Team Memento Mori: Predicting Unnatural Deaths in the U.S.

The Problem

The event of an unnatural death occurs with small probability within the United States general population. As these events are unexpected, life insurers can protect themselves from these losses by offering policies whose premia account for such a risk. The policies offered should not only compensate the insurer, but also compete with similar policies from other insurers. In order to be competitive and to increase customer conversion and acquisition after an initial advertisement of a policy's premium, ads can be made to reduce the sticker shock between the advertised premium and the final offered premium. One step in accomplishing this is to account for the likelihood of unnatural death for a particular individual based on publicly-available demographic information.

The Data

To this end, we created a binary classifier to predict whether an U.S. citizen will die of an unnatural cause versus a natural cause. Unnatural causes of death are classified as those falling into the categories of all-cause accidents, homicides, and suicides, while natural causes make up the remaining causes. The team used publicly-available Multiple Cause Mortality Data from the Center for Disease Control collected each year from 2005 to 2015. These data consist of demographic and all-cause mortality information for every death in the United States, totalling about 2.7 million deaths each year.

Exploratory Data Analysis

In order to select features on which to train a model, we explored the data. We found that the likelihood of unnatural death is small, about 8%, revealing a heavy imbalance of the two classes. There was a significant difference between the distributions of age depending on type of death; age of natural deaths were skewed left, while ages of unnatural deaths presented more symmetry. The data confirm the conventional wisdom that the sex of an individual has a significant role in type of death, showing us that males are more likely to die of unnatural death at all ages than females. Finally, we found significant interaction between age of death and marital status, with those who are single are much more likely to die of an unnatural death at a young age than those who are married.

Key Performance Metrics

To reduce the frequency of sticker shock, it is best for the model to prefer to make false positive errors instead of false negative errors. However, the model should be penalized for misclassifying too many training instances as unnatural deaths, since the overwhelming majority are natural deaths. Therefore, we measure model performance by the macro f_1 score. Our baseline model for comparison was a model which classified an instance as an unnatural death with probability equal to the training proportion of unnatural deaths.

Results and Conclusions

After training models like Logistic Regression, Naive Bayes, and XGBoost, and comparing their performance as described above, our best model, a cost-sensitive Random Forest Classifier, obtained a test macro f_1 score of around 0.65, while classifying 18% of the testing set as unnatural deaths. This gives a reasonable balance of predicted unnatural to natural deaths than the other tested models, and improves upon the baseline performance of 8% recall and 0.5 macro f_1. By using this model, we can offer a more competitive initially-advertised premium to a larger number of potential customers while preventing the insurer from offering an artificially low rate which does not adequately compensate for risk.