Final Executive Summary

Werner Enterprises, a transportation and logistics company headquartered in Omaha, has difficulty retaining freight drivers. This is not a company issue, but rather a problem faced industry wide. Sometimes Werner has very little notice that a driver will quit before it happens. This can interrupt Werner's workflow and can cost the company time and money. How can Werner know of these quitting drivers early enough to either persuade the driver to stay or find a replacement?

The goal of this analysis is to accurately predict which drivers are going to quit within 30 days. The model will focus on predicting drivers who quit voluntarily, rather than those who are fired.

First and foremost, it is important that we define what exactly will be predicted. To determine which drivers will quit within 30 days, we found the difference between the date the driver quit and each day the driver reported to work. If this difference was less than 30, the driver was flagged to quit within 30 days. This is the value that was predicted in the models.

Another important point to note is that the number of flagged drivers and the number of non-flagged drivers was not balanced; that is, there were about six times more non-flagged drivers than there were flagged drivers. Thus, a model would be skewed to predict a driver as non-flagged more often than flagged. To account for this imbalance, we used several metrics to validate our models including precision, recall, F-measure, G-means, and weighted accuracy. A good model would have all these measures as close to 1 as possible.

We built several models and found the "best" one to have a weighted accuracy of 0.95 in correctly classifying drivers as likely to quit within 30 days. In this model, we correctly classified 16649 of 17717 drivers who did not quit and 2944 of 3084 drivers who did quit. However, there was also some misclassification of drivers. The model predicted 1068 drivers would quit when they did not and 140 drivers to not quit when in fact they did. No model will ever be a perfect classifier to drivers who quit and do not quit. Given Werner's purposes, we prefer the error of predicting a driver will quit when truly the do not, rather than predicting a driver will not quit when in fact they do. We decided that the former error would be less detrimental to Werner overall. However, in practice, this is completely Werner's decision.

Going forward from our findings in this project, we think it would be beneficial to Werner Enterprises if an application or executable were built using this model. This way, Werner could stay in the know on their expected employee attrition and how many new drives to hire. An added benefit of an application would be that employees with no background in coding could easily run the prediction.