KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

XGBoost Unleashed: A Data-Driven Approach to Predicting Employee Churn

James Down Graduation: December 2023
Faculty Advisor: Dr. Paul Johnson

INTRODUCTION

Employee turnover is a costly endeavor for companies.

- Wide range between 30% and 500% of employee salary.
- Includes posting, interviewing, hiring and training costs.

Identifying the key reasons for employee departures within an organization can equip them with the knowledge needed to address voluntary turnover.

- Reduce turnover costs if they can retain employees
- Take preventative steps for identified at risk employees
- Explore influential factors and reduce their negative effect on employee turnover

Utilizing classification models can uncover the essential variables driving voluntary employee turnover and offer predictions for future turnover.

- Random Forest
  - Collection of Decision Trees, that are each slightly different, that make individual decisions which are then averaged to make a final decision.
- XGBoost
  - Starts with a simple decision tree and makes predictions. Then it checks where it made mistakes and pays more attention to those mistakes, trying to correct them in the next tree.

METHODS

Fictional data was provided by IBM on Kaggle.

- 1470 Observations
- 35 Variables
  - Target Variable: Attrition
- Performed Data Cleaning
  - Removed Department and Education Field due to Linear Dependency
- Split data into training and test sets.
  - 80% training, 20% test

Created XGBoost and Random Forest Models.

- 500 trees grown in Random Forest model
- 50-150 iterative trees grown in XGBoost model
- Tuned hyperparameters to increase accuracy/kappa
- Examined area under the Receiver Operating Characteristic (ROC) curve to determine best model
- Extrapolated important variables

All analysis were completed in R Studio using the caret, ROCR, ranger, randomforest, xgboost, mlbench, DiagrammeR packages.

Figure 2. Attrition Split

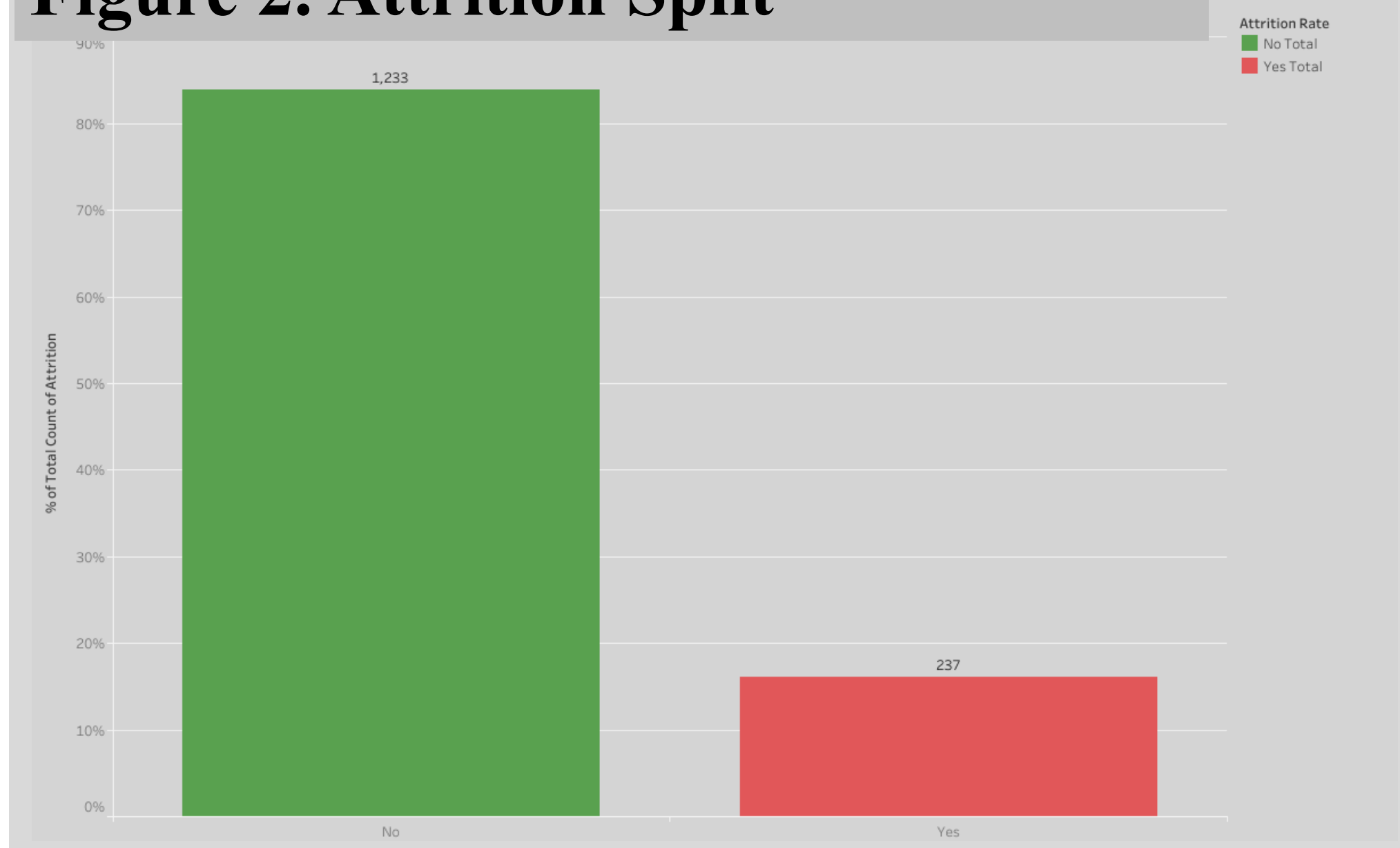


Figure 2. Overtime Attrition Split

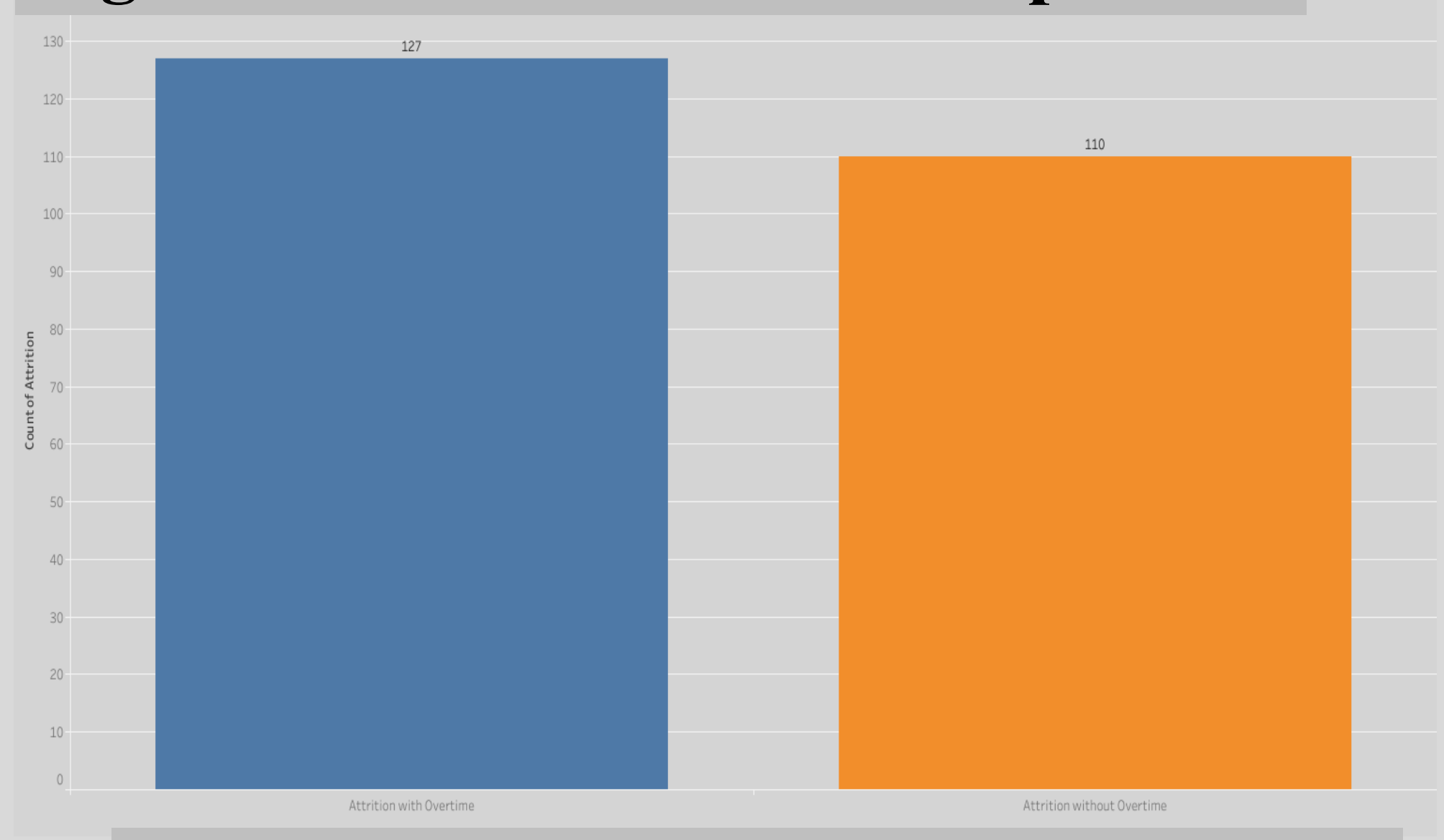


Figure 3. Sample of XGBoost Iterations

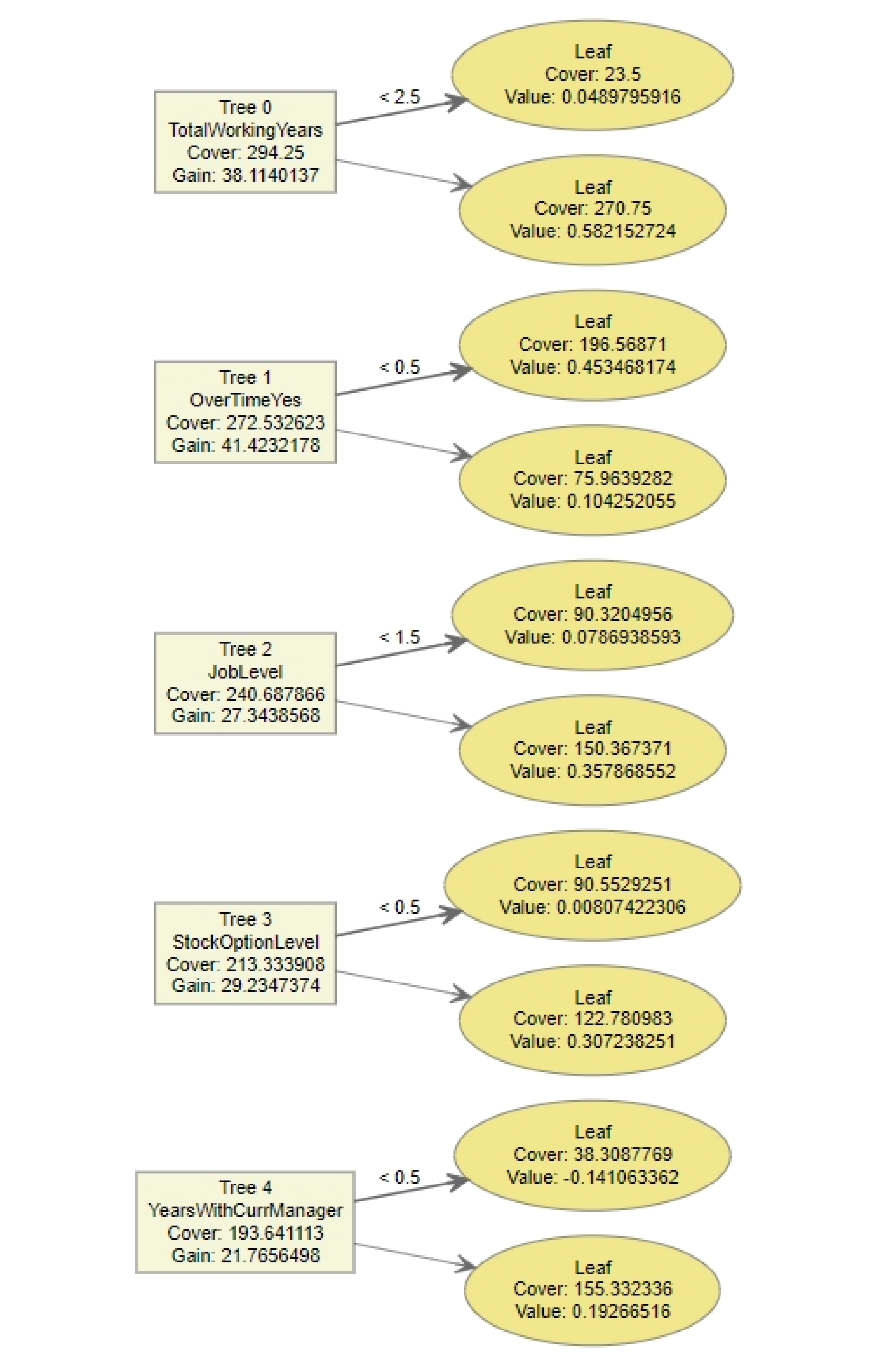


Figure 4. Random Forest ROC Curve

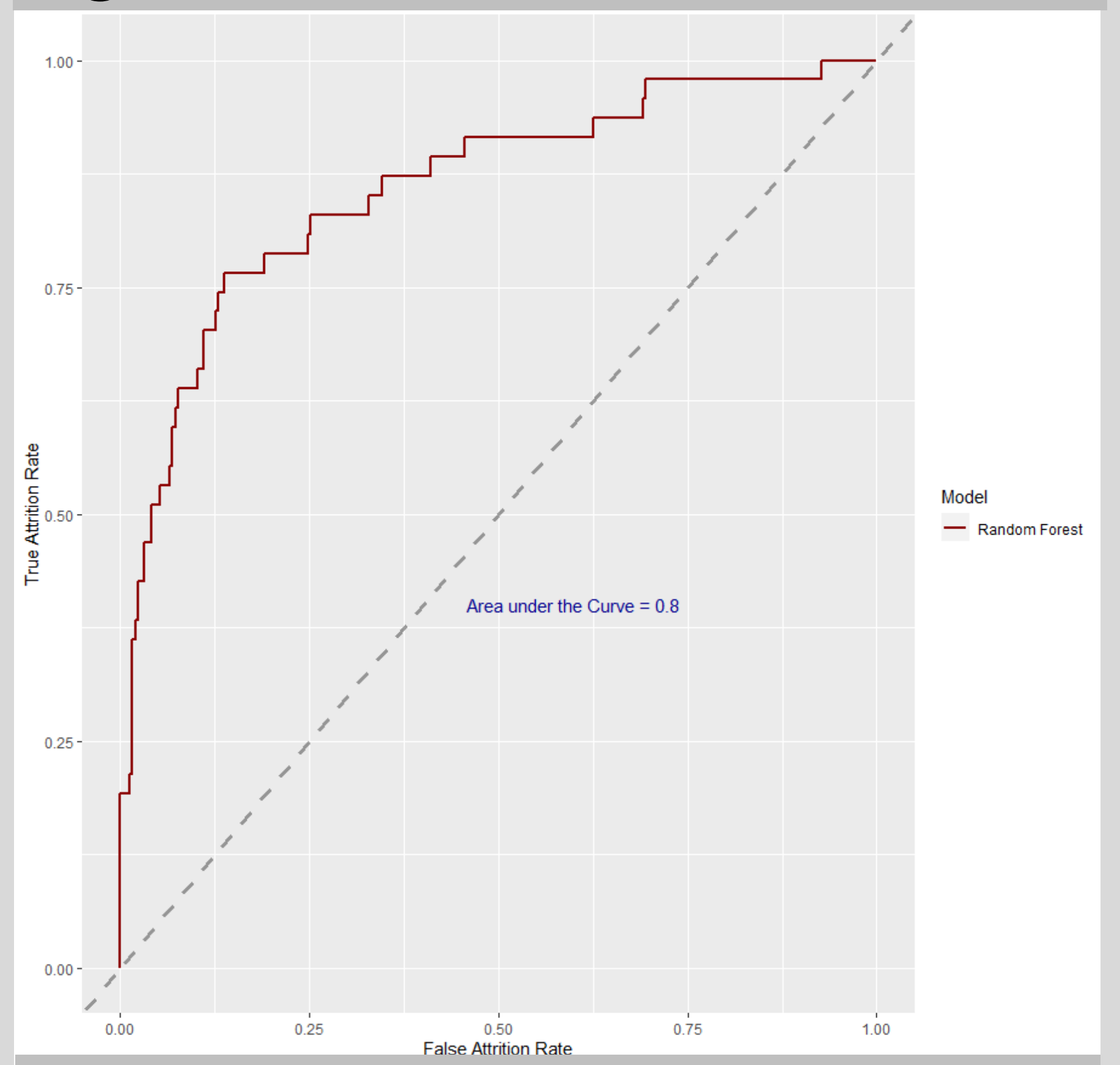


Figure 5. XGBoost ROC Curve

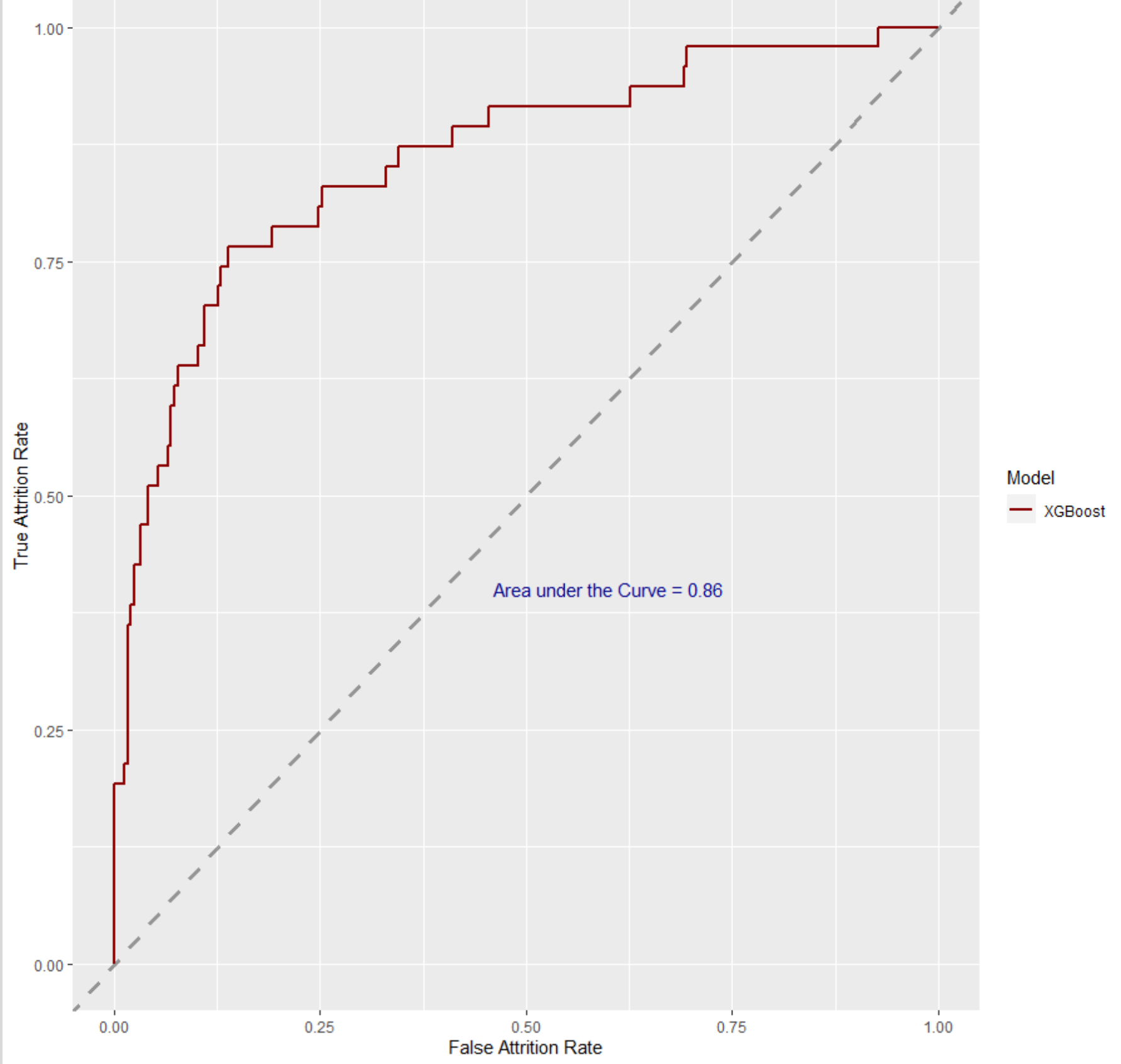
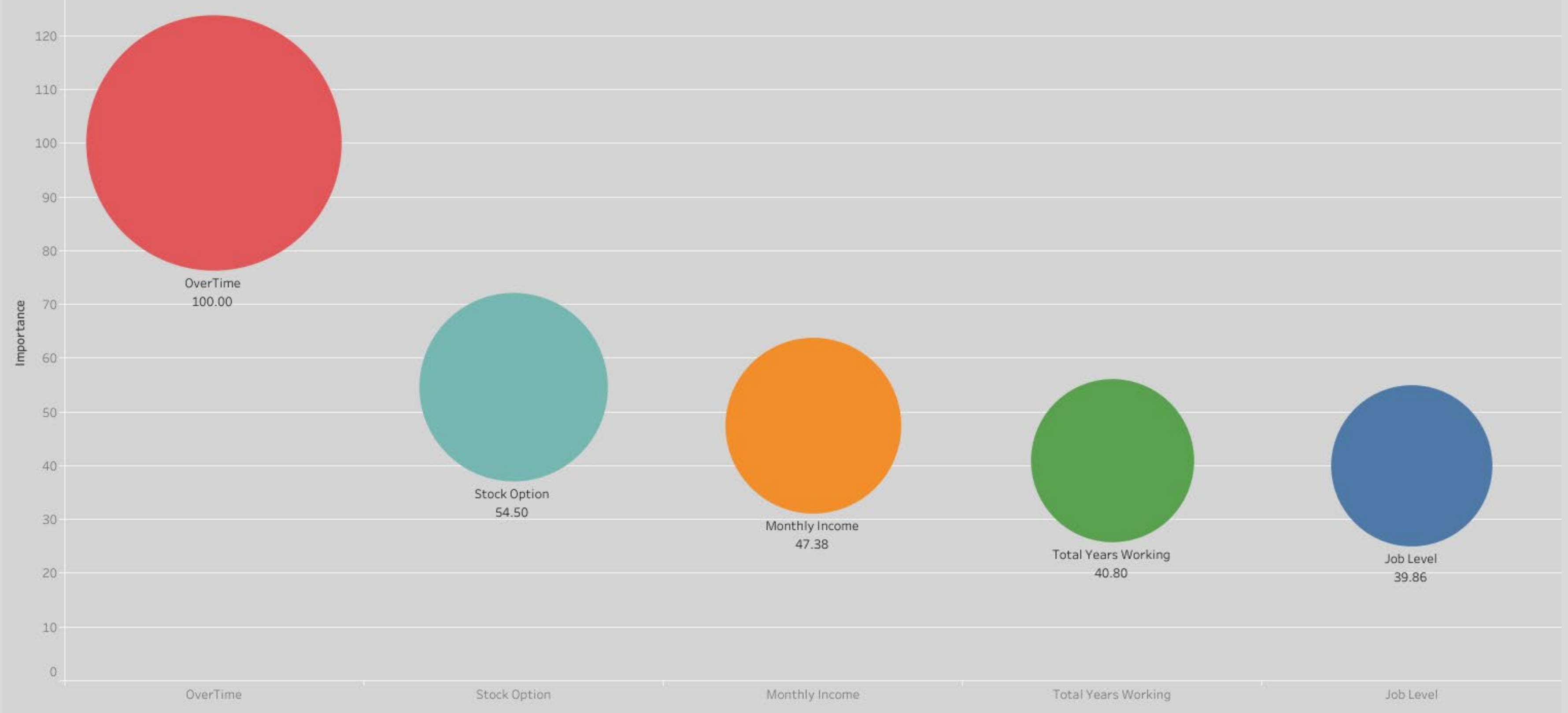


Figure 6. Scaled Feature Importance on Attrition



RESULTS

The data exploration in **Figure 1** reveals an imbalance between the number of employees who left the company and those who are still employed. Instead of attempting rebalancing or dataset sampling, two models were tested to determine which one could better manage this imbalance.

During the data exploration shown in **Figure 2**, a distinct group of employees stood out, constituting over 50% of all those who left – specifically, those who worked overtime. This factor's significance will be further highlighted in the variable importance analysis.

Random Forest:

Accuracy : 83.28% True Positive Rate: 95.12%

Kappa : 20.89% True Negative Rate: 21.28%

AUC: 80% **(Figure 4)**

The Random Forest model struggled with the imbalance of Attrition and thus suffered a low true negative rate percentage.

XGBoost:

Accuracy : 88.05% True Positive Rate: 96.75%

Kappa : 46.98% True Negative Rate: 42.55%

AUC: 86% **(Figure 5)**

The XGBoost model didn't encounter the same level of difficulty as the Random Forest model in handling the Attrition imbalance, but it still exhibited a low true negative rate.

The iterative process of XGBoost is illustrated in **Figure 3**. Factors with higher gains had a more pronounced influence on the model. Moreover, the higher the cover, the greater the number of observations impacted by the factor.

The scaled importance of features in **Figure 6** shows how influential OverTime is compared to all other features. The rest of the features exponentially decay.

DISCUSSION

The XGBoost model is a recommended tool for predicting employee attrition, assessing factor importance, and evaluating potential employees. While every organization possesses its unique dataset containing information about current employees or potential hires, the application of XGBoost can be a uniform and beneficial approach. By leveraging this data, organizations can personalize their talent acquisition strategies, design competitive employment offers, identify and address problem areas, and proactively plan for voluntary employee departures.

It is advisable to explore the application of higher penalties in conjunction with the investigation of various thresholds to minimize incorrect predictions. While XGBoost offers robust predictive capabilities, it can be computationally expensive. Organizations open to a slight trade-off in accuracy may find Random Forest a more suitable alternative.