What Makes an Employee quit?

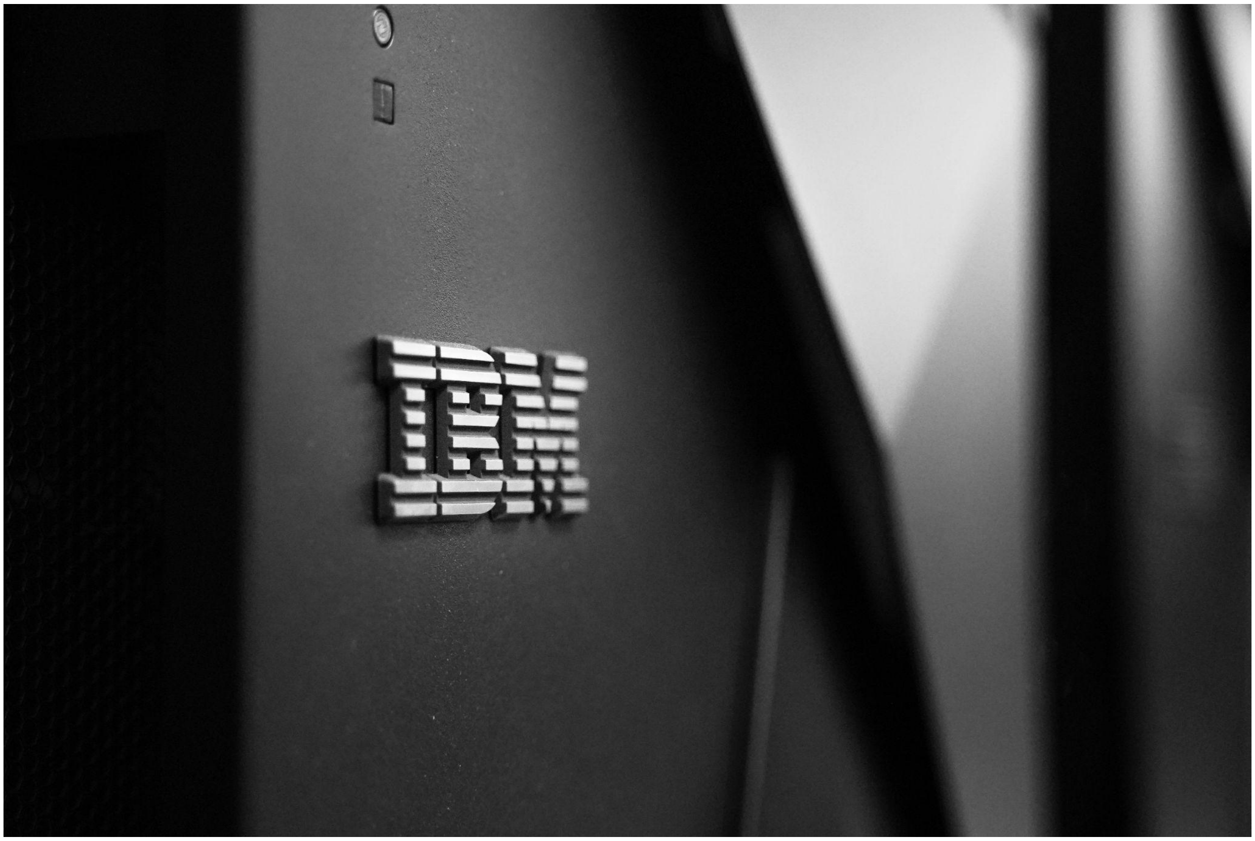


Photo by Carson Masterson on Unsplash

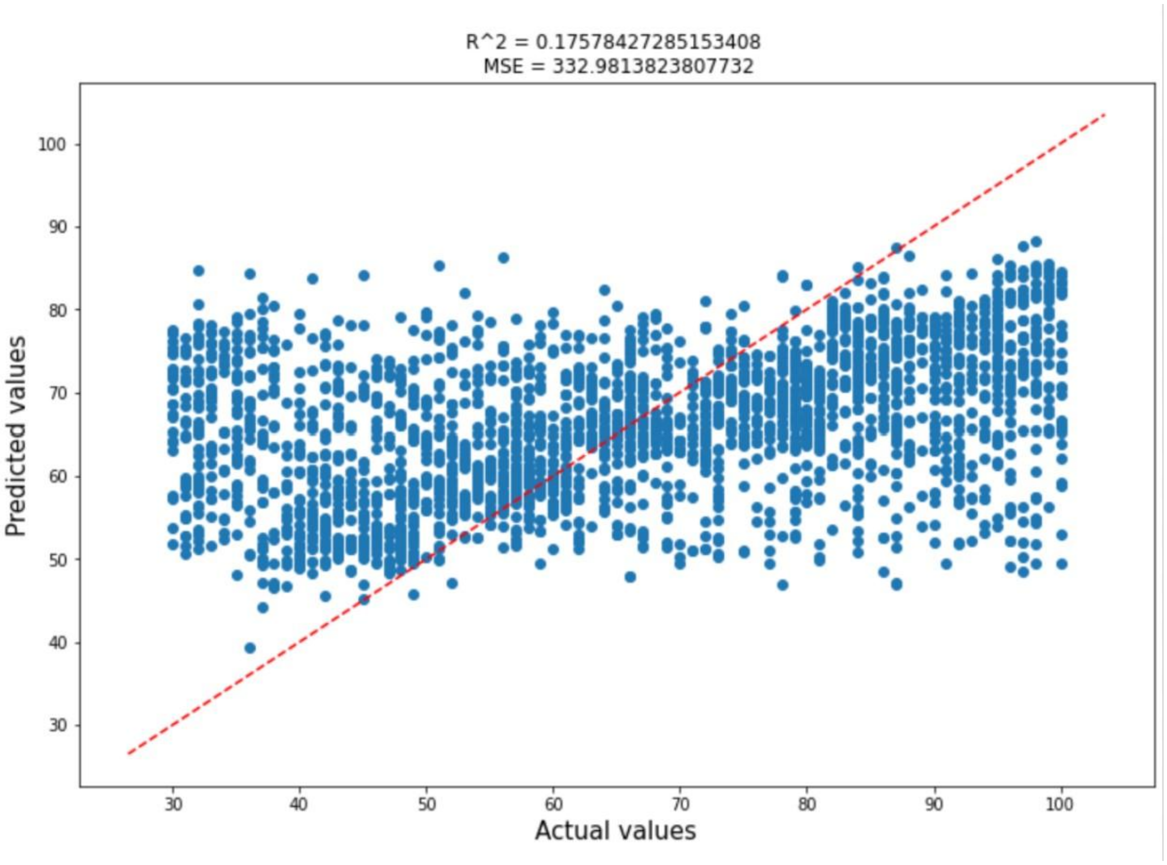
By: Anne Cuzeau, Eric Vandament, Andrew Nguyen

The objective of this project was to analyze employee attrition, which is known to be a costly issue for organizations as it involves hiring and training new employees, often at a cost of 3 or 4 times the position's salary. With this in mind, we set out to answer the question “What makes an employee likely to quit their job?”. To answer this question and build a model that would potentially be helpful to an HR professional, we found a hypothetical dataset created by IBM data scientists on Kaggle specifically targeting employee attrition.

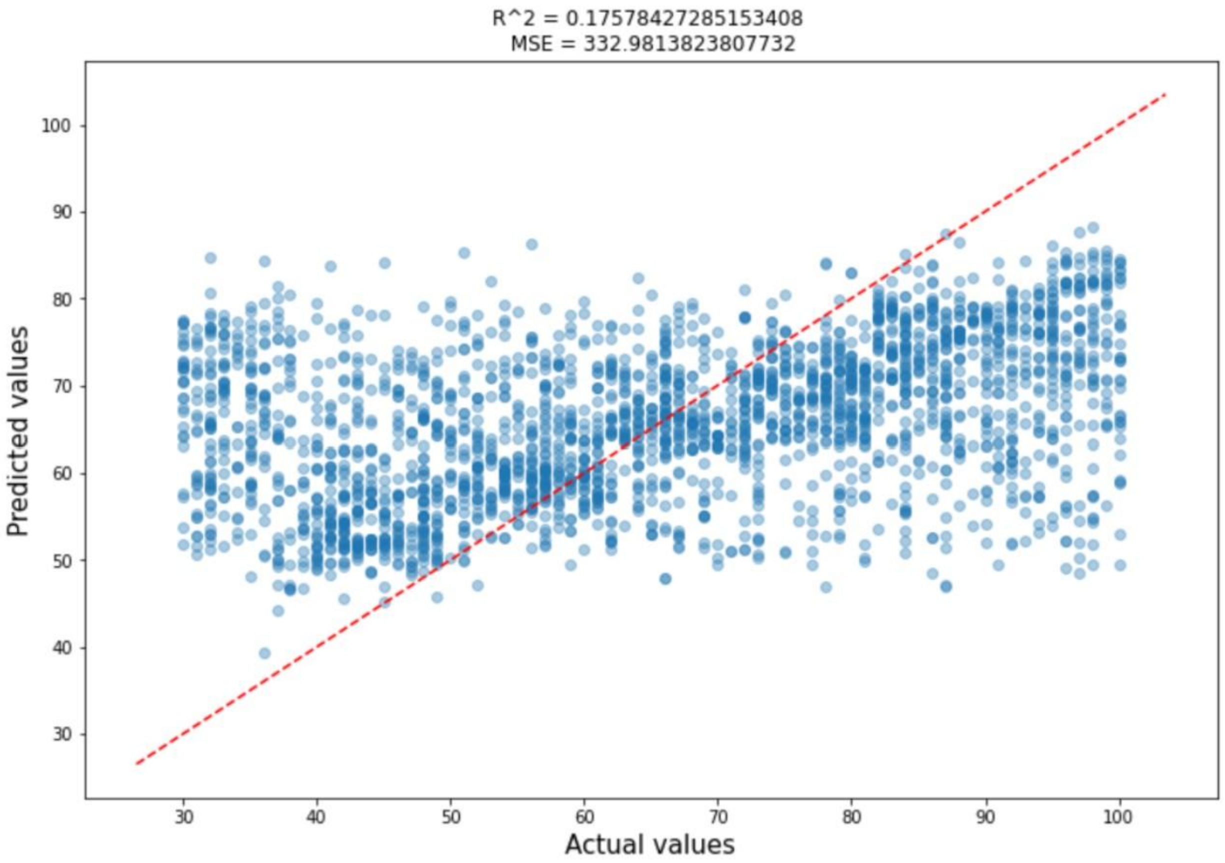
The variables within our dataset can be classified into four categories. The first category is Job Related, which includes an employee’s job role, department, job level, and travel frequency. The second category is Demographics, which comprises an employee’s age, gender, relationship status, educational background, and the distance between an employee's home and office. The third category includes career-related variables, such as the number of companies an employee has previously worked for, years with the current company, years in their current role, and years since their last promotion. Finally, the last category pertains to pay, which includes daily rate, hourly rate, monthly income, and monthly rate.

Pay rate

When considering why an individual quits their job, some people's initial thought is often “Were they getting paid enough?”. Hence, to understand attrition, we first decided to focus on hourly rate. Initially, when putting hourly rate to other variables in our dataset, no variable showed a strong correlation, with the highest correlation coefficient being around 0.4

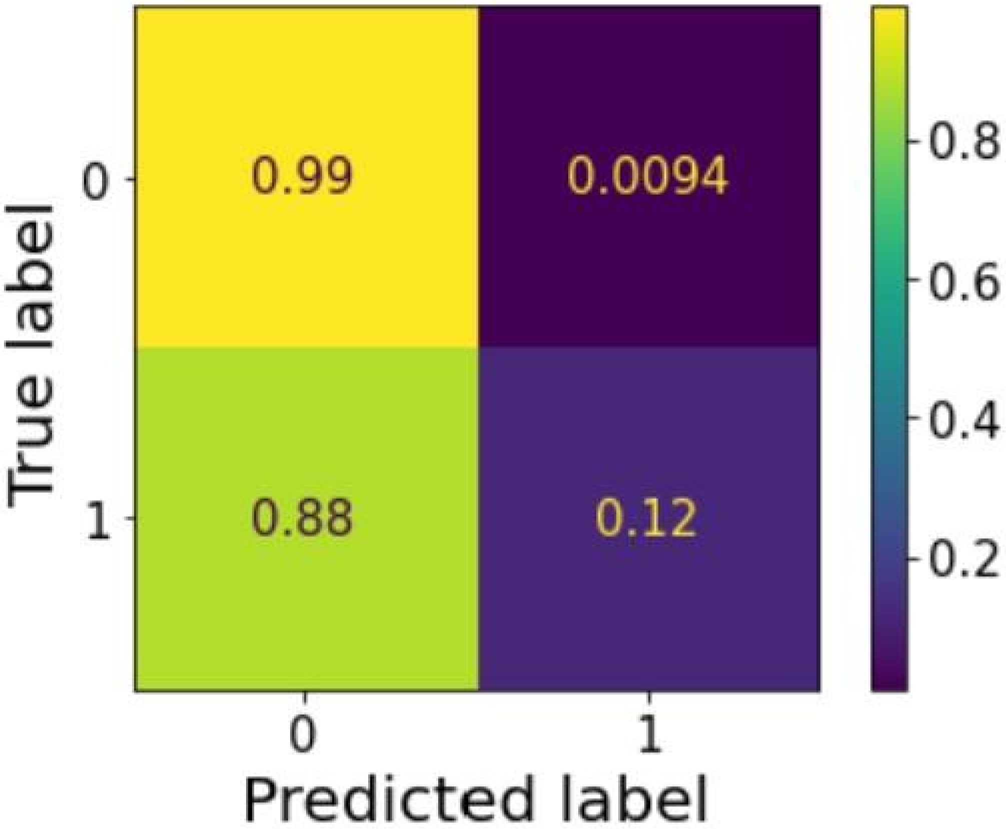


After examining the model above (with all variables being used), we observed that hourly rate was not being predicted accurately. Indeed, the model predicted values below 20 (not in our dataset) for the actual value without showing a strong relationship, which was not very helpful. Moreover, both the R^2 and MSE values were also pretty poor, proving that this model was not performing well.

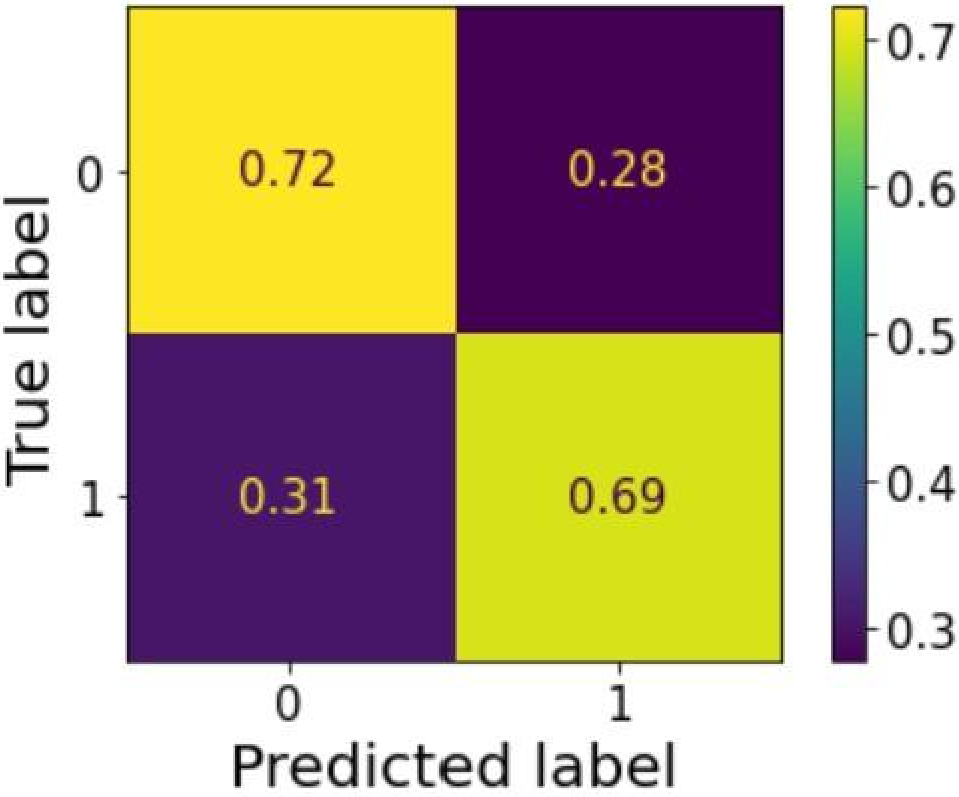


This graph is very similar to the one above, except for the lower alpha value. The lower alpha value indicates that there is a slight upwards trend, though our model was still not making accurate predictions: if we were to add a line based on the slope of the upward trend shown in the graph, we would observe that some of the values would be accurately predicted, but any value below and above this line would be poorly predicted.

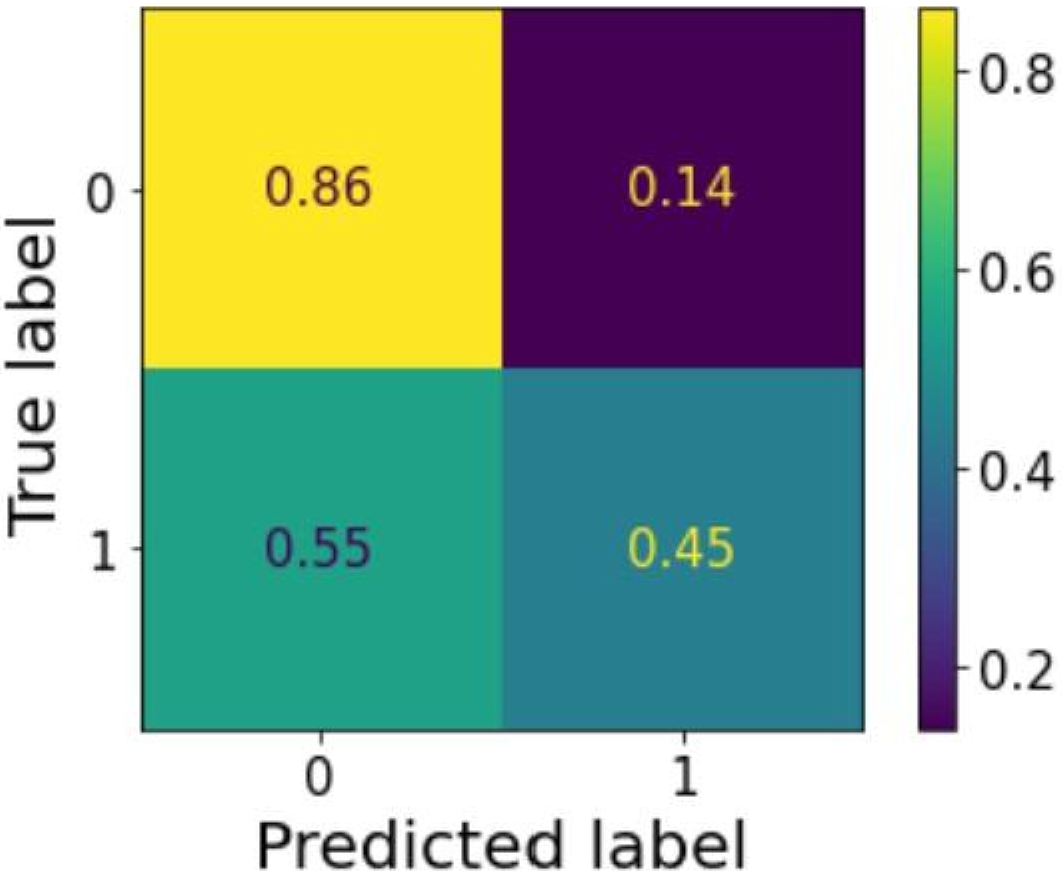
Attrition

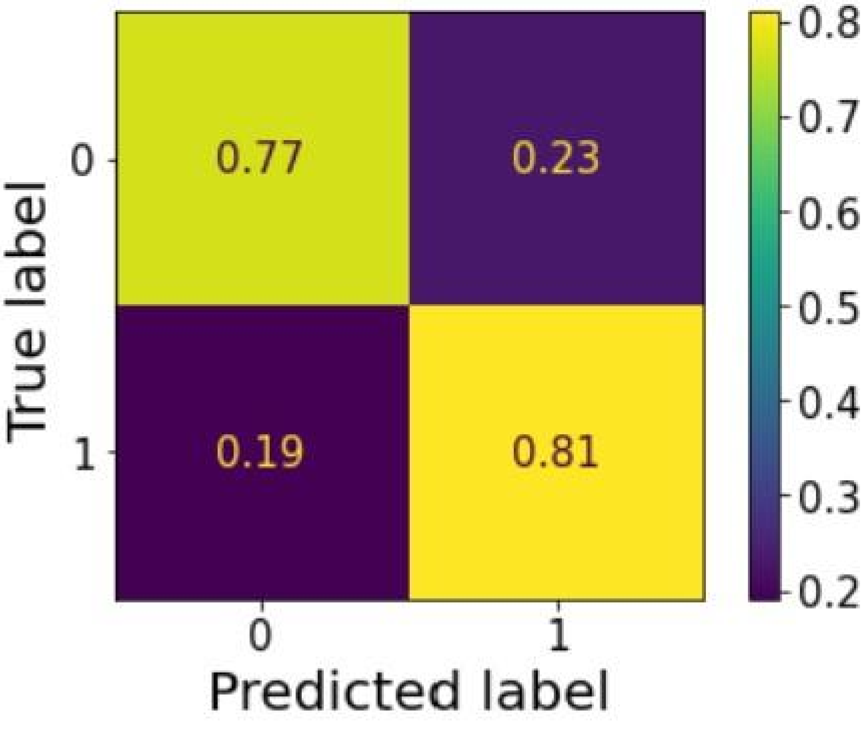


This first confusion matrix is the result of our baseline model, a logistic regression. This model used all variables and aimed to predict attrition (1 meaning employees left). The model's performance for employees who stayed was deceptive: while it appeared to be highly accurate for predicting employees who will remain in the company, it was terrible at predicting those who left, which was our original goal. The poor performance for predicting attrition can be explained by the fact that most employees in the dataset actually stayed. This imbalance made it difficult for the model to learn and predict the minority class (i.e., employees who left), and the logistic regression just predicted that almost all employees would stay.



To address this issue, we decided to balance our data in order to have an equal number of employees who stayed and who left. After this adjustment, we ran another logistic regression model which performed substantially better. The confusion matrix for this model is shown above. As you can see, this new model showed a slightly inferior performance for predicting employees who will stay. This was acceptable for us since our main goal was to predict attrition, not employees who stayed, and we saw a 57% increase in correct prediction for this category (1, employee who quit). This represented considerable progress in our journey to build an accurate attrition prediction model. However, we still aimed at improving our results further by exploring alternative models.


0.86
0.14
0.55
0.45
0
1
True label
0
1
Predicted label
0.8
0.6
0.4
0.2

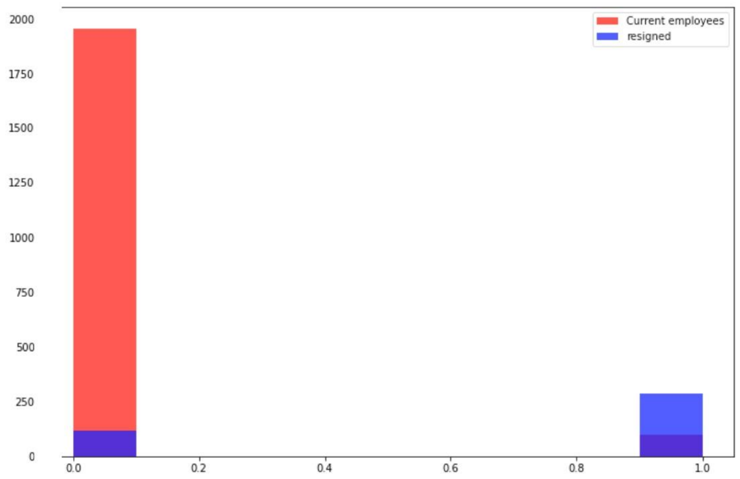

0.77
0.23
0.19
0.81
0
1
True label
0
1
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2

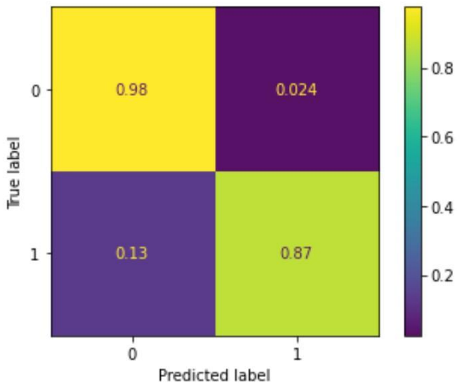
As stated previously, we wanted to see if other types of models would help us get better results when trying to predict employee attrition. Therefore, we experimented with a decision tree model. Initially, we created a baseline decision tree fitted with a maximum depth of 5, a minimum sample split of 10, and a minimum sample leaf of 5. The results of this model are shown in the matrix above. As you can see, it yielded better results than the logistic regression model without balancing for predicting if an employee will leave, but was still inferior to the logistic regression model with balanced data.

Subsequently, we decided to create another decision tree with balanced data and performed hyperparameter tuning. Based on our tuning, we concluded that the best parameters would be a maximum depth of 30, a minimum sample split of 5, and a minimum sample leaf of 2. With these new parameters we were able to get the second confusion matrixThe second confusion matrix shows the output of this optimized model. While there was a 9% reduction in correctly predicting employees who stayed compared to the first decision tree, there was a 36% increase in predicting those who left correctly, resulting in an overall accuracy of 81%. This model proved to be our strongest model to date, but we were still curious about the performance of other models.

Nearest Neighbor analysis

The next step for the team was to build a nearest neighbor model: flexible and powerful, we thought it could be a good fit for our goal. To optimize our model's performance, we used gridSearchCV, which allowed us to test any number of neighbors between 1 and 20 and choose the best metric between hamming, euclidean, and jaccard. We found that the most effective parameters were a neighbor of 1 and hamming's distance metric. In addition to fine-tuning our model, we combined high attrition job roles/levels and high attrition demographics and ran the model with stratified and balanced data. The nearest neighbor model proved to be able to accurately predict employee's turn-over with an impressive 87% accuracy. However, analyzing the 13% of employees who were predicted to stay but actually left revealed that our model was struggling to categorize employees working in the R&D department: 61% of wrong predictions involved R&D professionals. This was not surprising as the R&D department as very diverse job levels and pay scales. Indeed, it includes the second-highest attrition job role (Lab Technician) and low-attrition, high-paying positions. In addition, the model incorrectly predicted that executives in the Sales department would quit, swayed by the high attrition sales representative position.

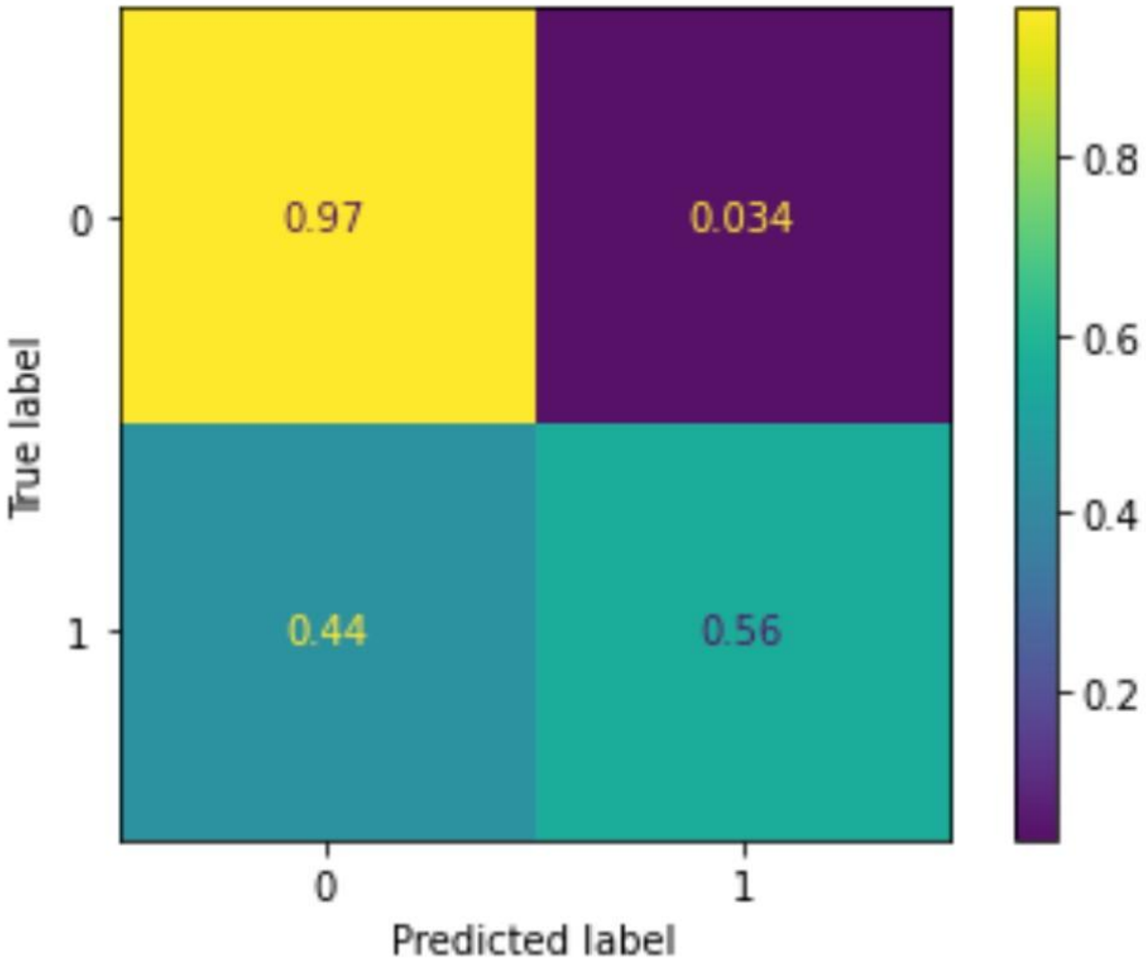




XGBoost

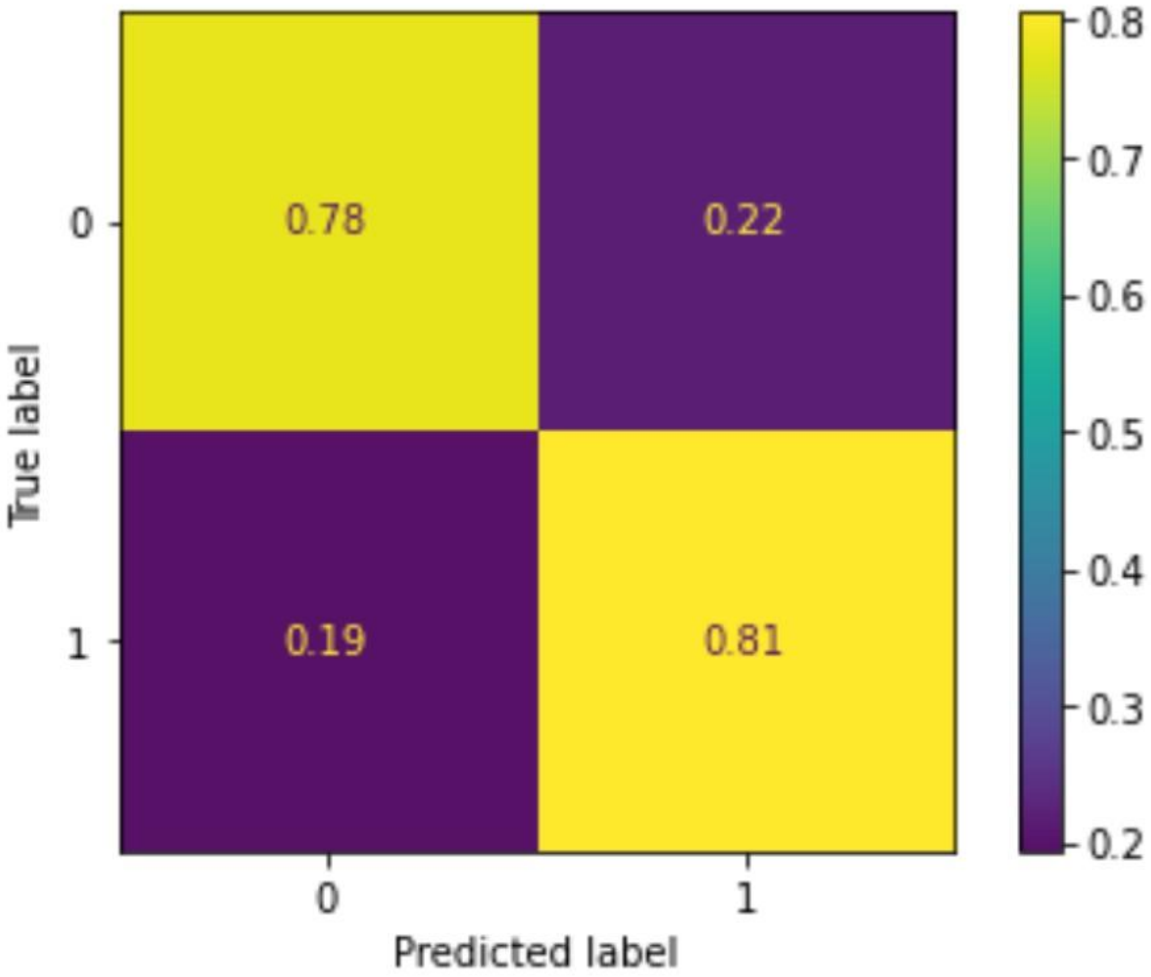
Stratified data

Finally, since the decision tree we built earlier proved to be pretty effective, we set up to build an XGBoost model. First, we simply fed this model stratified data. The resulting model demonstrated strong predictive capabilities in determining whether employees would stay in their current position, with an accuracy rate of 98%. However, it was not able to predict if employees would quit: with a 56% accuracy, this was only slightly better than guessing randomly. The R&D department remained a challenging area for the model, with 55.9% of incorrect predictions being attributed to this specific department. However (and unlike some of our previous models), this model was not predicting that most employees would quit, which was an encouraging sign.



Balanced data

Since the first XGBoost model showed some encouraging signs, we decided to train a second XGBoost model on balanced data (using the same variables as the model above). This model showed a significant improvement in predicting employee attrition, with a higher accuracy rate (81%) in determining whether employees would quit. However, it is worth noting that the dataset used to train the model was relatively small, consisting of only 1,875 rows. Despite the improvement in accuracy, the model still encountered similar challenges with the R&D department, with this department being the highest source of wrong predictions.



Conclusion

So after building and investigating many different models, we still needed to answer our original question: “What makes an employee likely to quit their job?”. Our findings suggest that the majority of employees who left had job levels of 1 and 2 (reserved for entry-level or recently hired employees). Additionally, age was a significant factor in all our models, with our dataset showing that people younger in age are more likely to leave their jobs. Interestingly, 96% of sales representatives with a job level of 1 are at high risk of leaving, which is why so many of our models wrongly predicted that sales executives (higher level, fewer employees) would quit. Moreover, sales representatives and laboratory technicians account for 46.33% of employees who quit, making them the primary areas of concern for businesses seeking to retain employees.