NUMSENSE!

DATA SCIENCE for the LAYMAN

(no math added)

Annalyn Ng & Kenneth Soo

10. Random Forests

10.1 Wisdom of the Crowd

Can several wrongs make a right?

Yes!

While counter-intuitive, this is possible—even expected—for some of the best prediction models.

This plays on the fact that while there are many possible wrong predictions, only one will be correct. By combining models of different strengths and weaknesses, those that yield accurate predictions tend to reinforce each other, while wrong predictions cancel out. This method of combining models to improve prediction accuracy is known as *ensembling*.

We observe this effect in a *random forest*, which is an ensemble of decision trees (see Chapter 9). To show how a random forest is superior to its constituent trees, we generated 1000 possible decision trees to each predict crime in a US city, before comparing their prediction accuracy to that of a random forest grown from the same 1000 trees.

10.2 Example: Forecasting Crime

Open data from the *San Francisco Police Department* provided us with information on the location, date, and severity of crimes that occurred in the city from 2014 to 2016. As research has shown that crimes tend to occur on hotter days, we also obtained the city's weather records for daily temperature and precipitation levels over the same period.

We hypothesized that it would not be feasible for the police force to implement extra security patrols for all areas predicted to have crime, given personnel and resource constraints. As such, we programmed our prediction model to identify only the top 30% of regions with the highest probability of violent crime occurring each day, so that these areas could be prioritized for extra patrols.

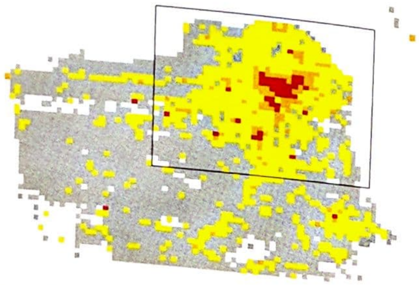

Figure 1. Heat map of San Francisco, showing frequency of crimes: very low (gray), low (yellow), moderate (orange), or high (red).

A preliminary analysis showed that crime occurred mainly in the north-eastern part of the city (as boxed up in Figure 1), and hence we divided it into smaller regions measuring 900ft by 700ft (260m by 220m) for further analysis.

To predict when and where a crime might occur, 1000 possible decision trees were generated based on crime and weather data, before combining them in a random forest. We used the data from 2014 to 2015 to train the prediction models, and tested their accuracy with data from 2016 (January to August).

So how well could we predict crime?

The random forest model successfully predicted 72% (almost three quarters) of all violent crimes. This proved superior to the average prediction accuracy of its 1000 constituent decision trees, which was 67% (see Figure 2).

With only 12 out of 1000 individual trees yielding accuracies better than that from the random forest, we could be 99% certain that predictions from a random forest would be better than those from an individual decision tree.

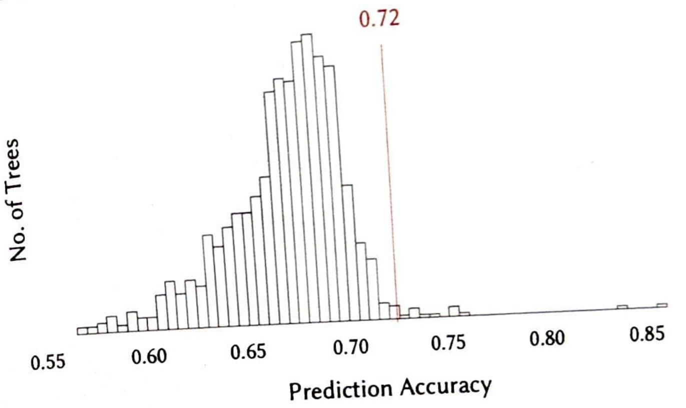

Figure 2. Histogram of prediction accuracies from 1000 decision trees (67% on average), compared to that from combining these trees into a random forest (72%).

Figure 3 shows a sample of the random forest's predictions over four days. Based on our predictions, the police should allocate more resources to areas coded red, and fewer to those coded gray. While it seems obvious that more patrols are required in areas with historically high crime, the model goes further and pinpoints the likelihood of crime in non-red areas. For instance, on Day 4 (lower-right chart), a crime in a gray area was correctly predicted, despite a lack of violent crimes occurring there in the previous three days.

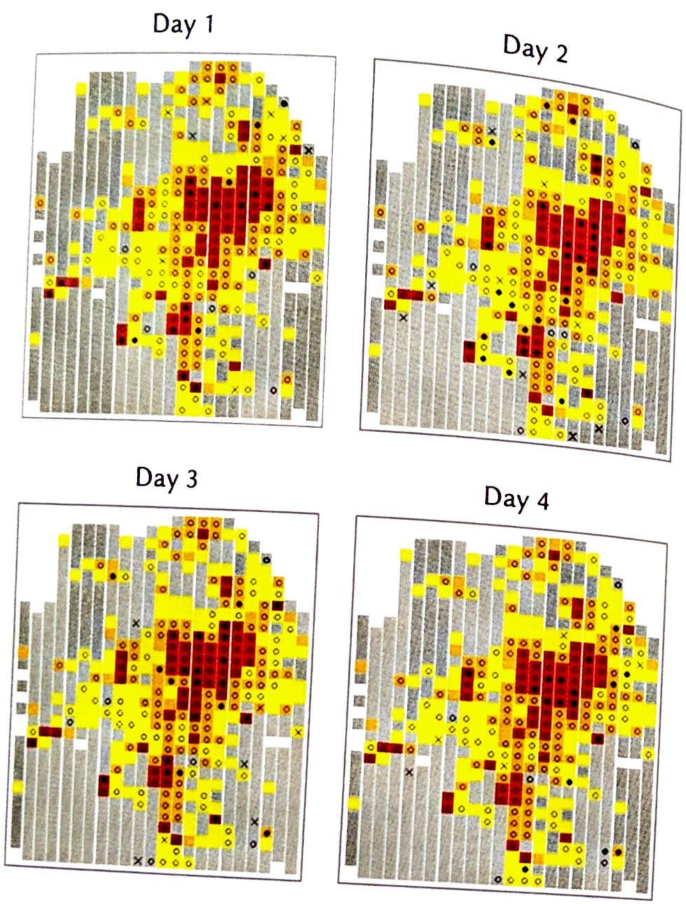

Figure 3. Crime predictions for four consecutive days in 2016. Circles denote locations where a violent crime was predicted to occur. Solid circles denote correct predictions. Crosses denote locations where a violent crime occurred, but was not predicted.

A random forest model also allows us to see which variables contributed most to its predictive accuracy. Based on the chart in Figure 4, crime appears to be best forecasted using crime frequency, location, day of the year and temperatures during the day.

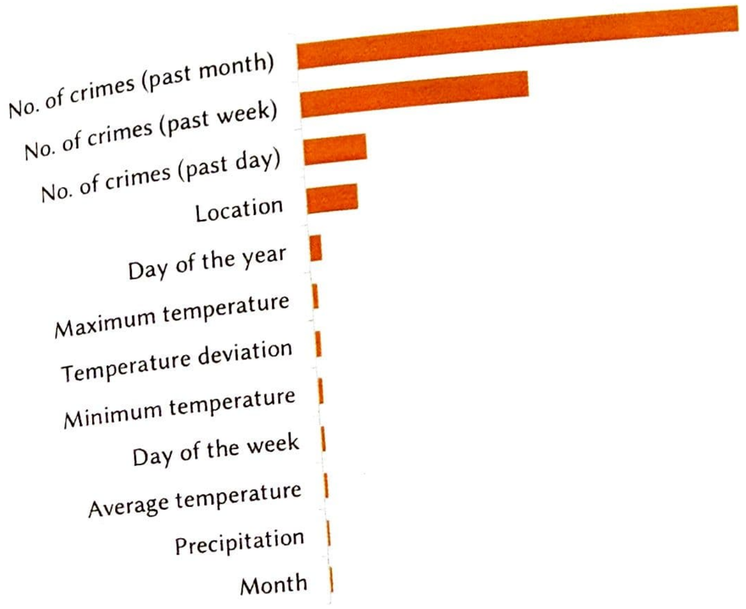

Figure 4. Top variables contributing to the random forest's accuracy in predicting crime.

We have seen how effective a random forest could be in predicting complex phenomenon such as crime. But how do they work?

10.3 Ensembles

A random forest is an *ensemble* of decision trees. An ensemble is the term for a prediction model generated by combining predictions from many different models, such as by majority voting or by taking averages.

We show in Figure 5 how an ensemble formed by majority voting could yield more accurate predictions than the individual models it was based on. This is because correct predictions reinforce each other, while errors cancel each other out. But for this effect to work, models included in the ensemble must not make the same kind of errors. In other words, the models must be uncorrelated.

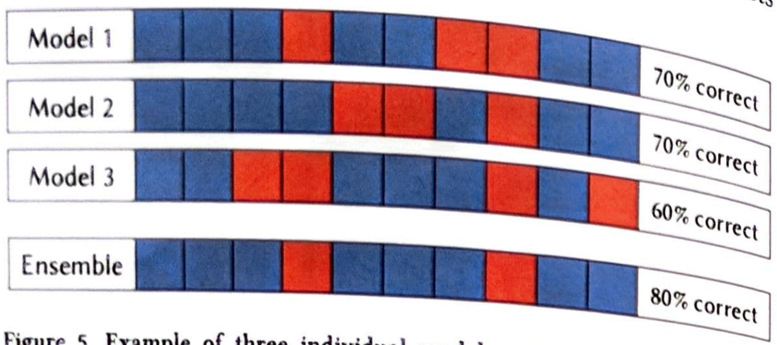

Figure 5. Example of three individual models attempting to predict ten outputs of either blue or red. The correct predictions were blue for all ten outputs. An ensemble formed by majority voting based on the three individual models yielded the highest prediction accuracy of 80%.

A systematic way to generate uncorrelated decision trees is a technique known as bootstrap aggregating.

10.4 Bootstrap Aggregating (Bagging)

We mentioned in the last chapter that, in constructing a decision tree, a dataset is repeatedly divided into subtrees, as guided by the best combination of variables. However, finding the right combination of variables can be difficult as decision trees are prone to the phenomenon of overfitting (explained in Chapter 1.3).

To overcome this, we could construct multiple decision trees by using random combinations and orders of variables, before aggregating the results from these trees to form a random forest.

Bootstrap aggregating (also termed as *bagging*) is used to create thousands of decision trees that are adequately different from each other. To ensure minimal correlation between trees, each tree is generated from a random subset of the training data, using a random subset of predictor variables. This allows us to grow trees that are dissimilar, but which still retain certain predictive powers. Figure 6 shows how the predictor variables, allowed for selection at each split on a tree, are restricted.

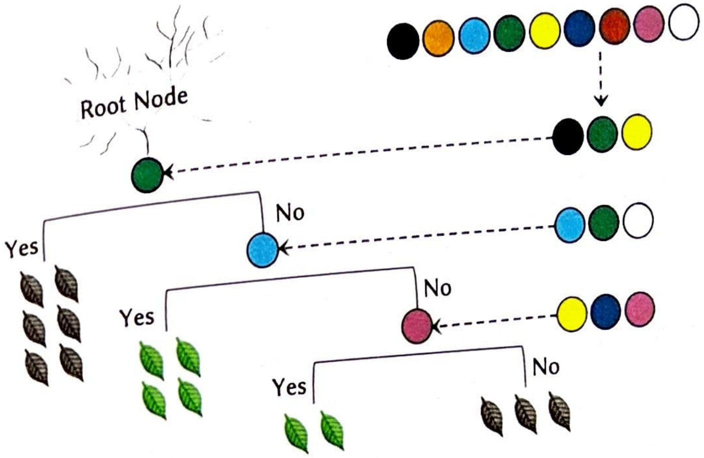

Figure 6. Creating a decision tree via bootstrap aggregating.

In Figure 6, there are a total of nine predictor variables, as represented by the different colors. A subset of predictor variables is randomly sampled from the original nine at each split, from which the decision tree algorithm then selects the best variable for the split.

By restricting the possible predictors for use at each split in the tree, we are able to generate dissimilar trees that prevent overfitting. To reduce overfitting even further, we could increase the number of trees in the random forest, which would result in a model that is more generalizable and accurate.

10.5 Limitations

No model is perfect. Choosing whether to use a random forest model is a trade-off between predictive power and interpretability of results.

Not interpretable. Random forests are considered *black boxes*, since they comprise randomly generated decision trees and are not led by clear prediction rules. For example, we could not know exactly how a random forest model reached its result—say, a prediction of a crime occurring at a specific place and time—except that a majority of its constituent decision trees came to the same conclusion. The lack of clarity on how its predictions are made could bring about ethical concerns when applied to areas such as medical diagnosis.

Nonetheless, random forests are widely used because they are easy to implement. They are particularly effective in situations where the accuracy of results is more crucial than their interpretability.

10.6 Summary

- A random forest often yields better prediction accuracy than decision trees because it leverages two techniques: *bootstrap aggregating* and *ensembling*.
- Bootstrap aggregating involves generating a series of uncorrelated decision trees by randomly restricting variables during the splitting process, while ensembling involves combining the predictions of these trees.
- While results of a random forest are not interpretable, predictors can still be ranked according to their contributions to prediction accuracy.