Introduction of problem - Driver inattentiveness results in considerable loss of life for drivers, passengers, and bystanders all around the world. Newer models of vehicles incorporate 'inattentiveness' countermeasures alerting drivers to changing conditions that should be responded to, whether that be staying on your own side of the center line, responding to another’s sudden stop, or the random animal that darts into a driver's path. In our increasingly safety-conscious world, having a top-notch predictive safety system can be a strategic marketing tool for the improvement of customer welfare as well as sales.

Interested parties - Outside of the obvious attention to this problem from automobile manufacturers, those interested in optimized alertness solutions also include international equivalents of the NTSB (National Transportation Safety Board) and publications like Consumer Reports. The former is charged with determining probable causes of transportation mishaps and promoting transportation safety while the latter employs scoring systems comparing performance across a wide range of consumer products.

Description of dataset and intended usage - Aside from trial ID and observation ID numbers, the datasets provided by Ford contain a binary response (Alert/Not Alert) in conjunction with three sets of predictive variables grouped under the headings of physiological data, vehicular data, and environmental data. There are eight variables within the physiological group and eleven each within the other two groups. Consistent with those group profiles, the true predictor source names have been masked by Ford and are labeled P1-P8, V1-V11, and E1-E11.

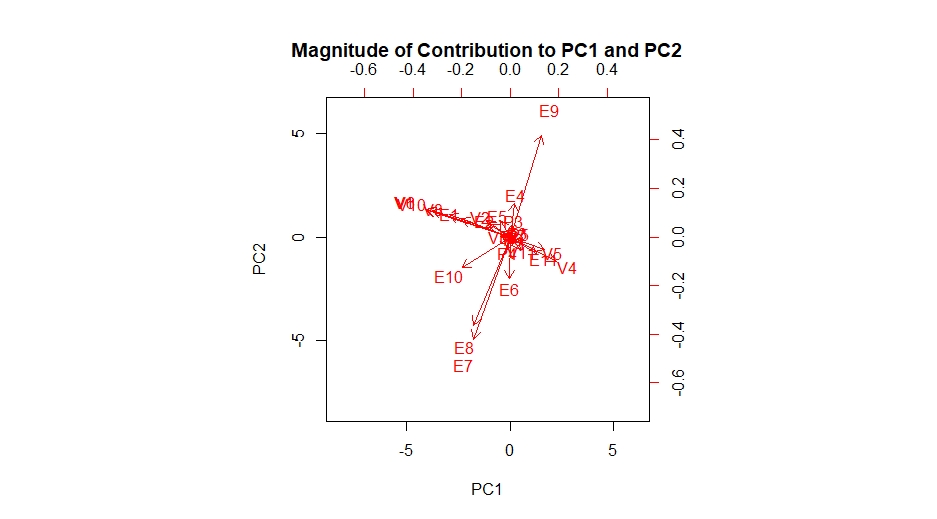
Three .csv datasets are available: **train**, consisting of just over 600k records, as well as **test** and **solution** files. Both contain over 121k records. **Test** has the binary response withheld and its value is provided in the **solution** file. For the interests of time, instead of a 600k train / 121k test analysis across which to do model parameter selection, the data from **train** was randomly split up into a 200k record train and 400k test validation. The decision to proceed in this fashion was due to the length of time to perform model selection over the much smaller 200k training set: over 10 hours of processing.

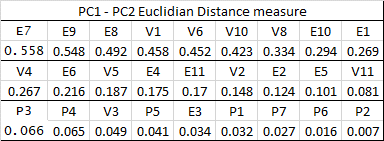
Basic exploratory data analysis - A summary of the train.csv data set revealed that there were no missing data values that would have otherwise resulted in either observation removal or in data imputation (providing an expected value of the missing data from the context provided by the data without missing values, assuming random and not systematic reasons for their absence). Further, three of the variables (P8, V7, and V9) had nothing but zeros in them, resulting in their removal from the data set. This left 27 predictors from which to base further efforts.

A correlation analysis between all the predictors in the 200k training set found that only four out of the 351 combinations (removing the diagonal and one of the halves of a 27x27 matrix) of predictors had an absolute correlation greater than .80. These combinations are:

P3 x P4: -0.94436, V6 x V1: 0.93797, V10 x V1: 0.91121, V10 x V6: 0.82412.

In the case of the first combination, as P3 trends strongly up, P4 strongly trends down and vice versa. For the remaining three combinations, both components trend strongly together in the same direction.

Results of exploratory Principal Component Analysis (PCA) – The goal of PCA is to identify the predictors contributing the most to variance by remapping them into a new coordinate plane. The first axis in the plane is the weighting of the predictors that results in the most variance, the second axis is the weighting that contributes the next-most amount of variance, and so forth. It is calculated for the smaller of p, the number of our (27) predictors or n-1, one less than the number of observations. Taking the contribution for our first two principal components, the contribution of our predictors becomes clearer. Variables E7, E8, and E9 have visibly longer vectors in PC1-PC2 space. This is borne out by calculating the Euclidian distance *(PC12 + PC22)1/2* from the origin to the vector’s endpoint:



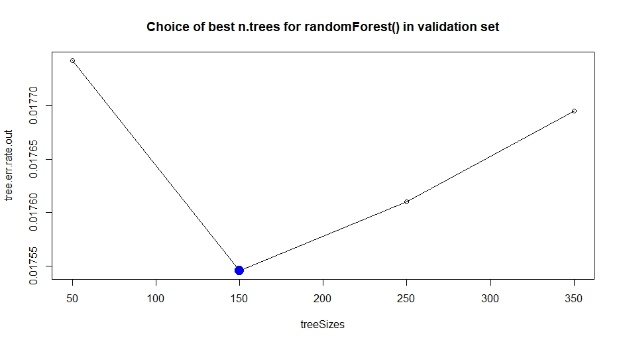
Some models will be made to ignore lesser-contributing predictors during model selection.

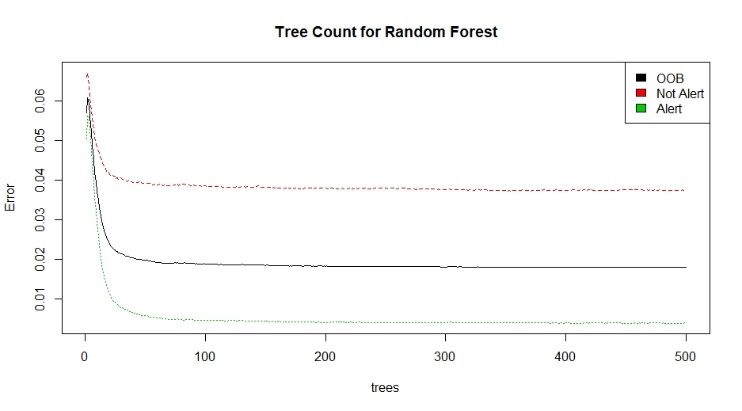
Model selection – There are three types of classification model considered in this analysis. All three use different methods to take the 27 predictors and identify an ‘alert’ or ‘not alert’ response variable. The first is logistic regression, a method commonly used to predict yes/no responses. It will predict well if the relationship between the predictors and the predicted log-likelihood of the response is linear. Two versions of logistic regression will be attempted: the first uses all predictors, and the other leaves out the six least-contributing predictors listed above (P5, E3, P1, P7, P6 and P2) as determined by PCA.

The second is k-nearest neighbors (k-nn), which will predict better when that relationship is non-linear. K-nn assigns predictive values in the test data that depend on the higher count of responses out of the k-nearest data in the training dataset. The value of k is varied (5, 7, 9, 11) during model selection to determine which value of k yields the least predictive error.

The last method is random forest, which creates a hierarchical tree-based display of which values of predictors lead to solutions. Much like with k-nn, the value of the number of trees that best reduces the error rate is varied (50, 150, 250, and 350 trees) during the model selection process.

For all methods, a CV(5) cross-validation evaluation is used. For each different value of varied parameter in each model type, one-fifth of each data set is repeatedly held out to test its error rate. Given the size of the data set and the 200k train and 400k test, a second outer cross-validation loop is not necessary.

Model performance – Cross-validation error rates during model selection show a distinct preference towards random forest. It had a much lower error rate than k-nn, and substantially lower than either of the two logistic regression alternatives. This leads one to believe that the relationship between the predictors and the response variable is highly non-linear. Of the random forest tree sizes considered, 150 trees yielded the smallest amount of error. Testing different ranges of tree sizes could potentially yield a better error rate.

Application of validation set – The approximately 400k observations used to validate the effectiveness of the 150-tree model trained on the remaining 200k observations resulted in the following confusion matrix, with accuracies illustrated in the random forest plot:

Predicted

Actual 0 1

0 163885 6271

1 860 231870

Overall Predictive Accuracy is 98.23 percent, or (163885+231870)/402886

Correctly predicted as not alert (row 1)? 96.31 percent, or 163885/(163885+6271)

Correctly predicted as alert (row 2)? 99.63 percent, or 231870/(231870+860)

While a 98.23% overall predictive accuracy speaks well to the performance of the model, the focus must be on the predictive power when the driver is not alert. While this is over 96 percent, its reliability must be improved upon and is the segment that most needs optimized in later iterations of the model. The ‘cost’ of predicting attentiveness when actually inattentive - as occurred 6271 times in this analysis - could result in a failed alert to the driver that they are in a situation risking loss of life. While the cost of predicting inattentiveness when actually attentive - 860 here - is not without its own risks (the vehicle potentially taking control over the driver), the costs are highly determinant on whether the system is alert-only or has the aforementioned control capabilities.

As such, the recommendation stemming from this analysis is that further efforts be launched to:

* Determine if additional data could be captured that serve as better predictors.
* Determine if existing data could be transformed such that it enhances predictive power.
* Thoroughly understand the costs of false negative and false positive predictions.