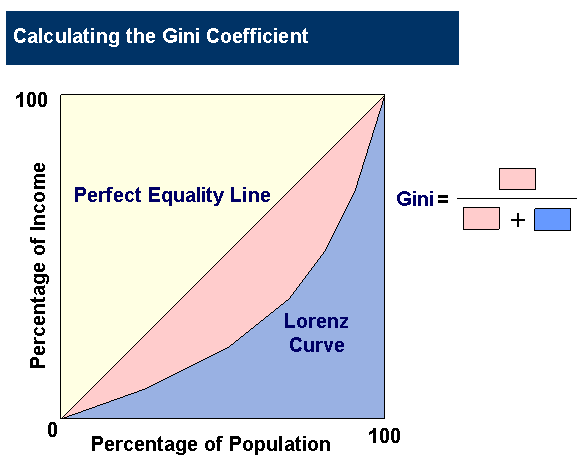
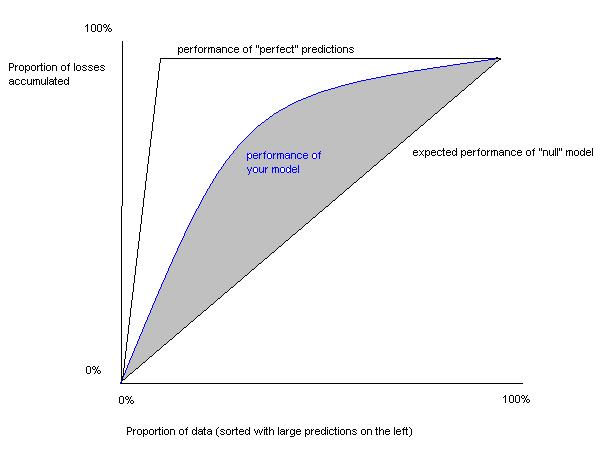
Liberty Mutual Hazard Score Prediction, Kaggle Competition

## Problem Statement.

Newly insured properties receive a home inspection. These inspections review the condition of key attributes of the property, including things like the foundation, roof, windows and siding. **Predict a transformed count of hazards or pre-existing damages using a dataset of property information** to more accurately identify high risk homes that require additional examination to confirm their insurability.

Contest Scoring is based on a normalized Gini coefficient. The Gini Coefficient measures the area between your models Lorenz curve and that of a completely random submission, similar to an AUC metric. This means that the goal is to build a model which predicts the correct order in terms of hazard. Since the scoring is based the hazard rank ordering, maybe calibration is not as important?

## Data set description.

The data is a Kaggle dataset presumably unaltered from Liberty Mutual. The columns are arbitrary names which forces the contestants to rely solely on intuitions from only the patterns in the data.

The home inspections review the condition of key attributes of the property, including things like the foundation, roof, windows and siding.

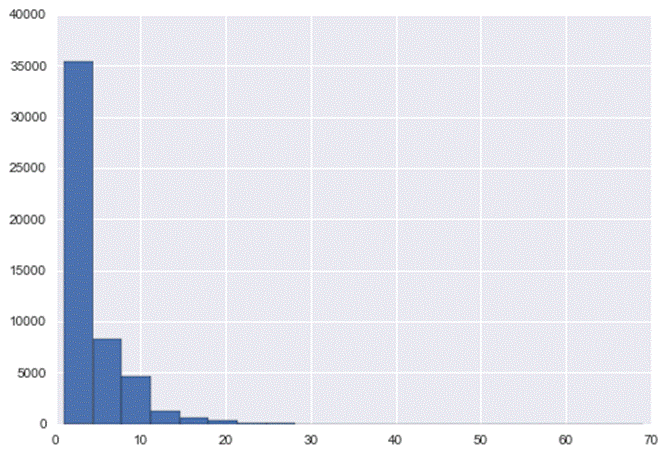
There are 32 features and all appear to be categorical values:

* 16 are numeric but all have 100 or less values
* 5 are binary character
* 11 are multinomial character

There are 51,000 observations each in training and test sets

## Pre-processing steps.

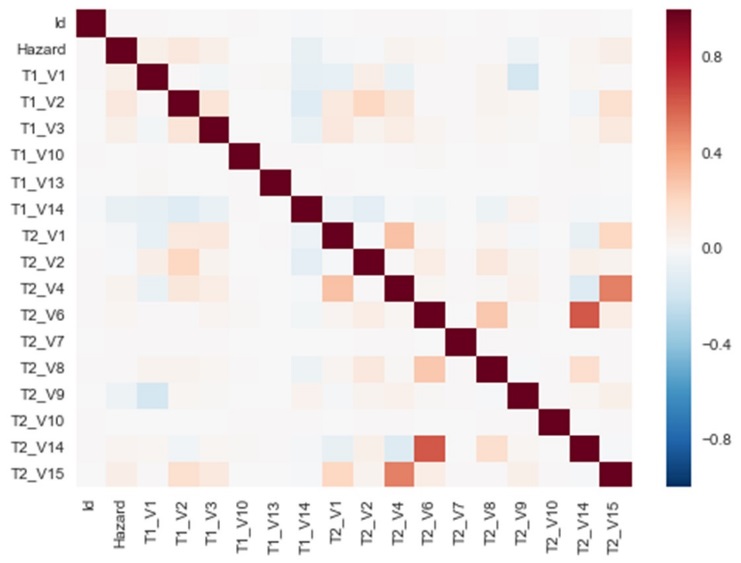
No pre-processing is required as all the data is furnished via Kaggle. All data is present there are no missing values. To facilitate linear regression, dummy variables were made which expanded the number of columns to 112.

**Data exploration and visualizations.**

One of the most notable features is the extreme skewness of the response variable, the hazard score.

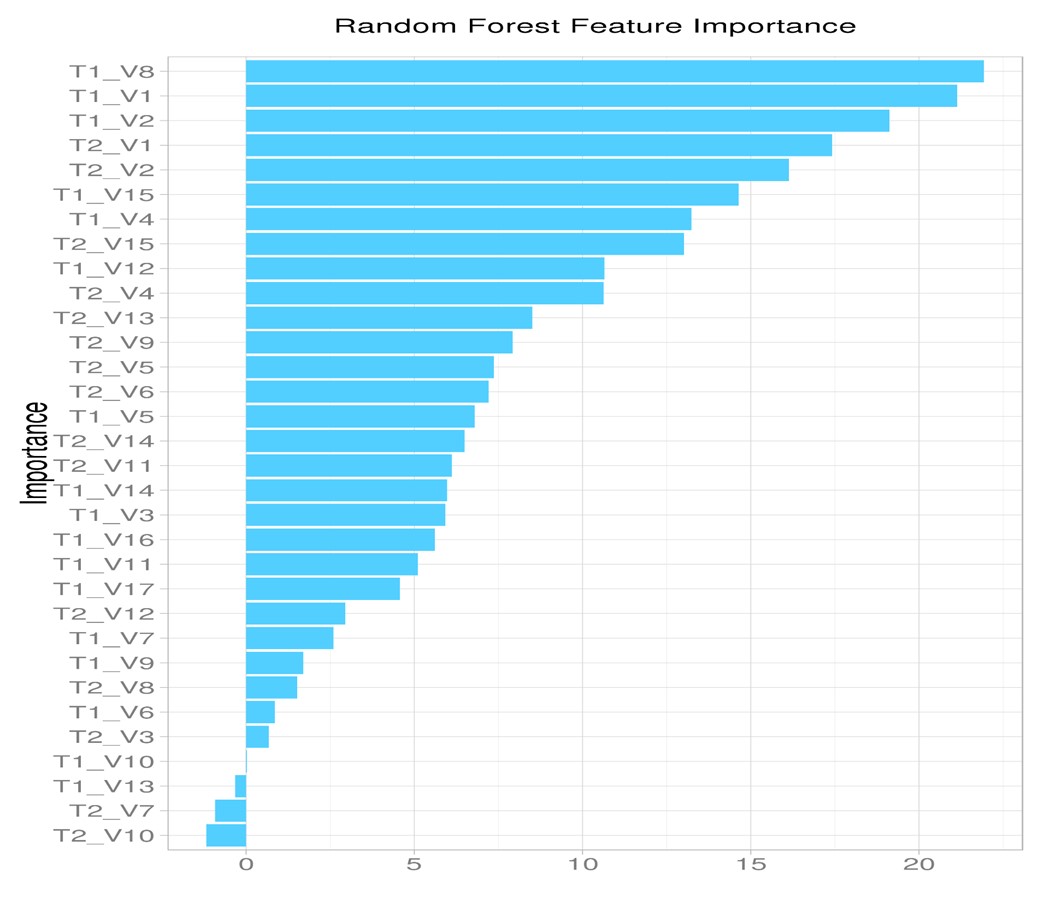
The 50 discrete scores, range from 1 to 69. 36% are no hazard, 1, and 75% < 5.

99.9% < 30, so maybe the best strategy is to concentrate on those since the scoring is based on Hazard ranking not error in the estimate.

Even before making the dummy variables, there is a small amount of direct collinearity in at least one pair of columns. Maybe they can be combined into a single variable or use the R style formula that automatically does this. For example see T2\_V4 + T2\_V6 in the correlation heat map at right:

## Feature selection

This is a major part of a Kaggle competition. With 112 columns, semi-automated means are very helpful.

The benchmark Random Forest model yields which features had the most importance in build the regression tree. The first linear regression used T2\_V1 (0.09 weight), T2\_V2 (0.06), T2\_V9 (0.06) and T1\_V2 (0.06).

Since 97.5% of the scores are less than 15, focus the analysis on those 14 unique Hazard values, discarding the 36 unique values from 15 to 59 as effective outliers.

The second way that variables could be selected is by look at the mean Hazard value for them and combining with their counts. That could give a hint as to which ones may influence the regression more. Looking the 97.5% under 15 hazard score, although only 5 occurrences of T1\_V5\_E had a value 1, the average hazard score for them was 5.8. Similarly T1\_V8\_C had 1326 occurrences of value 1 but the mean hazard was 5.22, both significantly higher than the mean of 3.62 for hazards under 15.

|  |  |
| --- | --- |
| column | hazard correlation |
| T1\_V15\_C | 0.060321 |
| T1\_V4\_N | 0.065470 |
| T2\_V15 | 0.066527 |
| T1\_V12\_C | 0.074483 |
| T1\_V11\_H | 0.079414 |
| T1\_V8\_C | 0.086275 |
| T1\_V5\_K | 0.094186 |
| T1\_V2 | 0.104895 |
| T1\_V9\_E | 0.108297 |

The third way to select variables is by looking at the correlation to the hazard score. One of the issues that makes this contest difficult for linear regression is that no single column correlates to the hazard score. But a first model for linear regression will use these top correlating columns. That results in 9 candidate columns for the regression:

|  |  |  |
| --- | --- | --- |
| column 1 | column 2 | correlation |
| T1\_V17\_N | T2\_V12\_N | 0.966224 |
| T1\_V17\_Y | T2\_V12\_Y | 0.966224 |
| T2\_V11\_N | T2\_V13\_A | 0.804619 |
| T1\_V9\_E | T1\_V11\_H | 0.628185 |
| T2\_V6 | T2\_V14 | 0.624224 |
| T1\_V5\_K | T1\_V9\_E | 0.589479 |
| T1\_V9\_D | T1\_V11\_B | 0.584554 |

Collinearity among the dummy variables. Another way to discern more of a hazard signal is to combine collinear columns. These 7 are significantly to highly collinear:

## The modeling process.

Random Forest appears to be the most successful and resilient to hazard outliers. A random forest with the dummy variables ran just as well with the hazard outliers included as excluded. The benchmark Gini coefficient was 0.263387. The major insight RF gives is order of importance of the variables but it can be difficult to get intuition from the model other than that.

|  |  |  |
| --- | --- | --- |
| column | metric | why included |
| T1\_V5\_E | 60% | diff from mean |
| T1\_V8\_C | 44% | diff from mean |
| T1\_V15\_C | 0.0603 | hazard correlation |
| T1\_V4\_N | 0.0655 | hazard correlation |
| T2\_V15 | 0.0665 | hazard correlation |
| T1\_V12\_C | 0.0745 | hazard correlation |
| T1\_V11\_H | 0.0794 | hazard correlation |
| T1\_V8\_C | 0.0863 | hazard correlation |
| T1\_V5\_K | 0.0942 | hazard correlation |
| T1\_V2 | 0.1049 | hazard correlation |
| T1\_V9\_E | 0.1083 | hazard correlation |
| T2\_V1 | 0.09010 | RF significance |
| T2\_V2 | 0.06229 | RF significance |
| T2\_V9 | 0.06144 | RF significance |
| T1\_V2 | 0.06069 | RF significance |

However, in order to better understand the variables a linear regression was built from 2 variables that had high differences from the mean hazard, RF contributors and variables correlated to the hazard score.

## Challenges and successes

The linear model built with these 15 columns actually was an improvement over the random forest benchmark, improving from .262 to .286. While still far behind in the competition, this gives promise that feature engineering even with a linear regression model could be competitive.

Next steps are to print the p-value of the features and eliminate or include based on that statistic. Regularization will be critical for feature selection, probably LASSO in order to quickly eliminate most of the 112 columns so the model will generalize well on the test data.

Random forest and other ensemble or stack models seem to be a definite candidate. Random forest parameter setting to limit the number of features may help.

Feature engineering will start by eliminating or combining collinear variables, attempting to make them more discriminatory. A closer inspection of the high and low hazard scores and the mean values of the features at the high and low ends may help identify features or creating a new feature.

Andrews plots and parallel coordinates are ways to view structure in high-dimensional data. Those will be tried as well.