# Nurse Disengagement Analysis

Jacquie Furbish, Aiyana Weekes-Tulloch, Chau Nguyen
Wentworth Institute of Technology
550 Huntington Ave, Boston, MA 02115
Professor: Dr.Semere Gebresilasie
Industrial Sponsor: IntelyCare, Inc.
{furbishj, nguyenc23, weekestullocha}@wit.edu
April 17, 2020

## Abstract

IntelyCare is a platform that provides on-demand nurse for healthcare facilities, aiming at giving registered nurses the flexibility in their schedule. While the thought of taking shifts to their liking sounds encouraging, IntelyCare is having trouble keeping their nurses consistently engaged after the 6-month period, some for even shorter periods of time. This study focuses on the potential reasons that could be causing their nurses to become disengaged. Some methods that will be used are Non-Linear Least Squares with a testing and training group, Decision Trees, Logistic Regression, and Survival Analysis using Kaplan Meier estimator.


## Introduction

IntelyCare is a fast-growing company made to assist nurses in finding shifts that work for their schedule. In 2018, IntelyCare hired approximately 1,000 nurses a month. Since hiring registered nurses is a costly procedure, and with such a large number hiring's every month, IntelyCare is always finding ways to provide their staff with the best experience to keep them engaged. In order to provide them with the best experience, they need to understand the reason their nurses become disengaged after 6 months (roughly 27 weeks). This study focuses on the factors that may disengage the nurses, such as cancellation, periodic bonuses, customer service, number of messages received etc. Attention to these factors will help create a model to predict when such event happens and encourages nurses to stay engaged and active in the application.

Similarities as well as differences in the nurses that became disengaged are analyzed and compared to the nurses that are still engaged after 6 months of employment using statistics and probability. In order to test these nurses and predict their disengagement, we defined active as cancelling a shift, completing a shift, accepting a new shift, and being cancelled on last minute by the facility. This idea will be tested using different methods in order to find trends and predict disengagement.

Non-linear least squares (NLS) is a great method to use because it can get fitted to the data set easily while allowing us to input various different categories for analysis. This is a prediction model to help determine whether a nurse will become disengaged after six months based on their activity, with completed shifts and released shifts being the explanatory variables, while activity is the response variable. The coefficients will be determined using a random sample of 60% of the nurses. The remaining 40% is used for testing the model's accuracy. These results will show whether it is predicted that the nurse will be disengaged after 6 months or not. The predicted results of inactivity after 4-months and 6-months will then be compared with their true results to determine how accurate the model is.

Decision trees create models that represent the possible paths that can be taken in order to reach specific outcomes. For this problem it is important to gain an understanding of the paths a nurse takes before becoming inactive. This will assist it determining the warning signs that a nurse is about to become disengaged from the company.

Kaplan-Meier estimators are a form of survival analysis. This provides visuals as well as probabilities that a subject will survive or not. In this case, nurses will be analyzed in order to predicted whether they will survive (remain active with the company), or die (become disengaged with the company).

## Non-Linear Least Squares

Non-linear least square is an analysis method used to fit non-linear observations based on several parameters.

Since our goal is to predict disengagement, we create a column in the data set that is only filled with 1's and 0's, with 1 being if the nurse matched our definition of active, and 0 otherwise. We choose our response variable to be activity, and simple linear regression would not give us the results we want, which are 1's and 0's. For that reason alone, we choose a special nonlinear model – logistic regression.

Activity has already been classified prior to this process. To test out our hypothesis, we picked a random sample of 60% of our subset of nurses and used the logistic regression model,

to obtain the coefficients of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, with $X_1$ being completed shifts and $X_2$ being released shifts. This was then tested on the remaining 40% of our nurses. As a result, we were able to acquire some accurate predictions using the sections of data with the best p-value. The probability of each nurse being active after 6 months was obtained. This was then compared to whether the nurse was active or not after the six months. We have 2 ways of testing accuracy – where less than 50% is considered disengaged and vice versa, or less than or equal to 25% is considered disengaged and greater or equal to 75% is considered engaged. The test was run multiple times and the average accuracy was calculated from that.

We subset a 6-month period of the data set given, from 10/07/2018 to 04/07/2018, which gives us a total of 420 nurses. We sum up this data set and start testing it out with the next 6-month and the next 2-week period. With the 50% case, we get 68% and 63% accuracy. With the 25% and 75% case, we get 78% and 68%.

We also subset a 4-month period of the data set given, from 10/07/2018 to 02/07/2018, which gives us a total of 307 nurses. Much like the 6-month data set, we sum this data set before testing it out with the same periods. With the 50% case, we get 66% and 75%. With the 25% and 75% case, we get 66% and 81%.

While the $25^{th}$ and $75^{th}$ percentile case yields better accuracy, it is good to keep in mind that there are fewer probabilities that lie in these two areas, as there are in the middle $50^{th}$ percentile. Since our aim was to correctly predict disengagement, we can't overlook this middle section of the data. We use the same logistic regression model, but the more relevant explanatory variables for this model are completed shifts and cancelled shifts. While they have a low p-value, indicating their significance to the model, they are only moderate at predicting activity. Using the 50% method, this model is only 50% accurate. Using the 25% and 75% method, we encounter the same problem as last time, where there are not a lot of value within these two areas, sometimes none, that the test is inconclusive; the times that there are probabilities within, it is 70% correct.

| pid | future_shifts | completed_shifts | cancelled_shifts | released_shifts | active_after_6(1-Yes, 0-No) | disengage_probability | true_or_not |
|---|---|---|---|---|---|---|---|
| 5214106721 | 0 | 0 | 0 | 12 | 0 | 0.044964984 | 1 |
| 5214584725 | 4 | 1 | 0 | 1 | 1 | 0.542888903 | 1 |
| 5217726727 | 1 | 1 | 0 | 0 | 0 | 0.503705648 | 0 |
| 5219187721 | 0 | 0 | 0 | 1 | 0 | 0.354980066 | 1 |
| 5219762727 | 0 | 0 | 0 | 1 | 1 | 0.354980066 | 0 |
| 5224164721 | 24 | 1 | 3 | 16 | 0 | 0.441987476 | 1 |
| 5225674726 | 50 | 25 | 6 | 1 | 1 | 0.804239941 | 1 |
| 5225865728 | 0 | 2 | 0 | 4 | 0 | 0.360171206 | 1 |
| 5226853728 | 10 | 1 | 1 | 2 | 1 | 0.644611683 | 1 |
| 5228458724 | 27 | 35 | 3 | 8 | 1 | 0.804239752 | 1 |

All the data was imported from Microsoft Excel and modeled using RStudio.

## Decision Trees

Decision trees provide visuals as well as probabilities of various events taking place. These visuals are represented as diagrams that start with one action that branches off to two other possible actions that depend on what had taken place. These actions then split again, just as branches on a tree. The result is the predicted number of outcomes in each category that were in question.

In this problem, the outcome in question is whether a nurse will be active. Engagement will be compared to various categories in order to determine the probability of each nurse being active after six months. Two different types of decision trees will be used in this problem. One being a conditional decision tree and the other being traditional. A conditional decision tree calculates the strongest given factors in order to eliminate bias. This is done through significance tests. Traditional decision trees calculate all the variables given and does not alter anything. Conditional decision trees are generally more accurate however with this problem, the opposite was discovered.

In order to obtain the decision tree diagram, nurses' activity over the first six months of future shifts, completed shifts, cancelled shifts, and released shifts were compared to whether they were active after 6 months or not. The following diagram was produced.

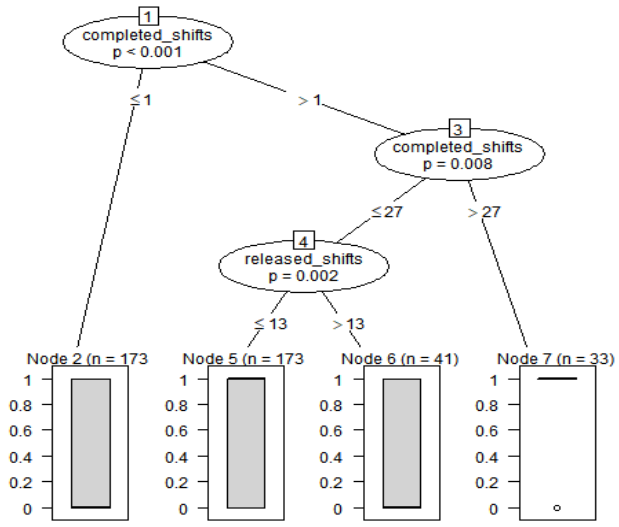



*Figure 1 Image of Conditional Decision Tree*

This diagram is a conditional decision tree and shows the first two options as completing more than one shift and less than or equal to one shift. It then states the P value of this occurrence a less than 0.001. If the nurse completes one or no shifts, then they are predicted to be active after six months while if they complete more than one shift it moves to another possible option. This option is again completing a shift however the p value is slightly higher at 0.008. Having a slightly higher p value means there should be a stronger ability to predict the outcome. It then branches off to if a nurse completes more than 27 shifts and completing 27 shifts or less. If the nurse completes more than 27 shifts, then it is predicted that the nurse will become inactive after six months. If the nurse completes 27 or less shifts, then it moves onto analyzing released shifts. This has a p value of 0.002 however both outcomes result in the nurse being active. These results were then used in order to predict whether a nurse would be active or not which a table was then created from. The first column displays the probability of whether a nurse will be active or not, the second column displays if they were active after six months, and the third column displays whether the prediction was accurate or not.

| active_after_six | decision_tree.active_after_six | True? |
|---|---|---|
| 0.96969697 | 1 | 1 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 1 | 0 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 1 | 0 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 1 | 0 |
| 0.375722543 | 1 | 0 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 0 | 1 |
| 0.375722543 | 1 | 0 |
| 0.375722543 | 0 | 1 |

*Figure 2 Data Table Result*

The results from this table were calculated to be 68.81% accurate in predicting whether a nurse will be disengaged after 6 months. Another model of decision trees was created in order to make a more accurate prediction.

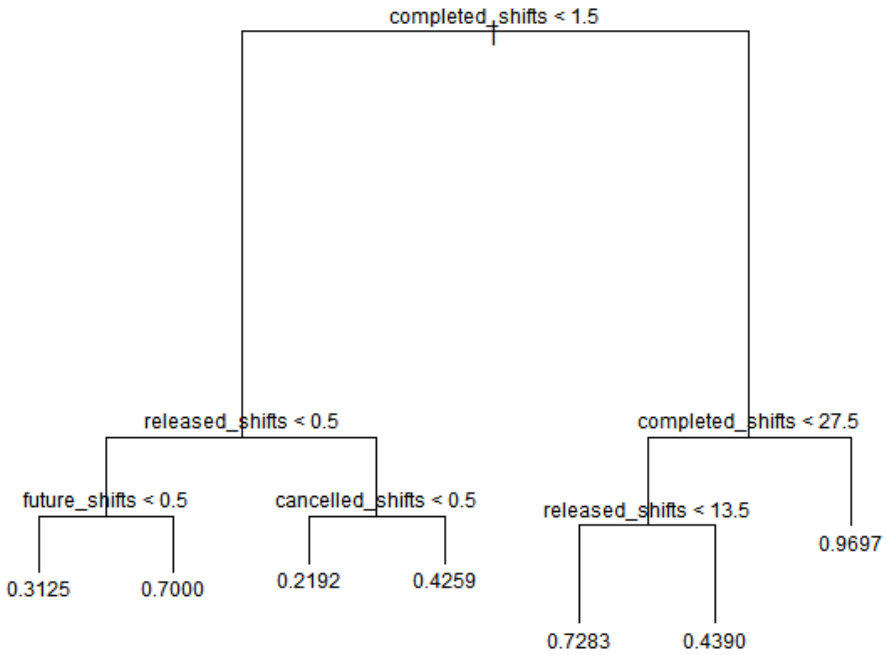



*Figure 3 Decision Tree*

This model is a traditional decision tree and represents all possible actions that are being analyzed in the current data set. The possible actions to take start with completed shifts and split onto released shifts and completed shifts and a deciding factor of 1.5. This means if the nurse completes less than 1.5 shifts (which can be viewed as one or less because there is no existing data with partially completed shifts) then the next option is released shifts, while the other option is completed shifts. Released shits has a deciding factor of 0.5 with the two options as future shifts and cancelled shifts. Completed shifts has a deciding factor of 27.5 with the options as released shifts and being inactive. Both future shifts and cancelled shifts has a deciding factor of 0.5 with the two options as active and inactive. Released shifts has a deciding factor of 13.5 with two options as active and inactive. These end results all display the probability of it occurring. From these results, another table was created in or to calculate the accuracy of the new predictions. Just as the other chart, the first column displays the probability of whether a nurse will be active or not, the second column displays if they were active after six months, and the third column displays whether the prediction was accurate or not.

| predict.output. | decision_tree.active_after_six | True? |
|---|---|---|
| 0.96969697 | 1 | 1 |
| 0.219178082 | 0 | 1 |
| 0.219178082 | 1 | 0 |
| 0.7 | 0 | 0 |
| 0.425925926 | 1 | 0 |
| 0.219178082 | 0 | 1 |
| 0.219178082 | 0 | 1 |
| 0.425925926 | 1 | 0 |
| 0.219178082 | 1 | 0 |
| 0.3125 | 0 | 1 |
| 0.425925926 | 0 | 1 |
| 0.425925926 | 0 | 1 |
| 0.219178082 | 0 | 1 |
| 0.7 | 1 | 1 |
| 0.425925926 | 0 | 1 |

*Figure 4 Data Table Results*

This model was discovered to be slightly more accurate with a 71.67% accuracy.

Since the traditional decision tree had more accurate results, this model was focused on for further analysis. I obtain a more accurate model; the results were split into 3 sections based on the probabilities. Probabilities less than or equal to .25, greater than or equal to .75, and the area between. If the accuracy of the two end margins were then analyzed and it was calculated to be 83.96% accurate.

Seeing that the accuracy in predicting a nurse outcome was increasing, a new data set was then made. This data set sums up the first two months of the nurse's data instead of the first six. The idea of this new data set is to try and predict the nurse's inactivity sooner. It was then split into two groups for testing and training. The training group is a randomly selected 60% of the nurses while the testing is the remaining 40%. The training group was then fit to a traditional decision tree. Results from this were then used to predict the remaining 40% in order to test the accuracy. The output was as follows:

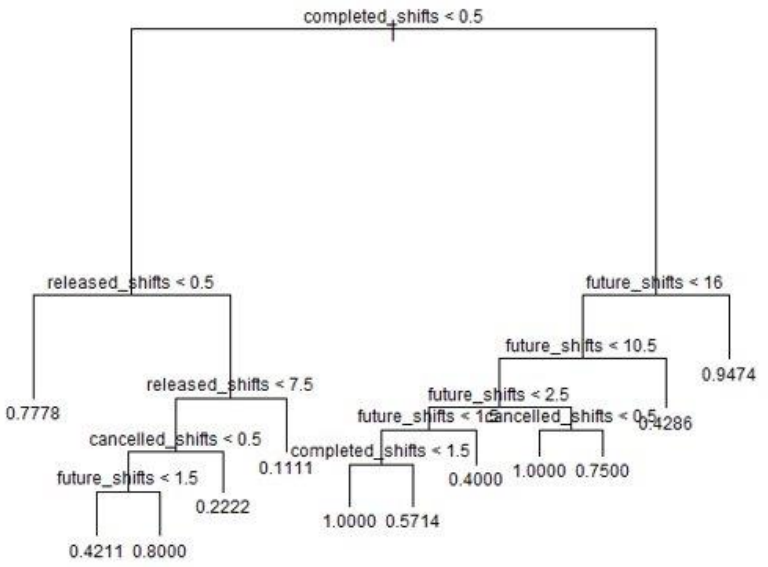



*Figure 5 Traditional Decision Tree from 2 months of Data*

A sample of the results:

| | | | |
|---|---|---|---|
| 2 | 0.96969697 | 1 | 1 |
| 3 | 0.375722543 | 0 | 1 |
| 4 | 0.375722543 | 1 | 0 |
| 5 | 0.375722543 | 0 | 1 |
| 6 | 0.375722543 | 1 | 0 |
| 7 | 0.375722543 | 0 | 1 |
| 8 | 0.375722543 | 0 | 1 |
| 9 | 0.375722543 | 1 | 0 |
| 10 | 0.375722543 | 1 | 0 |
| 11 | 0.375722543 | 0 | 1 |
| 12 | 0.375722543 | 0 | 1 |
| 13 | 0.375722543 | 0 | 1 |
| 14 | 0.375722543 | 0 | 1 |
| 15 | 0.375722543 | 1 | 0 |
| 16 | 0.375722543 | 0 | 1 |
| 17 | 0.728323699 | 1 | 1 |
| 18 | 0.728323699 | 1 | 1 |
| 19 | 0.728323699 | 0 | 0 |

*Figure 6 Sample Table of Results from 2 months of Data*

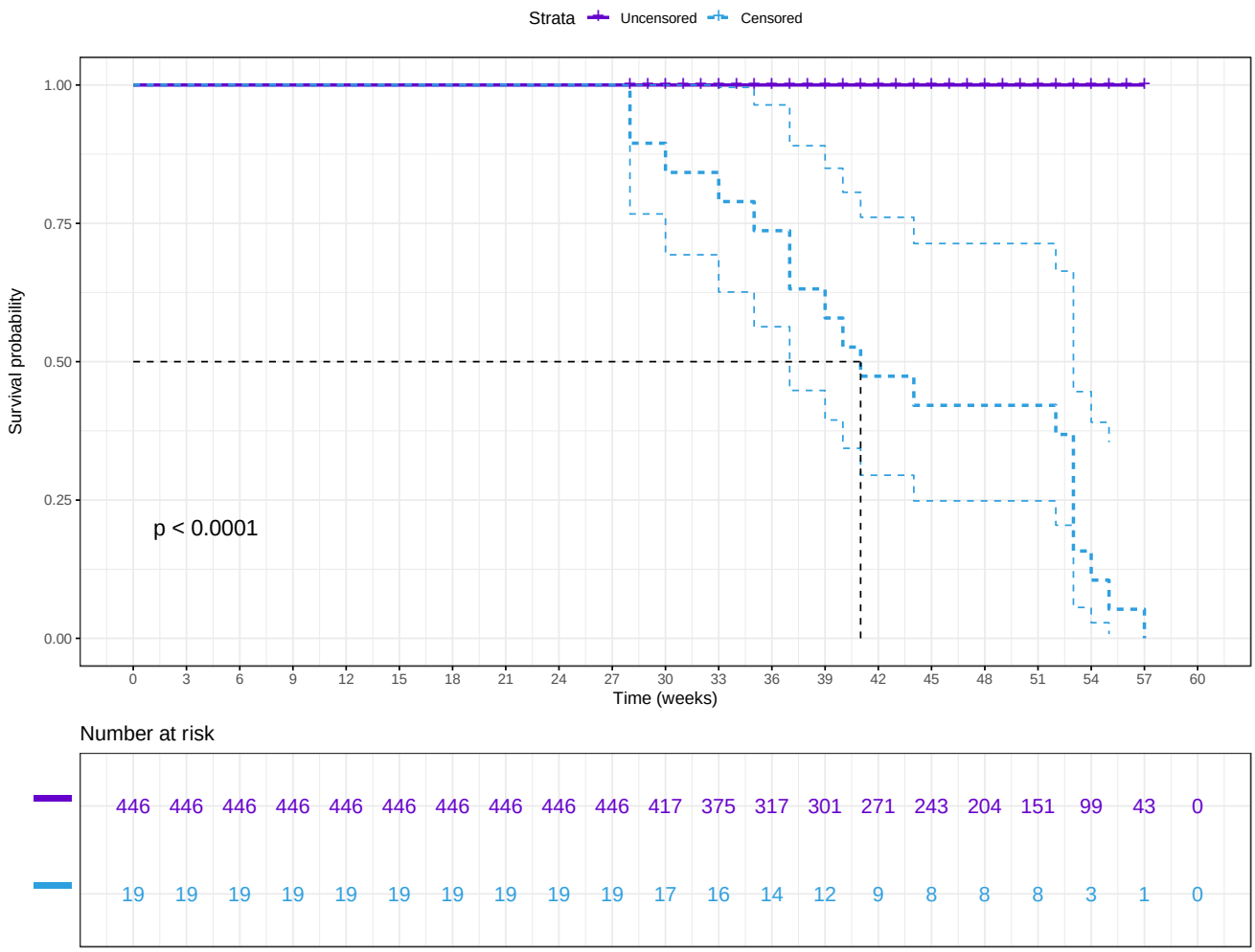
The accuracy was calculated by deciding that a probability greater than or equal to .5 mean active, while a probability below .5 is inactive. This was compared to if the nurse had become inactive or not. A percent was then calculated, and it was determined that the decision tree created is 83.67% accurate in predicted disengagement after just two months.

## Kaplan-Meier Estimator

The next method we used was the Kaplan-Meier Estimator. The Kaplan-Meier Estimator is a survival analysis function that provides information on the probability of a subject or subjects of interest surviving beyond any specific time interval. We used this estimator to approximate the number of nurses that were continuously active throughout the 6-months and were most likely to remain active after 6-months (roughly 27 weeks) on the app. The independent variable being time and the dependent variable being status, i.e. censored or uncensored. A nurse was considered censored if they completed their entire first 6 months on the IntelyCare app without being inactive. On the other hand, a nurse was considered uncensored if they were inactive for any amount of time within that 27-week period. The survival function used to obtain these results was:

- ✓ $S(t_i) = \mathrm{S}(t_{i-1})(1 - \frac{d_i}{n_i})$ .
  - $S(t_i - 1) =$ the probability of being alive at $t_i - 1$
  - $n_i$ = the number of patients alive just before $t_i$
  - $d_i$ = the number of events at $t_i$
  - $t_0 = 0,\ S_0 = 1$

In order to arrive at these results, we had to collect data on the total weeks the nurses were on the app based on the data provided by IntelyCare. We collected data on which nurses were continuously active throughout the 6-months and which nurses were not active the entire 6-month interval. That data is what we called censored vs uncensored. When implementing all the data into the estimator, using the function $Survfit()$ and $Surv()$, in RStudio, two survival functions were provided, and the following graph is what we got as a result:




# Results

Overall, the best method uses was the decision trees as it predicted the most accurate results out of all the methods used. For future testing, we would continue the path of the Decision tree and focus on strengthening the accuracy of the predictions. In addition, we would use random forests to help predict inactivity as well as accurately predicting what factors cause inactivity to give IntelyCare the means to provide the best long-term work experience for their nurses.

# Acknowledgements

Support for this Mathematical Association of America (MAA) program is provided by the National Science Foundation (NSF grant DMS-1722275) and the National Security Agency (NSA).

# References


"(Tutorial) Survival ANALYSIS in R For BEGINNERS." *DataCamp Community*, www.datacamp.com/community/tutorials/survival-analysis-R.

"Decision Tree." *R*, www.tutorialspoint.com/r/r_decision_tree.htm.

Gupta, Prashant. "Decision Trees in Machine Learning." *Medium*, Towards Data Science, 12 Nov. 2017, towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

Hertzog, Lionel. "First Steps with Non-Linear Regression in R." *DataScience* , 30 Jan. 2018, datascienceplus.com/first-steps-with-non-linear-regression-in-r/.

James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2017.

Paemel, Ruben Van. "Kaplan Meier Curves." *Medium*, Towards Data Science, 26 July 2019, towardsdatascience.com/kaplan-meier-curves-c5768e349479.

Wollschlaeger, Daniel. "R Examples Repository." *Survival Analysis: Kaplan-Meier*, dwoll.de/rexrepos/posts/survivalKM.html.

"Stats." *Function | R Documentation*, www.rdocumentation.org/packages/stats/versions/3.6.2/topics/nls.