**Problem**

Inaccuracies in a car insurance company’s claim predictions most directly concerns drivers. Good drivers would have to pay for their car insurance more than they should and the insurance cost for bad drivers would likely be less. Naturally, that would also be a problem for the car insurance companies too, because from a business standpoint, it is in their best interest to provide the most accurate and reliable claim predictions. They would not want to lose their customers. Likewise, they would not want to lose the payments that the bad drivers should be making.

**Challenges**

Dealing in big data in general can be very time consuming as running and cleaning all of them together takes long to load. For both the linear and non-linear models, there is a lot of data to be cleaned. That is true for both the training data set and the test data. When we combined them, we found 2116753 cells that have -1 is a value, cells with missing values, and data we did not need to use.

To add to the complications, our data includes numbers such as binary data that need to be treated as categorical data. So before cleaning the data, we would need to define which variables are supposed to be numerical and which variables to be categorical. And then, cells with missing values for numerical data need to be filled with the means for that column while categorical inputs of -1 need to be imputed.

As another result of having so many features, we needed to explore which of them to use for a viable linear model and which to exclude, because a linear model certainly would not be able to represent that data using all of the features. In order to see which variables work best, a lot of different linear models (with different combinations of features) needed to be run, compared, and validated.

**Similar Problems**

In one project, we needed to predict the final house prices of every home in a city. The data we were provided contained 79 explanatory variables to work with. A non-linear model could incorporate all of them, but like with this current project, we needed to find a select number of features to use in our linear model. Likewise, the variables included integer features that we needed to specify as categorical (as intended) or it would be read in as numerical.