Nidhish Nerur
Michael Crosson
Yue Taira
Megha Sengupta
Erin Kim

Predict Nurse Attrition

1. Description of Project Goals

1.1 Dataset Description

The dataset contains employee and hospital data to predict attrition of nurses in the U.S. healthcare system. There are a useful range of features including the nurse's personal characteristics (e.g., age, relationship satisfaction), workplace satisfaction (e.g., environment satisfaction, job satisfaction), and potential for job growth (e.g., salary growth, years since last promotion). The questions we aim to answer include 1) Are there patterns to better understand which nurses are likely to leave a healthcare facility?, and 2) What features are the most important in predicting whether a nurse is likely to leave a healthcare facility?

1.2 Importance of the Problem

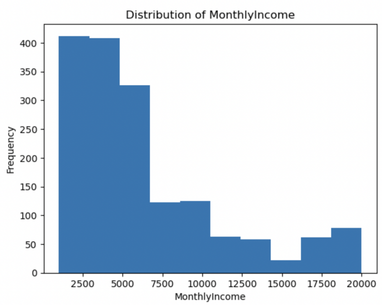
Nurse attrition has risen dramatically with the on-set of COVID-19 and the U.S. is expected to have a shortage of nurses in the coming years. The dataset has enough variety to identify specific reasons nurses may leave a healthcare facility. Our work will benefit hospitals focused on retaining talent, nurses looking for greater fulfillment, and board of directors acting in the hospital's best interest. We will determine features with the greatest predictive power to help healthcare facilities retain nurses, thereby minimizing re-hiring costs while improving brand reputation. The 2022 National HealthCare Retention & RN Staffing Report argues the cost of nurse attrition is roughly $50,000 per nurse which contributes to significant losses for healthcare organizations. The broader healthcare industry will appreciate our analysis and interpretation, as it could pinpoint key factors furthering nurse attrition and improve retention rates.

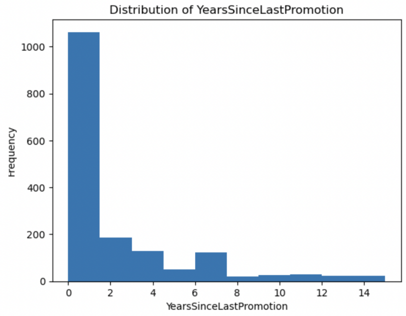
2. Exploratory Analysis

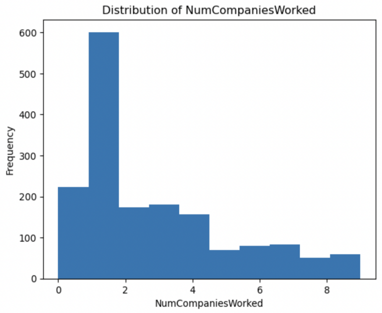
2.1 Descriptive Statistics and Intuition of Dataset

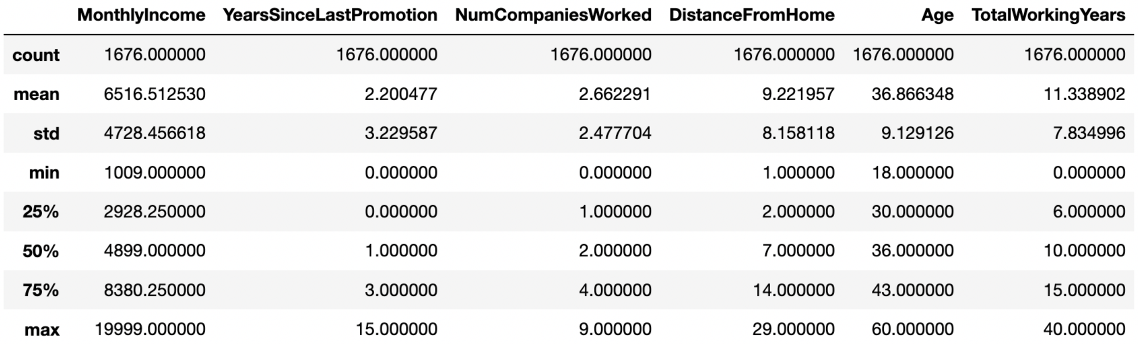
Several of our numeric features are right-skewed with means greater than the medians, and we standardized our variables to ensure our models do not overweight features with wider scales and improve interpretability of our results. The distributions and descriptive statistics of Monthly Income, Years since Last Promotion, and Number of Companies Worked are shown below:

Nidhish Nerur
Michael Crosson
Yue Taira
Megha Sengupta
Erin Kim



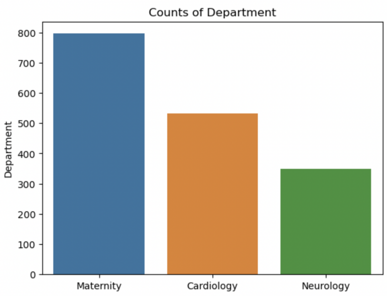


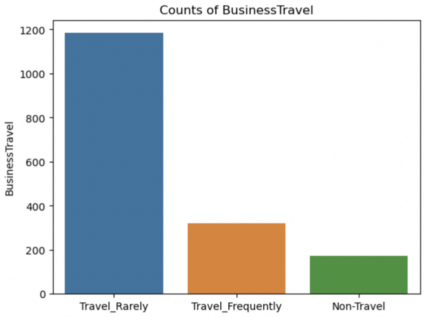


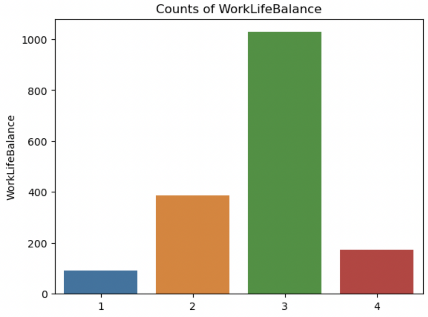


	MonthlyIncome	YearsSinceLastPromotion	NumCompaniesWorked	DistanceFromHome	Age	TotalWorkingYears
count	1676.000000	1676.000000	1676.000000	1676.000000	1676.000000	1676.000000
mean	6516.512530	2.200477	2.662291	9.221957	36.866348	11.338902
std	4728.456618	3.229587	2.477704	8.158118	9.129126	7.834996
min	1009.000000	0.000000	0.000000	1.000000	18.000000	0.000000
25%	2928.250000	0.000000	1.000000	2.000000	30.000000	6.000000
50%	4899.000000	1.000000	2.000000	7.000000	36.000000	10.000000
75%	8380.250000	3.000000	4.000000	14.000000	43.000000	15.000000
max	19999.000000	15.000000	9.000000	29.000000	60.000000	40.000000

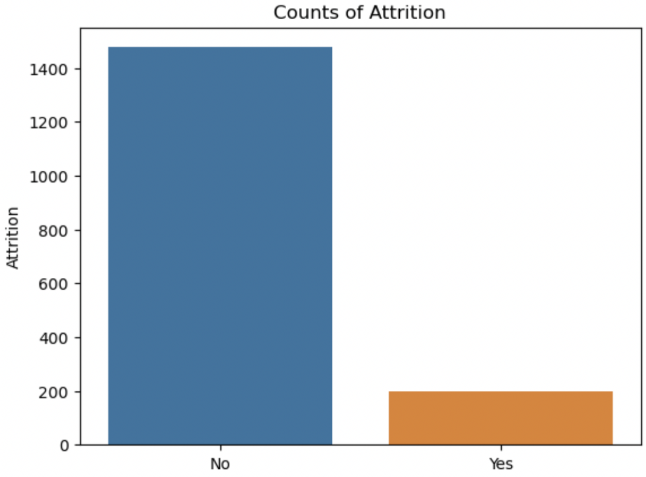
The averages indicate most nurses are in their late 30s, live within 10 miles of the hospital, are paid around $65 an hour, and have worked for more than one company with around 11 years of experience. These features could impact attrition rates as nurses may seek higher pay or greater satisfaction with a different employer. We explored the categorical features as well to find patterns with attrition rate. The variables had multiple levels and we applied one-hot encoding to create binary dummy variables and improve our predictions. For instance, there were several Departments employing nurses, levels of Business Travel, and ratings for Work Life Balance with 4 being the highest and 1 the lowest seen here:







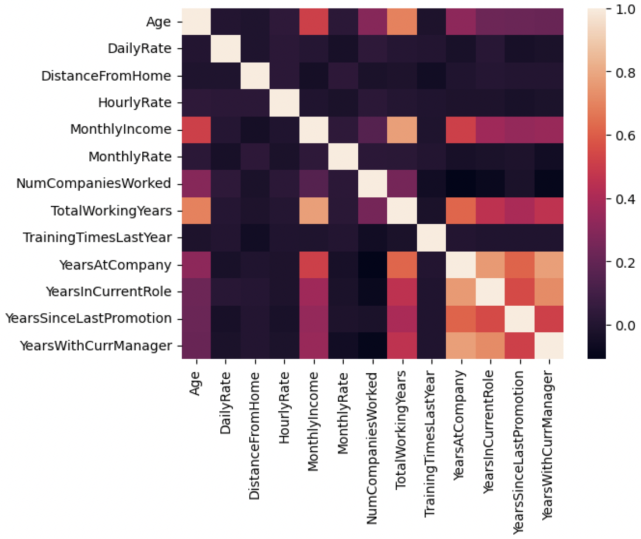
Our outcome variable of nurse attrition is imbalanced with 199 nurses leaving the hospital and 1477 staying, yielding a roughly 12% attrition rate.



We applied the Synthetic Minority Oversampling Technique (SMOTE) to generate new examples for the minority class and balance the outcome variable. SMOTE was applied only to the training set examples, as the test set must remain completely unseen. SMOTE improved the model performance and interpretation of our results.

2.2 Correlation Matrix between Features

We created a heatmap with the Seaborn package where the lighter boxes correspond with greater correlations between the numeric features of the dataset.



We observe that MonthlyIncome paired with TotalWorkingYears, YearsAtCompany paired with YearsAtCurrentRole, and YearsAtCompany paired with YearsWithCurrentManager had high correlations among our features. Consequently, certain features relate to the amount of time nurses will spend at a given hospital and these variables may add value to predict nurse attrition.

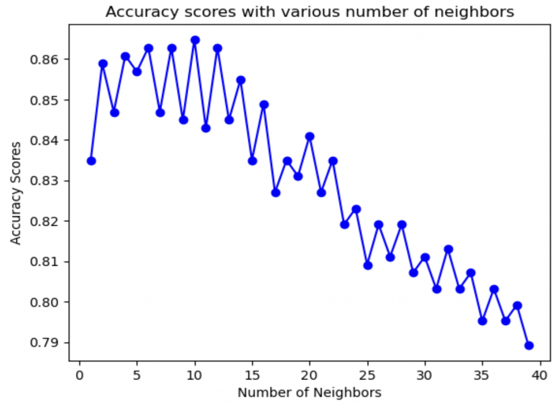
3. Solution and Insights

3.1 Features Rationale

We decided to remove the features containing values which were the same for each row as they would not add any predictive value. Consequently, the EmployeeCount, Over18, and StandardHours were removed. Additionally, we dropped the EmployeeID column as it was unique for each nurse and will not add any predictive power. We used all the remaining features to predict nurse attrition as they related to the nurse’s personal life, work characteristics, and education level. We standardized all features, applied one-hot encoding to the categorical variables, and handled class imbalance before model development.

3.2 Classifiers Tested and Summary of Results

The baseline accuracy when predicting all nurses will not leave the hospital is approximately 87%, suggesting we want our model to have higher accuracy to provide predictive value. We tested the K-nearest neighbors, Logistic Regression, Decision Tree, Random Forest, and XGBoost classifiers. For the KNN classifier, we determined the optimal number of neighbors with this plot.



The plot suggests 9 nearest neighbors yield a testing accuracy of approximately 87%, which is equal to the baseline accuracy, thereby not adding any value. In contrast, the Logistic Regression, Decision Tree, Random Forest, and XGBoost classifiers had testing accuracies above 90%, meaning they were able to identify patterns as to why nurses may leave their place of work. Logistic Regression is our optimal model with a training accuracy of ~96% and test accuracy of ~94%. For the positive class, the precision is low at 76% conveying the percent of nurses who were predicted to turnover out of those who actually left their place of work. Similarly, the recall score is also low at 73% reflecting the percent of nurses who actually left their place of work out of those predicted to turnover. Results for the Logistic Regression model:

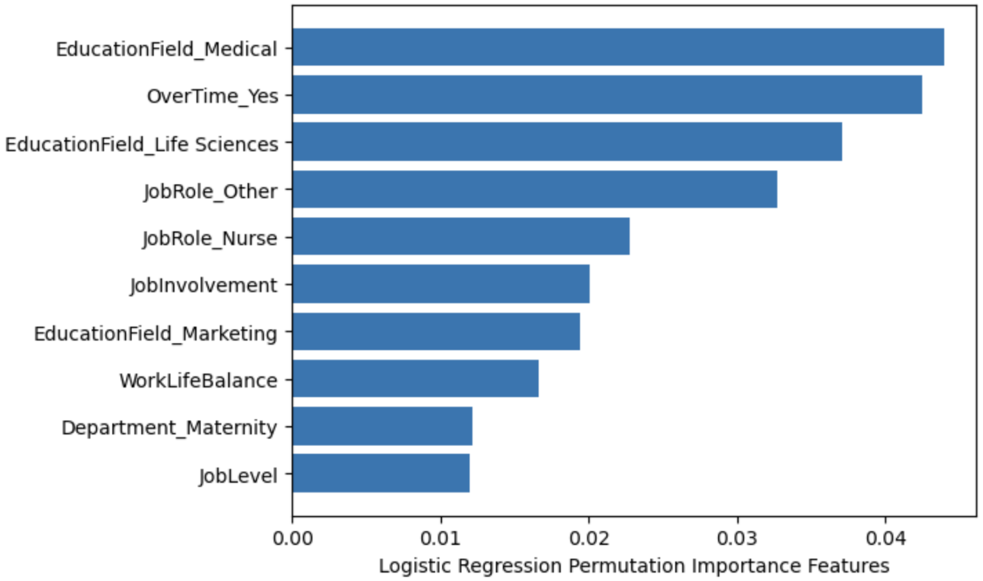
	Precision Score	Recall Score
0 (Nurse stays in hospital)	0.96	0.97
1 (Nurse leaves hospital)	0.76	0.73

Confusion Matrix for Logistic Regression	Predicted 0	Predicted 1
0 (Nurse stays in hospital)	429	14
1 (Nurse leaves hospital)	16	44

3.3 Model Analysis

We created a Permutation Importance plot for the Logistic Regression model by looking at how metrics such as accuracy, precision, and recall change when each variable is dropped independently:

Nidhish Nerur
Michael Crosson
Yue Taira
Megha Sengupta
Erin Kim



Nurses who received an education in the medical or life sciences field, tend to work overtime, and are heavily involved in their job tend to be predictors of nurse attrition. Perhaps nurses with a background in medicine are more familiar with other career opportunities and growth potential, so they may decide to pursue a different career path. Additionally, nurses who work long hours and night shifts may feel burnout, contributing to higher attrition rates. The least valuable features were training time, marital status of single, monthly rate, and gender likely since single female nurses have committed a significant amount of time to nursing and do not wish to learn skills for another occupation. Our team was surprised that Logistic Regression is our optimal model; however, this is likely because we need more data and the outcome variable was highly imbalanced. In the future, we aim to analyze aggregate data from various hospitals to assess the validity of our results and identify more nurse attrition patterns.

Nidhish Nerur
Michael Crosson
Yue Taira
Megha Sengupta
Erin Kim

Reference Page

"7 Ways Employee Attrition Rate Affects Business Advancement."

Whatishumanresource.com, https://www.whatishumanresource.com/7-ways-employee-attrition-rate-affects-business-advancement.

Brownlee, Jason. "Smote for Imbalanced Classification with Python."

MachineLearningMastery.com, 16 Mar. 2021, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

Gamble, Molly. "The Cost of Nurse Turnover in 23 Numbers." *Becker's*

Hospital Review, https://www.beckershospitalreview.com/finance/the-cost-of-nurse-turnover-in-23-numbers.html#:~:text=The%20average%20cost%20of%20turnover%20for%20a%20staff%20RN%20is,range%20averaging%20%2433%2C900%20to%20%2458%2C300.

JohnM. "Employee Attrition for Healthcare." *Kaggle*, 15 Feb. 2023,

https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare?select=watson_healthcare_modified.csv.

"Permutation Importance¶." *Permutation Importance - ELI5 0.11.0*

Documentation, https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html.