Data Science Report: Employee Attrition Analysis and Prediction

Jakub Ćwięka[1]

[1] Jagiellonian University, Department of Sociology, Cracow

Abstract :

Employee attrition is the graudal reduction of a company's workforce due to employees leaving (resignation, retirement, death) and *not being replaced*. High attrition may induce cost implications, knowledge loss, decreased morale and productivity. Based on HR Employee Data I conduct an exploratory data analysis to understand the key aspects of company's problem and fit three predictive models to help prevent future employee departures. The report showcases my data analysis and results presentation skills, including key steps, methods, results and learnings from this data science project.

Key Terms: Employee Attrition, Classification Trees, Machine Learning, Logistic Regression, R, LaTeX

1 Introduction

Employee attrition remains a critical challenge for organisations, impacting operational stability, knowledge continuity and overall producitivity. This report presents a data-driven analysis and predictive modelling of employee attrition based on *HR Analytics Dashboard: Employee Attrition* 2025 dataset. By identifying key factors contributing to employee exits and developing predictive models, the analysis aims to:

- Support strategic HR decisions.
- Improve retention efforts.
- Minimise the costs associated with unwanted attrition.

The insights derived will enable proactive interventions to sustain a stable, engaged, and high-performing workforce.

2 Data Overview and Preparation

The data set used for analysis consists of 35 variables and has 1470 observations, which allows for precise and effective predictive modelling. To prepare data for the exploration, some key steps in data analysis were taken. As there was no missing values in the data base, I focused on identifying unnecessary variables or with zero variation. Following variables were constant across observations: Employee Count, Age Over 18, Standard Working Hours, and they were deleted from data set. Additionaly, the employee ID was removed, as this information is non-informative.

Out of four income indicators only one was left - Monthly Income, as it was the most intuitive and sensible one. Because correlations between these variables were low and the values did not convert (e.g. monthly wage divided by number of working hours gave different values to hourly wage) I treated them as incorrect.

Some of the variables were scales, measuring work-life balance score, satisfaction with the job and other factors that may affect employee attrition. One of them - rating of employee's performance, was measured on a 1-4 scale. Because it only took 2 highest values and were most probably given by managers, I also excluded this variable from analysis.

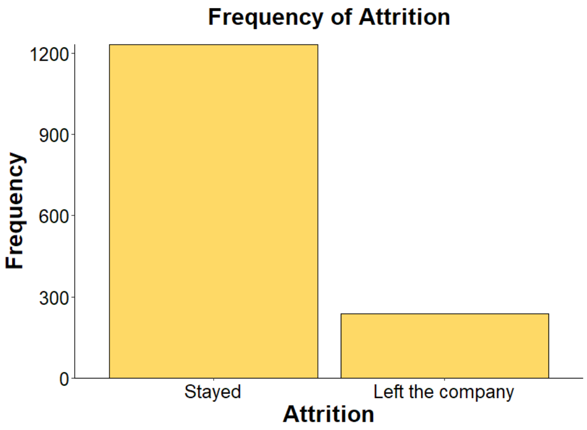
Dychotomous factor variables were transformed into numeric type, and factors with ordered categories into ordered factors.

This way the data set became ready for an exploratory analysis, clear of reduntant variables and with appropriately marked variable types.

3 Exploratory Data Analysis

Data exploration is crucial in any data analysis, it enables for understanding the problem and the data itself, creating initial hypotheses and planning futher steps. By checking simple frequency tables, distributions and correlations we gain an overview of the problem and possible explanations.

What it turns out, the main problem in upcoming analysis will be disproportion in dependent variable - employee attrition. Out of 1470 employees, only small fraction of them actually left the company. Such imbalance will be a significant challenge for predictive models, because simply assigning all observations to the majority class will result in a high accuracy of the model.

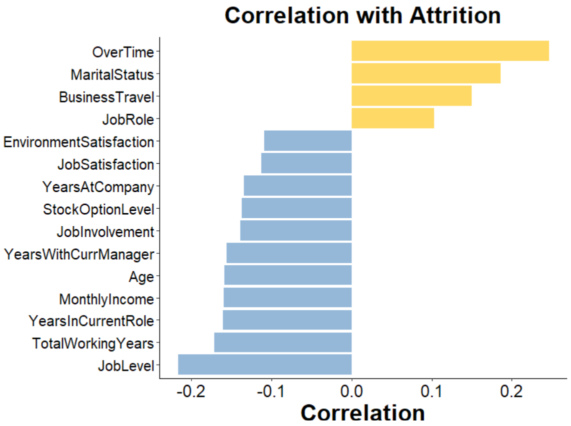


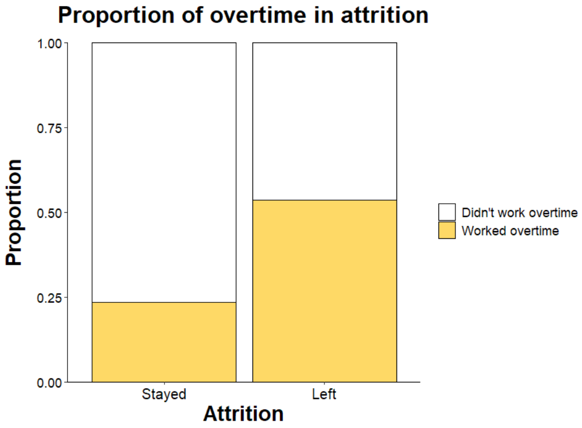
Analysis of averages and distributions shows that the company is composed mainly of permanent employees who have been working there for years and are constantly developing.

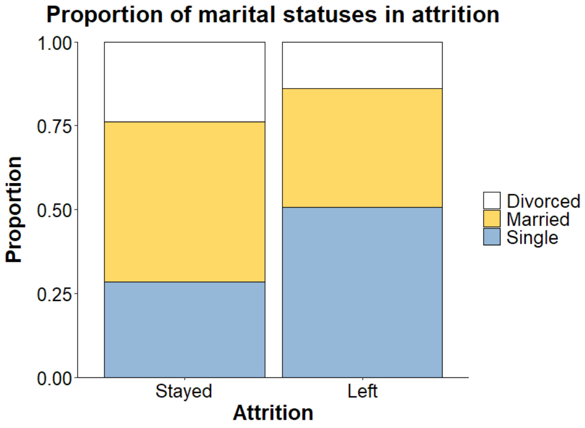
Descriptive statistics of selected variables

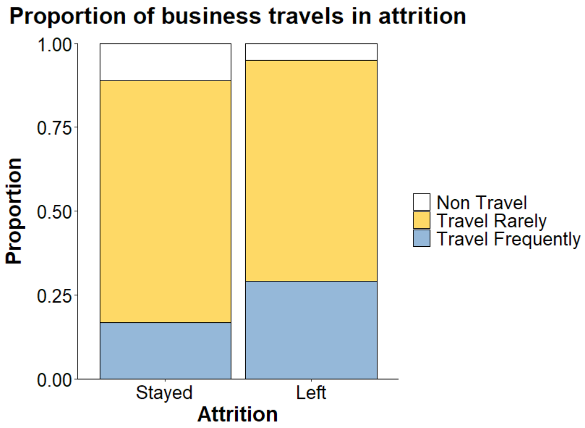
Variable	Mean	SD	IQR	Min	Max
Age	36,9	9,14	13	18	60
Male	0,6	0.49	1	0	1
Monthly Income	6502,9	4707,95	5486	1009	19999
Years At Company	7	6,12	6	0	40
Years Since Last Promotion	2,18	3,22	3	0	15

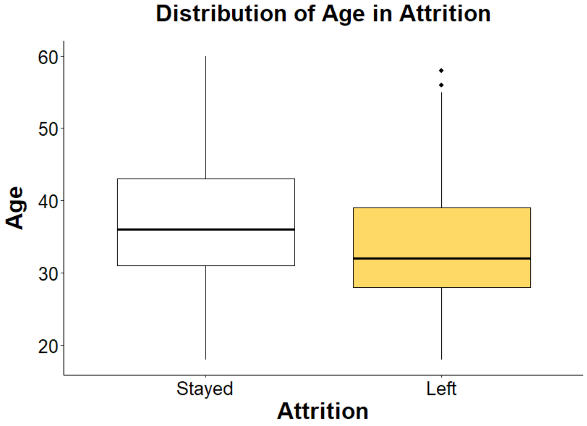
To gain insight into the specific characteristics of employees who left the company, I conduct a heterogenic correlation matrix from R's *polycor* package, which "computes a heterogenous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables". This way we can compute a correlation matrix between all variables in the data set and make sure they are unbiased (Olkin et al., 1958; Drasgow, 2014). Taking over time, businness travels, long commute to work along with being single and working in the sales department show the strongest correlation with employee attrition. On the correlation plot variables with correlation between -0.1 and 0.1 were hidden for improved visibility.











From various theories we know that attrition is mainly explained by either better economic external opportunities or poor work-life balance; many studies show that being over worked and having low job satisfaction leads to increased chance of mental health crisis and leaving the workplace (Iqbal, 2010; Alam et al., 2019; Self et al., 2011). In the case of studied company, most of the employees who left were early into their carrers, yet overworked and unsatisfied. Company struggles to retain young, fresh employees and relies on staff that is already established.

4 Predictive Modelling

Main goal of this analysis is to find the best model to predict employee attrition. For this purpose, I've decided to fit three different models:

- Conditional Random Forest
- Extreme Gradient Boosting
- Logistic Regression

Each of these models has it's strenghts and weaknesses, and my goal is to find the best parameters for the machine learning classification trees, as well as the best cut-off point for logistic regression. Then, I will compare the results and precision of classification, to choose the best out of three.

4.1 About Methods Used

Conditional Random Forest and Extreme Gradient Boosting are both machine learning algorithms based on decision trees, which goal is to classify into categories or predict values based on provided indicators. In general, decision tree makes a statement, and then makes a decision based on whether or not the statement is true. Because basic decision trees are very simple and straightforward, there is many different algorithms that enhance performance and allow to build a robust and precise models.

Random Forests are a collection of thousands of decision trees. The algorithm randomly select only a subset of indicators for each tree, and calculates the estimates on a portion of dataset (Breiman, 2001). Then, the accuracy of the prediction is tested on the rest of the data, the *Out of Bag Data*, which results in a *Out of Bag Error* (OOB) - the rate at which model falsely predicts observation on a new data. The phrase "Conditional" in Conditional Random Forest means that the algorithm conditionally tests whether each variable is truly useful before using it to split. This leads to fairer and more reliable trees, especially with highly correlated predictors or mixed-type data (Hothorn et al., 2006).

Random Forests are great for detecting interactions between predictors or non-linear effects, they have a built-in cross-validation in a form of OOB Error and provide information about variable importance in explaining the target variable. Because of this, they excel in hypothesis generation. The biggest disadvantage of Random Forests is that they are not transportable. Because there is no equation behind prediction, the results are based on data provided to the model for training. Considering this, Random Forests can be used to construct hypotheses and understand the problem, what can be used later to build a linear model that will capture the most important features.

Extreme Gradient Boosting (XGBoost) (Chen et al., 2016) is another algorithm, similar to Random Forest in it's base principles. The main idea is to build model in stages, where each new tree tries to fix the errors made by the previous ones. It does so by learning from differences between predicted and actual values, also called *residuals*. Because XGBoost is a very flexible and powerful algorithm, it has many parameters like learning rate, regularization, tree depth etc. that need to be adjusted. Because of this tunabilty, XGBoost can perform better and provide more accurate predictions, especially in a case of imbalanced target variable.

4.2 Random Forest

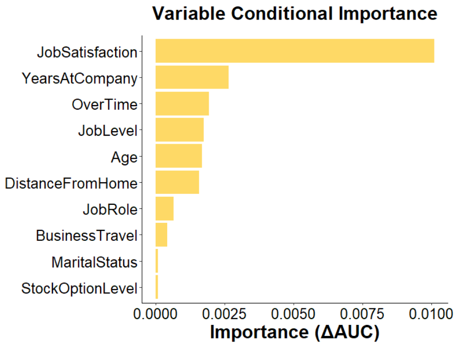
Package *cforest* in R is an implementation of the random forest and bagging ensemble algorithms utilizing conditional inference trees as base learners. Generally speaking, by using statistical tests for splits, *cforest* should reduce bias towards variables with many possible splits or categories and be less prone to overfitting due to hypothesis testing at each node. The disadvantage of this model is that it doesn't allow to assign weights to the less frequent target variable category, as can be done in *XGBoost*.

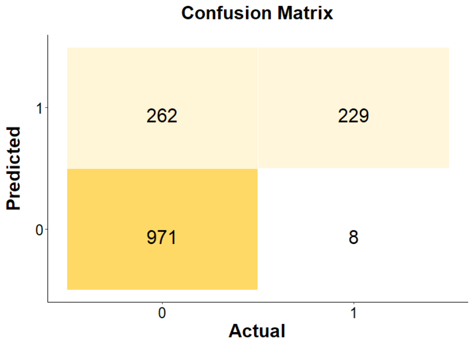
First, the model with complete data was fitted. By evaluating OOB Error rate, AUC-based variable importance and confusion matrices, I gradually excluded unnecessary variables, which had near-zero or negative importances. Each step improved the OOB Error Rate, Kappa value, No Information Rate and sensitivity, which proves that many variables in dataset where actually noise and distorted the classification process. The final model is composed of twelve most important predictors out of twenty-seven available. As expected, the model was classifying too much negative cases (employees who did not leave). One possible solution is to find threshold that maximizes sensitivity - a measure of how well the model predicts positive cases (employee attrition). For this purpose, the Youden's J statistic was used (Fluss et al., 2005).

To understand the process completely, consider that Random Forest can produce results in a form of either numeric, precise classification or probability of belonging to a particular class. By default, the threshold discriminating classes is set to 0.5, which means that employee with estimated probability above 0.5 will be classified as the one who left the company. But it doesn't have to be the most optimal point. In a case of employee attrition, the goal should be to maximize correct classification of positive occurence - attrition, just like in cases of diseases or costly events. Youden's J statistic finds threshold like this, which in this case is 0.15. After calculating class belonging, the new classification resulted in 96% sensitivity at the cost of missclassifying workers who stayed in a company.

Evaluation statistics of full and final model

Statistic	Full Model	Final Model
OOB	0,85	0,86
AUC	0,96	0,94
Accuracy	0,88	0,82
Balanced Accuracy	0,63	0,88
No Information Rate	0,84	0,84
Kappa	0,37	0,52
Sensitivity	0,27	0,96
Specificity	0,99	0,79
Precision	0,95	0,46





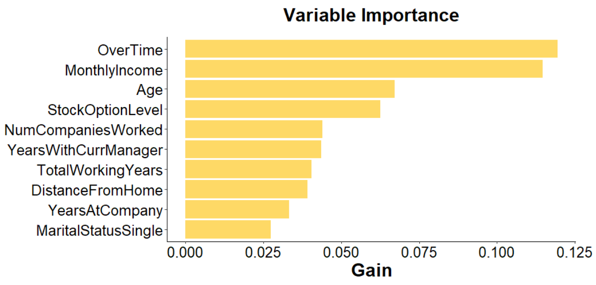
4.3 XGBoost

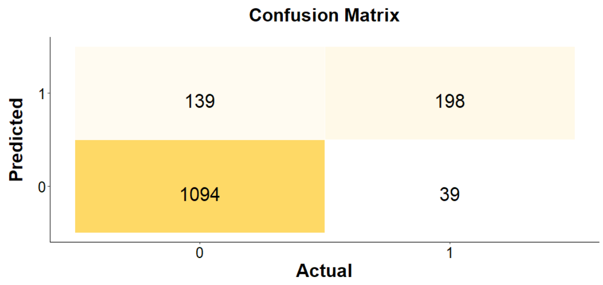
Major downside of the final Conditional Random Forest model is low precision - the proportion of true attrition in overall predicted attrition. Because of unbalanced target variable, the threshold for classification had to be set at a very low value, which results in biased estimates. In theory, the XGBoost model should perform better with adding weights to less frequent class and finding the best hyperparameters. For this purpose dataset was split into training and testing sets, containing random sample of successively 80% and 20% of observations. Then, training and testing sets were transformed into special format of *DMatrix* from *XGBoost* package. Because it's necessary to fine-tune model's parameters, I used grid search to find the most optimal ones. By specifying possible values each parameter can take, algorithm tries every possible combination, uses 5-fold cross-validation to well each performs, and chooses the best one. AUC-PR evaluation metric was used because of imbalanced dependent variable (Saito et al., 2015). Found parameters were used as a base value, from which parameters where further tuned until a compromise between under and overfitting was achieved. This was done by evaluating classication error, AUC and confusion matrices at each step.

After finding the optimal threshold cut-point, final model performs differently compared to Conditional Random Forest. The overall accuracy is higher at the cost of correct predictions of employee attrition. Variable importances also differ, suggesting complexity of the data.

Comparison of Conditional Random Forest and Extreme Gradient Boost models

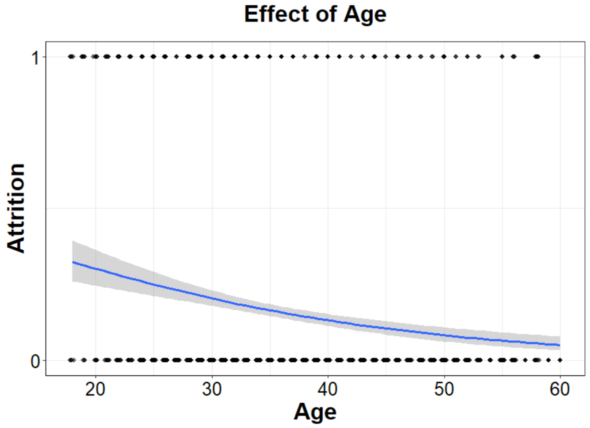
Statistic	CForest	XGBoost
AUC	0,94	0,93
Accuracy	0,82	0,88
Balanced Accuracy	0,88	0,86
No Information Rate	0,84	0,84
Kappa	0,52	0,62
Sensitivity	0,96	0,84
Specificity	0,79	0,89
Precision	0,46	0,59

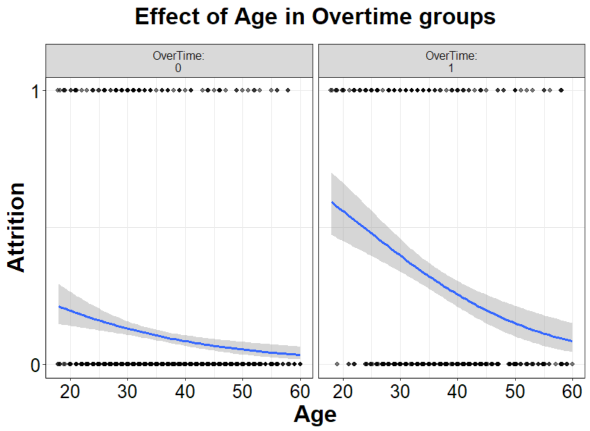


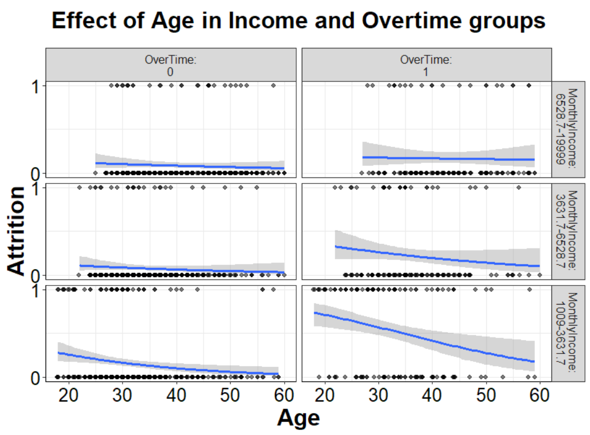


4.4 Learnings from previous analysis

Machine learning algorithms are very usefull in detecting non-obvious dependencies in the data. Because algorithms are not restricted by analysts equations, they can find numerous interactions between variables or non-linear effects, which won't be visible in correlations, since they measure only linear relationships. Both *CForest* and *XGBoost* used different predictors to achieve similar results. By using R package *Flexplot*, I investigated possible relations deducted from previous analyses to gain further insight.







Main conclusions reinforce those derived from initial exploratory analysis, that most employee attrition seen in the data comes from young, overworked workers. This can be seen in the graphs, as probability of attrition raises drastically when control of overtime and income is added. This suggest three-way interaction between these variables, which should be included in logistic regression model.

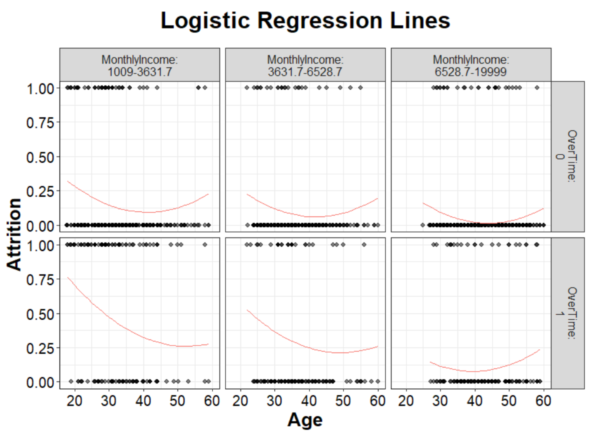
4.5 *Logistic Regression*

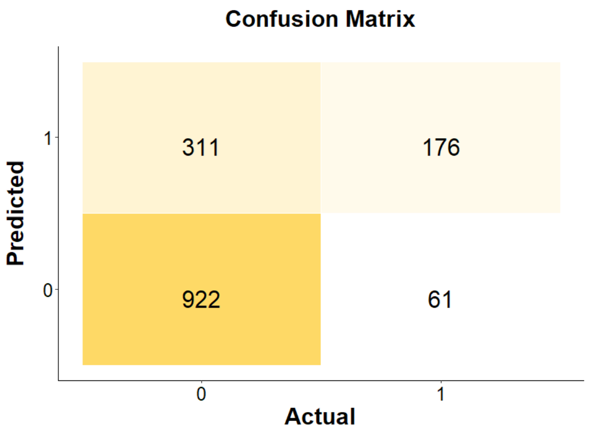
Final part of my analysis is to build a logistic model and check if the results are good enough for the model to be retained and used on future data. Because logistic regression is build on a specific equation, it provides a formula to solve future data. Since overspecyfing linear models in not desirable, at first only the most important and sensible predictors where used - taking overtime, monthly income and age, as seen in previous section. Then, the model was gradually expanded with new interactions and quadratic effects. Each step was evaluated with AIC, BIC, Bayes Factor, R-Squared and Adjusted R-Squared to make sure that complicating the model results in better estimation. At last, four variables were added: marital status, distance from home, job satisfaction and total working years, as they provided additional, important information.

Unfortunately the logistic regression is outperformed by machine learning models in this particular case. The lack of strong predictors causes even complicated model - containing three-way interaction, quadratic effects and additional variables - to explain only 17% variability of the employee attrition. As seen in correlation plot and variable importances before, even though there is many variables in the data, only a small fraction of them provide a substantive explanatatory power.

Comparison of simple and full logistic regression models

	AIC	BIC	Bayes Factor	R-Squared	Adj. R-Squared
Simple Model	1091.53	1117.99	0	0.096	0.095
Full Model	975.36	1070.634	19243147351	0.180	0.171





5 Conclusions and learnings

5.1 *Conclusions*

The aim of this analysis was to thoroughly understand the nature of employee attrition and fit a model that would be able to accuraty predict such events. To meet this goal, dataset was cleaned and prepared for work, exploratory analysis was conducted, and three classification models fit. Conditional Random Forest provides very high sensitivity of attrition, at the cost of missclassifying workers who did not leave the company. On the other hand, the XGBoost model yields more balanced results, maintaining a trade-off between sensitivity and overall accuracy. Quality of the models was evaluated using ROC Curves, classification error, confusion matrix, variable importance and cross-validation. Both models used slightly different variables to produce similar output, though related to each other in the context of theory. In the case of analysed company, employee attrition is highly connected to over work. Young employees with lower wages are the ones who make most of the attrition in the database. The lack of other strong pedictors makes the logistic regression model insufficient.

Future actions could inlude further hyperparameter tuning, repeated cross-validation, exploring alternative models, feature engineering, and generating recommendations for data gathering to acquire more of the usefull information.

5.2 Learnings

All of the analysis, graphs and tables were made using R Studio, and the article itself was assembled in LaTeX. For the purpose of preparing this sample analysis, multiple packages were used, including *caret*, *cforest*, *XGBoost*, *polycor*, *flexplot*, and many more. As the machine learning was new to me, I had to learn both the basics and more advanced aspects of classification trees, adaptive boosting, gradient boosting, hyperparametrization, regularization, and translating results into substantive knowledge. This introduced me to fitting robust models that can handle imbalanced target variable and get the most out of poor quality predictors.

Having gained new experience and knowledge, I am excited to tackle future problems and provide value to projects.

References

Alam, Aliya, Muhammad Asim, et al. (2019). "Relationship between job satisfaction and turnover intention". In: *International Journal of Human Resource Studies* 9.2, p. 163.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Drasgow, Fritz (2014). "Polychoric and polyserial correlations". In: *Wiley statsRef: statistics reference online*.

Fluss, Ronen, David Faraggi, and Benjamin Reiser (2005). "Estimation of the Youden Index and its associated cutoff point". In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47.4, pp. 458–472.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3, pp. 651–674.

HR Analytics Dashboard: Employee Attrition (2025). `https://www.kaggle.com/datasets/anubhav761/hr-analytics-dashboard-employee-attrition`. Accessed: 2025-07-01.

Iqbal, Adnan (2010). "Employee turnover: Causes, consequences and retention strategies in the Saudi organizations". In: *The Business Review, Cambridge* 16.2, pp. 275–281.

Olkin, Ingram and John W. Pratt (1958). "Unbiased estimation of certain correlation coefficients." In: *Annals of Mathematical Statistics*.

Saito, Takaya and Marc Rehmsmeier (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3, e0118432.

Self, John T and Ben Dewald (2011). "Why do employees stay? A qualitative exploration of employee tenure". In: *International Journal of Hospitality & Tourism Administration* 12.1, pp. 60–72.