Background

Coal mining plays a large role in Kentucky’s economy and politics, and accounts for roughly 10% of the nation’s coal production. Approximately one quarter of that coal is consumed here, with the rest exported to other states and even other countries. Coal mines in Kentucky can be separated into two regions: those in the eastern part of the state and those in the western part of the state. The two regions differ in a number of ways, both economically and in terms of the coal they produce. The goal of this project is to explore and identify differences between Eastern Kentucky and Western Kentucky coalfields in terms of the coal they produce and how it is consumed.

Data

The Kentucky Department for Energy Development and Independence collected and processed receipts of all utility-scale shipments of coal in the United States from 1990 to the present. From 2002 onwards, some – but not all – receipts have included the origin of the coal shipment, thus making it possible to distinguish shipments involving Eastern versus Western Kentucky coal. Of the approximately one million coal shipments in the United States from 1990 to the present, you have been provided (Train) with the 26,157 shipments between 2003 and 2008 whose origins can be tracked to either Eastern or Western Kentucky. This data set contains a number of variables describing both the plant and the coal itself, as well as Region, which can be either “EKY” or “WKY”, depending on whether the coal originated in Eastern or Western Kentucky, respectively. The meaning of these variables is briefly described on the attached “Kentucky Coal Delivery Database Codebook.”

Prediction problem

The previous section makes reference to the fact that coal origins can be tracked in some but not all coal shipments. One specific goal of this project is to develop a model capable of predicting the origin of coal shipments that cannot be tracked. To that end, you are also provided with two additional data sets, Test1 and Test2, which contain all the same variables as Train with the exception of Region, which you are asked to predict. Test1 contains shipments from 2002, while Test2 contains shipments from 2009 onwards.

I will award special commendation to the student whose model is able to offer the best predictions. The Department for Energy Development and Independence (DEDI) may also be interested in speaking with the winner to discuss his/her findings and how he/she was able to predict so well. As a benchmark, the current method used by DEDI predicted 97.2% of origins correctly for Test1 and 97.3% of origins correctly for Test2.
Report and grading

You must turn in two things to complete your final project: (a) your predictions for \texttt{Test1} and \texttt{Test2}, and (b) a written report summarizing your findings. The report should contain no more than 5 pages of text (\textit{i.e.}, not counting figures and tables). You do not need to go into great detail concerning coal and/or its role in Kentucky’s economy, but please provide a brief introduction to your report.

Be creative! Hunt for interactions! Try various ideas – even if they do not seem to produce a better model, you may wish to discuss the approach and the reasons it didn’t work. Do not focus exclusively on the main trends: if there are interesting outliers or small subgroups in the data, please comment on them. Also, do not focus exclusively on your final model. Descriptive statistics – especially if they involve methods introduced in class – are always helpful in communicating trends in the data. I also recommend spending some effort studying whether or not trends seem to be changing with time – this is interesting in and of itself, but also very relevant to the prediction problem. Although of course, beware of overfitting to the sample data, as the validity of your model will be judged by its out of sample predictions.

Your grade on this project will be determined by four factors, all of roughly equal weight:

- Your understanding of the methods introduced in the course and the statistical concepts described in lecture
- The insights you are able to provide concerning distinctions between Eastern and Western Kentucky coal
- The clarity of your writing and communication
- The success of your model in predicting coal origins for the two test sets you are given

These go hand in hand, of course. For example, if you observe a number of interesting trends, but your model is very bad at predicting coal origins, your trends will seem dubious. On the other hand, a “black-box” model that can predict origins but provides no insight is of limited value. Finally, you may have a model that is both incredibly insightful and accurate, but if you cannot write about it clearly, the reader will probably not appreciate all of its wonderful qualities. An “A” project will accomplish all four of the above goals.

The project takes the place of a final in this course, and is due Thursday, December 16.