Variable importance

Random Forests can output a list of predictor variables that they believe to be important in predicting the outcome. If nothing else, you can subset the data to only include the most "important" variables, and use that with another model.
The randomForest package in R has two measures of importance.

- One is "total decrease in node impurities from splitting on the variable, averaged over all trees."
- The other is based on a permutation test. The idea is that if the variable is not important (the null hypothesis), then rearranging the values of that variable will not degrade prediction accuracy.

```
# simulate the data
x1=rnorm(1000); x2=rnorm(1000,x1,1)
y=2*x1+rnorm(1000,0,.5)
df=data.frame(y,x1,x2,x3=rnorm(1000),x4=rnorm(1000),x5=rnorm(1000))

# run the randomForest implementation
library(randomForest)
rf1 <- randomForest(y~., data=df, mtry=2, ntree=50, importance=TRUE)
importance(rf1,type=1)

# run the party implementation
library(party)
cf1 <- cforest(y~.,data=df,control=cforest_unbiased(mtry=2,ntree=50))
varimp(cf1)
varimp(cf1,conditional=TRUE)
```

For the randomForest, the ratio of importance of the the first and second variable is 4.53. For party without accounting for correlation it is 7.35. And accounting for correlation, it is 369.5. The higher ratios are better because it means that the importance of the first variable is more prominent

Variable importance evaluation functions can be separated into two groups: those that use the model information and those that do not. The advantage of using a model-based approach is that is more closely tied to the model performance and that it may be able to incorporate the correlation structure between the predictors into the importance calculation. Regardless of how the importance is calculated:

- For most classification models, each predictor will have a separate variable importance for each class (the exceptions are classification trees, bagged trees and boosted trees).
- All measures of importance are scaled to have a maximum value of 100, unless the scale argument of varImp.train is set to FALSE.

Model Specific Metrics

The following methods for estimating the contribution of each variable to the model are available:

Linear Models the absolute value of the t-statistic for each model parameter is used.

Random Forest from the R package: "For each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all trees, and normalized by the standard error. For regression, the MSE is computed on the out-of-bag data for each tree, and then the same computed after permuting a variable. The differences are averaged and normalized by the standard error. If the standard error is equal to 0 for a variable, the division is not done."