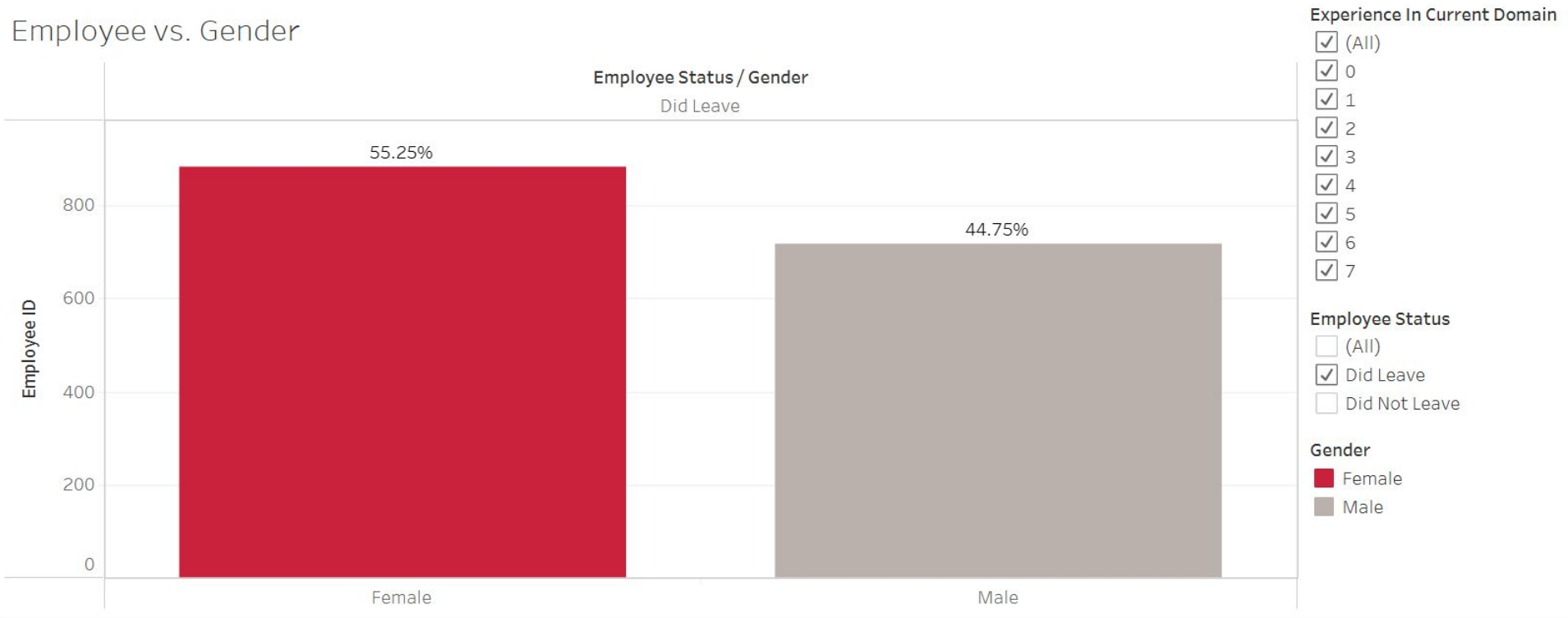


#2 – Education

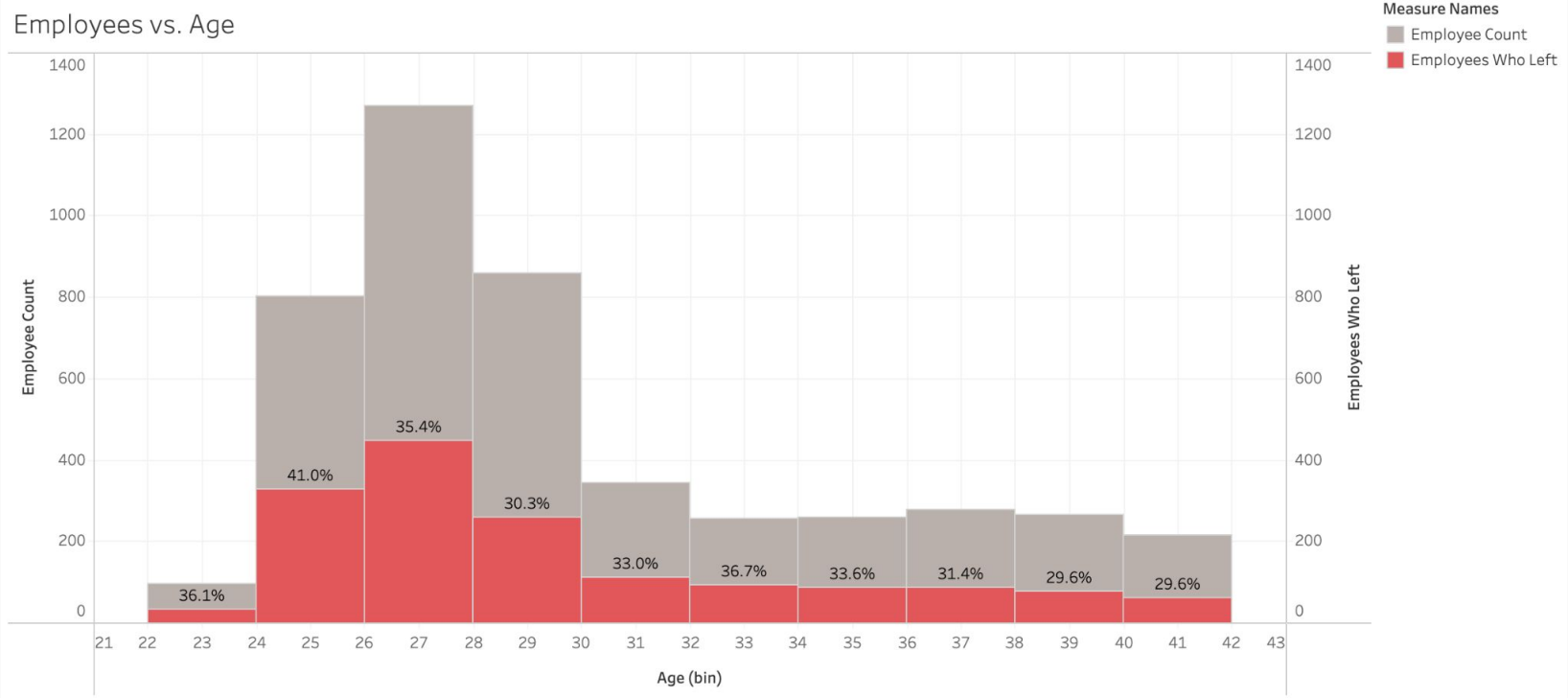

Employees who left vs stayed by Education
Education
Bachelors
1,129
31.35%
2,472
68.65%
Masters
426
48.80%
447
51.20%
PHD
45
25.14%
134
74.86%
employee status
Did Leave
Did Not Leave
CNT(Row Number)
179
1,000
2,000
3,000
3,601


employees that left in all payment Tiers by education
Educat..
Payment Tier
Bachel..
Masters
PHD
employee count
800
600
400
200
0
67
391
671
19
152
255
3
7
35
1
2
3
Education
Bachelors
Masters
PHD

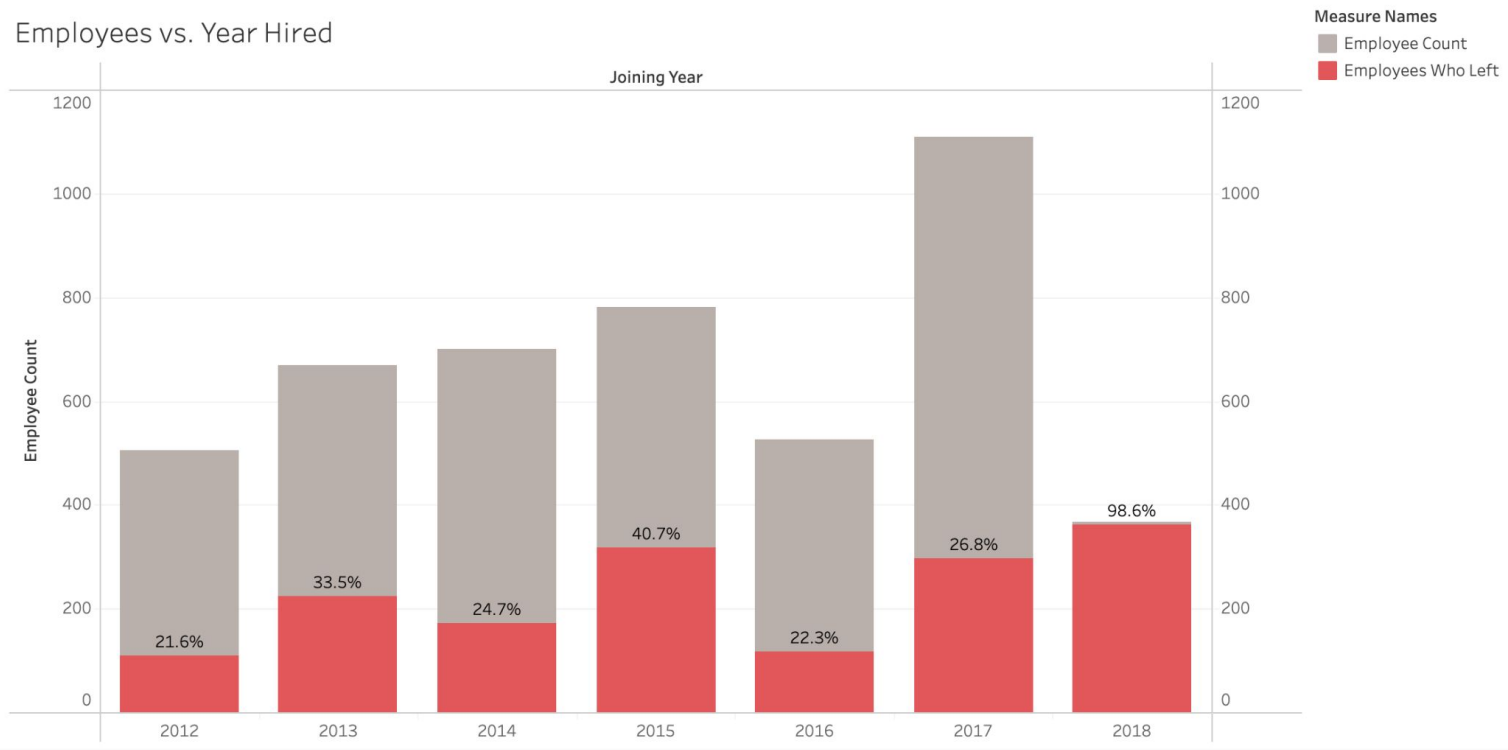
#3 – Gender


Employee vs. Gender
Employee Status / Gender
Did Leave
55.25%
44.75%
Employee ID
800
600
400
200
0
Female
Male
Experience In Current Domain
(All)
0
1
2
3
4
5
6
7
Employee Status
(All)
Did Leave
Did Not Leave
Gender
Female
Male

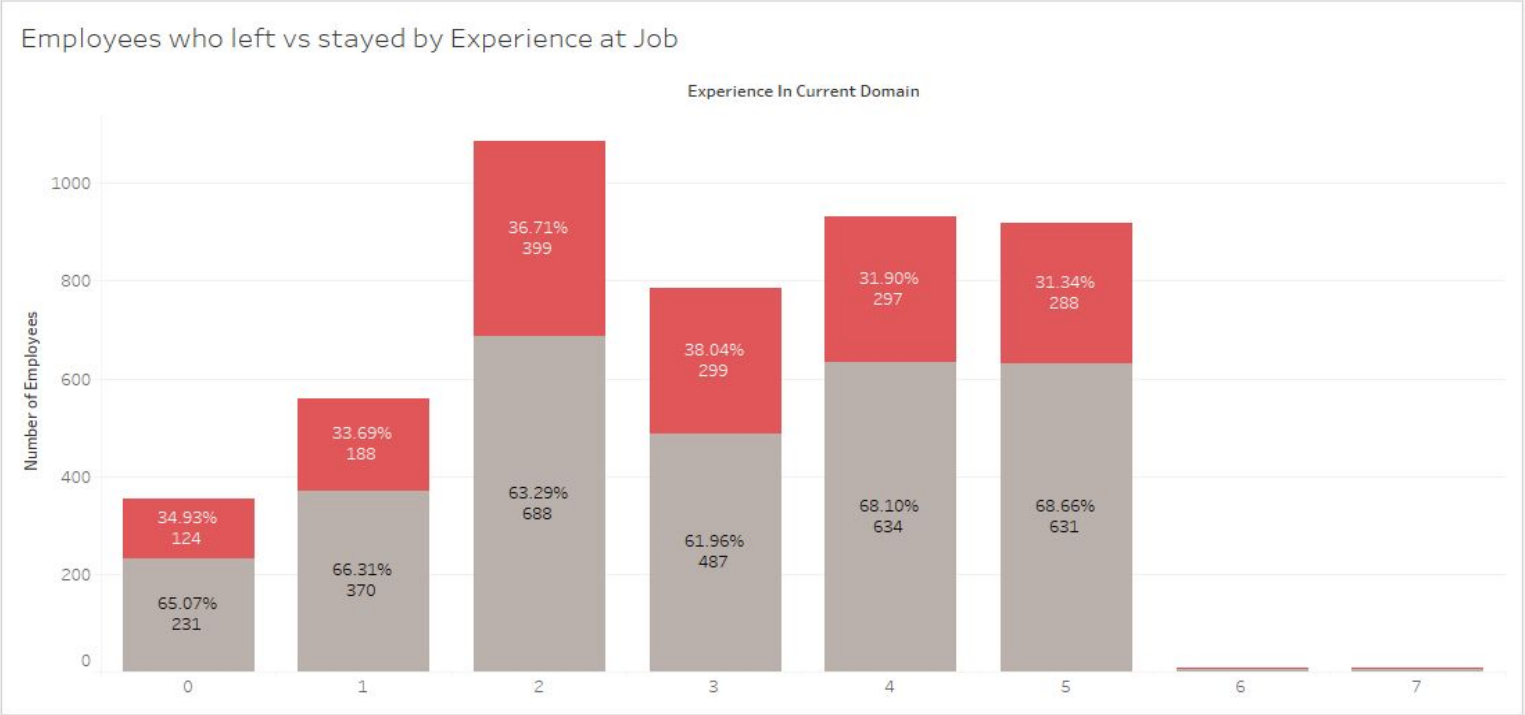
#4 – Age

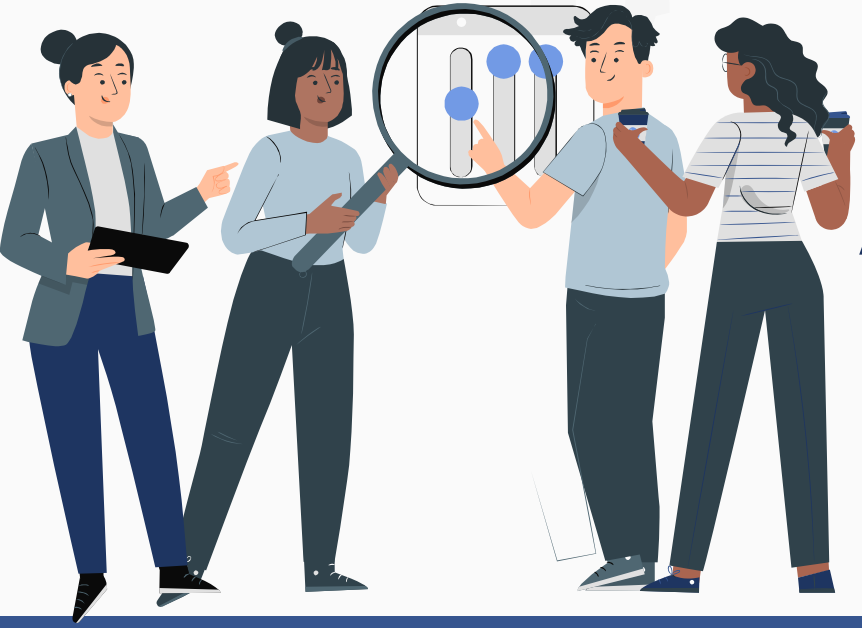

Employees vs. Age
Measure Names
Employee Count
Employees Who Left
Employee Count
Employees Who Left
Age (bin)
0
200
400
600
800
1000
1200
1400
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
36.1%
41.0%
35.4%
30.3%
33.0%
36.7%
33.6%
31.4%
29.6%
29.6%

#5 – Joining Year



#6 – Experience in Current Field





3

A PREDICTIVE MODEL

Target Variable

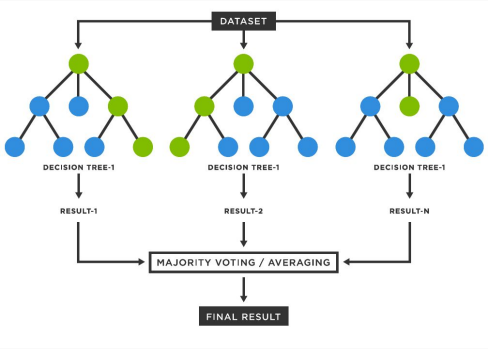
Employee Retention Status

- Employee stays with the company
- Employee leaves the company

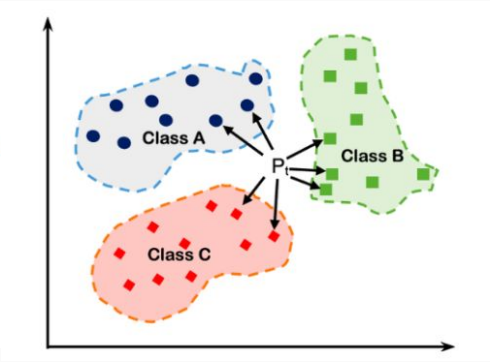
A classification model will be used to predict if an employee might leave or stay

Classification Models Tested

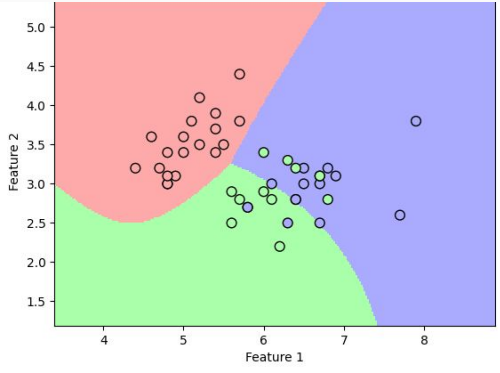
Random Forest



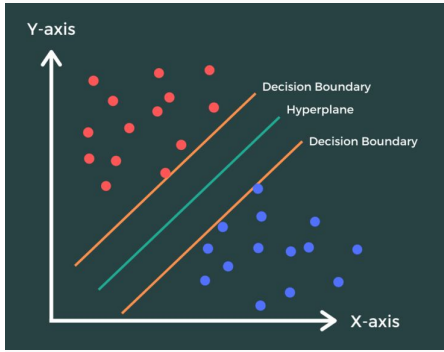
K-Nearest Neighbors



Gaussian NB



Support Vector Machines



Optimization Process

#	Model Type	Accuracy	Precision	Recall	Notes
1	Random Forest	85%	83%	71%	skewed by outlier 2018 data
2	Random Forest	81%	75%	56%	2018 data removed
3	Random Forest	77%	69%	43%	2018 data, EverBenched data, and City data removed
4	Random Forest	84%	87%	54%	2018 data removed + now using GridSearch / Cross Validation
5	Random Forest	75%	73%	56%	adjust GridSearch to focus on recall rather than accuracy
6	Random Forest	69%	50%	77%	reduced decision threshold from 0.5 to 0.35
7	Random Forest	77%	60%	71%	corrected code error, adjusted decision theshold to 0.415
8	Random Forest	75%	58%	65%	applied SMOTE
9	Random Forest	74%	56%	66%	applied SMOTE-EEN
10	Random Forest	74%	56%	65%	applied SMOTE-EEN, removed GridSearch
11	Support Vector Machines	75%	72%	47%	skewed by outlier 2018 data
12	Support Vector Machines	76%	71%	34%	2018 data removed
13	K-Nearest Neighbors (KNN)	83%	81%	66%	skewed by outlier 2018 data
14	K-Nearest Neighbors (KNN)	80%	73%	55%	2018 data removed
15	K-Nearest Neighbors (KNN)	81%	77%	53%	applied SMOTE
16	K-Nearest Neighbors (KNN)	77%	61%	64%	applied SMOTE-ENN
17	GaussianNB	100%	100%	100%	Target varable was not removed from the X factors
18	GaussianNB	62%	66%	85%	Removed target varable from X factors # note this is better at recalling 0 then 1

Optimized Classification Results

Accuracy	Precision	**Recall**
77%	**60%**	**71%**

Optimized Classification Results

What This Means

Higher Recall Rate = Less False Negatives

Recall

71%

Optimized Classification Results

What This Means

Higher Recall Rate = Less False Negatives

When the model thinks an employee is NOT at risk of leaving, when in fact, they ARE at risk.

Recall

71%

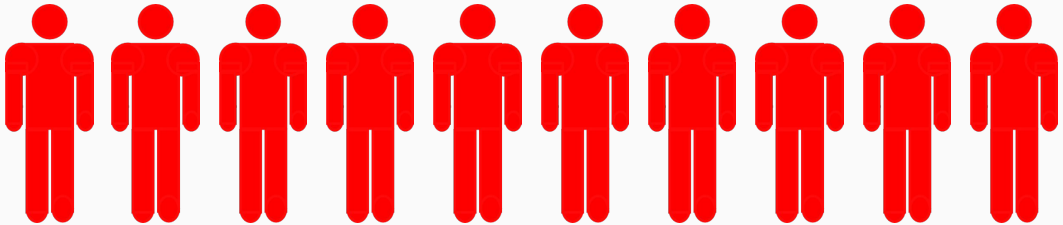
Optimized Classification Results

EXAMPLE

Let’s take 10 employees who are at risk of leaving.

Recall

71%



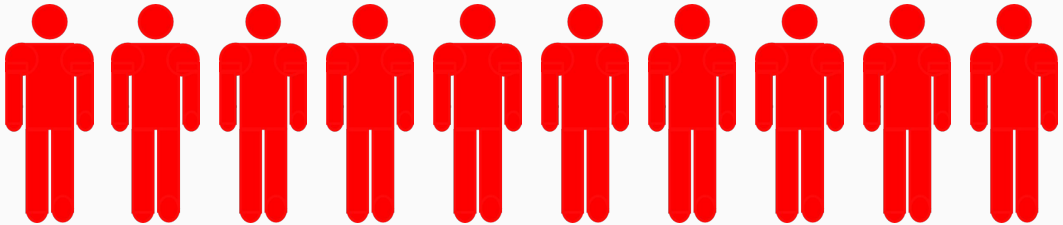
Optimized Classification Results

EXAMPLE

Employees at risk of leaving would be flagged by the model as "at risk" about 7 out of every 10 instances of employees considering leaving.

Recall

71%



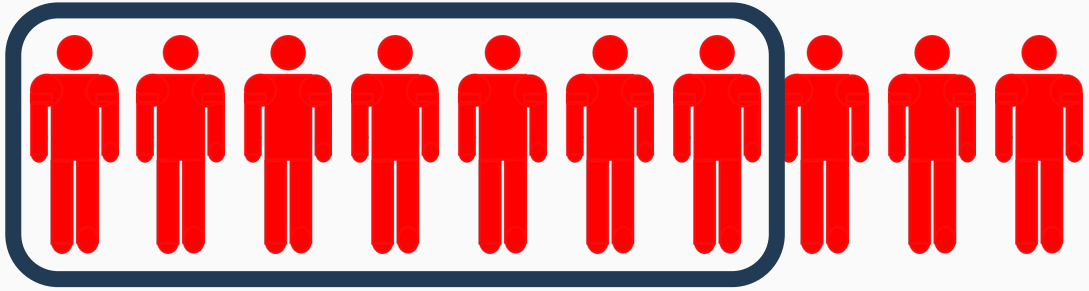
Optimized Classification Results

EXAMPLE

Employees at risk of leaving would be flagged by the model as "at risk" about 7 out of every 10 instances of employees considering leaving.

Recall

71%



Optimized Classification Results

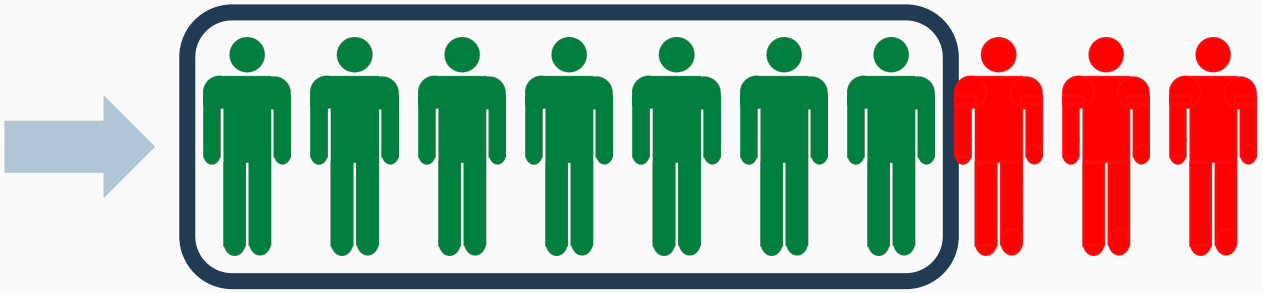
EXAMPLE

Employees at risk of leaving would be flagged by the model as "at risk" about 7 out of every 10 instances of employees considering leaving.

Recall

71%

Provide outreach and support;
Increase employee satisfaction;
Retain employees 😄



Optimized Classification Results

What This Means

Higher Precision Rate = Less False Positives

Precision

60%

Optimized Classification Results

What This Means

Higher Precision Rate = Less False Positives

When the model thinks an employee IS at risk of leaving, when in fact, they are NOT at risk.

Precision

60%

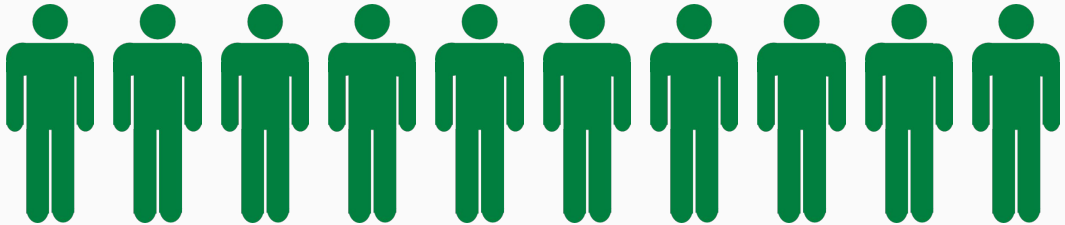
Optimized Classification Results

EXAMPLE

Let’s take 10 employees who are not at risk of leaving.

Precision

60%



Optimized Classification Results

EXAMPLE

Employees NOT at risk of leaving would be flagged as "at risk" by the model about 4 out of every 10 instances of an employee NOT considering leaving.

Precision

60%

Optimized Classification Results

What This Means

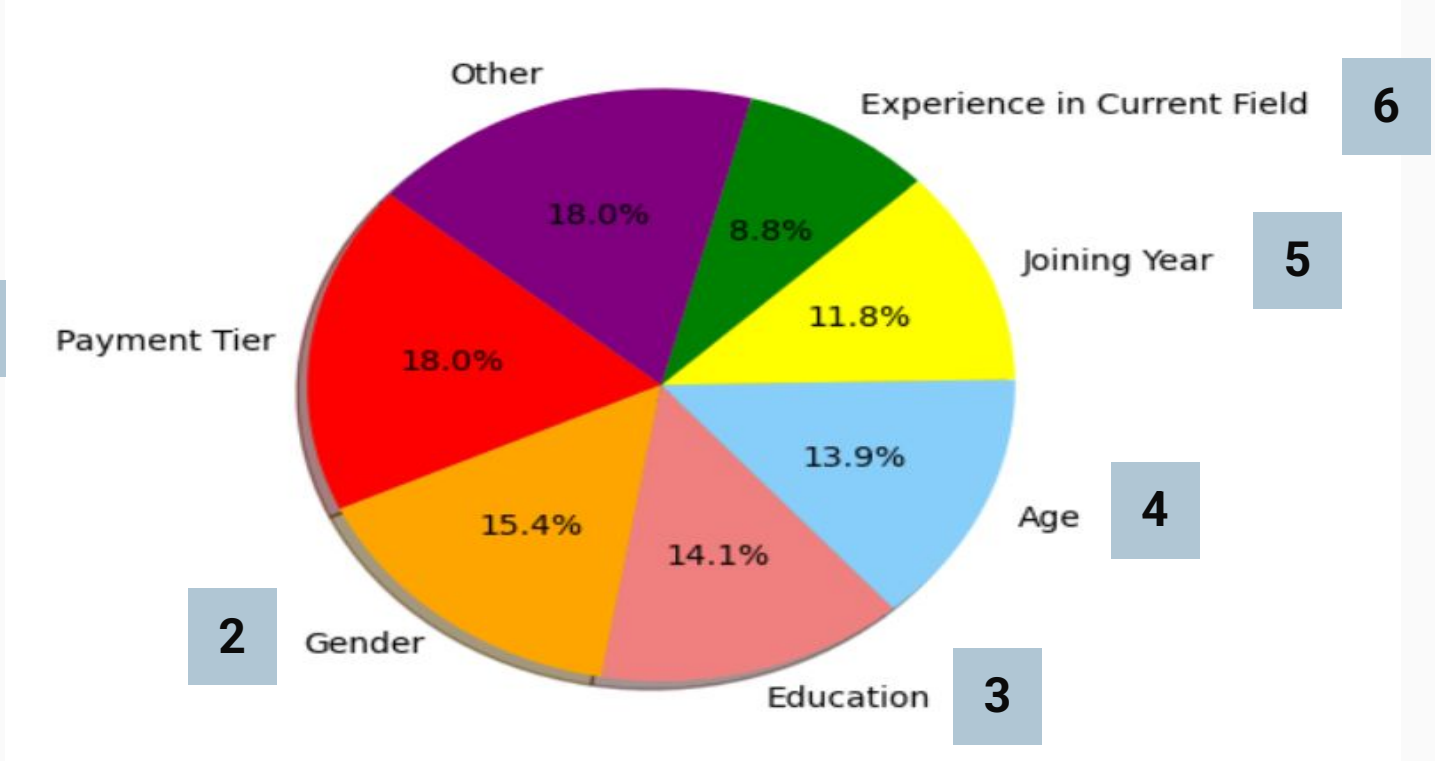
Employees NOT at risk of leaving would be flagged as "at risk" by the model about 4 out of every 10 instances of an employee NOT considering leaving.

Precision

60%

Provide outreach and support; Oops, you're already satisfied! Employees say "wow HR is really looking out for me" 😄

Breakdown of Feature Importance



Optimized Classification Results

Accuracy	Precision	**Recall**
77%	**60%**	**71%**


I love my job!

THANKS!