Who are You to Compromise (and Release)?

Kristy Cheng, Megan Muller, and Ming Zhang
Faculty Advisors: Janet Duncan, Xiyue Liao, and Ian Duncan
Sponsors: California State Association of Counties-Excess Insurance Authority
Department of Statistics and Applied Probability, University of California, Santa Barbara

Abstract

Workers' Compensation (WC) insurance covers employees when they get injured, or ill on the job. WC claims cost more the longer they are open, which is why the California State Association of Counties - Excess Insurance Authority (CSAC-EIA) asked us to investigate a way to predict Compromise and Release (C&R) settlement classification of a claim to facilitate cost-effective closure.

Initially, we balanced the data to enable us to reduce distortion as we built and analyzed eight predictive models. In the end, we selected the random forest and logistic regression models for their accuracy and interpretability respectively.

Introduction

C&R settlements are one way CSAC-EIA closes claims. This type of settlement allows for a given claim to be closed and the insurance company to be free of any future liability related to that claim. C&R settlements can only be made when the claimant is permanently disabled, the percent of disability has been determined, and the claimant has reached a stable health condition. These three conditions can take months or years to be met, but if a claim can be closed as soon as possible, it can over time save a company a great deal of money.

Utilizing both machine learning and statistical modeling methods we sought to find an effective model for categorization of claims as C&R and to find the criteria most influential to that categorization.

Methods

The WC claim data used was provided by CSAC-EIA, and after data cleanup the master dataset used for modeling included claims from accident years 1994-2018 evaluated as of June 30, 2018.

Data Balancing With ROSE

The data to be used for model building was unbalanced, containing only about 7% C&R claims. For modeling data are split into training and test data sets and only the training set needs to be balanced. The Random Over-Sampling Examples (ROSE) Method, a smoothed bootstrap method, was used to balance the data.

Analysis and Results

Accuracy and **True Positive Rate (TPR)** are the most popular model evaluation tools for balanced data, but our data are unbalanced, so the **Positive Predictive Rate (PPR)** must be used to find the lowest **False Negative Rate (FNR)**.

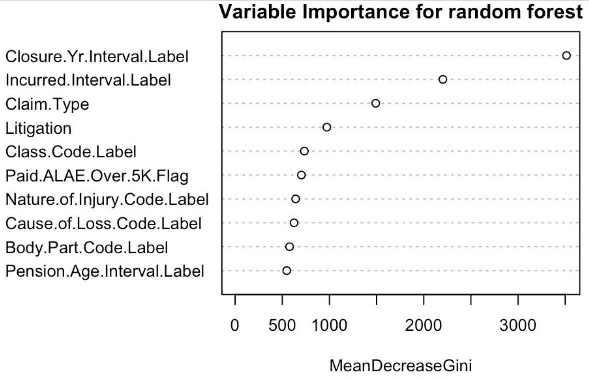
FNR for PPR=33%

Model	FNR
Decision Tree	.17
Logistic Regression	.16
Bagging	.11
Random Forest	.09
Boosting	.24
Support Vector Machine	.16
Lasso Regression	.16
Elastic Net	.16

Analysis and Results continued

Based on the FNR when the PPR is fixed at 33%, the Random Forest (RF) Model is found to be the best model for predicting a claim being C&R.

Variable Importance Plot for RF



Random Forest can show which variables are important for prediction of C&R such as Closure Year Interval, Incurred Interval, and Claim Type. However, the major drawback of this model is, it cannot tell which values of the important variables are more likely to lead to a C&R settlement.

Logistic Regression

Logistic Regression is a form of Generalized Linear Model (GLM), using a logit function. This Model is the most interpretable, and gives a way to directly compare the importance of each level of each variable.

Conclusions

Random Forest

Confusion Matrix		Target	
		All Other	C&R
Model	All Other	5435	50
	C&R	1070	537

We will sacrifice some accuracy by over classifying as C&R to avoid missing classifying any potential C&R claims.

Logistic Regression

The three most important criteria for predicting a C&R settlement is the claim coming from a Special Districts entity type (e.g. Police, or Fire), the claim being litigated, and legal expenses for the claim being more than $5,000.