

# **Employee Attrition Predictor: Executive Summary**

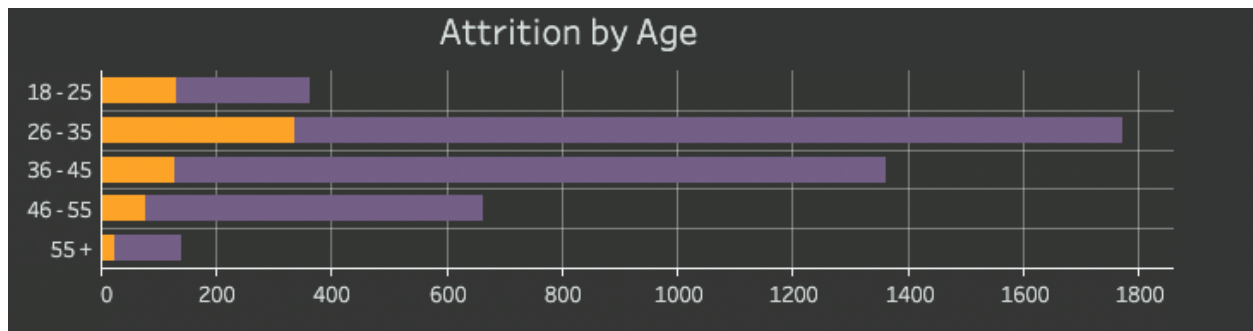
## **Overview**

This project consists of a full-stack web application which incorporates a machine learning model to predict employee attrition. A user may interact with multiple pages, allowing them to see the data from a variety of angles. A Plotly page graphs attrition by user-defined age groups; a SQLite database page allows a user to query the database by attrition, gender, and age ranges; a Tableau page gives a user an interactive dashboard on which various filters can be applied and the data immediately visualized; finally, a prediction page allows a user to input parameters which are run through a balanced random forest machine learning model and the attrition likelihood is displayed.

## **Summary**

Data was cleaned and encoded for logistic regression analysis. After running a number of machine learning models, we chose the supervised balanced random forest model because it resulted in the highest accuracy of classification (95 %). This model maintained a high-level of accuracy (93 %) even when pared down to fewer input features. We chose 10 features based on their usefulness to contribute to predicting the attrition classification target. These features were: age, job distance from home, total working years, monthly income, years at the company, years with current manager, number of other companies worked, percent salary hike (previous year), job satisfaction, and environment satisfaction.

## Results



*Income was the second most important factor in determining if someone would leave their job or not. Significant attrition rates were seen within the 26-35 year age group, as well as among those with lower monthly incomes.*

## Limitations and Further Study

The limitations of the predictions garnered based on this data are substantial, albeit not such that they're totally worthless. The dataset is for a single year: 2015. While the dataset size of over 4,000 rows is good for prediction, because it only covers a single year it is inherently limited in its predictive power. For example, based on the prediction output, which is classified as attrition or not and given as a percentage, does this output mean the employee is likely to leave this year? Next year? Because the data is drawn from a single year it's not clear whether the employees leave immediately or the following year. Multiple years of attrition data would clarify trends and patterns over time to increase predictive granularity and power. Currently however, the predictive features allow an important insight into the reasons employees leave and what conversations management needs to be having if such attrition is to be slowed. Further, it would be even more useful for employers if they could see the average amount of time an employee stays with their company given these factors, which would require multiple years of data. This would allow for companies to be more proactive with their hiring and retention efforts if the results were timebound.