Auto Insurance Claim Fraud Predictors

Mary Donovan Martello

DSC 530 Final Paper

Statistical/Hypothetical Question

What factors in auto insurance claims indicate that the claims are fraudulent claims? The data set has a variable that indicates whether fraud was reported on each observation (i.e., each auto insurance claim instance). This variable is either Y or N. I created subsets of the dataframe based on the fraud_reported variable. The hypothetical question is whether the fraud_reported subsets can help identify which variables can be used as predictors to identify fraudulent auto insurance claims.

Outcome of EDA

I was hoping that the analysis would show certain variables were strongly indicative of predicting fraudulent auto insurance claims. For example, perhaps most of the fraud reported cases might have very high umbrella_limit values because bad actors acquired umbrella insurance with high limits so that their payouts could be higher when they made fraudulent claims. However, I do not see any very strong indications of any one or few variables predicting fraud_reported.

However, a few of the variables did repeatedly show distinction or significance in the various steps of my analysis. Total claim amounts for the fraud reported = Y cases showed higher claim amounts in the histograms, descriptive statistics, CDF and hypothesis testing. The umbrella limit histogram shows a distinction at $4-$5 million limit vs no umbrella coverage between the two subsets. Witnesses for the fraud reported = Y cases showed higher witnesses in the histograms and descriptive statistics. The biggest distinction between the two subsets appears to be in incident severity variable. The fraud reported = Y cases have a significantly higher count of Major Damage than the no fraud reported cases.

None of the point-biserial correlations for the fraud_reported variable is strong. However, the p-value for umbrella limit appears significant. The logistic regression model indicates that the best accuracy would come from including most all of the variables instead of narrowing down some key indicating variables.

Some of the individual variables seem more significant than others, such as incident_severity, total_claim_amount, umbrella_limit, and witnesses. However, none of these show very strong correlations or indications of influence in the predictive models.

What do you feel was missed during the analysis?

I tried to use the data mining function for my logistic regression model like we did for linear regression in the assignments. However, I never got passed this error: “PerfectSeparationError: Perfect separation detected, results not available”. After trying many times to solve this error, I simply ran the logistic model separately on each variable to find the test statistic.

Were there any variables you felt could have helped in the analysis?

No.

Were there any assumptions made you felt were incorrect?

I assumed that some of the variables would be identified to be predictive.

What challenges did you face, what did you not fully understand?

I think the analytical plots are still very difficult to understand and employ.