UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING

MIE368: Final Report

Predicting Employee Attrition with Machine Learning: Analyzing Key Factors

Prepared by:
Si Han (Catherine) Xiong 1008935404
Zichen (Zach) Liu 1008789152
Seunghyun (Joe) Lee 1006882651
Michael Abou Zeid 1008279388

Date:
December 4th, 2024

Instructor:
Jangwon Park

TA Mentor:
Rachel Wong

Abstract

Background: Global healthcare personnel shortages are forecasted to reach 14 million by 2030, mostly due to employee attrition [1]. Healthcare employee attrition is often difficult to predict. High attrition rates place financial burdens on hospitals, reduce the quality of patient care, and cause understaffing, which negatively impacts the mental health and productivity of remaining employees [2].

Objective: This study aims to develop and evaluate multiple machine learning (ML) classifiers to predict whether a healthcare employee is likely to leave and to identify significant factors influencing attrition.

Methods: Simulated data of 1,500 healthcare employees with 30 features from IBM Watson was used for training and testing ML classifiers [3]. The CART model served as a baseline for its simplicity and interpretability. Seven additional models, including tree-based models (Random Forest, Bagged Trees, XGBoost, CatBoost), Support Vector Machines (SVM), and Logistic Regression with L1 and L2 Regularization, were compared using AUC, F1, and Recall scores. Feature importance across all models was determined by a Borda Count System, which assigns weighted votes to the top five features identified by each model and aggregates the scores across all models to determine the most significant features [4].

Results: Logistic Regression with L2 Regularization was the best-performing model, achieving an AUC score of 0.93, outperforming the baseline by 30%, an F1 score of 0.69, exceeding the baseline by 50%, and a Recall score of 0.79, surpassing the baseline by 70%. Using the Borda Count system, the aggregation of all models identified four key features as most significant: marital status (single), overtime, age, and years of working.

Conclusion: Given the use of an ML classifier to predict employee attrition and identify contributing features, we can help healthcare organizations define interventions for preventing employee attrition. Some limitations include the simulated data being based on the U.S. healthcare system which may not generalize to the trends within Canada. Further research is needed to validate model performance through real hospital data. Incorporating additional features, such as time series for attrition, could improve model performance and accuracy.

1.0 Introduction

Healthcare worker personnel shortages are projected to increase over the next 10 years leading to a workforce crisis. There are forecasts for 2025 that there will be a shortage of more than 400,000 health aides and 29,400 nurse practitioners in North America[5]. This shortfall is caused by multiple factors: demographic changes, increasing healthcare demands, and organizational work environments. The aging healthcare workforce itself presents a significant challenge, with a projected 17% retirement rate for nurses in North America over the next decade and approximately half of practicing physicians expected to retire by 2030 without the same rates of healthcare workers entering the field [6]. As a result, these shortages impact healthcare delivery, leading to increased mortality rates, medical errors, and compromised care. Moreover, the average financial impact of a healthcare worker turnover could be as high as 200% of the employee's annual salary" [7].

To address this issue, our study employs machine learning classifiers to analyze data from a synthetic dataset of 1,470 healthcare employees with 30 distinct features from IBM Watson, aiming to predict employee attrition and identify key retention factors. This analysis aims to develop a classifier to predict employee attrition based on personal attributes and provide healthcare organizations with organizational recommendations to mitigate high attrition rates and their associated impacts.

2.0 Methods

2.1 Data Pre-Processing

The preprocessing steps consisted of three steps to prepare the training data [Figure 1]. Appendix A provides an overview of features and their values. For feature encoding, ordinal encoding was used to preserve the natural order of categories while converting them into numerical values. SMOTE was used to generate synthetic records of the minority class (Attrition) to balance the training dataset, preventing model bias and improving the reliability of predictions.

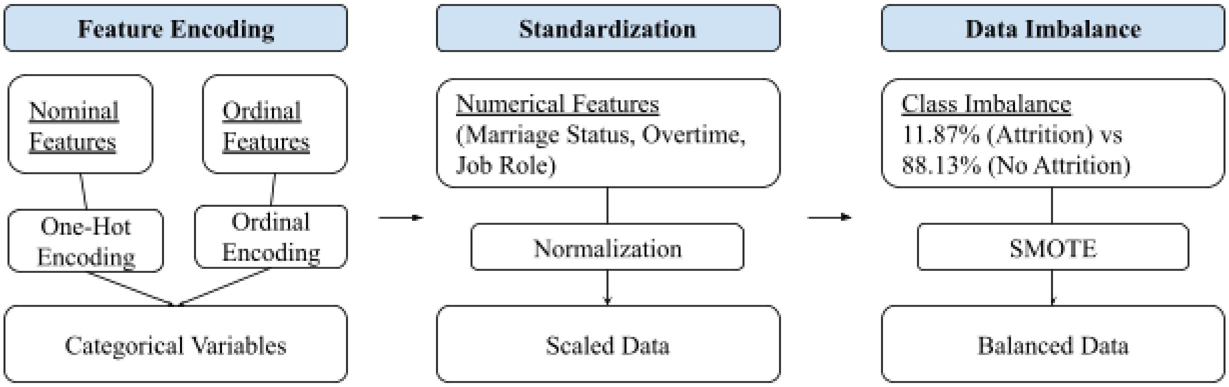


Figure 1: The Data Pre-Processing Steps

2.2 Model Selection

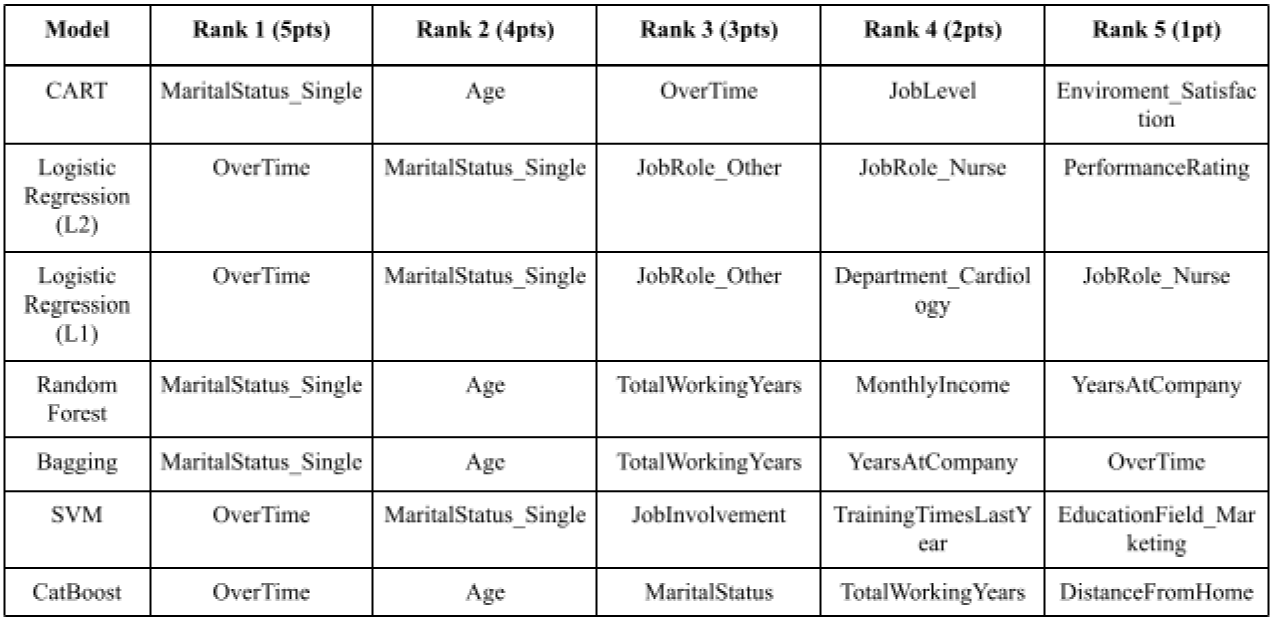
The following predictive models were used for evaluation. Please view Appendix Section B Figure 1 for model parameter information.

Model	Advantages
Classification and Regression Trees (CART)	- Simple and highly interpretable baseline.
Random Forest (RF)	- Improves classification accuracy compared to CART by combining multiple decision trees.
Bagged Trees	- Reduced model variance compared to RF through bootstrap aggregation.
Logistic Regression (L1 Regularization)	- Penalizes the sum of absolute values of the regression coefficient and selects only the important features contributing to attrition. - Avoids overfitting.
Logistic Regression (L2 Regularization)	- Penalizes the square of values of the regression coefficient and selects important features contributing to attrition. - Avoids overfitting.
Support Vector Machines (SVM)	- Accounts for the high cardinality of the features for employee attrition.
XGBoost	- Fast Execution Speed. - Proven model performance in solving data science problems.
CatBoost	- Categorical feature support without one-hot encoding. - Automatic hyper-parameter tuning.

Figure 2. Models and corresponding advantages for selection.

2.2 Borda Count Feature Selection

For employee attrition factors, we utilized an ensemble-based feature selection method incorporating the Borda count, a rank aggregation technique originally developed for voting systems [8]. For each model, we identified the top five features influencing employee attrition based on the model-specific feature importance metrics (e.g., coefficients for logistic regression, feature importance scores for tree-based models). We assigned points to features according to their rank within each model [Figure 3]. Then, we aggregated points across models. Features were ranked based on their total aggregated points. The features with the highest total points were considered the most influential predictors of employee attrition. This was able to reduce individual model biases and enhance the robustness of feature selection.



Model	Rank 1 (5pts)	Rank 2 (4pts)	Rank 3 (3pts)	Rank 4 (2pts)	Rank 5 (1pt)
CART	MaritalStatus_Single	Age	OverTime	JobLevel	Enviroment_Satisfaction
Logistic Regression (L2)	OverTime	MaritalStatus_Single	JobRole_Other	JobRole_Nurse	PerformanceRating
Logistic Regression (L1)	OverTime	MaritalStatus_Single	JobRole_Other	Department_Cardiology	JobRole_Nurse
Random Forest	MaritalStatus_Single	Age	TotalWorkingYears	MonthlyIncome	YearsAtCompany
Bagging	MaritalStatus_Single	Age	TotalWorkingYears	YearsAtCompany	OverTime
SVM	OverTime	MaritalStatus_Single	JobInvolvement	TrainingTimesLastYear	EducationField_Marketing
CatBoost	OverTime	Age	MaritalStatus	TotalWorkingYears	DistanceFromHome

Figure 3. Top 5 features of each model are ranked based on the Borda Count system.

3.0 Discussion

3.1 Model Performance

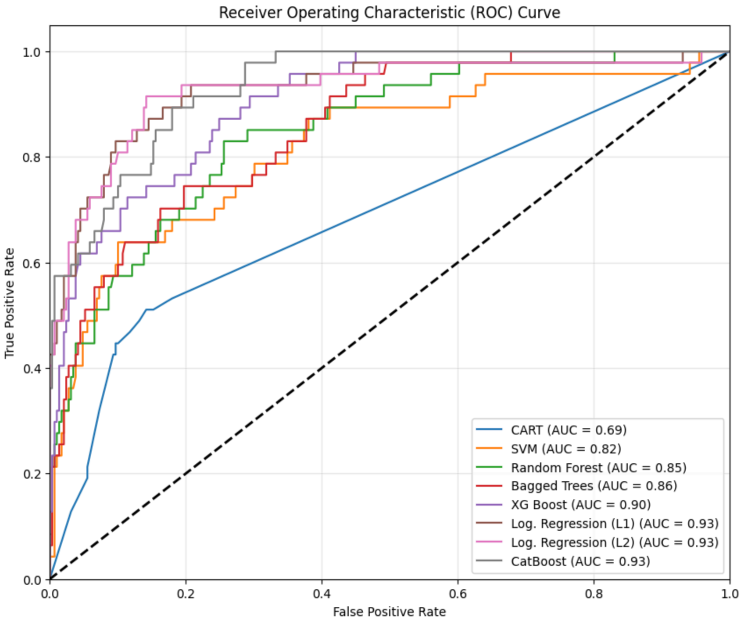


Figure 4. ROC curve that evaluates and visualizes AUC Scores for all models.

Evaluating the models based on AUC scores, Logistic Regression with Regularization (L1, L2) and the CatBoost model demonstrated the best performance, achieving an AUC score of 93%, indicating a strong capability for the models to distinguish between positive and negative classes—AUC score reflects performance in terms of recall and fall-out. Although this metric is valuable for evaluating the model's ability to rank predictions, it does not account for precision.

Therefore, we also considered the F1 score [Figure 5], which combines both precision and recall. The F1 score confirmed that these three models achieved the best overall performance.

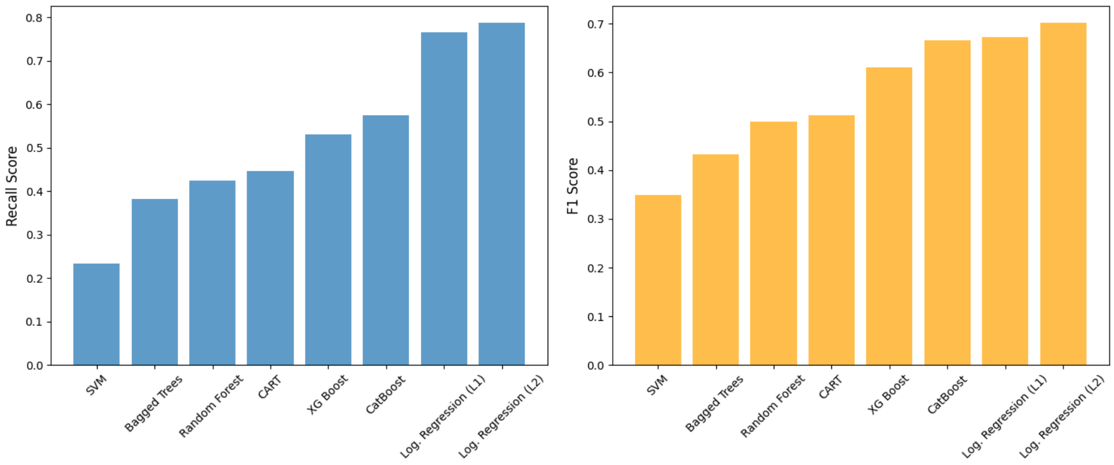


Figure 5. Recall Scores for all models (Left), F1 Scores for all models (Right).

The primary goal of the project is to accurately identify employees likely to leave among those who were observed to leave, rather than focusing on maximizing the overall predictive accuracy of the model. Therefore, recall is the most important metric for this objective, as it measures the model's ability to correctly identify employees who are truly likely to leave.

Among the top three models selected based on AUC and F1 scores, we found that Logistic Regression with L2 Regularization demonstrated the best recall performance. Additionally, logistic regression is a relatively simple model with high interpretability and greater computational efficiency compared to other models in our evaluation. These advantages make it an ideal choice for effectively predicting employee attrition, enabling us to identify approximately 80% of employees likely to leave.

3.2 Borda Count Feature Importance

After applying the Borda Count system to aggregate total feature importance scores across the models [Figure 6], we identified marital status, overtime, age, total working years, and job level as the five most significant features. Among these, marital status, overtime, and age stood out with substantially higher aggregate scores, indicating their consistent importance across all models compared to other features. This highlights their strong predictive value.

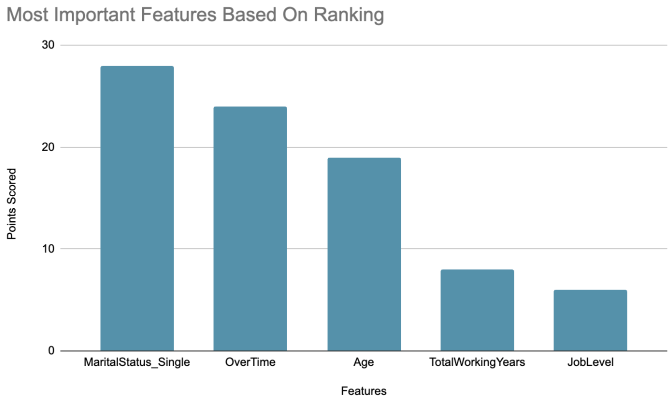


Figure 6. Top 5 most important features for healthcare employee attrition prediction

3.3 Literature

The following section will evaluate evidence to interpret our findings. Please view the Appendix Section C for the graphs of the data within our dataset with each feature.

3.3.1 Marital Status (Appendix C Section C.1)

Based on our feature importance ranking, marital status is the most important feature that predicts healthcare worker attrition rate. Research depicted this trend as well, stating that "single healthcare professionals are significantly more likely to leave their jobs compared to their married counterparts" [9]. This could be explained by the fact that married healthcare workers, despite additional responsibilities, were more likely to benefit from "psychosocial buffers such as family engagement and structured routines that promote work-life balance" [9].

3.3.2 Overtime (Appendix C Section C.2)

Workers who regularly reported excessive overtime were also associated with higher attrition rates with adverse outcomes for both the patients and the employees. However, moderate overtime (1–11 hours) correlates with lower turnover compared to no overtime, potentially due to the additional compensation and job engagement it provides [10].

3.3.3 Age (Appendix C Section Section C.3)

Age was also a very predictive feature with our data exploration highlighting that employees between 25-40 years with the highest associated attrition rates. This is corroborated by research from Canadian Healthcare Technology which claims that 40% of nurses are leaving their jobs before age 35 [11]. The most common concerns were due to " a lack of control over their work schedules, including mandatory overtime and a lack of shift flexibility as principal sources of workplace stress" [11].

3.3.4 Job Levels and Years of Working (Appendix C Section C.4)

On the other hand, attrition rates were inversely proportional to the number of years an employee stayed at a company and the years within their careers. Longer tenure was associated with greater commitment and job satisfaction [12]. This may reflect alignment of values, job satisfaction, and access to professional development opportunities, which motivate workers to remain in their roles.

3.3.5 Interactions of Features

Within our research, we were able to identify correlation between the following features:

1. <u>Marital Status & Overtime:</u> Single healthcare workers were more likely to take on overtime hours due to fewer family obligations which could lead to excessive stress and a lack of familial support to regulate these feelings [13].
2. <u>Age and Job Levels</u>: Younger workers (25–40 years) face higher attrition due to dissatisfaction with scheduling, flexibility and compensation, which is mitigated for those with longer tenure or higher job levels [13].

4.0 Conclusion

4.1 Summary

The project evaluated 8 different models, including tree-based models, Logistic Regression with L1 and L2 regularization, Support Vector Machines (SVM), and boosting-based models. The top 3 performers were selected based on their AUC and F1 scores. These models were then further assessed using recall as the primary metric, given its importance in identifying employees likely to leave based on the subset of employees observed to leave, rather than the entire population. Logistic Regression with L2 regularization emerged as the best-performing model, offering the highest recall, simplicity, and strong interpretability. The top 5 most important features are marital status, overtime, age, total working years, and job level.

4.3 Recommendations to Organizations.

Given our findings within this project, we can recommend the following things for healthcare organizations to mitigate attrition rate's negative effect on their organizations:

1. *Anticipate Turnover by Worker Demographics:* Plan within their budget the unavoidable attrition rates due to individual characteristics.
2. *Limit Overtime with Fair Compensation:* Reduce excessive overtime with better staffing and scheduling to prevent burnout and errors in addition to offering better benefits for the necessary overtime [14].
3. *Support New Employees:* Strengthen onboarding, mentorship, and transition programs to retain early-career workers [15].

5.0 Future Directions

5.1 Limitations

The limitation of the project includes the synthetic nature of the data, geographic discrepancies and the generalization of healthcare professions. Synthetic datasets may generalize or contain inherent relationships might not accurately reflect actual attrition trends among healthcare workers. Nevertheless, the use of the synthetic dataset was due to the lack of accessibility and privacy concerns within the healthcare industry. Additionally, since the data represents U.S. healthcare workers, insights may differ for Canada due to variations in healthcare systems and policies, such as the Canada Health Act's universal coverage compared to the U.S.'s mixed system [16]. Finally, attrition rates likely vary by profession; for example, nurses have been reported to be up to three times more likely to quit compared to physicians [17].

5.2 Next Steps

To enhance our analysis, we aim to collect additional datasets on employee attrition rates in Canadian healthcare, incorporating time-series data and exploring differences across healthcare professions. This includes examining unique attrition patterns among roles such as nurse practitioners compared to other positions.

Reference

[1] “Why do people stay in healthcare?,” *National Center for Biotechnology Information*, Jul. 01, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9243774/ [Accessed: Nov. 25, 2024].

[2] “Employee turnover rates in the healthcare industry,” *DailyPay*, Nov. 18, 2023. [Online]. Available:https://www.dailypay.com/resource-center/blog/employee-turnover-rates-in-the-healthcare-industry/ [Accessed: Nov. 25, 2024].

[3] J. P. Miller, "Employee Attrition for Healthcare," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare. [Accessed: Nov. 26, 2024].

[4] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," in Machine Learning and Knowledge Discovery in Databases, W. Daelemans, B. Goethals, and K. Morik, Eds., Lecture Notes in Computer Science, vol. 5212, Berlin, Heidelberg: Springer, 2008, pp. 313–325. doi: 10.1007/978-3-540-87481-2_21.

[5] "The Shortage of Healthcare Workers: A Crisis Requiring an Innovative Solution," Duquesne University School of Nursing. [Online]. Available: https://onlinenursing.duq.edu/post-master-certificates/shortage-of-healthcare-workers/ [Accessed: Dec. 3, 2024]

[6] "Global health care worker shortage projection analysis," Nat. Commun., vol. 13, no. 1034129, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10341299/

[7] "Employee retention: What is the true cost of losing an employee," Simply Benefits, 2023. [Online]. Available: https://www.simplybenefits.ca/blog/employee-retention-what-is-the-true-cost-of-losing-an-employee

[8] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in Machine Learning and Knowledge Discovery in Databases, W. Daelemans, B. Goethals, and K. Morik, Eds. Berlin, Heidelberg: Springer, 2008, pp. 313–325.

[9] A. Nkwate et al., "Exploring associations between personal characteristics and turnover intentions among healthcare workers during COVID-19," Frontiers in Health Services, vol. 2, 2022. [Online]. Available: https://www.frontiersin.org/journals/health-services/articles/10.3389/frhs.2022.918843/full

[10] J. Lee et al., "The relationship between overtime work and nurse turnover: A systematic review and meta-analysis," PMC, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10164927/

[11] "40% of nurses leaving profession before age 35," Canadian Healthcare Technology, Sep. 2024. [Online]. Available: https://www.canhealth.com/2024/09/25/40-of-nurses-leaving-profession-before-age-35/

[12] D. Jackson et al., "Why do people stay in or leave nursing?" PMC, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9243774/

[13] "Why Nurses Quit: Causes of Turnover & Retention Strategies," American Nurses Association, 2024. [Online]. Available: https://www.nursingworld.org/content-hub/resources/nursing-leadership/why-nurses-quit/

[14] "5 Health Care Workforce Shortage Takeaways for 2028," American Hospital Association Center for Health Innovation Market Scan, Sep. 2024. [Online]. Available: https://www.aha.org/aha-center-health-innovation-market-scan/2024-09-10-5-health-care-workforce-shortage-takeaways-2028

[15] "Understanding Overtime Rules in Healthcare," HR for Health Blog, 2024. [Online]. Available: https://hrforhealth.com/blog/overtime-rule

[16] "U.S. vs. Canadian Healthcare: What Are the Key Differences?" Ross University School of Medicine Blog, 2023. [Online]. Available: https://medical.rossu.edu/about/blog/us-vs-canadian-healthcare

[17] "The State of the Health Workforce in Canada 2022," Canadian Institute for Health Information, 2022. [Online]. Available: https://www.cihi.ca/en/the-state-of-the-health-workforce-in-canada-2022

Appendix

Section A: Pre-Processing

Figure 1. Example overview of Nominal features: Education, MartialStatus, JobRole

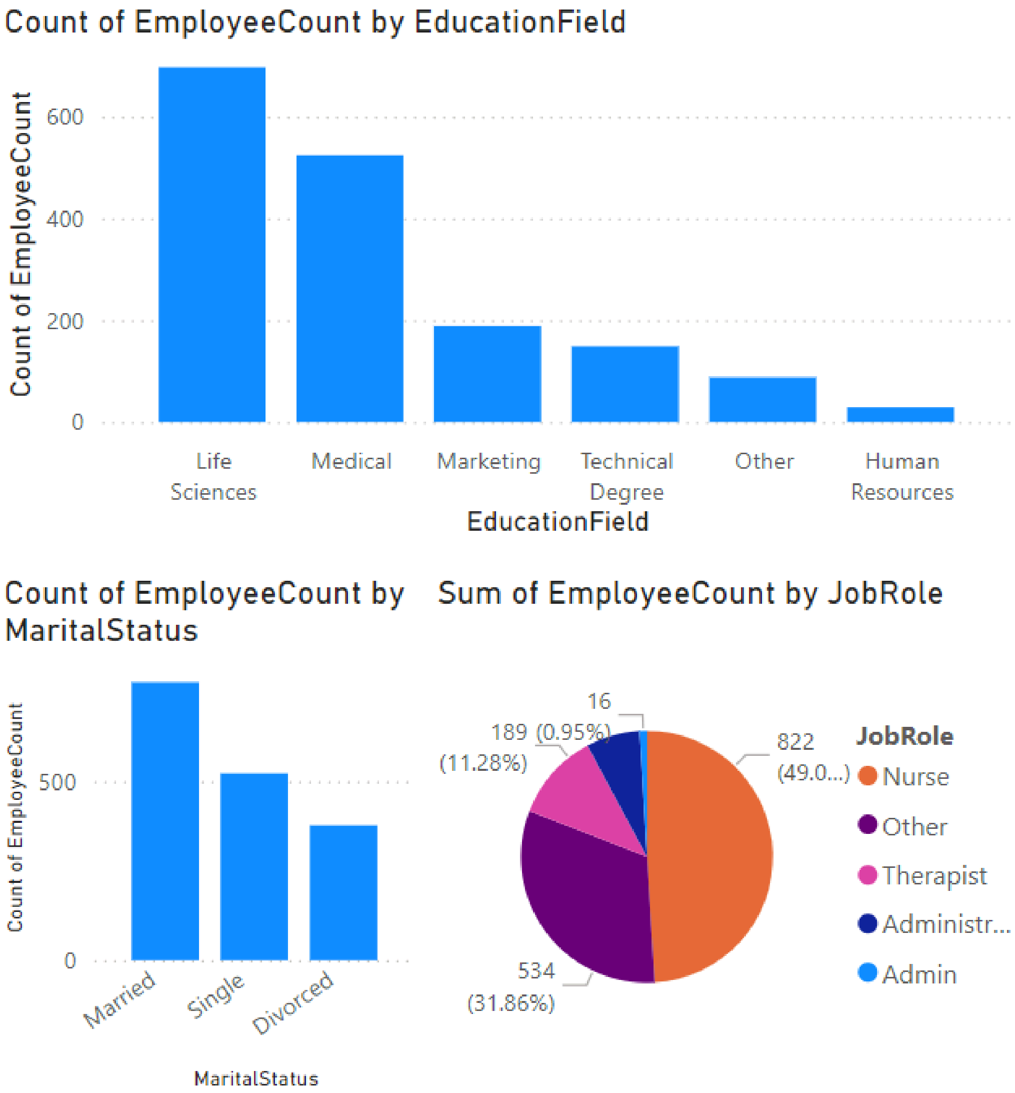
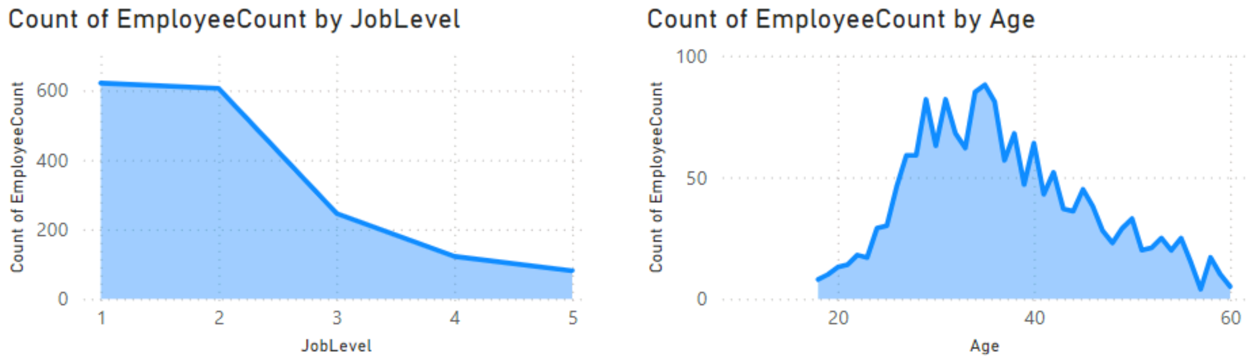


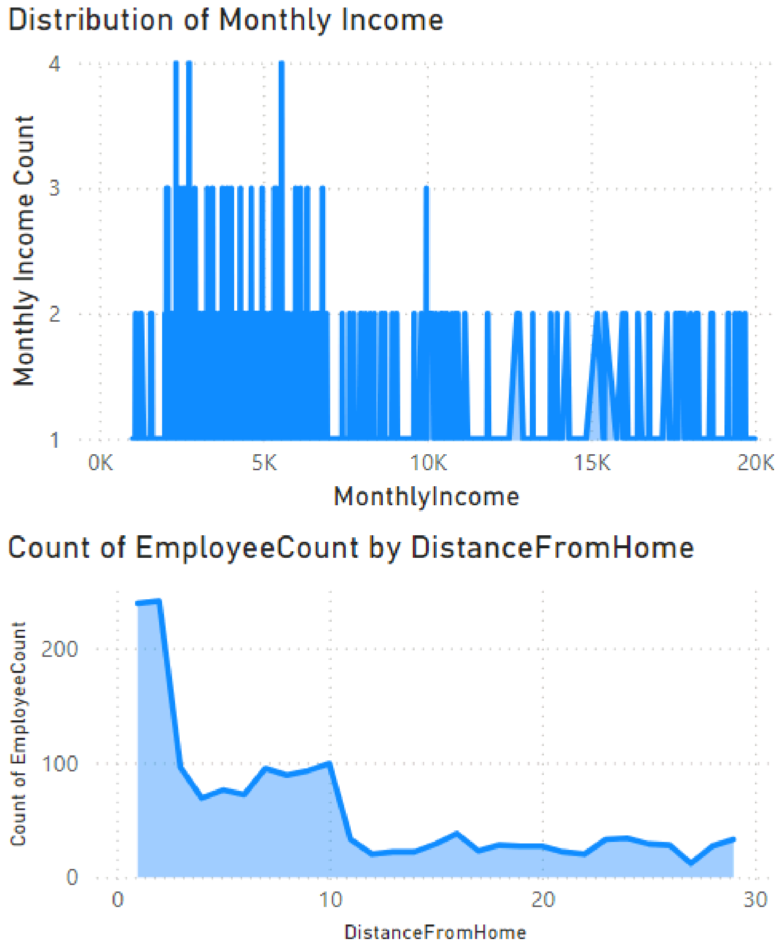
Figure 2. Example overview of ordinal features: Job level, age



Count of EmployeeCount by JobLevel

Count of EmployeeCount by Age

Figure 3. Example overview of numerical features: Monthly Income, Distance from Home



Distribution of Monthly Income

Count of EmployeeCount by DistanceFromHome

Section B: Methods

Figure 1. Methods and Parameters

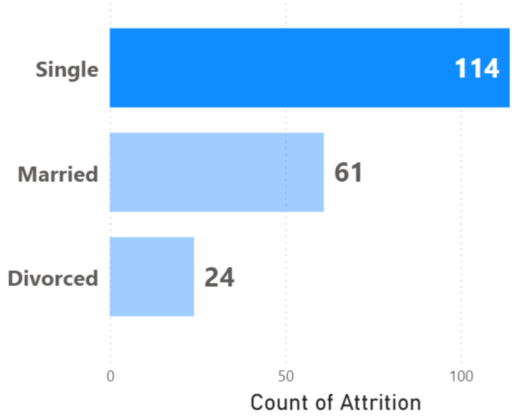
Model	Model Parameters
Classification and Regression Trees (CART)	{'max_depth': 9, 'min_samples_split': 9}
Random Forest (RF)	{'bootstrap': True, 'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}
Bagged Trees	{'max_depth': 10, 'max_features': 0.5, 'max_samples': 1.0, 'n_estimators': 200}
Logistic Regression with L1 Regularization	{'C': 1.0, 'intercept_scaling': 1, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear’}
Logistic Regression with L2 Regularization	{'C': 100, 'intercept_scaling': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'liblinear'}
Support Vector Machines (SVM)	{'svc__C': 10, 'svc__gamma': 0.1, 'svc__kernel': 'rbf'}
XGBoost	{‘max_depth’=3, ‘learning_rate’=0.1, ‘n_estimators’=100}
CatBoost	{'learning_rate': 0.7111, 'depth': 1, 'max_ctr_complexity': 8, 'boosting_type': 'Plain'}

Section C: Literature

The following graphs represent the counts of different features and attrition based on our data. This was created during the EDA stage.

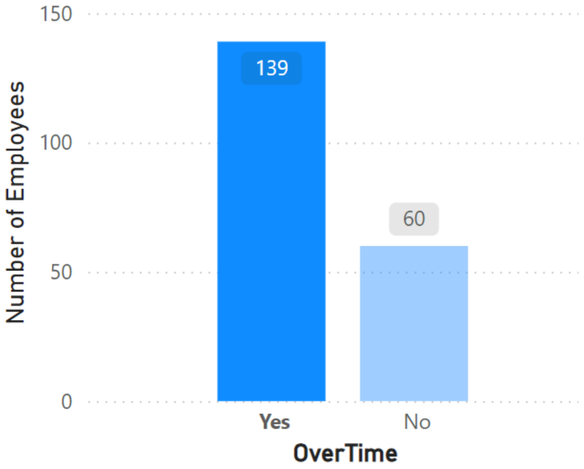
Section C.1

Marital Status: Single Healthcare workers had higher attrition rates compared to their married co-workers.



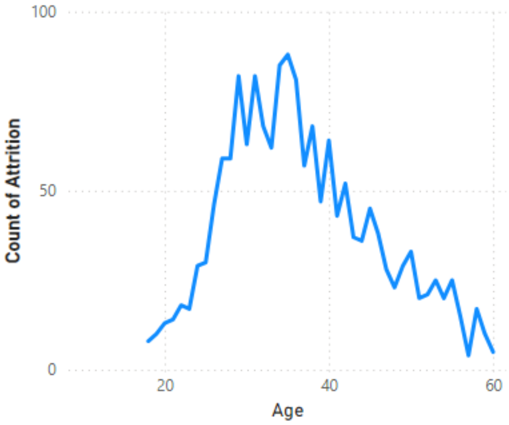
Section C.2

Overtime: Attrition rates were higher given the employee regularly reported overtime.



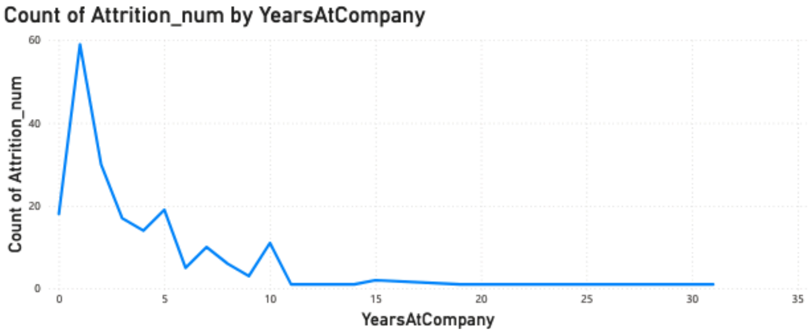
Section C.3

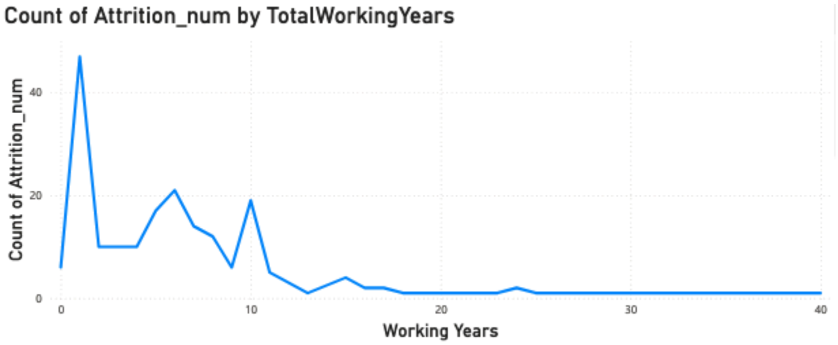
Age: Workers who were aged 25-40 were associated with higher attrition rates:



Section C.4

Years at company: Higher attrition rates were associated with employees who were new to the company and earlier within their careers.


Count of Attrition_num by YearsAtCompany


Count of Attrition_num by TotalWorkingYears