Minimal Mistakes Final Report

Employee Attrition

AMAT 465/565 - Applied Statistics - Fall 2019

Erin Gozalkowski

Enzo Rodriguez

Matthew Zimmer

Nicole Zvorsky

Table of Contents

Problem Statement

A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons:

1. The former employees' projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners

2. A sizeable department has to be maintained, for the purposes of recruiting new talent

3. More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

Hence, the management has contracted us to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

Goal of our Case Study

We aim to model the probability of attrition using logistic regression. The results thus obtained will be used by management within an organization to understand what changes they should make to their workplace, in order to get most of their employees to stay.

Uncover the factors that lead to employee attrition and explore important questions such as `show me a breakdown of distance from home by job role and attrition` or `compare average monthly income by education and attrition`. This is a fictional data set created by IBM data scientists.

Education 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'

EnvironmentSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

JobInvolvement 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

JobSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

PerformanceRating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'

RelationshipSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

WorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

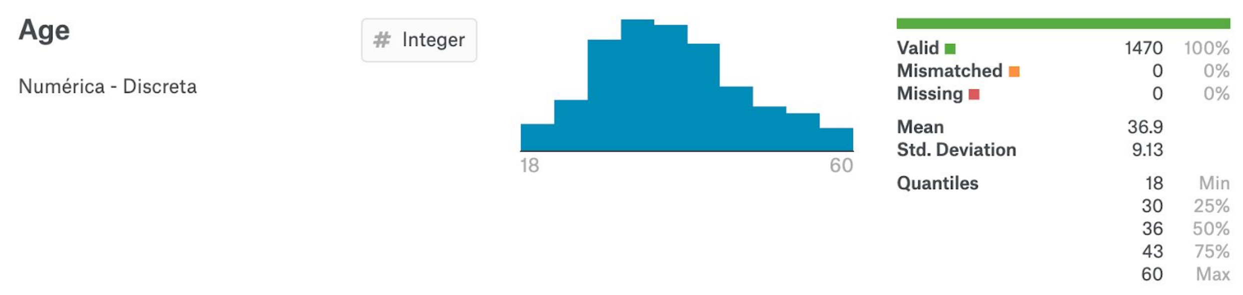
Data Sources

All data sources are available for download here.

WA_Fn-UseC_-HR-Employee-Attrition.csv (35 columns) Contains employee attrition data

Columns

Age



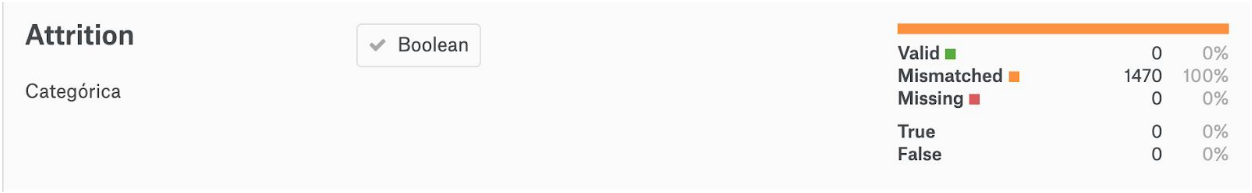
Age

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	36.9	
Std. Deviation	9.13	
Quantiles	18	Min
	30	25%
	36	50%
	43	75%
	60	Max

Attrition



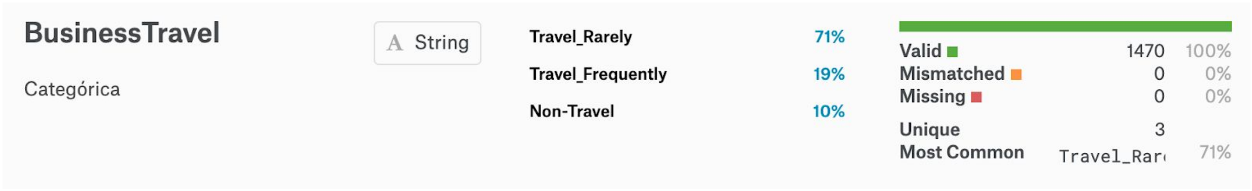
Attrition

✓ Boolean

Categórica

Valid 🟩	0	0%
Mismatched 🟧	1470	100%
Missing 🟥	0	0%
True	0	0%
False	0	0%

Business Travel Categorical



BusinessTravel

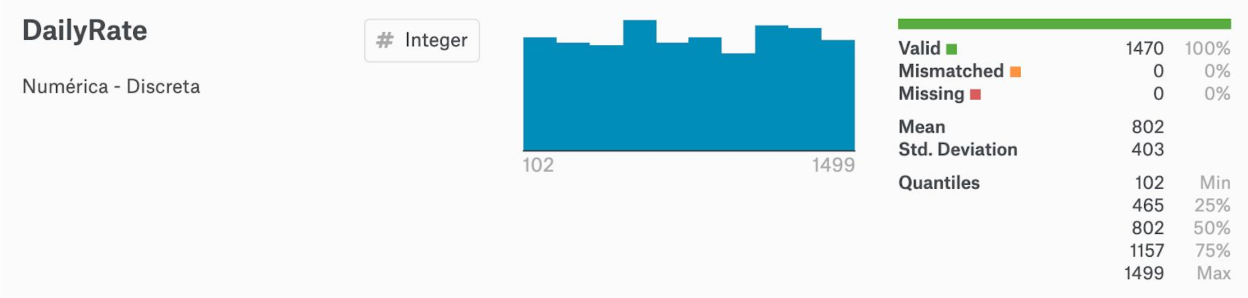
A String

Categórica

Travel_Rarely	71%
Travel_Frequently	19%
Non-Travel	10%

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	3	
Most Common	Travel_Rar	71%

Daily Rate



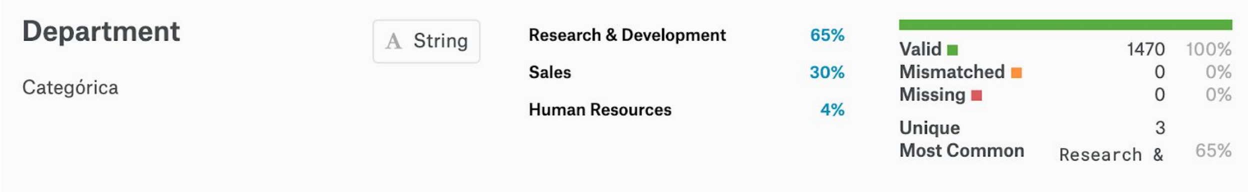
DailyRate

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	802	
Std. Deviation	403	
Quantiles	102	Min
	465	25%
	802	50%
	1157	75%
	1499	Max

Department



Department

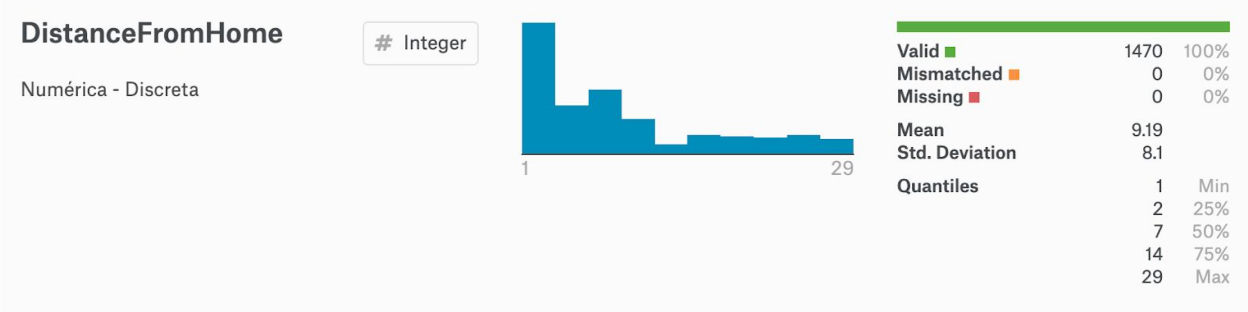
A String

Categórica

Research & Development	65%
Sales	30%
Human Resources	4%

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	3	
Most Common	Research &	65%

Distance From Home



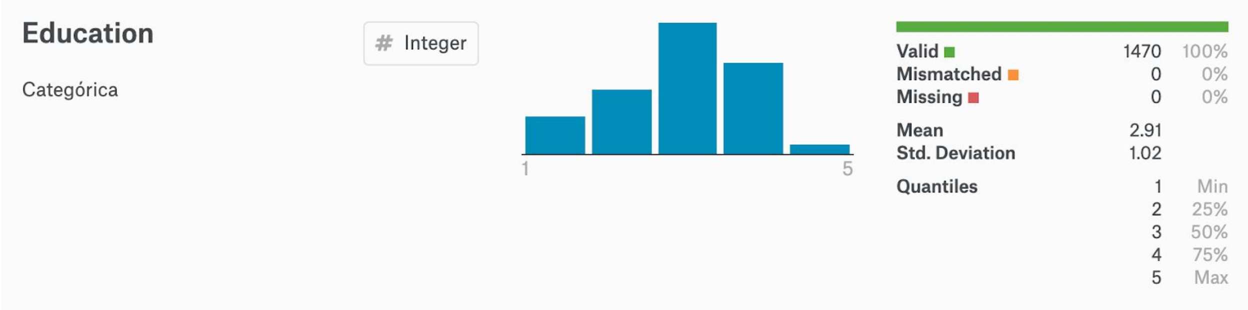
DistanceFromHome

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	9.19	
Std. Deviation	8.1	
Quantiles	1	Min
	2	25%
	7	50%
	14	75%
	29	Max

Education



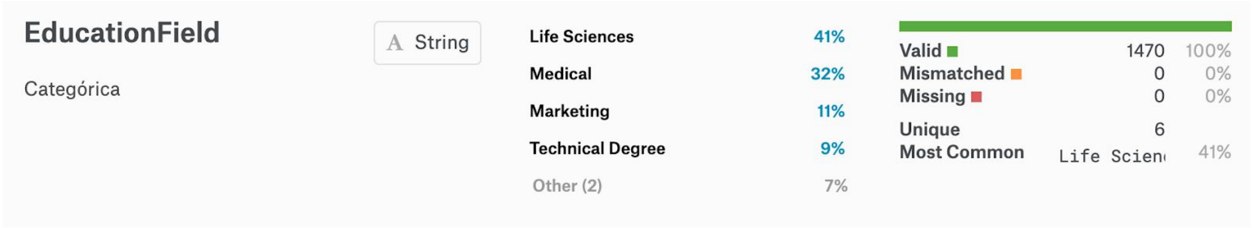
Education

Integer

Categórica

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.91	
Std. Deviation	1.02	
Quantiles	1	Min
	2	25%
	3	50%
	4	75%
	5	Max

Education Field



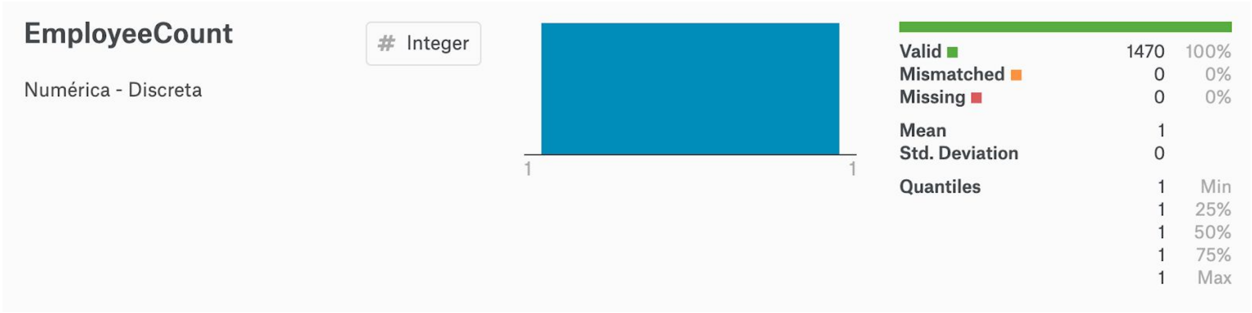
EducationField A String

Categórica

Life Sciences	41%
Medical	32%
Marketing	11%
Technical Degree	9%
Other (2)	7%

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	6	
Most Common	Life Scien	41%

Employee Count

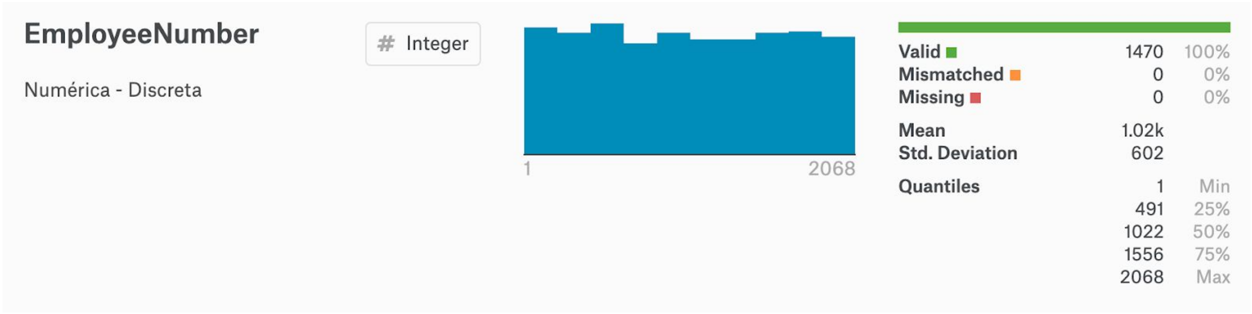


EmployeeCount # Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	1	
Std. Deviation	0	
Quantiles	1	Min
	1	25%
	1	50%
	1	75%
	1	Max

Employee Number

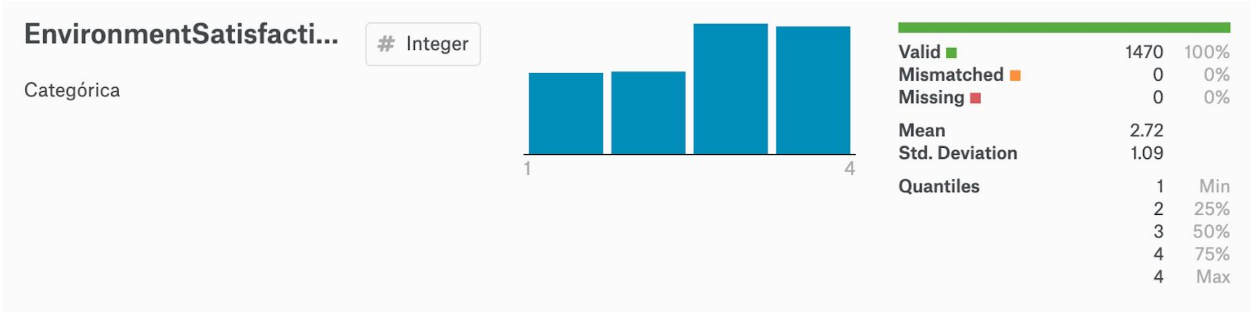


EmployeeNumber # Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	1.02k	
Std. Deviation	602	
Quantiles	1	Min
	491	25%
	1022	50%
	1556	75%
	2068	Max

Environment Satisfaction

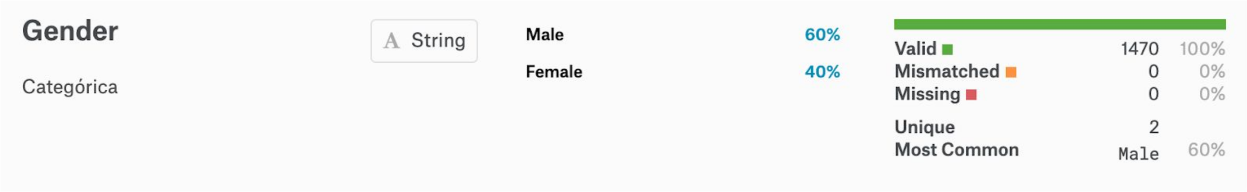


EnvironmentSatisfacti... # Integer

Categórica

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.72	
Std. Deviation	1.09	
Quantiles	1	Min
	2	25%
	3	50%
	4	75%
	4	Max

Gender



Gender

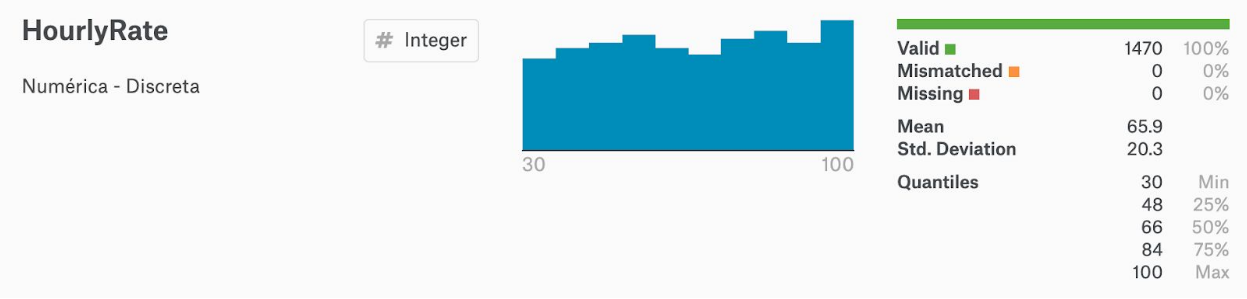
Categórica

A String

Male	60%
Female	40%

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	2	
Most Common	Male	60%

Hourly Rate



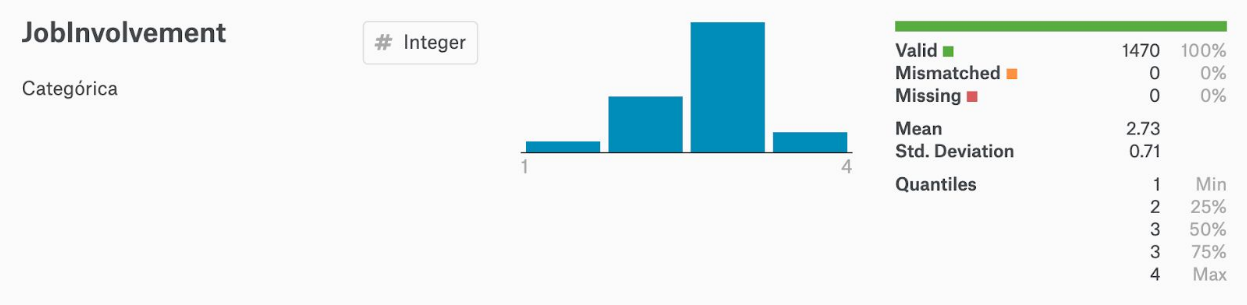
HourlyRate

Numérica - Discreta

Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	65.9	
Std. Deviation	20.3	
Quantiles	30	Min
	48	25%
	66	50%
	84	75%
	100	Max

Job Involvement



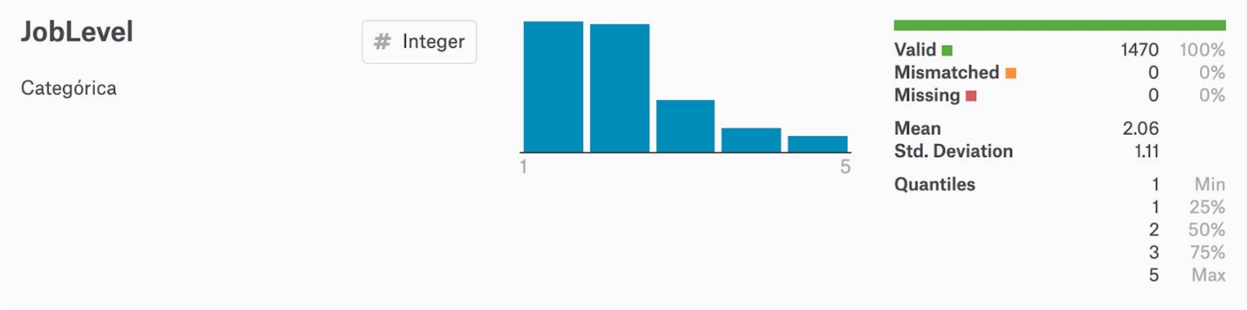
JobInvolvement

Categórica

Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.73	
Std. Deviation	0.71	
Quantiles	1	Min
	2	25%
	3	50%
	3	75%
	4	Max

Job Level



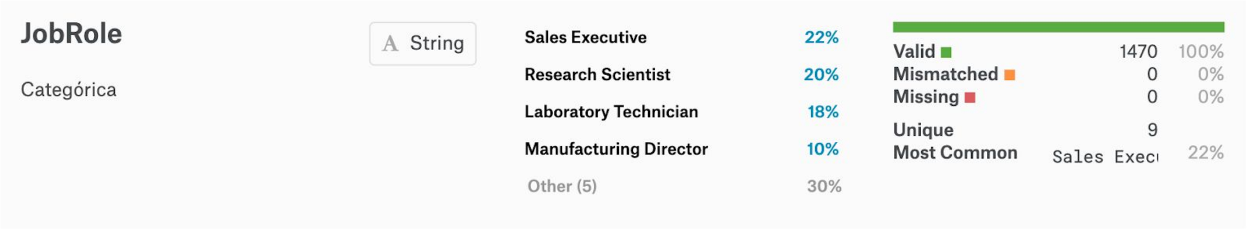
JobLevel

Categórica

Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.06	
Std. Deviation	1.11	
Quantiles	1	Min
	1	25%
	2	50%
	3	75%
	5	Max

Job Role



JobRole

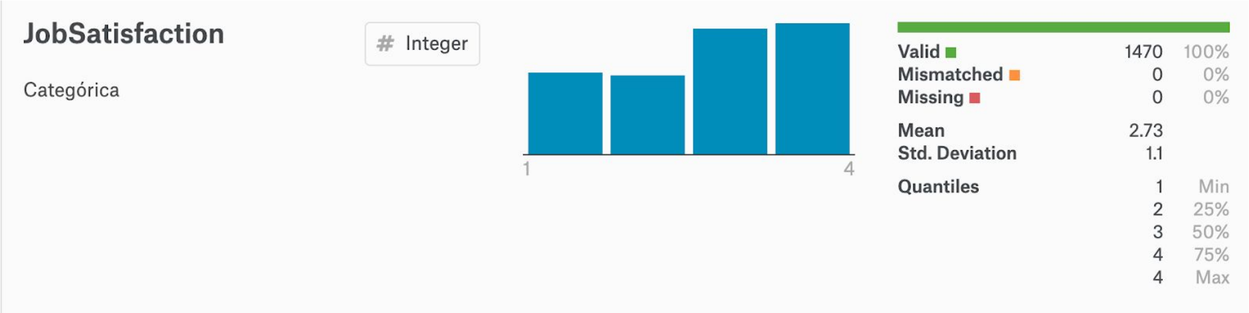
Categórica

A String

Sales Executive	22%
Research Scientist	20%
Laboratory Technician	18%
Manufacturing Director	10%
Other (5)	30%

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	9	
Most Common	Sales Exec	22%

Job Satisfaction



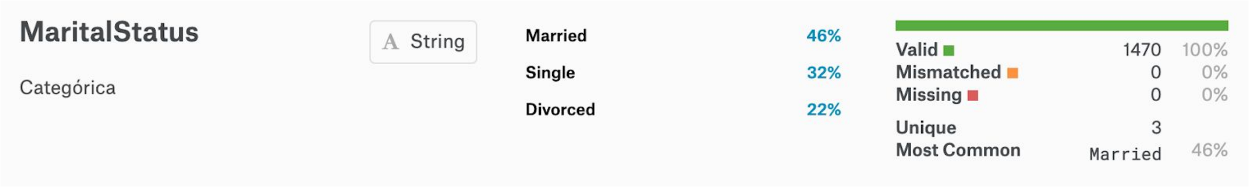
JobSatisfaction

Categórica

# Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.73	
Std. Deviation	1.1	
Quantiles	1	Min
	2	25%
	3	50%
	4	75%
	4	Max

Marital Status



MaritalStatus

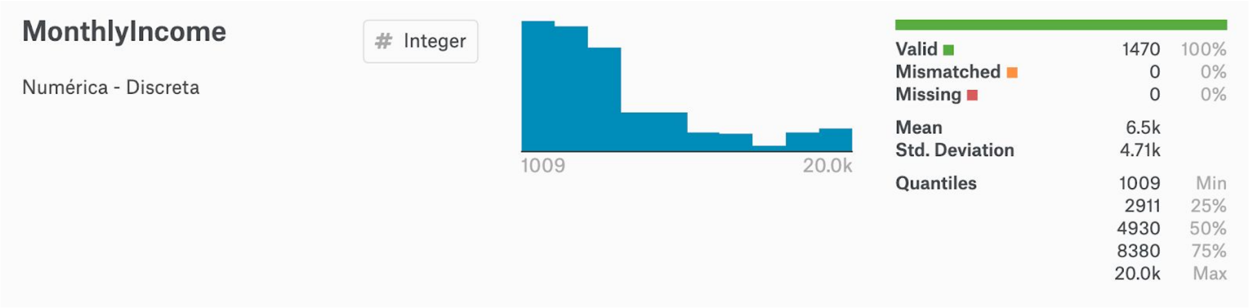
Categórica

A String

Married	46%
Single	32%
Divorced	22%

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	3	
Most Common	Married	46%

Monthly Income



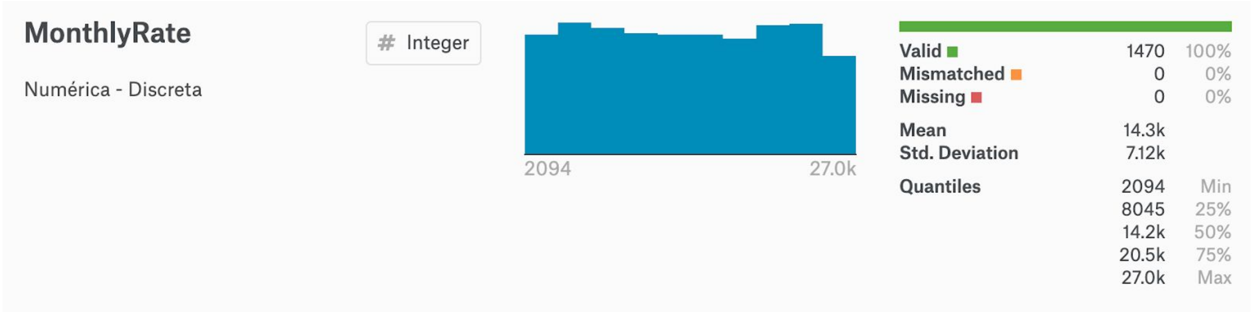
MonthlyIncome

Numérica - Discreta

# Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	6.5k	
Std. Deviation	4.71k	
Quantiles	1009	Min
	2911	25%
	4930	50%
	8380	75%
	20.0k	Max

Monthly Rate



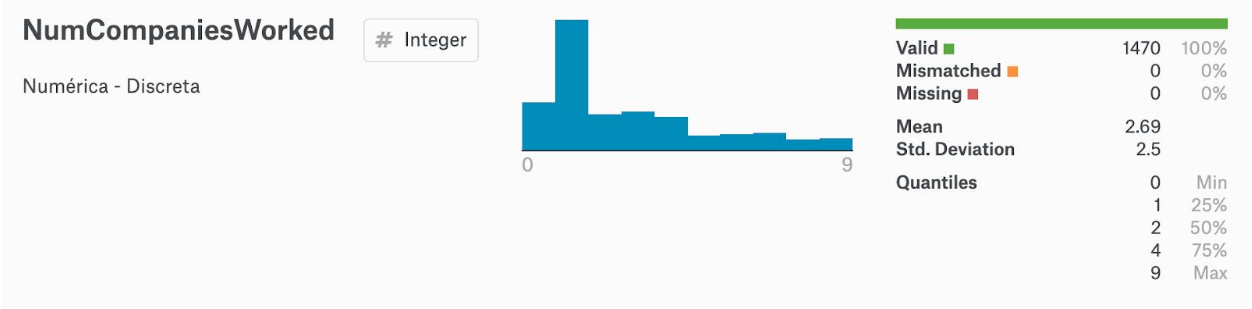
MonthlyRate

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	14.3k	
Std. Deviation	7.12k	
Quantiles	2094	Min
	8045	25%
	14.2k	50%
	20.5k	75%
	27.0k	Max

Num Companies Worked



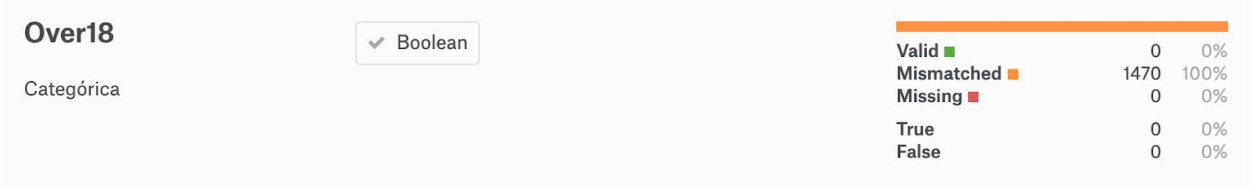
NumCompaniesWorked

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.69	
Std. Deviation	2.5	
Quantiles	0	Min
	1	25%
	2	50%
	4	75%
	9	Max

Over 18



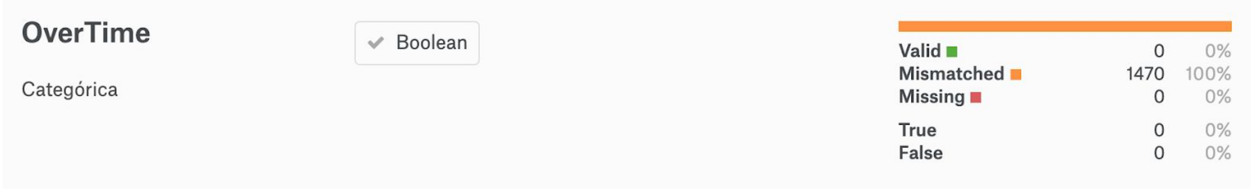
Over18

✓ Boolean

Categórica

Valid 🟩	0	0%
Mismatched 🟧	1470	100%
Missing 🟥	0	0%
True	0	0%
False	0	0%

Over Time



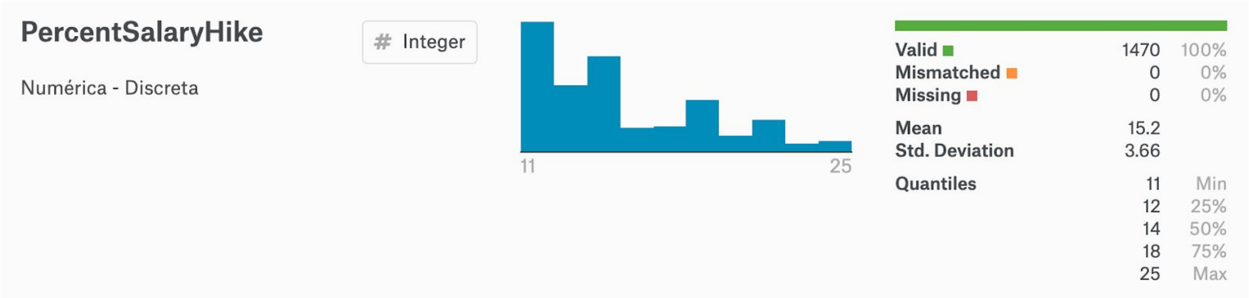
OverTime

✓ Boolean

Categórica

Valid 🟩	0	0%
Mismatched 🟧	1470	100%
Missing 🟥	0	0%
True	0	0%
False	0	0%

Percent Salary Hike



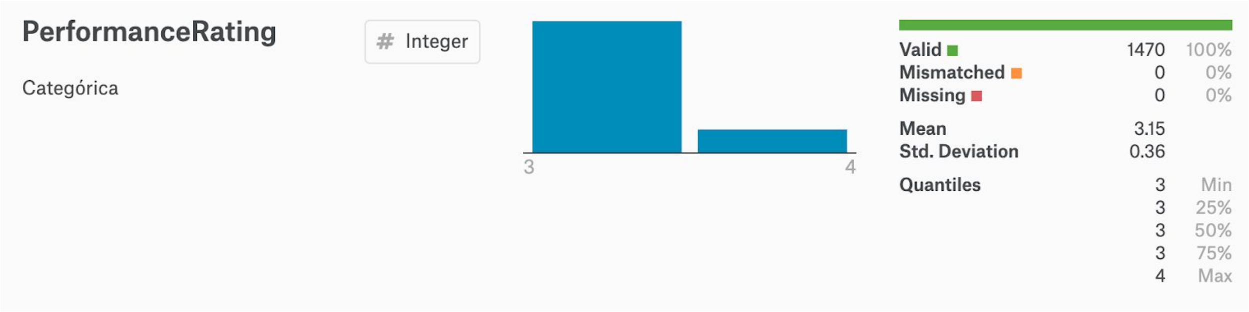
PercentSalaryHike

Numérica - Discreta

# Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	15.2	
Std. Deviation	3.66	
Quantiles	11	Min
	12	25%
	14	50%
	18	75%
	25	Max

Performance Rating



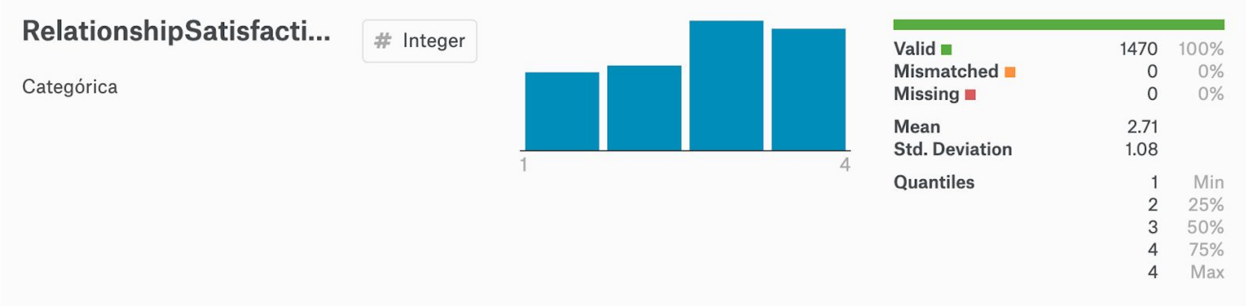
PerformanceRating

Categórica

# Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	3.15	
Std. Deviation	0.36	
Quantiles	3	Min
	3	25%
	3	50%
	3	75%
	4	Max

Relationship Satisfaction



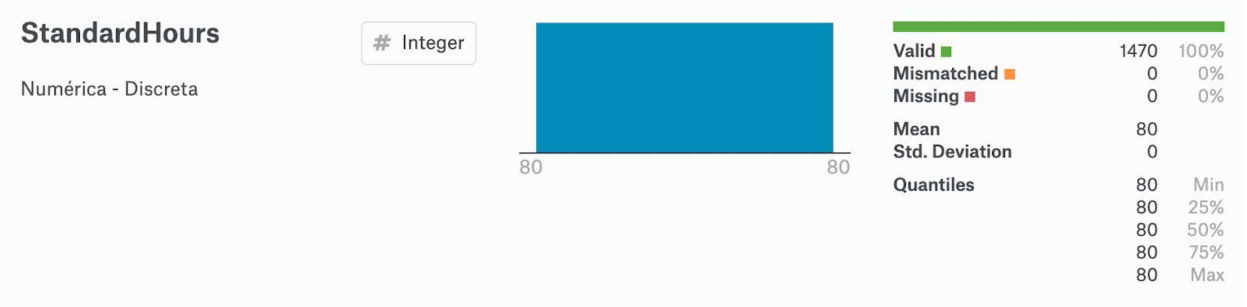
RelationshipSatisfacti...

Categórica

# Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.71	
Std. Deviation	1.08	
Quantiles	1	Min
	2	25%
	3	50%
	4	75%
	4	Max

Standard Hours



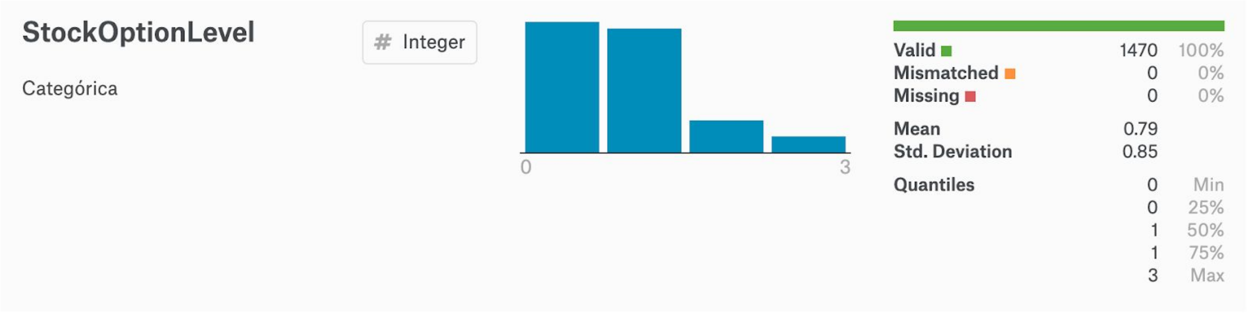
StandardHours

Numérica - Discreta

# Integer

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	80	
Std. Deviation	0	
Quantiles	80	Min
	80	25%
	80	50%
	80	75%
	80	Max

Stock Option Level

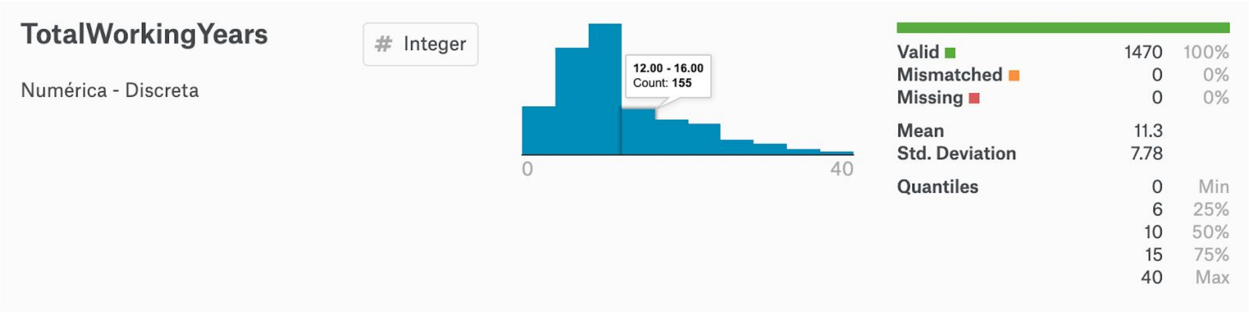

StockOptionLevel

Integer

Categórica

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	0.79	
Std. Deviation	0.85	
Quantiles	0	Min
	0	25%
	1	50%
	1	75%
	3	Max

Total Working Years

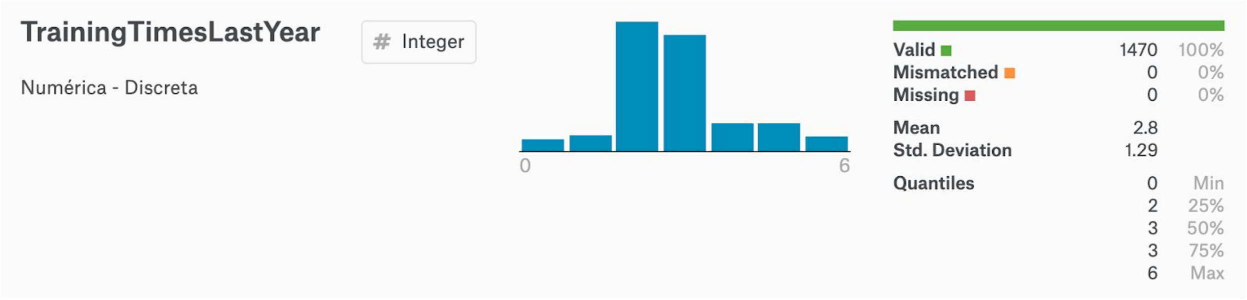

TotalWorkingYears

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	11.3	
Std. Deviation	7.78	
Quantiles	0	Min
	6	25%
	10	50%
	15	75%
	40	Max

Training Times Last Year

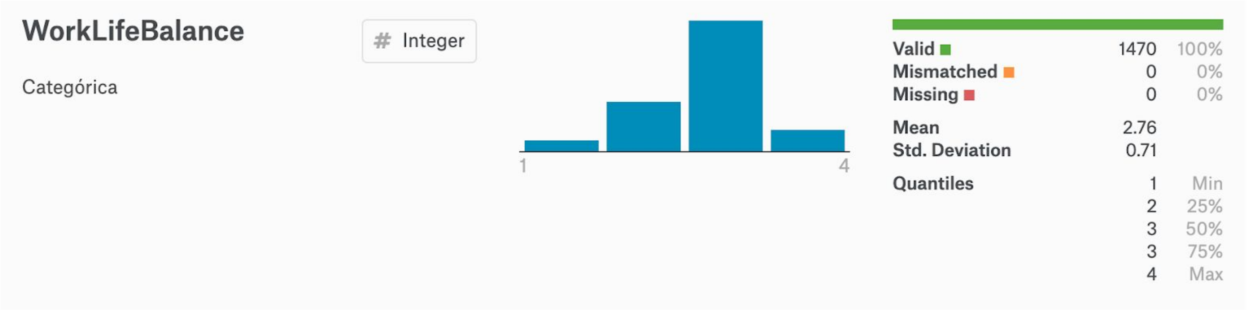

TrainingTimesLastYear

Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.8	
Std. Deviation	1.29	
Quantiles	0	Min
	2	25%
	3	50%
	3	75%
	6	Max

Work Life Balance

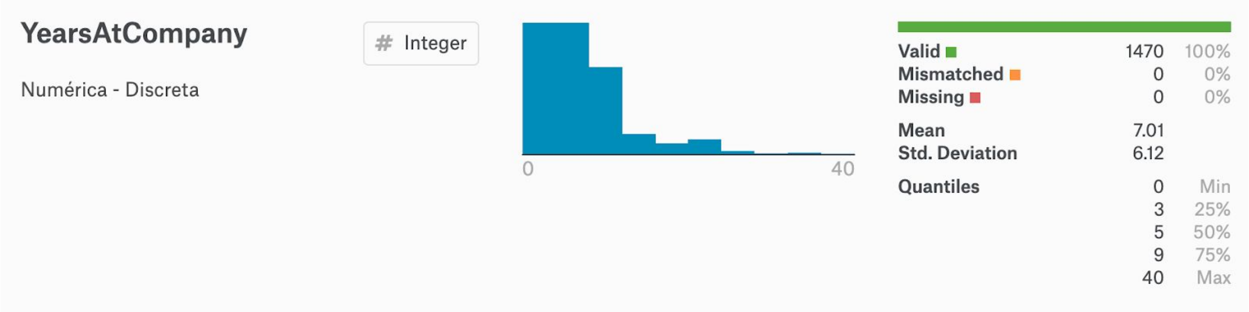


WorkLifeBalance

Categórica

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.76	
Std. Deviation	0.71	
Quantiles	1	Min
	2	25%
	3	50%
	3	75%
	4	Max

Years At Company

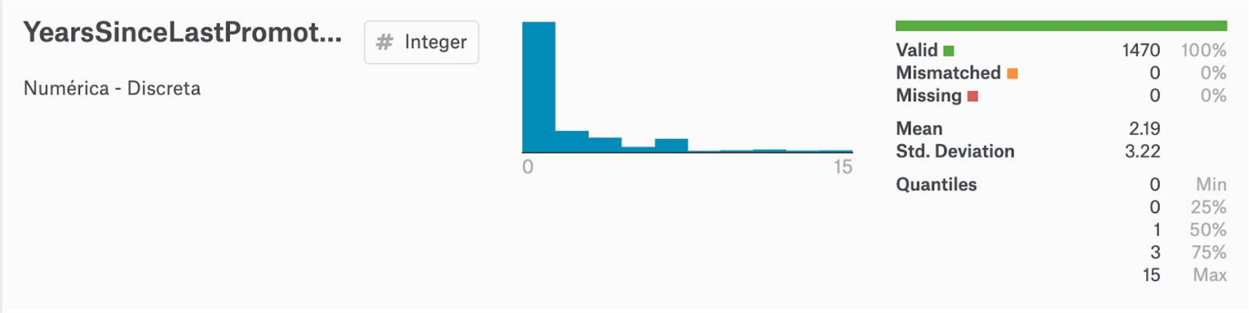


YearsAtCompany

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	7.01	
Std. Deviation	6.12	
Quantiles	0	Min
	3	25%
	5	50%
	9	75%
	40	Max

Years Since Last Promotion

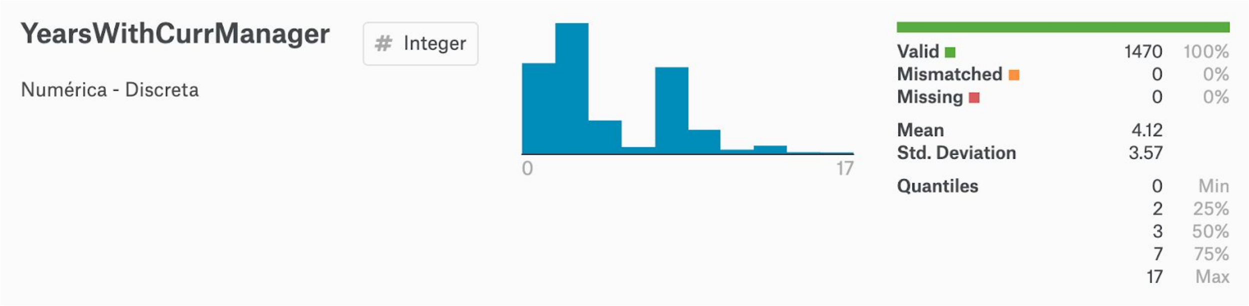


YearsSinceLastPromot...

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	2.19	
Std. Deviation	3.22	
Quantiles	0	Min
	0	25%
	1	50%
	3	75%
	15	Max

Years With Curr Manager

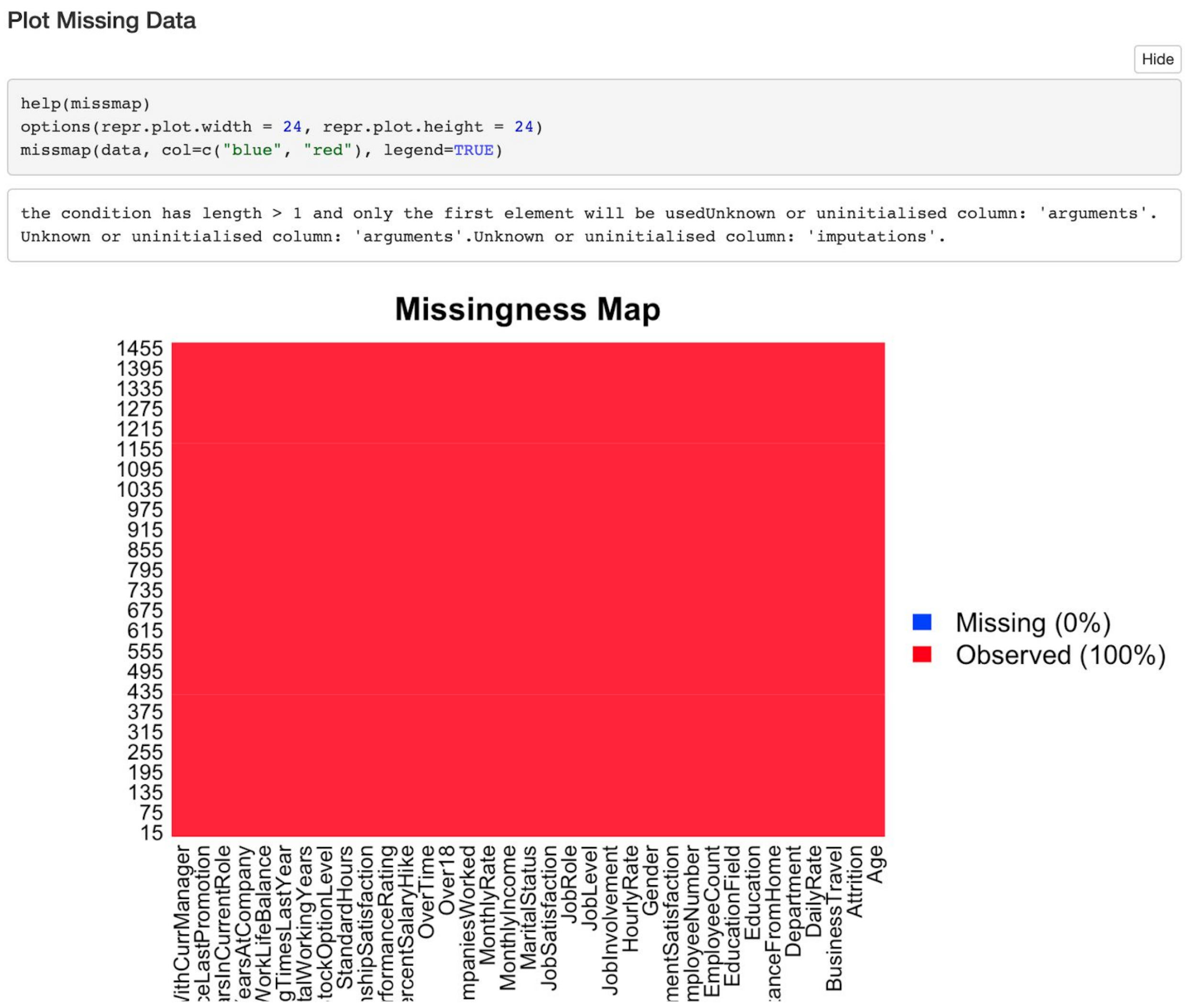


YearsWithCurrManager # Integer

Numérica - Discreta

Valid 🟩	1470	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Mean	4.12	
Std. Deviation	3.57	
Quantiles	0	Min
	2	25%
	3	50%
	7	75%
	17	Max

Data Quality Assessment



Plot Missing Data

```
help(missmap)
options(repr.plot.width = 24, repr.plot.height = 24)
missmap(data, col=c("blue", "red"), legend=TRUE)
```

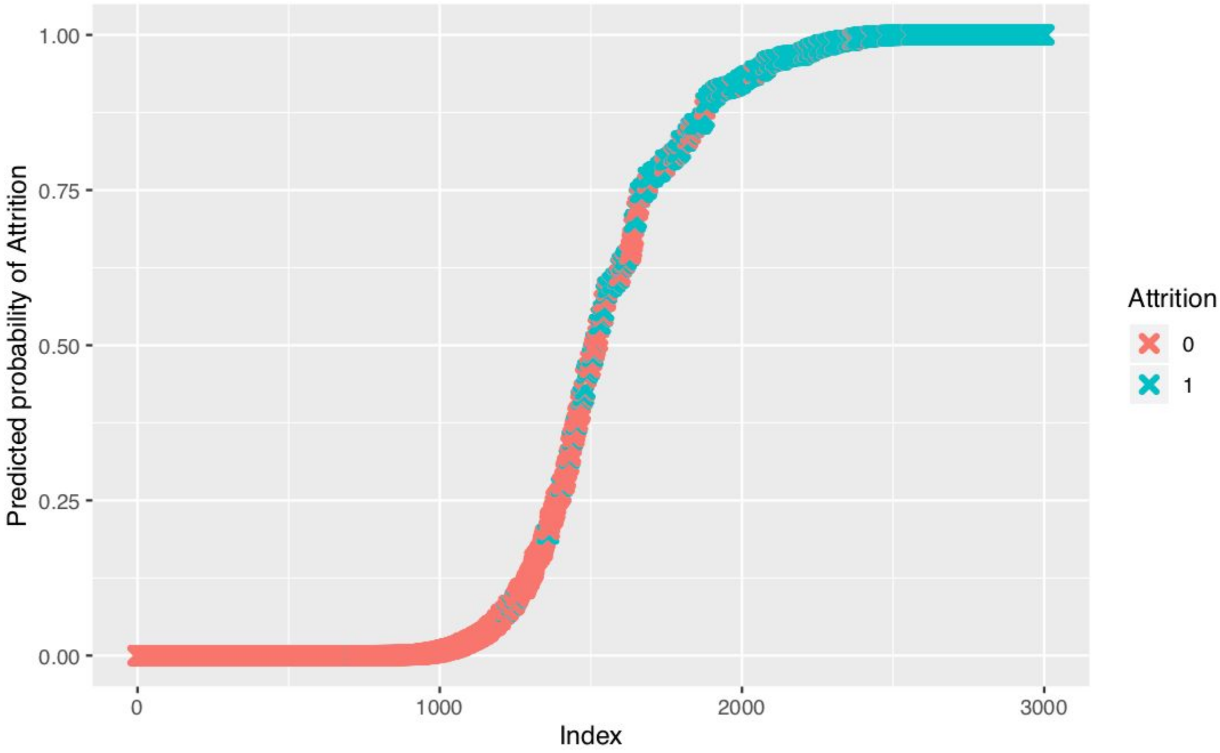
```
the condition has length > 1 and only the first element will be usedUnknown or uninitialised column: 'arguments'.
Unknown or uninitialised column: 'arguments'.Unknown or uninitialised column: 'imputations'.
```

As can be seen by the **Missingness Map**, we did not have to consider removing rows or interpolating our data. All values were available for our model to work with at the onset.

Logistic Regression Model Building in R Studio

Once we built our Logistic Regression model using the techniques used in lecture such as AIC drop1 evaluation, interactions, and feature elimination of items having a lower p-value than desired indicating an insignificant regressor, we were ready to plot the results of the performance of our model.

After plotting our predicted values for all 3000 observations in our dataset, we see a nice “S” shaped curve which is expected for a logistic regression model. Notice how most items below 0.50 are orange whereas the majority of items above 0.50 are green. Of course, it is not perfect as we can see some green items below 0.50 and some orange items above 0.50 indicating misclassifications which will end up in the False Positive and False Negative buckets of our confusion matrix.



Sensitivity and Specificity

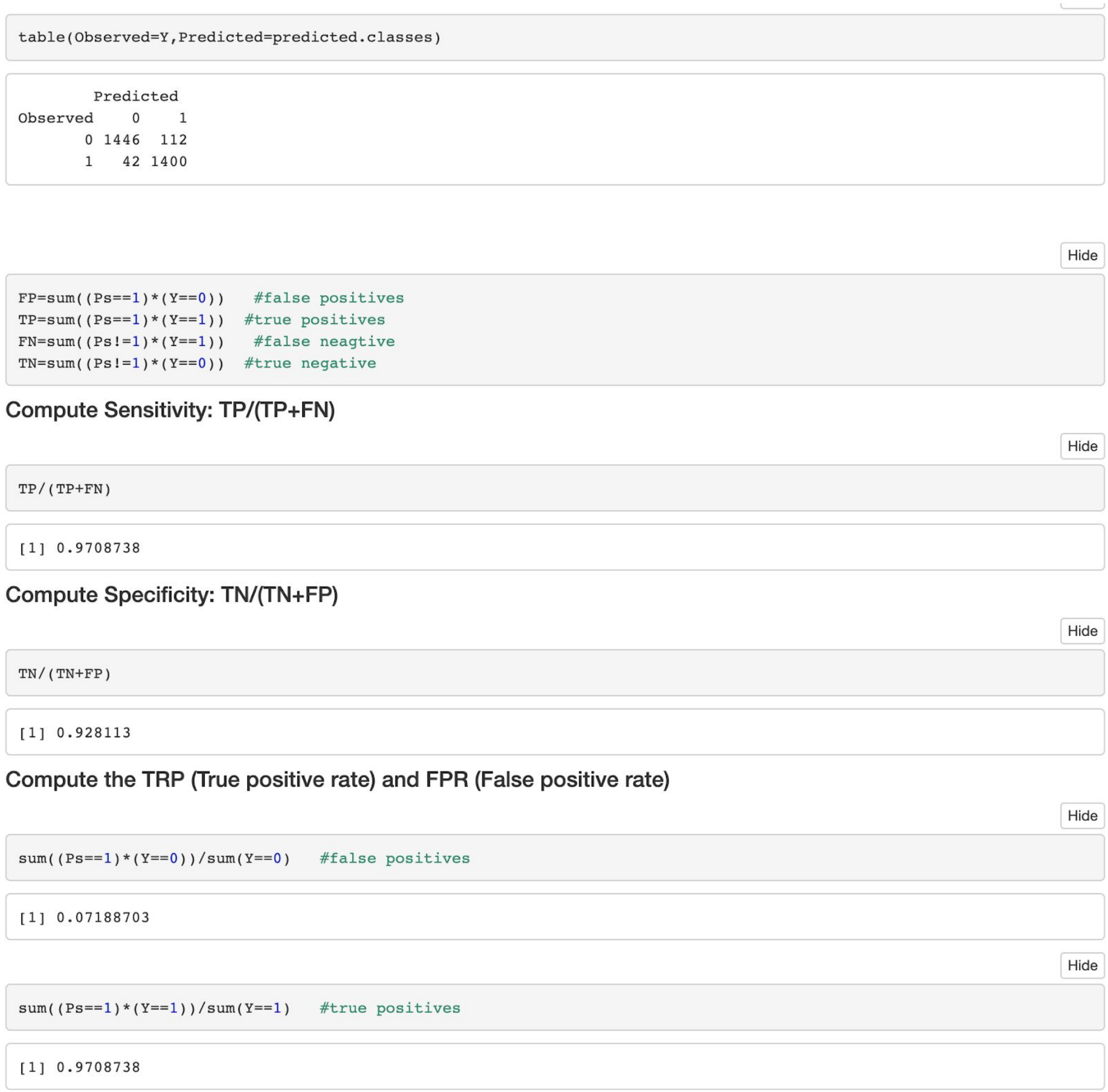
There are a number of methods of evaluating whether a logistic model is a good model. One such way is sensitivity and specificity.

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function:

Sensitivity (also called the *true positive rate*, or the recall in some fields) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate. Sensitivity= true positives/(true positive + false negative)

Specificity (also called the *true negative rate*) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. Specificity=true negatives/(true negative + false positives)

Here are the results of our predictions relative to the data’s true observed values:



```
table(Observed=Y,Predicted=predicted.classes)
```

```
        Predicted
Observed    0    1
       0 1446  112
       1   42 1400
```

```
FP=sum((Ps==1)*(Y==0))   #false positives
TP=sum((Ps==1)*(Y==1))  #true positives
FN=sum((Ps!=1)*(Y==1))   #false neagtive
TN=sum((Ps!=1)*(Y==0))  #true negative
```

Compute Sensitivity: TP/(TP+FN)

```
TP/(TP+FN)
```

```
[1] 0.9708738
```

Compute Specificity: TN/(TN+FP)

```
TN/(TN+FP)
```

```
[1] 0.928113
```

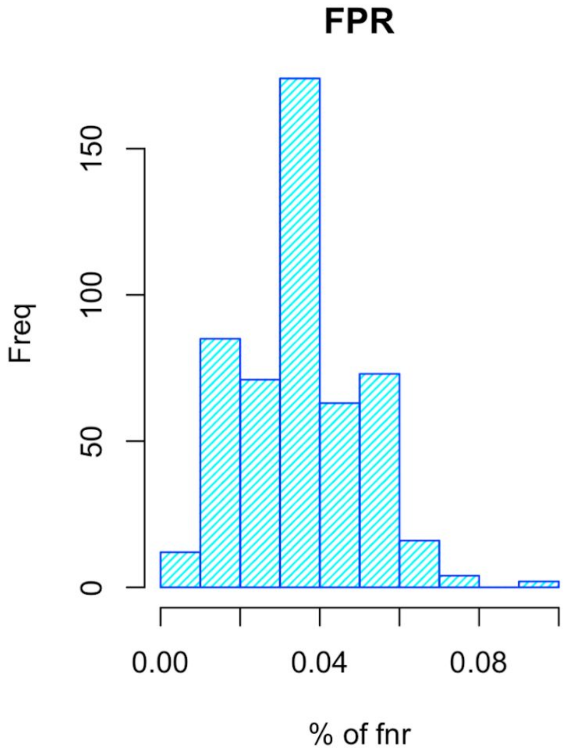
Compute the TRP (True positive rate) and FPR (False positive rate)

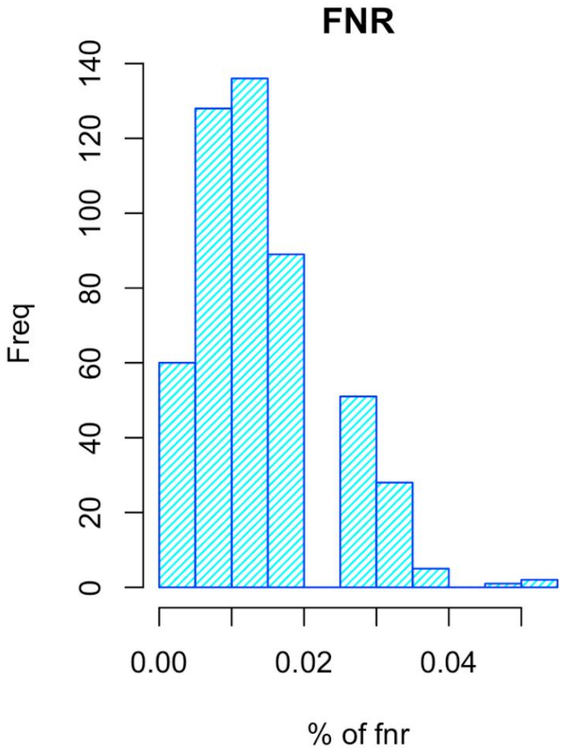
```
sum((Ps==1)*(Y==0))/sum(Y==0)   #false positives
```

```
[1] 0.07188703
```

```
sum((Ps==1)*(Y==1))/sum(Y==1)   #true positives
```

```
[1] 0.9708738
```

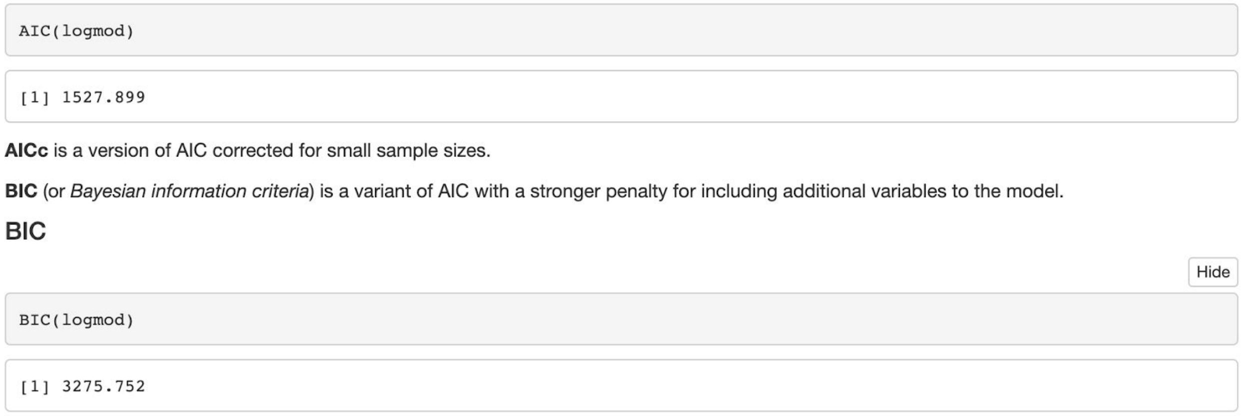

FPR
Freq
0
50
100
150
0.00
0.04
0.08
% of fnr


FNR
Freq
0
20
40
60
80
100
120
140
0.00
0.02
0.04
% of fnr

AIC and BIC

Additionally, there are four other important metrics - **AIC, AICc, BIC and Mallows Cp** - that are commonly used for model evaluation and selection. These are an unbiased estimate of the model prediction error **MSE**. The lower these metrics, the better the model.

AIC stands for (Akaike's Information Criteria), a metric developed by the Japanese Statistician, Hirotugu Akaike, 1970. The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model.



```
AIC(logmod)
```

```
[1] 1527.899
```

AICc is a version of AIC corrected for small sample sizes.

BIC (or *Bayesian information criteria*) is a variant of AIC with a stronger penalty for including additional variables to the model.

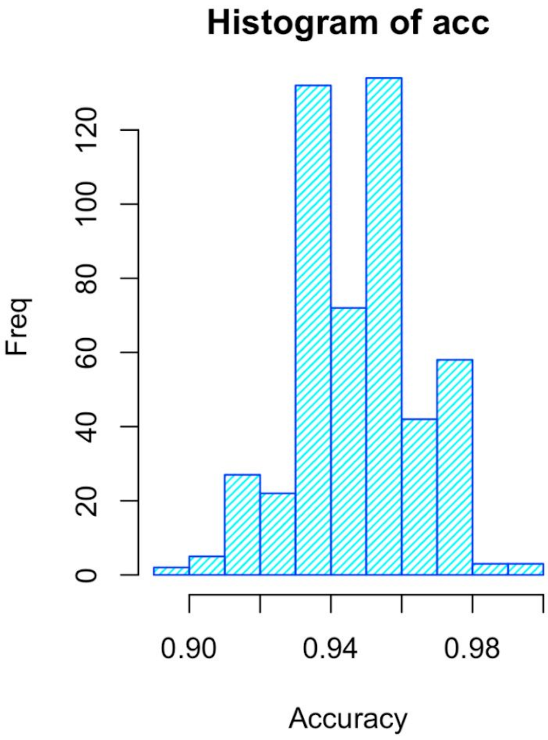
BIC

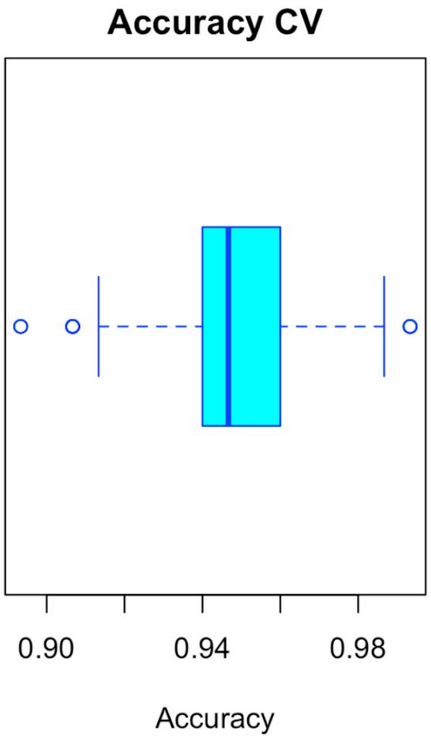
```
BIC(logmod)
```

```
[1] 3275.752
```

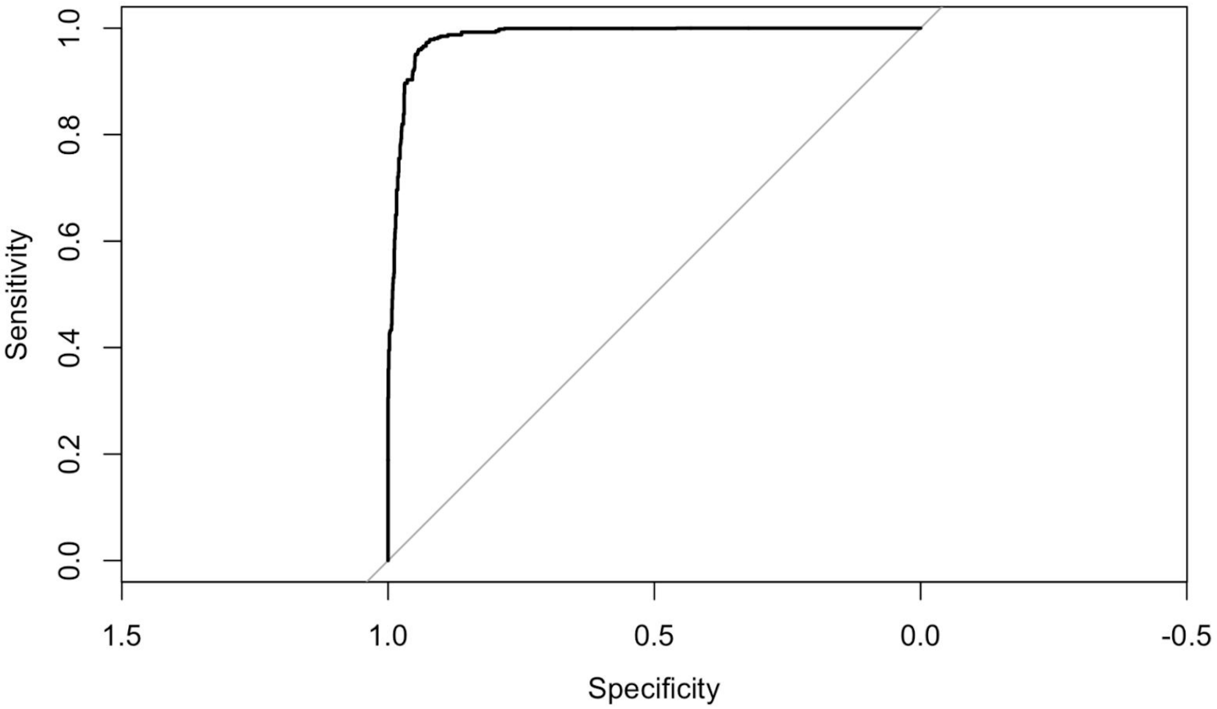
Cross-Validation

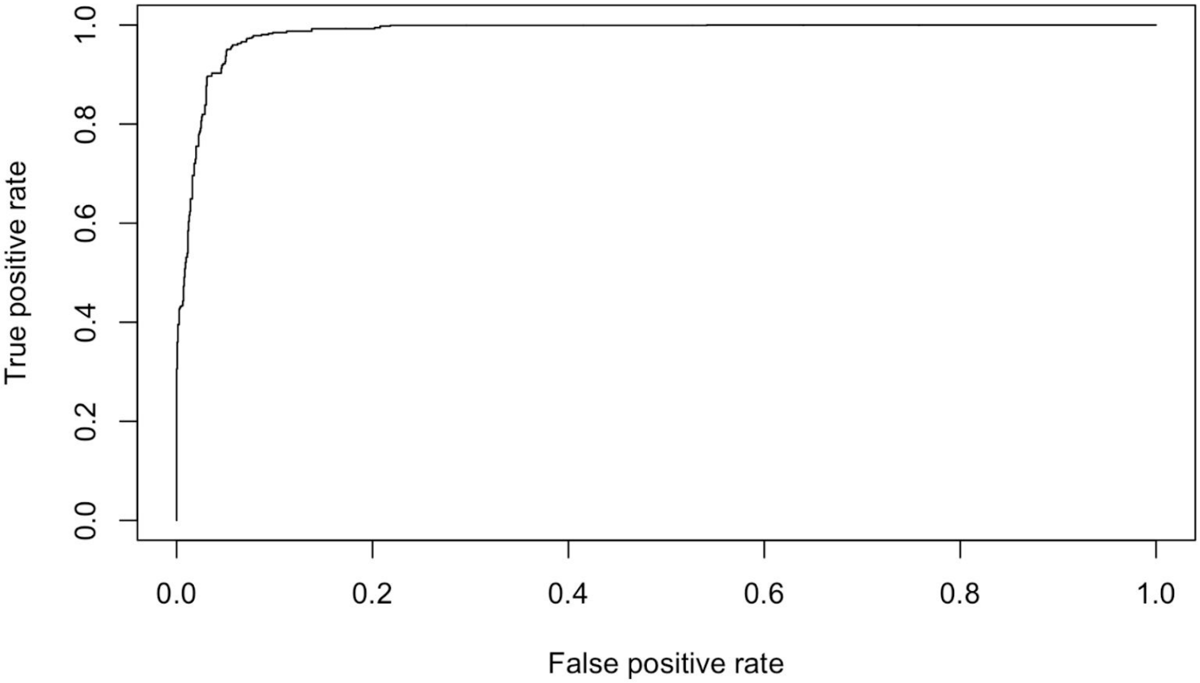
After performing 500 epochs of test/train Cross-Validation testing of our model, the next visualization shows that our logistic regression model is on average 94.86% accurate at predicting the correct True Positive or True Negative value corresponding to our response variable, Attrition.


Histogram of acc

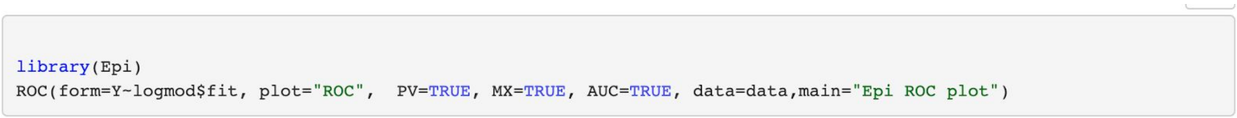

Accuracy CV

Receiver Operating Characteristic (ROC) Curve

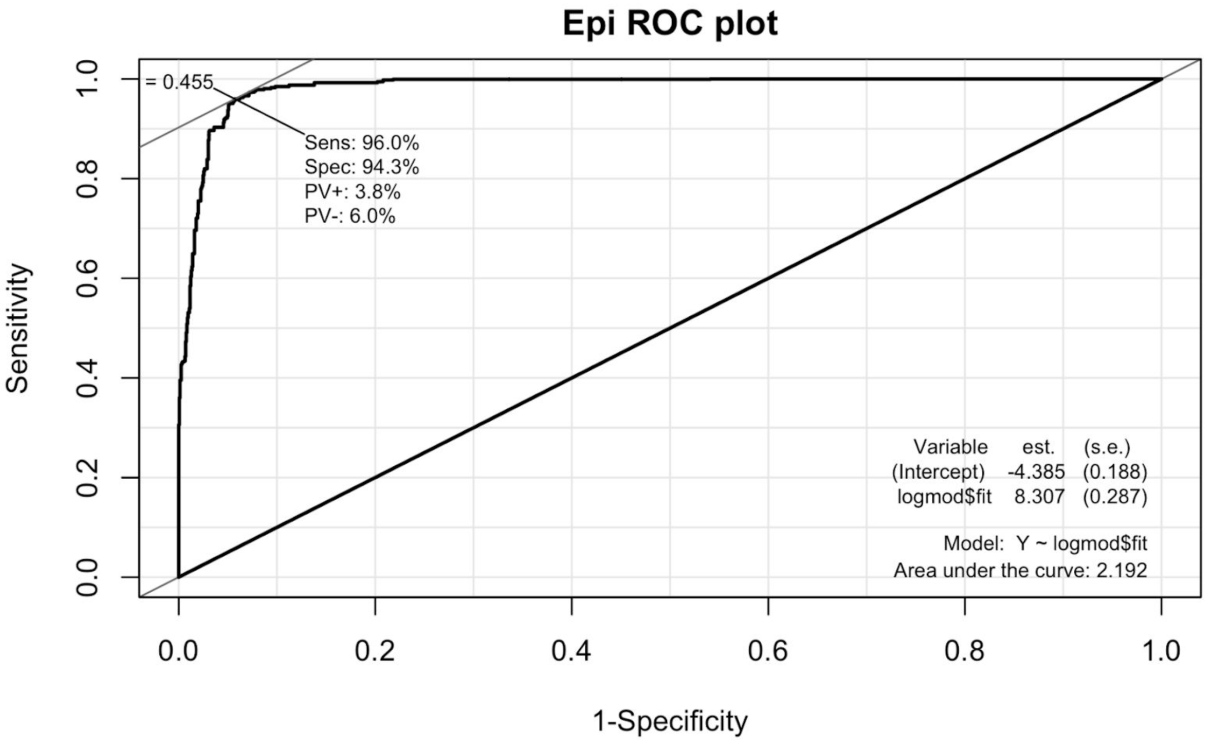




A good ROC curve will have a steep vertical line hugging the left side of the graph, a slight bend which then turns into a strong horizontal line at the top of the chart. Our graph depicts those characteristics indicating a good ROC curve. This equates to a larger AUC (Area Under the Curve) where 1 is the best AUC achievable.



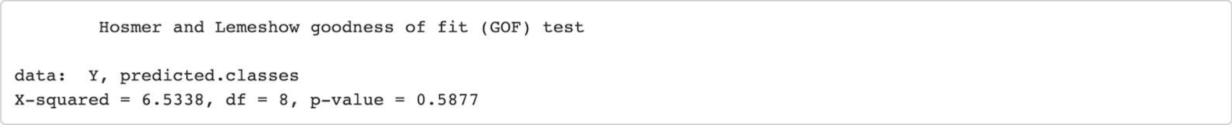
```
library(Epi)
ROC(form=Y~logmod$fit, plot="ROC",  PV=TRUE, MX=TRUE, AUC=TRUE, data=data,main="Epi ROC plot")
```



Using R’s ROC() function, the previous visualizations shows an AUC of 2.192. This value was unexpected and appears to be an error on the R library but indeed our ROC plot looks solid and aligns well with our model’s actual Specificity and Sensitivity performance metrics.

Hosmer-Lemeshow Goodness of Fit

How well our model fits depends on the difference between the model and the observed data. One approach for binary data is to implement a Hosmer Lemeshow goodness of fit test.



```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  Y, predicted.classes
X-squared = 6.5338, df = 8, p-value = 0.5877
```

Our model appears to fit well because we have no significant difference between the model and the observed data (i.e. the p-value is above 0.05).

As with all measures of model fit, we'll use this as just one piece of information in deciding how well this model fits. It doesn't work well in very large or very small data sets, but is often useful nonetheless.

Model Selection and Improvements for Best Fit

We used several techniques to build a robust Logistic Regression model to predict the Attrition of an employee at a given company.

First, we loaded the data into R Studio. Then we converted some of the variables to factors taking into consideration specific variables that are ordinal in nature. The dataset was not missing any data. We knew we were going to build a logistic model, as Attrition was a binary with outcomes of 0 and 1 representing leaving the job or staying, respectively.

After summarizing and visualizing our data regressors with respect to Attrition using various common libraries and techniques, we discovered that our Attrition response variable had a minority value of 1 resulting in a significantly unbalanced dataset. If not treated properly, our model would more frequently predict 0. Therefore, we used a common under/oversampling technique to balance our dataset.

We then spent a significant amount of time looking at all of the 35 regressors using our intuition and empirical analysis to determine those variables that had significant interactions. Our primary heuristic used to gauge a better model was by minimizing our AIC to a value of 1527.899 as we adjusted various interactions within the model. The end result was a pseudo-adjusted R^2 value of 0.77 and a p-value of 0 indicating a good model.

We conducted the likelihood ratio test to determine if the model with all of the variables is appropriate. The test statistic, Chi-square= deviance – residual deviance=4154.4-945.9=3208.5. The degrees of freedom=2999-2709=290. So, then we found the p value by using 1-pchisq (3208.5,290) to be 0. Which means we can conclude that the full model, with all of the variables, is significant.

We then computed the confusion matrix of our model which consisted of our precision and sensitivity corresponding to our true positive and true negative rates, respectively. Our recall was 0.97 where we predicted 97% of all True Positive samples correctly and we predicted 93% of all True Negative samples correctly.

For example, the following interactions we used in our model make intuitive sense:

- **factor(BusinessTravel) * DistanceFromHome * factor(MaritalStatus) * factor(WorkLifeBalance)**

  One interpretation for this is that married employees with a family that have a bad work/life balance could lead to attrition as they seek positions that offer a better work/life balance. Conversely, this could also indicate that they stay because the employer values a work/life balance.

- **factor(JobInvolvement) * factor(JobLevel) * factor(JobSatisfaction)**

  If you are not involved in your job, you may not be satisfied.

- **factor(EnvironmentSatisfaction) * factor(Gender) * factor(JobRole)**

  If you are not happy in your environment due to various social reasons based on say gender imbalance, for example, this can be a key motivator to leave.

- **MonthlyIncome * HourlyRate * factor(OverTime)**

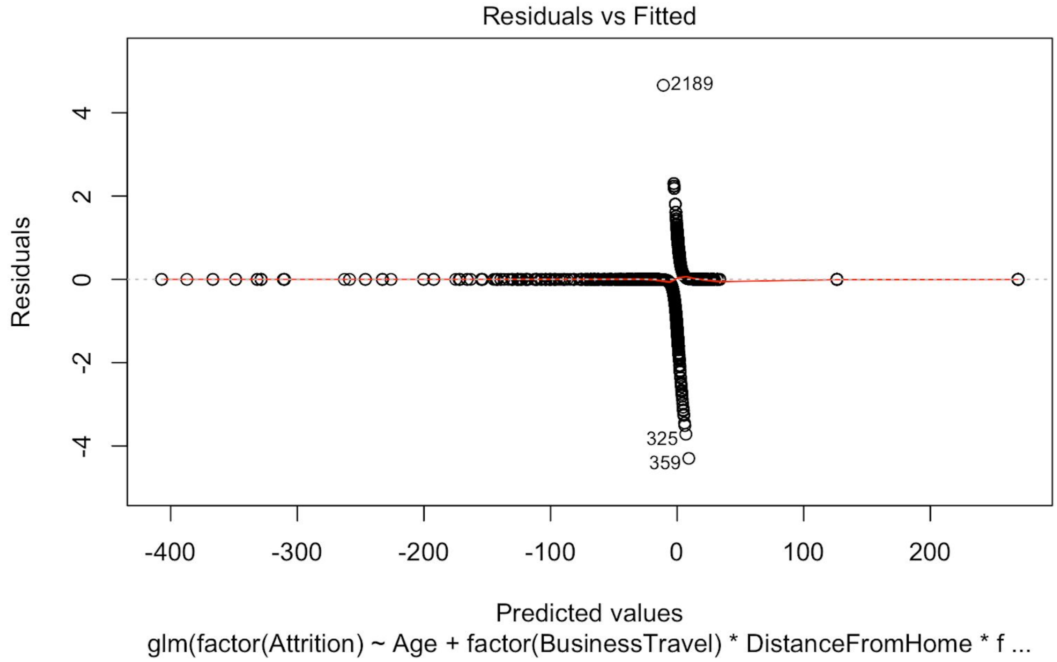
  Sometimes employees that feel they should make more will leave for higher wages. Likewise, if you get a lot of overtime, this can typically be a positive aspect for some.

- **factor(StockOptionLevel) * TotalWorkingYears * TrainingTimesLastYear**

  The longer you work at a company, the more stock options you are able to obtain. Likewise, if you are not able to receive training, you may become stagnant or bored doing the same thing over and over again.

- **PercentSalaryHike * YearsAtCompany * YearsSinceLastPromotion * YearsWithCurrManager**

  Again, depending on how long you have worked at a company, these interactions can lead to leaving if your manager is very bad or they could stay with their manager is great and they stay with the same manager for a long time. Often if you get along with your manager, the relationship gets better over time.

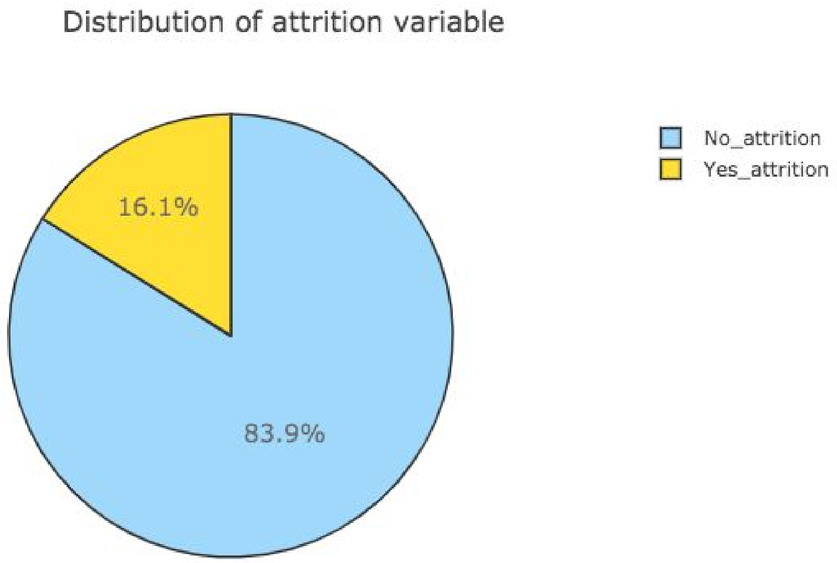


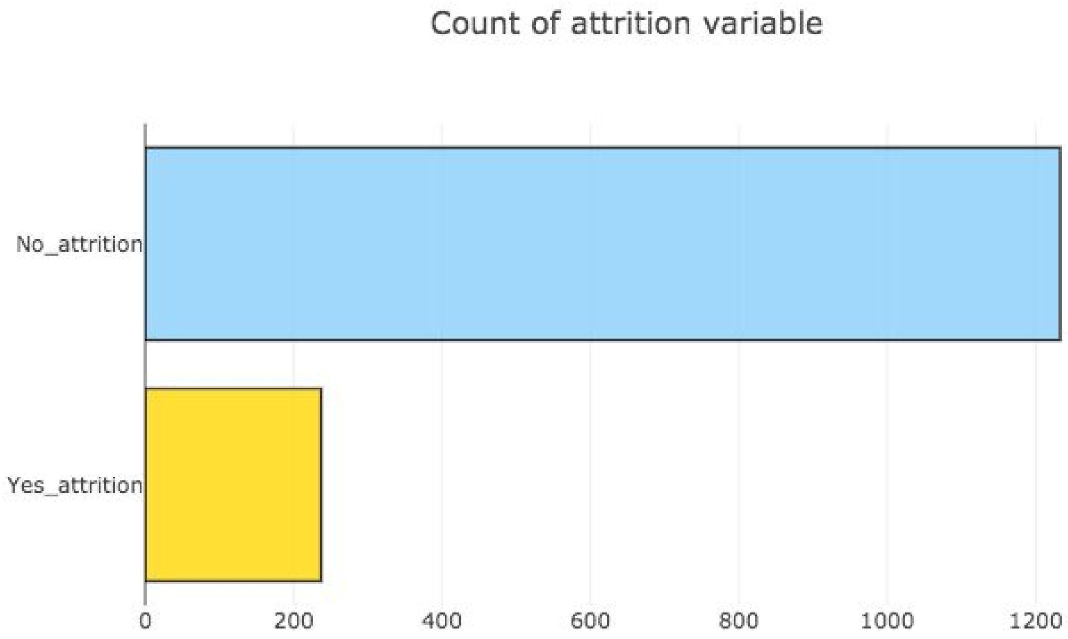
When we plotted our logistic model, the scatterplot of residuals and predicted values followed a pattern and showed 3 outliers.

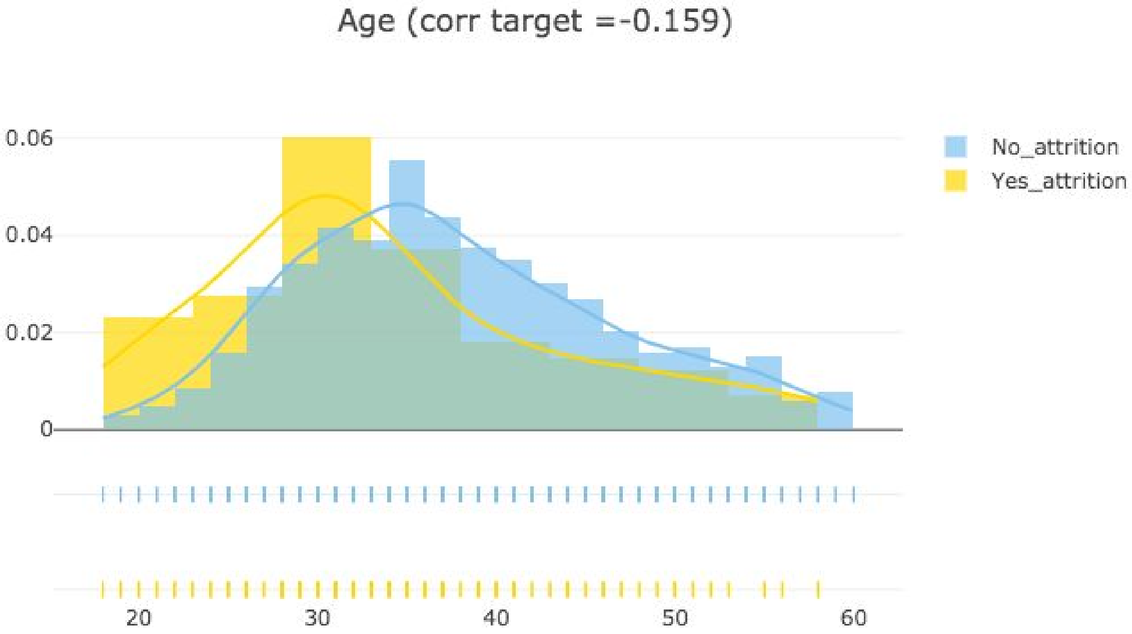
Advanced Modeling Techniques

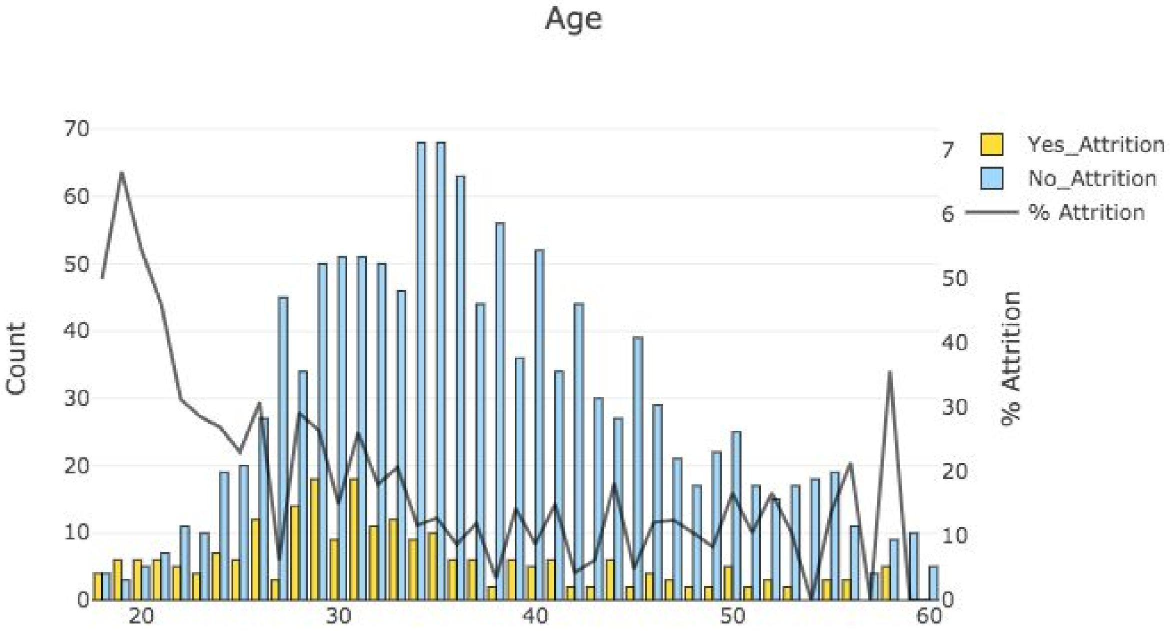
Utilizing Machine Learning algorithms in Python to compare my model generated utilizing backward-propagation to what Matthew has created.

- Performing a count on the attrition variable, it is obvious that the occurrence of attrition is around 16% yes to 84% no.
- Lets see how some of the variables correspond to attrition as we begin the feature selection of the model.

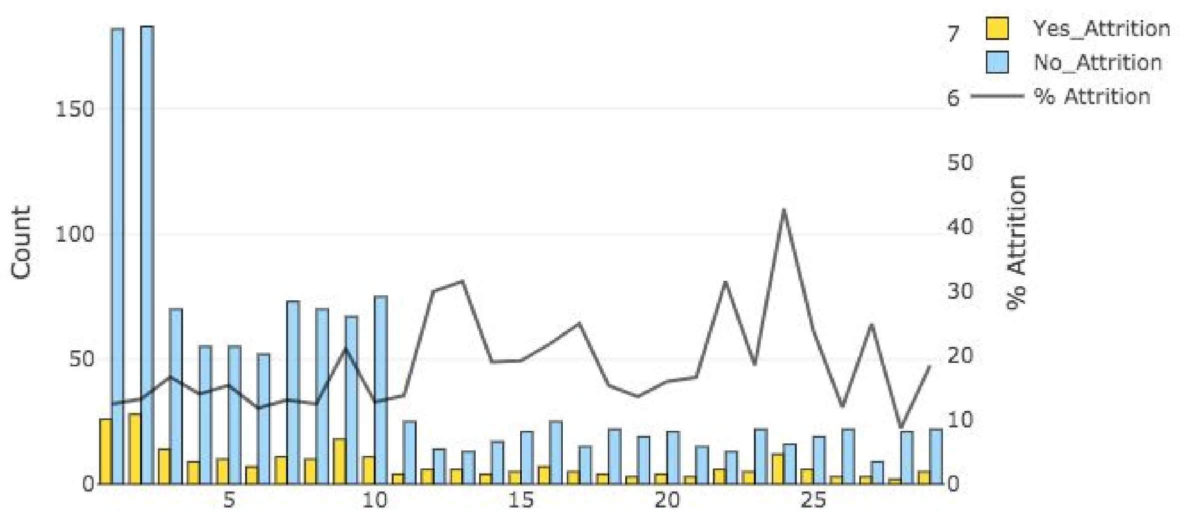



Count of attrition variable
No_attrition
Yes_attrition
0
200
400
600
800
1000
1200

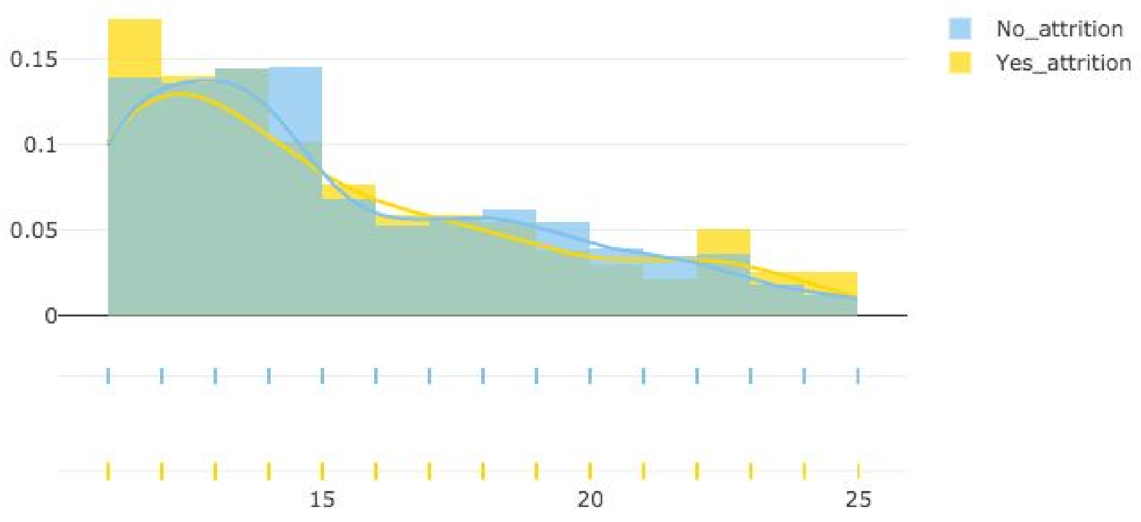

Age (corr target =-0.159)
0.06
0.04
0.02
0
No_attrition
Yes_attrition
20
30
40
50
60


Age
Count
70
60
50
40
30
20
10
0
20
30
40
50
60
% Attrition
7
6
50
40
30
20
10
0
Yes_Attrition
No_Attrition
% Attrition

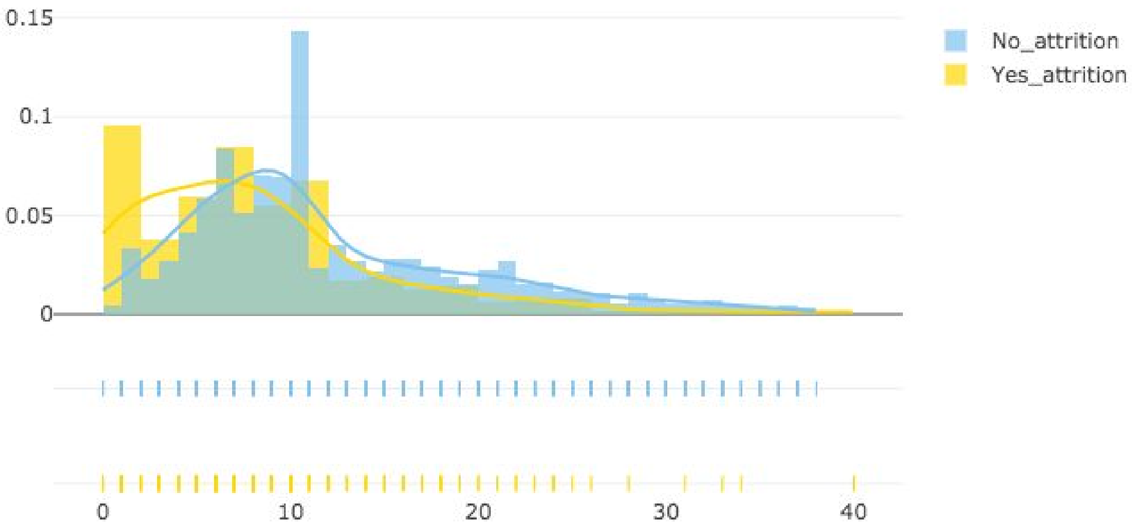
DistanceFromHome



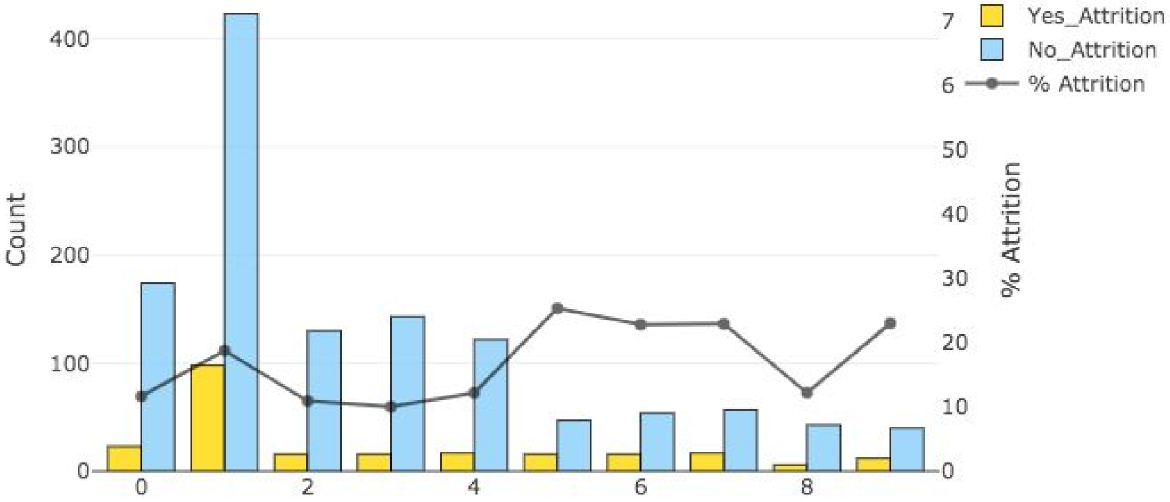
PercentSalaryHike (corr target =-0.013)

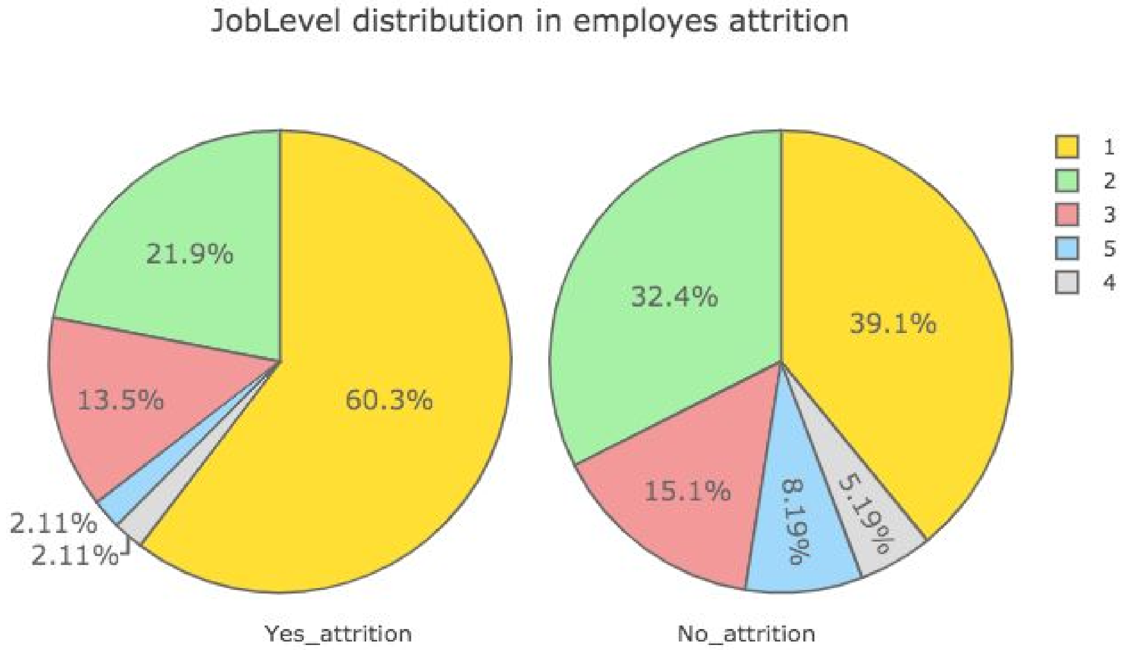


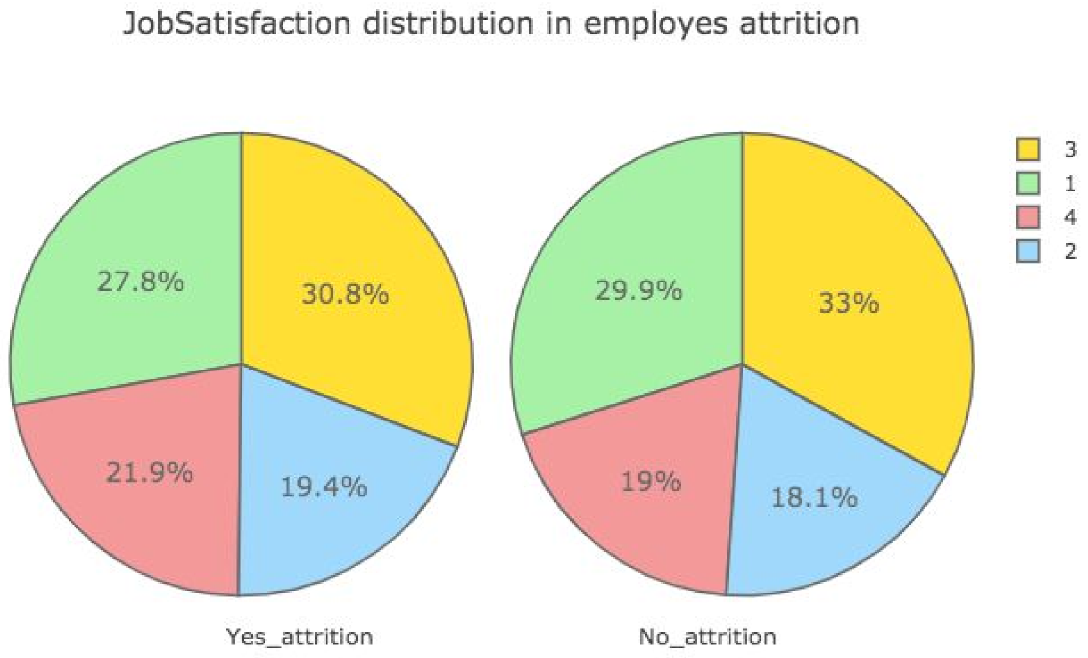
TotalWorkingYears (corr target =-0.171)



NumCompaniesWorked

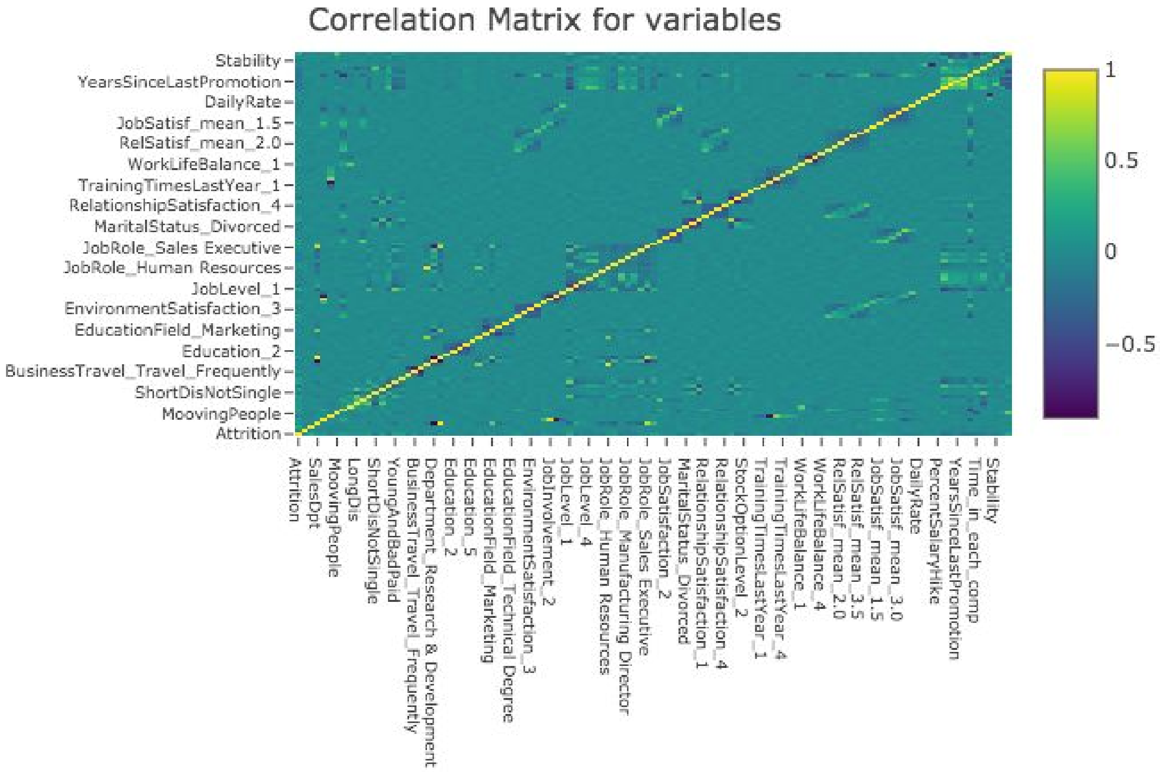



JobLevel distribution in employes attrition
21.9%
13.5%
60.3%
2.11%
2.11%
Yes_attrition
32.4%
39.1%
15.1%
8.19%
5.19%
No_attrition
1
2
3
5
4

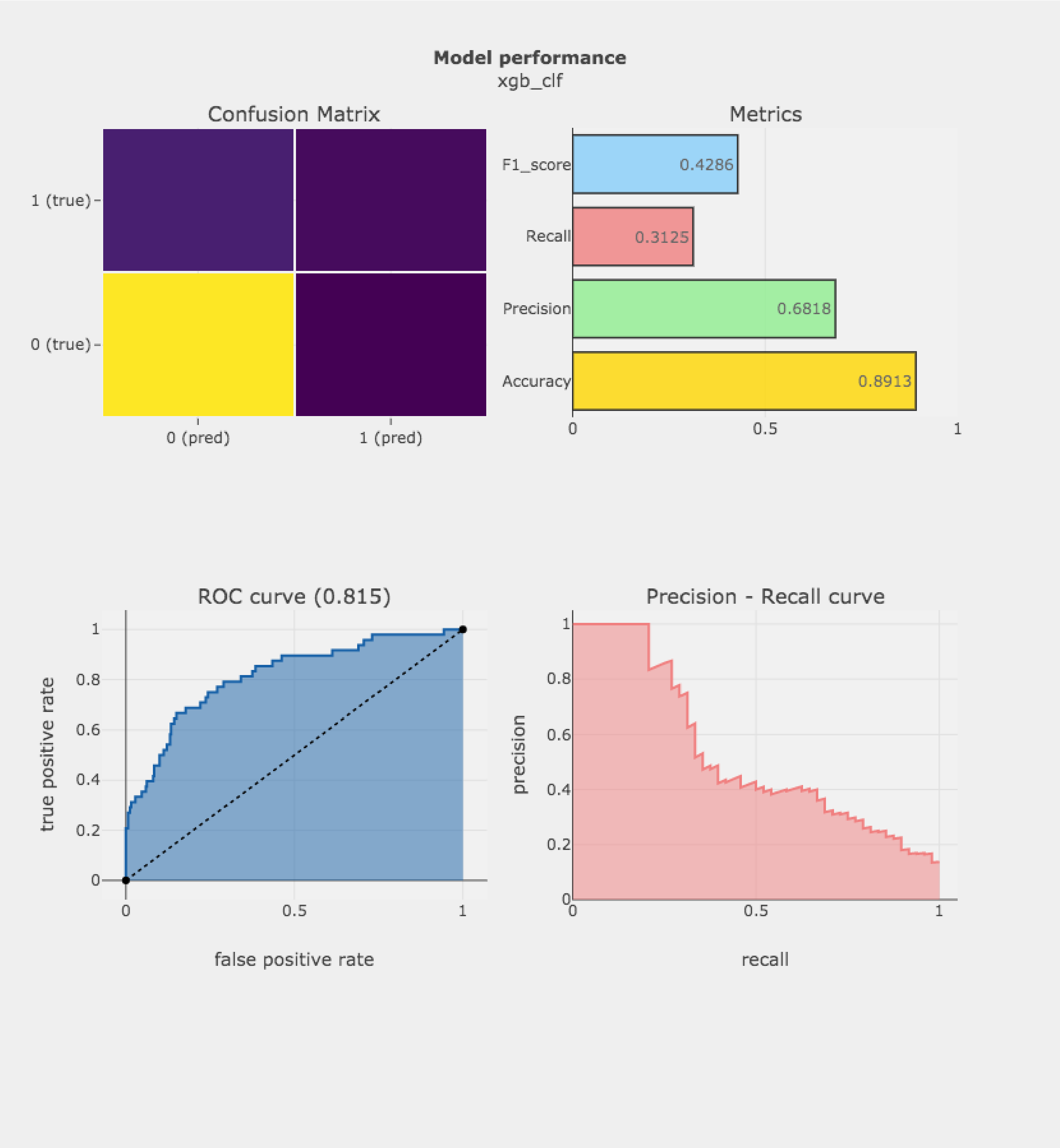


Time to utilize the XGB Classifier and perform a RandomizedSearchCV to optimize hyperparameters.

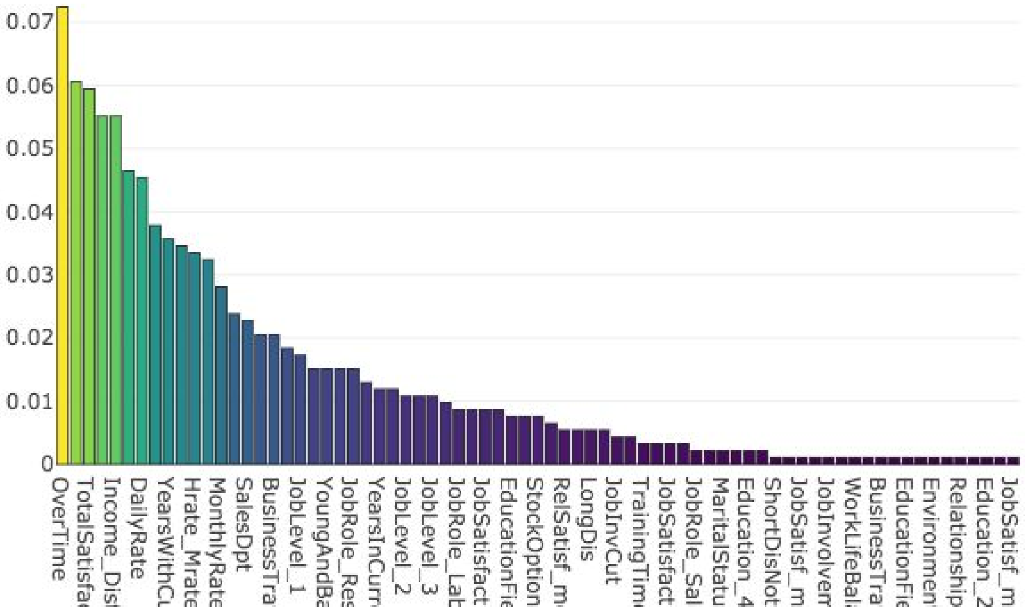
- Visualizing the correlation matrix
- Best accuracy for 5-fold search with 800 parameter combinations: 89%
- Understanding the importance of all of the features in a high dimensional space.
- Using ROC curves to understand the relationships between variables
- Understanding cross-validation scores.



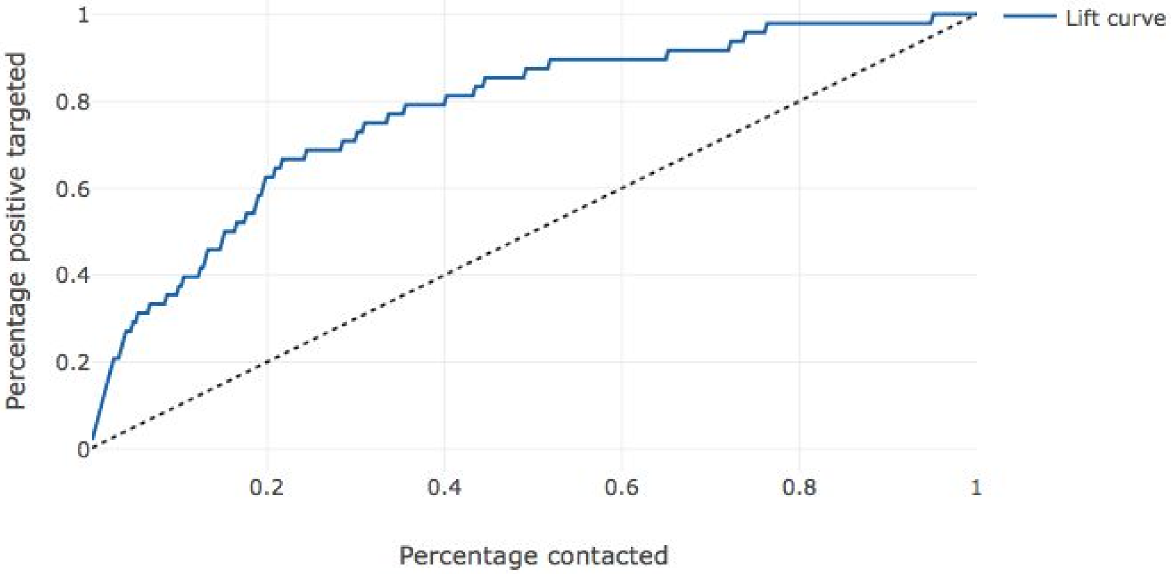
Correlation Matrix for variables


Model performance
xgb_clf
Confusion Matrix
1 (true)
0 (true)
0 (pred)
1 (pred)
Metrics
F1_score
0.4286
Recall
0.3125
Precision
0.6818
Accuracy
0.8913
0
0.5
1
ROC curve (0.815)
true positive rate
false positive rate
0
0.2
0.4
0.6
0.8
1
0
0.5
1
Precision - Recall curve
precision
recall
0
0.2
0.4
0.6
0.8
1
0
0.5
1

Feature Importances xgb_cfl



Cumulative gains curve xgb_cfl



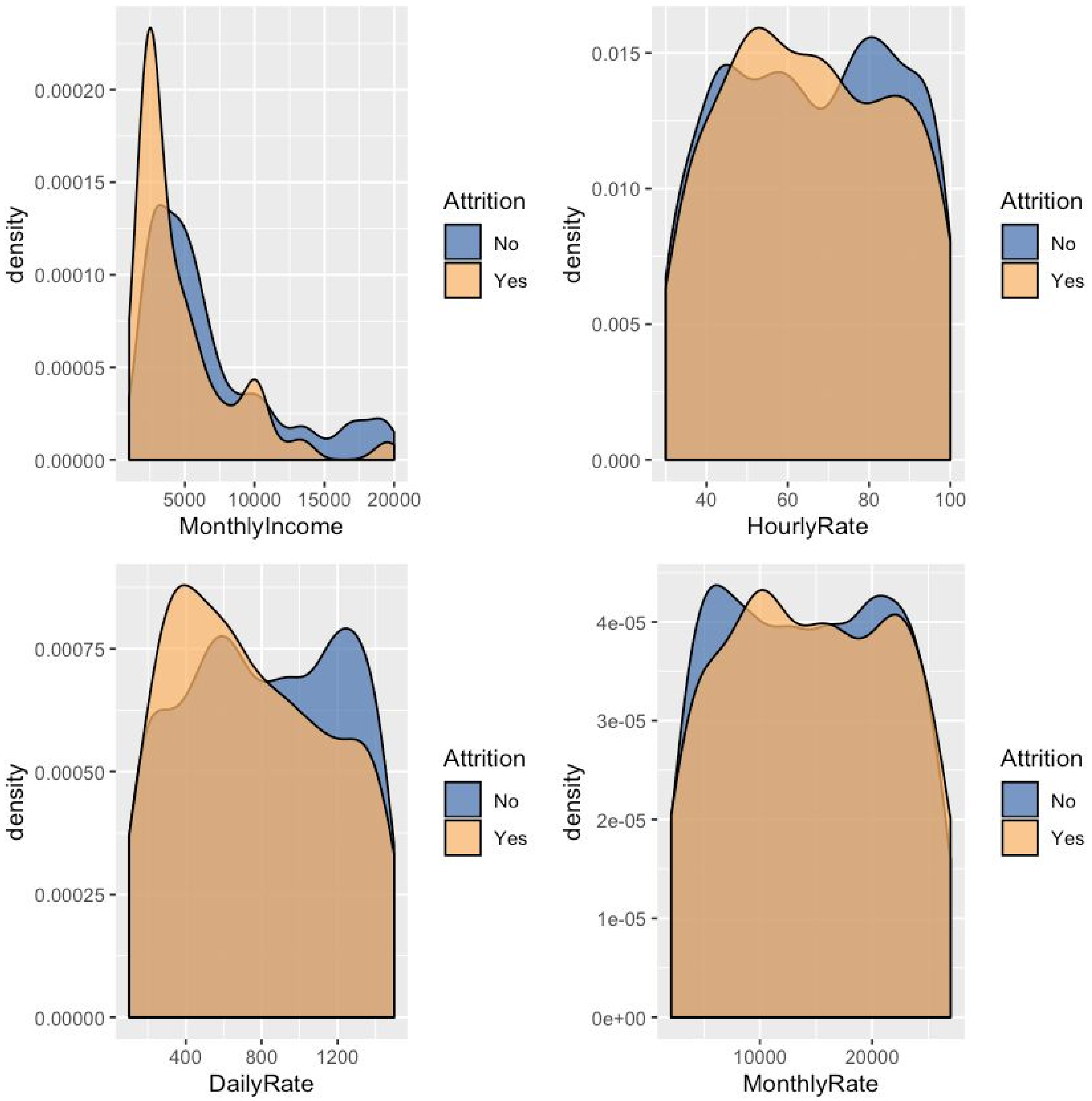
[accuracy] : 0.89115 (+/- 0.00621)

[precision] : 0.82588 (+/- 0.04748)

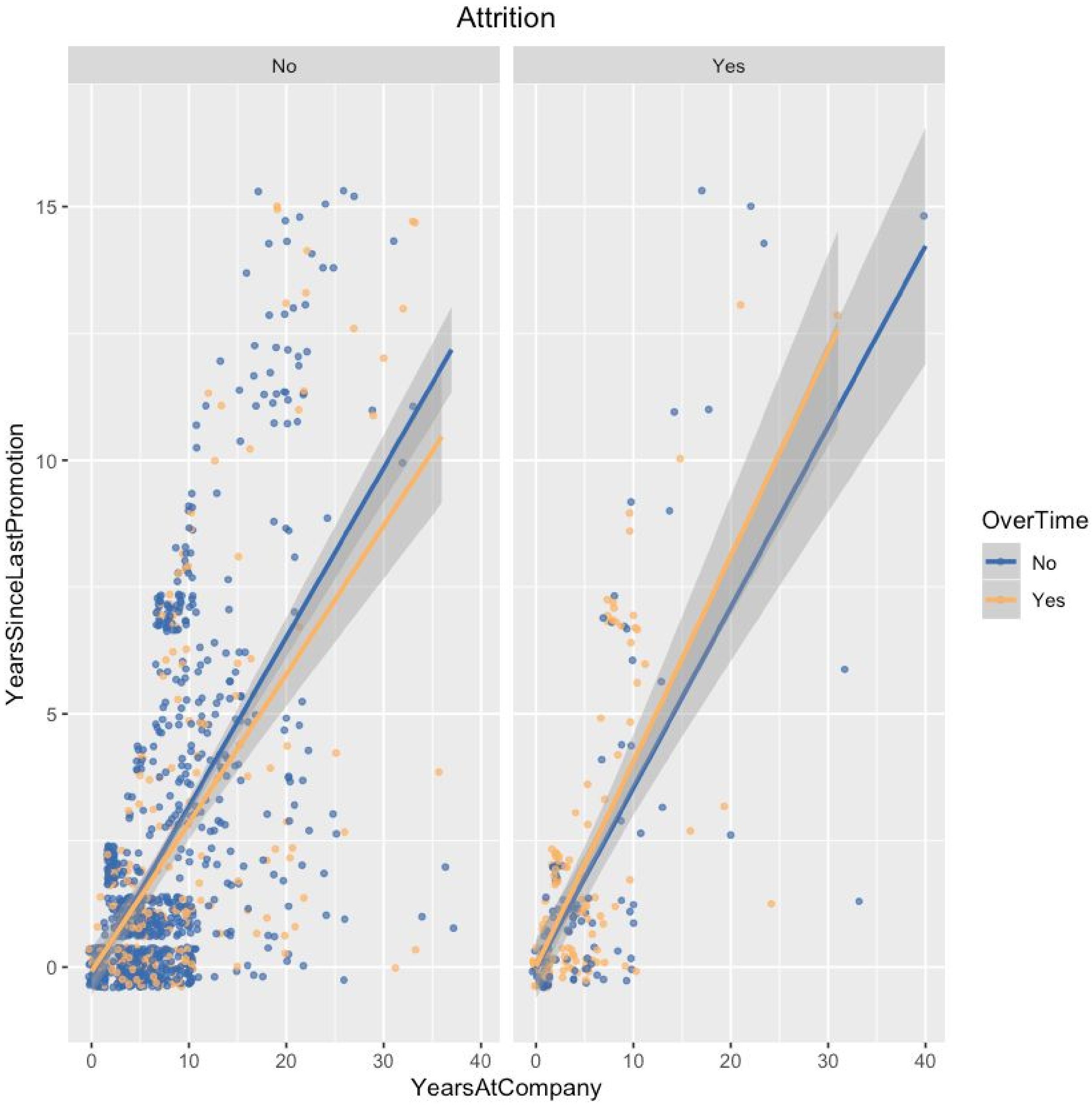
[recall] : 0.41348 (+/- 0.03070)

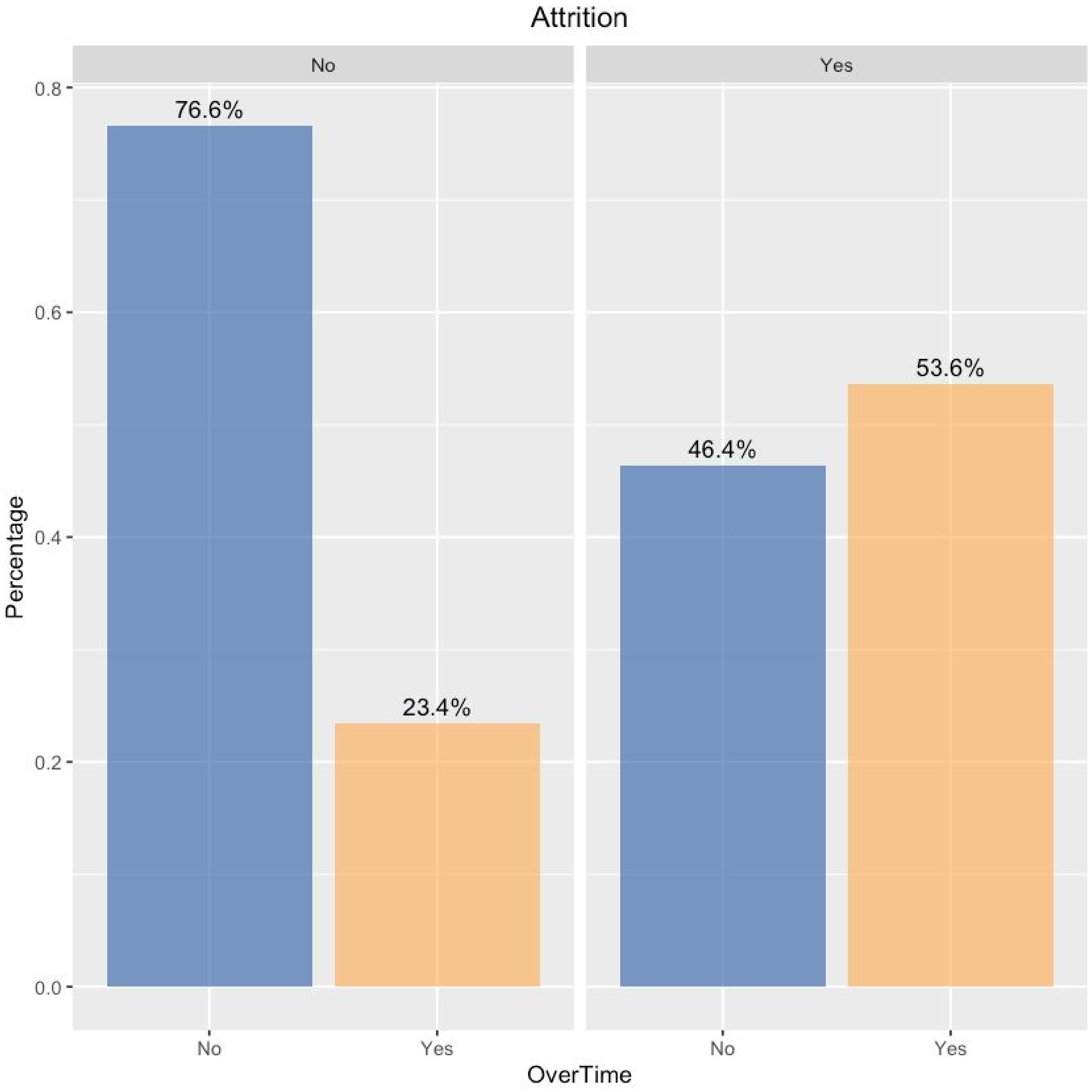
Performing the same analysis in R utilizing ROC curves.

- Creating a quick exploratory data analysis.
- Visualizing the relationship between variables.
- Understand how attrition occurs over time.



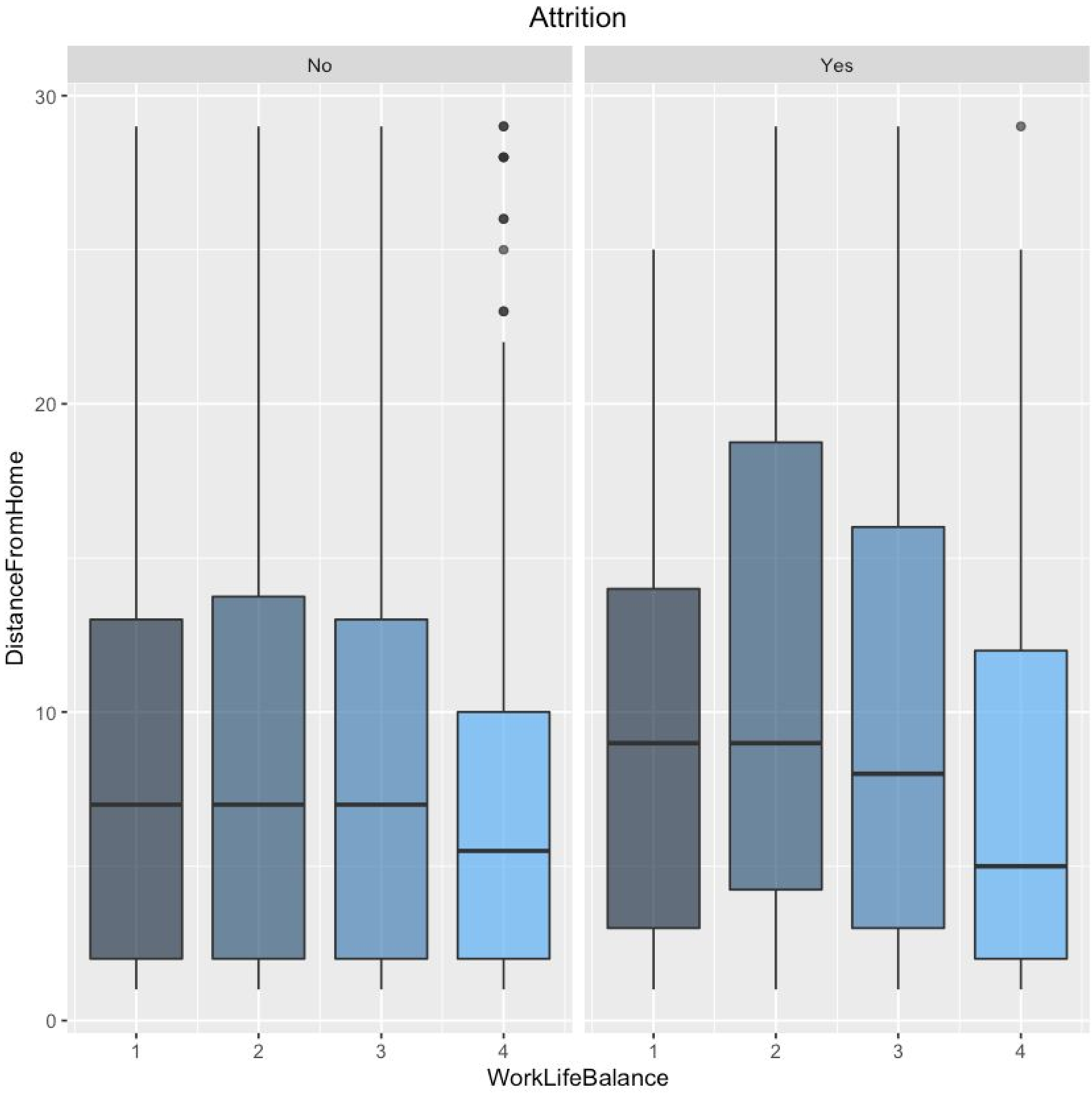
Monthly income shows something interesting in the data

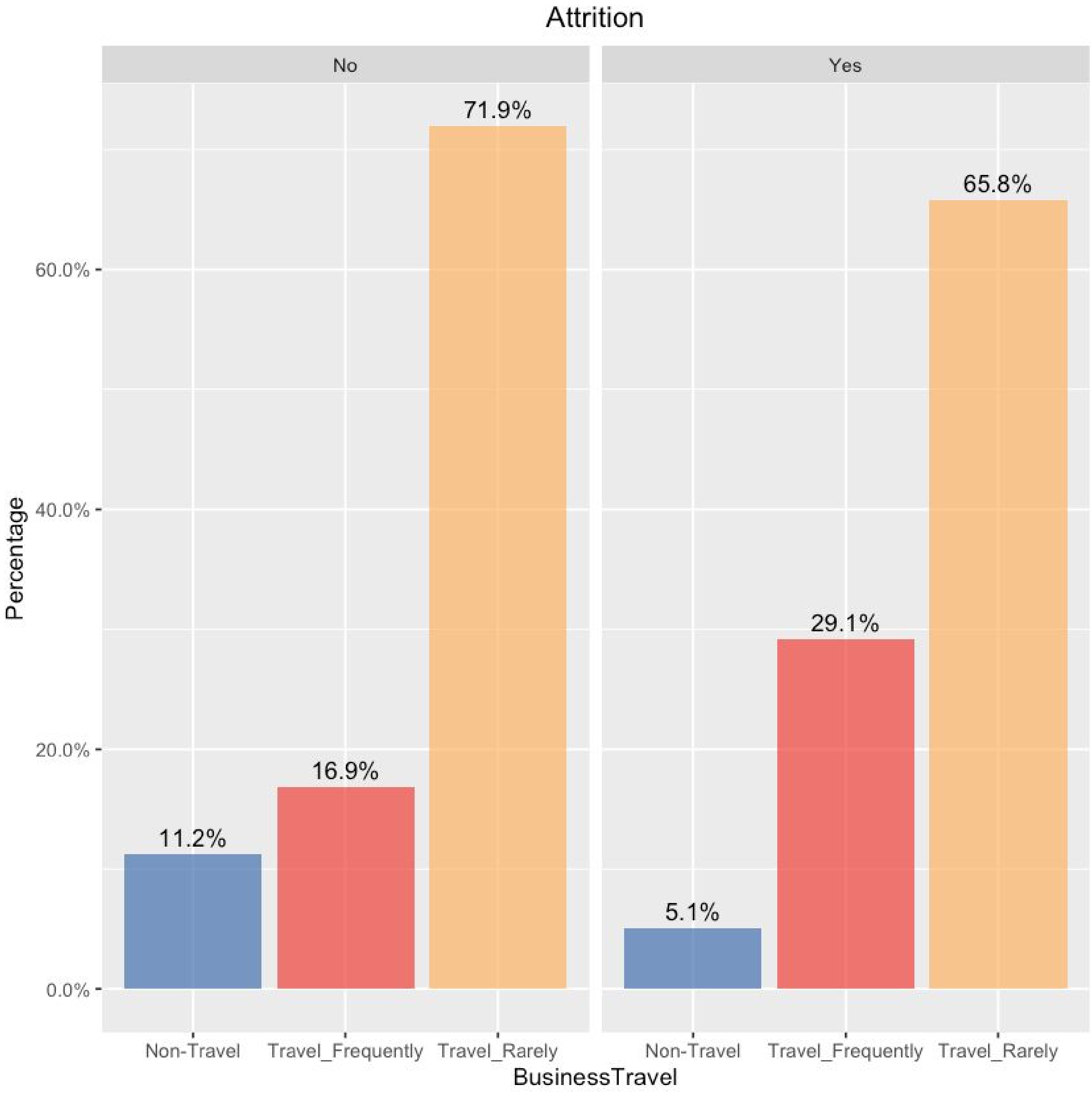

Attrition
No
Yes
YearsSinceLastPromotion
0
5
10
15
YearsAtCompany
0
10
20
30
40
OverTime
No
Yes



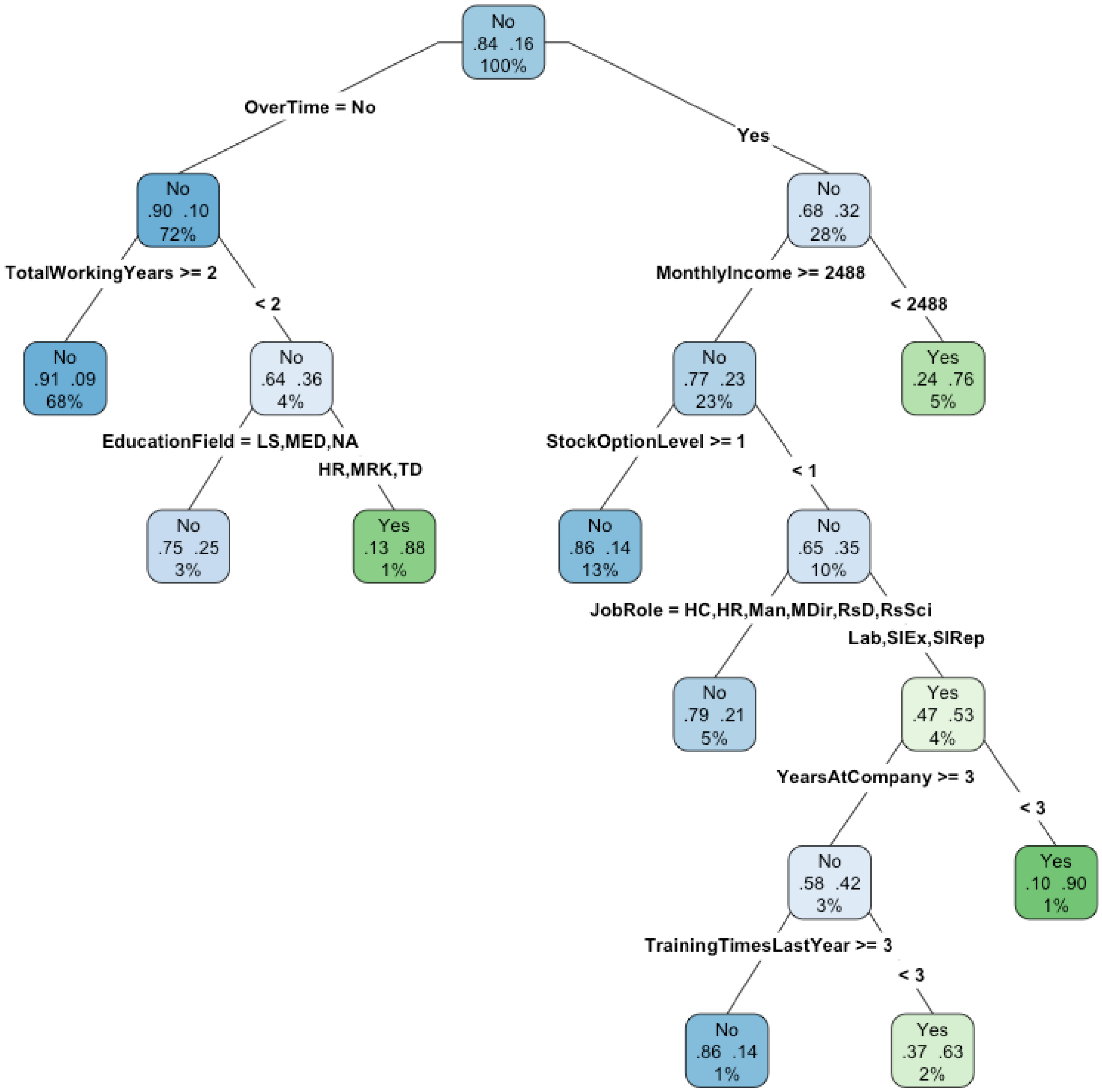
The first graph shows Years at a company in relation to Years since last promotion, and I grouped both variables attrition and overtime. This is an interesting issue, since a high correlation between these two variables (the longer you are in the company, and the probability your less likely to have a promotion so to speak) may mean that people are not really growing within the company. However, since this is a simulated dataset we cannot compare it with some norms outside it of the data set. We can compare certain groups within our dataset, for instance. those who are working overtime and those who are not.

Here we can note two things. Firstly, there is a relatively higher percentage of people working overtime in the group of those who left, and this observation was confirmed by our barchart. Secondly, while things seem to be going in the right direction for the group of people who still work for IBM (higher correlation between years since last promotion and years at company for those who don't work overtime), you can see that the opposite is happening in the other group. It seems that there may be a pattern of people leaving because they are not promoted although they work hard (as promotion is a viable variable in our data). This is only an assumption at this point, since the confidence intervals (gray area around the lines) are getting wider, meaning there is not that much certainty about this, especially at higher values of X and Y (probably due to the lack of data).


Attrition
No
Yes
DistanceFromHome
30
20
10
0
1
2
3
4
1
2
3
4
WorkLifeBalance

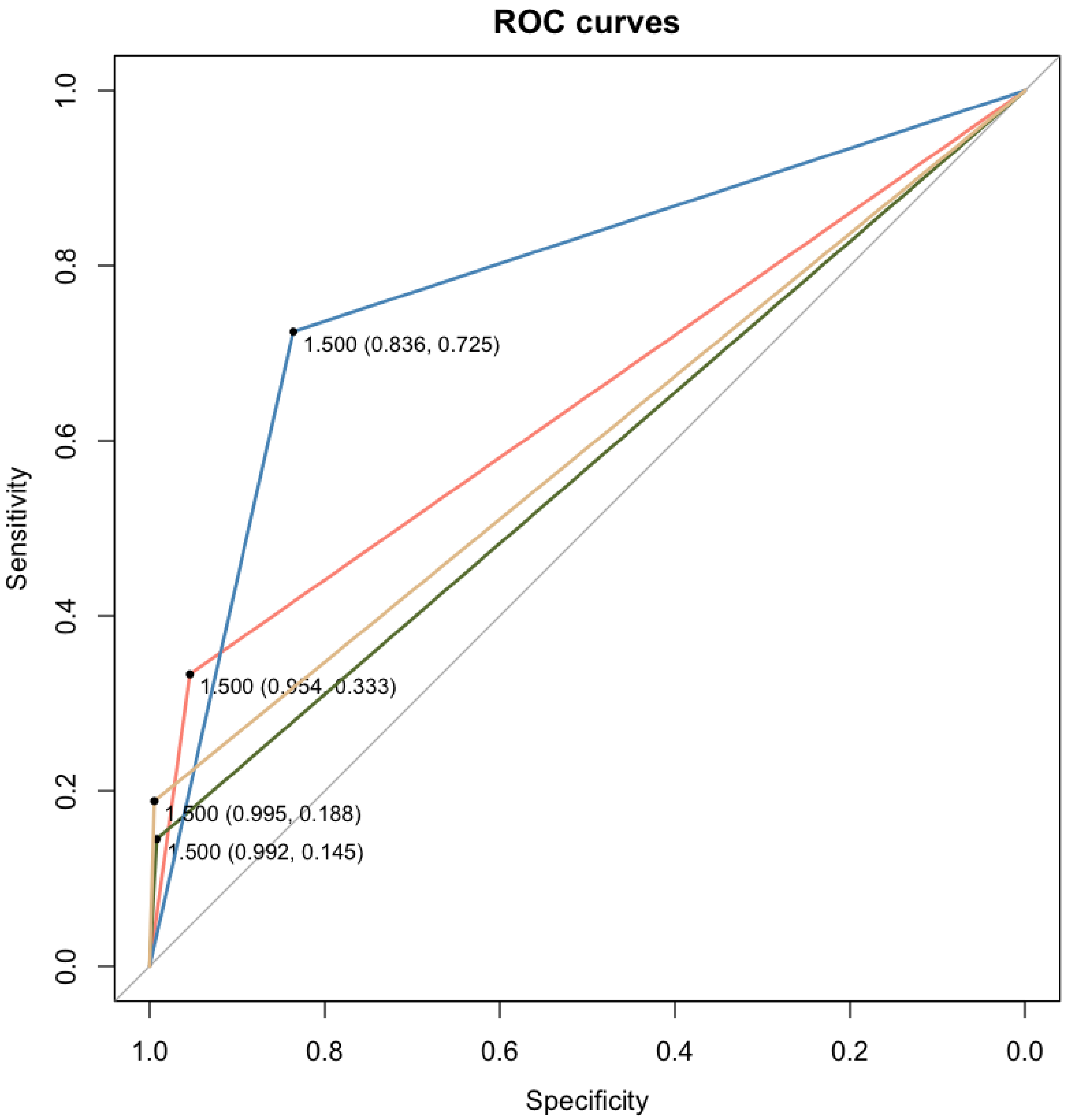

Attrition
No
Yes
71.9%
16.9%
11.2%
65.8%
29.1%
5.1%
60.0%
40.0%
20.0%
0.0%
Percentage
Non-Travel
Travel_Frequently
Travel_Rarely
Non-Travel
Travel_Frequently
Travel_Rarely
BusinessTravel

Modeling (Decision Tree)



You can see that the most important variables seem to be overtime and monthly income - something we have already discerned through our graphic EDA. Remember, the sensitivity of this model is quite low, which is why we would in principle advise against any general interventions on this basis.

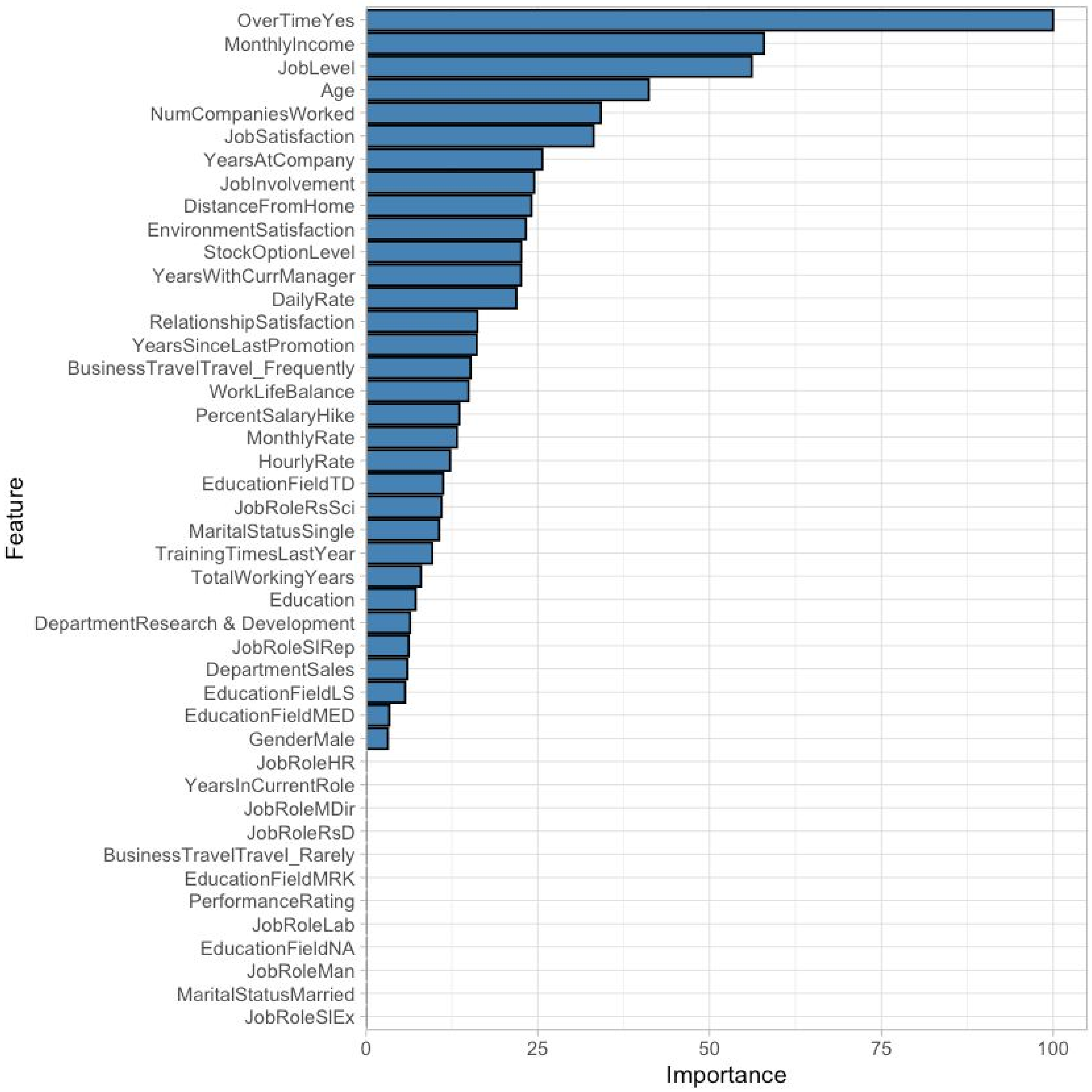
However, we can see that a major percentage of those who left can be relatively reliably identified using the criteria of combined overtime and monthly income. If we consider them jointly, this points to another factor: effort-reward imbalance. This is why it is very useful to plot a decision tree, it makes you see some patterns that you might have forgotten during your EDA and it gives a visual representation of our data.



Making sense of our models & analyses

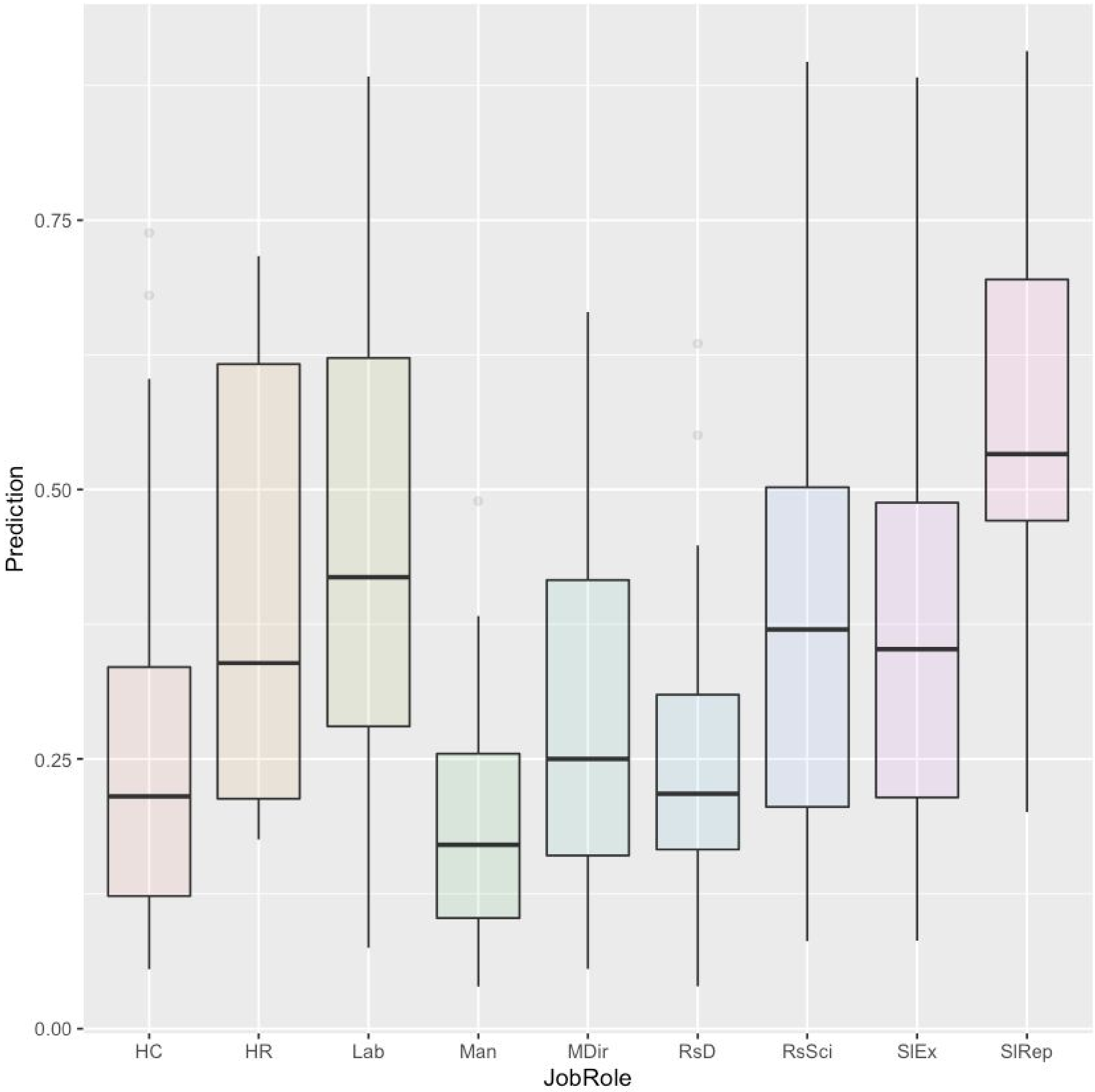
How can we help ourselves understand employee attrition with these findings? Complex algorithms aren't easy to interpret, but there are several ways in which they can be useful:

We can examine the variable importance list, and see which factors in general are helpful in determining the outcome (employee attrition); this can be useful in determining where should we carry out our (HR) audit first, We can use our model to calculate the probabilities of leaving for each and everyone of our new employees; we can also make new variables from these probabilities, determining who one has the highest possibility to leave and has at the same time high performance rating, working several hours, and contributes in a meaningful way to our company. We can then convey this information to our management team who can perhaps assess the situation, and speak to the person in a tactful way. We can evaluate our organisational tree with regards to these probabilities;. we can assess which department or role has the highest probability of leaving, and then focus our efforts there, or do additional analyses on that department/role (either quantitative or through focus groups). Let us plot the variable importance list from our best model.



The top 5 factors that influence the attrition seem to be:

Overtime, Monthly income, Job level, Age, Number of companies worked for. Two of these are already familiar to us from our EDA and decision tree plot - it seems that we should indeed do something about those who work overtime and then leave and those who have a low monthly income (which is probably also linked to the job level).



This graph shows the distribution of various responsibilities that fall under the category “Job Role.” You can see here that being required to perform certain responsibilities leads to a higher probability of affecting attrition. One of those variables is Sales Representatives, as shown in pink on the far right. The probability of this having an affect on attrition is higher than 50%, being the category with the highest percentage out of all the job roles.

Conclusion

After applying various analyses and tests to our data, we can say we have produced the optimal logistic regression model. Our model showed several features that played a significant role in building a more robust model. These variables are Overtime, Monthly Income, Job Level, Age, and Number of Companies Worked For. These variables, more so than the others, influence attrition.

We plotted our 3000 predicted values which resulted in an S-shaped curve which is to be expected for a Logistic Regression model with a large number of samples. The S-Shaped curve shows that our model increases gradually in the beginning, then increases rapidly in the middle of the growth period, then very slowly at the end. The S-shaped curve we have modeled has a threshold of 0.45.

It was very apparent to us from day one when we decided to tackle a Logistic Regression project without having any prior experience that this was going to be a great challenge for each of us. In addition, we technically only learned how to apply the various statistical methods and techniques to build a quality Logistic Regression model with only 3-4 lectures remaining in the semester.

Overall, we are very proud of the work we ultimately put into our project. We are most interested to learn whether our analysis was indeed insightful and could be beneficial to an actual corporate.

Appendix A

Variables used in the dataset:

1. Age represents the age of each employee at IBM
2. Attrition is whether the employee stayed or left the company, either ‘Yes’ being 1 or ‘No’ being 0
3. Business Travel is a categorical variable based on if the employee has to travel as part of their job, with the responses being no travel, travel rarely, or frequent traveling
4. Daily Rate is the numeric amount an employee gets paid
5. Department is a categorical variable that represents which areas the employees work in, such as, Research and Development, Sales and Human Resources
6. Distance from Home is the number of miles the person has to travel to work
7. Education is a categorical variable measuring the level of education the employee has reached represented by values 1 to 5 with 1 being ‘Below College,’ 2 being ‘College,’ 3 being ‘Bachelor’s Degree,’ 4 being ‘Master’s Degree,’ 5 being ‘Doctor’
8. Education Field is a categorical variable based on major

9. Employee Count is a numeric variable
10. Employee Number is an identification number that is specific to each employee
11. Environment Satisfaction is a categorical variable on the scale of how much employees like their working environment with 1 being ‘Low,’ 2 being ‘Medium,’ 3 being ‘High’ and 4 being ‘Very High’
12. Gender of the employee with results of ‘Male’ and ‘Female’
13. Hourly Rate is an integer value of how much an employee makes per hour
14. Job Involvement is a categorical variable representing how actively involved an employee is in their work with 1 being ‘Low,’ 2 being ‘Medium,’ 3 being ‘High’ and 4 being ‘Very High’
15. Job Level is categorical with numbers 1 to 5 representing the level an employee is at where 1 is entry level and 5 is a boss
16. Job Role is a categorical variable describing with the name of the position
17. Job Satisfaction is a categorical variable with values 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
18. Marital Status is a categorical variable of the relationship status of employees responses of ‘Divorced,’ ‘Married,’ or ‘Single’
19. Monthly Income measures the salary an employee earns per month
20. Monthly Rate is numerical
21. NumCompaniesWorked measures the number of companies an employee has worked at in their career
22. Over 18 is a categorical variable measuring whether or not the employee is an adult responses of ‘Yes’ or ‘No’
23. OverTime is a categorical variable recording if employees work overtime
24. PercentSalaryHike is a numerical variable measuring the difference in salary between 2017 and 2018
25. Performance Rating is a categorical variable with values 1 representing 'Low,' 2 representing 'Good,' 3 representing 'Excellent' and 4 representing 'Outstanding'
26. Relationship Satisfaction is a categorical variable with values 1 being 'Low,' 2 being 'Medium, ' 3 'High' 4 'Very High'
27. Standard Hours measures the hours an employee typically works
28. StockOptionLevel is a categorical variable on whether or not an employee owned company stock of IBM
29. TotalWorkingHours is time employee work per week
30. TrainingTimesLastYear records the time spent in training sessions employees attended
31. WorkLifeBalance is a categorical variable with values 1 being 'Bad,' 2 representing 'Good,' 3 representing 'Better' and 4 representing 'Best'
32. Years at Company records how long an employee has worked at IBM
33. Years in Current Role records how long an employee has been at the same job level
34. Years Since Last Promotion measures the time in years
35. Years With Current Manager is time the employee has worked with the same boss

Residuals vs Fitted shows the graph of residual values vs the fitted values

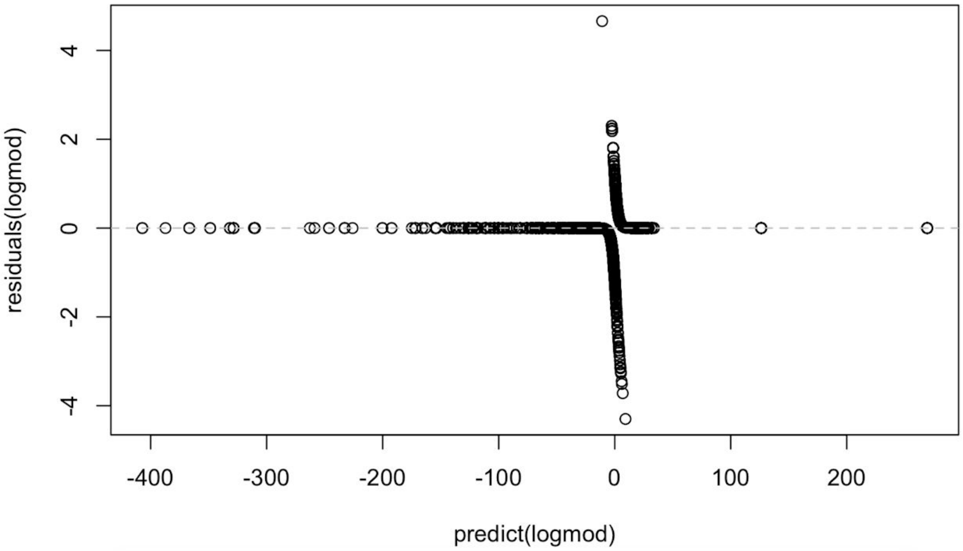
Appendix B

Packages/Libraries used throughout this project:

- Dplyr

- Ggplot2
- Pastecs
- Psych
- Amelia
- Mlbench
- Corrplot
- Caret
- Readr
- gridExtra
- Grid
- Lattice
- Leaps
- Rpart
- Rpart.plot
- RandomForest
- Gbm
- Survival
- pROC
- DMwR
- scales

Appendix C


residuals(logmod)
predict(logmod)