Auto Insurance Claim Fraud Predictors

Mary Donovan Martello

DSC 520 Final Paper

Auto Insurance Claim Fraud Predictors

Automobile insurance fraud results in upward of $7.7 billion a year in losses (Croll, 2018). Therefore, automobile insurance claim fraud detection and prevention are important endeavors for businesses. Identifying factors that indicate fraudulent claims will assist in fraud detection and prevention.

Problem Statement

What factors in auto insurance claims indicate that the claims are fraudulent claims?

Data and Methodology

Data

The data I used to address the auto insurance claim fraud problem is a data set from Kaggle: https://www.kaggle.com/patilk1/fraudulentinsuranceclaim. The data set has 1,000 observation and 39 variables. The following are the variables:

- months_as_customer
- age
- policy_number
- policy_bind_date
- policy_state
- policy_csl (CSL is a single number that describes the predetermined limit for the combined total of the Bodily Injury Liability coverage and Property Damage Liability coverage per occurrence or accident)
- policy_deductable
- policy_annual_premium
- umbrella_limit
- insured_zip
- insured_sex
- insured_education_level
- insured_occupation
- insured_hobbies
- insured_relationship
- capital-gains
- capital-loss
- incident_date

- incident_type
- collision_type
- incident_severity
- authorities_contacted
- incident_state
- incident_city
- incident_location
- incident_hour_of_the_day
- number_of_vehicles_involved
- property_damage
- bodily_injuries
- witnesses
- police_report_available
- total_claim_amount
- injury_claim
- property_claim
- vehicle_claim
- auto_make
- auto_model
- auto_year
- fraud_reported

Methodology.

Data Cleaning

My first step was to explore the data in order to identify any aspects that needed addressed. I first used str() to see the structure of the data frame. str() showed 1000 observations and 40 variables. However, the last variable, labeled "X_c39", was an empty variable with just the logical type NA for all 1000 observations. I removed the "X_c39" empty column. str() shows the data type of each of the variables. R is treating all of the variables as either int/num types or factors. For the factors, str() shows how many levels and the label R is using for each of the factor variables.

Two of the variables are dates, “policy_bind_date” and “incident_date”. R treated each of these as variables as factors. I did not see how these variables were useful as factors, but I decided that it would be useful to know the time period between the incident date and the policy bind date. Thus, I converted these data types to dates and then created a new variable, called “weeks_bf_incident”, by computing the difference in these two date variables.

The variable “policy_csl” has two values in each column. I separated the “policy_csl” column into two columns with one value in each column ("cslBodily" and "cslProp"). The “policy_csl” variable was a factor, and after the separation the "cslBodily" and "cslProp" variables were factor variables. However, because these columns contain numeric data I converted the data types to integers. I removed the original incident_date and policy_bind_date variables because they were no longer needed. The original data set included a variable called “months_as_customer”. I converted these values from being measured in months to being measured in weeks to be consistent with the time measurement in the new weeks_bf_incident variable. Note: When I created the new weeks_bf_incident variable I tried to use months as the measurement but the difftime() function did not have months as a measurement option.

I searched for missing values with the countInf() function from the VIM library. The results showed no missing values in any of the variables (notwithstanding the empty variable). I used the count() function on each variable to get a better idea of what was in each variable. In doing so, I found that three of the variables had “?” as a value instead of NA or being empty. This is why it was not discovered in the countInf() function.

I ran histograms on the variables, and studied each variable’s count, to look for outliers. I only found one outlier in the “umbrella_limit” variable. There was a negative 1,000,000 value and thus it looks like an erroneous value. I filtered out this one outlier.

Finally, in order to make use of the several factored variables in the data set I needed to convert the factor type variables to numeric type variables. However, for some of the analysis I still needed the categorical variables to be factor data types. Thus, instead of converting these categorical variables back and forth between factors and numeric I created a copy of the dataframe and converted the categorical variables to numeric in the copied dataframe.

Exploration and Visualization

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 both dataframes based on the fraud_reported variable. I used the subsets to explore the counts and histograms of variables to try to identify differences in variables in the fraud_reported == "Y" subset versus those same variables in the fraud_reported == "N" subset. Using the subsets was useful in spotting differences in counts and histograms.

I used histograms with normal curves and colored by the fraud_reported variable to look for variables that might be an indication of a fraudulent auto claim. Many of the histograms showed the same curve for the fraud_reported variable, and thus did not appear to identify a difference for fraudulent claims. However, some of the histograms show a difference in the curve for fraud_reported == "Y" versus fraud_reported == "N". These may indicate variables that can identify fraudulent claims.

The fraud_reported == "N" subset has 752 observations but the fraud_reported == "Y" subset has only 247 observations. If each subset had the same proportional number of values for a variable, then the histogram's red curve (for N) would always be proportionally higher than the green curve (for Y). However, for some variables the green curve (for Y) was much higher than the red curve. The histograms for incident_severity, total_claim_amount and weeks_bf_incident demonstrate this discrepancy.

Correlation

My outcome variable is fraud_reported. In order to determine which variables will explain this outcome variable, I ran correlation functions to try to discover some correlation between fraud_reported and other variables. Fraud_reported is a discrete dichotomous variable. Point-biserial correlation is used when one of the variables is a discrete dichotomous variable (Klopper, n.d.). The point-biserial correlation is equivalent to calculating the Pearson correlation between a continuous and a dichotomous variable (stats.stackexchange.com, n.d.). Therefore, I can use the standard cor.test function in R, which will output the correlation, a 95% confidence interval, and an independent t-test with associated p-value.

I tested the correlation of each variable against the fraud_reported variable. None of the variables had a very high correlation number. However, some of the p-values show significance. The p-values from the correlation test is the significance level of the t-test. If the p-value is less than 5% then we can conclude that the correlation is significant (Correlation Test Between Two Variables in R, n.d.). The variables that have p-values that are 6% or less are (1) total_claim_amount, (2) incident_severity, and (3) umbrella_limit.

Given the results of the correlation tests, I ran partial correlation on the fraud_reported and incident-severity variables, with total_claim_amount and umbrella_limit variables as controlled variables. These correlations are stronger than any of the two-variable correlations.

Logistic Regression

My outcome variable is a binary categorical variable and therefore I am using logistic regression to try to determine explanatory variables to predict the fraud_reported variable. Note: I am able to use the categorical variables in the regression model as I converted them to numeric data types.

I first included all of the variables in the formula with the exception of (1) the three variables with a "?" as a value because I have no means of deciding on meaningful replacement values and (2) the variables that caused multicollinearity. There was a multicollinearity problem with including all the variables. I tested for multicollinearity with the vif() function. When the vif() function shows a multicollinearity problem, you can see which variables are the cause of the problem with the alias() function. I removed the variables causing the multicollinearity problem and the three "?" variables, and ran a logistic regression model with the remaining variables (the "Permissible Variables Model"). The accuracy result of the Permissible Variables Model is 80.58%.

I looked at the z-statistics for the variable coefficients of the variables in the Permissible Variables Model, which tells us whether the coefficient for that predictor is significantly different from zero. If the coefficient is significantly different from zero, then we can assume that the predictor is making a significant contribution to the prediction of the outcome (Field, 2012). umbrella_limit and incident_severity showed significance because the significance of their z-factors is less than .05. incident_severity in particular has a significant z-value.

I also looked at the odds ratio of the variables in the Permissible Variables Model. Odd ratio is an indicator of the change in odds resulting from a unit of change in the predictor. If the odds-ratio is greater than 1, then the coefficient predicts an increased chance of an event occurring (as the predictor increases, the odds of the outcome occurring also increases). There were several variables with odds ratios greater than one, but none of them had an odds ratio greater than two. This tells me that no variable particularly increases the outcome of the fraud_reported variable.

Next, I tried to improve on the accuracy of the logistic regression model by using only some of the Permissible Variables in the formula. I reviewed the odds ratio from the Permissible Variable Model and only included the variables that had an odds ratio of 1 or greater. However, the accuracy from this model decreased to 75.18%.

Because incident_severity showed to be significant in the counts, histogram and correlation analysis I added incident_severity back to the model even though it only showed an odds ratio of 0.273. The accuracy results of this model (the "Odds Ratio + Severity Model") is 81.48%.

I tried various other combinations of variables for the logistic regression model (including only the variables I found significant from other parts of the analysis) and the highest accuracy I found is the 81.48% for the "Odds Ratio + Severity Model".

Classification Model

Classification algorithms can be used predict categorical outcomes. My research question is ultimately determining the outcome of the categorical fraud_reported variable and thus I am using a k nearest neighbor model to try to predict the fraud_reported variable. Because the dataframe has many categorical variables, I used one-hot encoding to convert the factors to dummy variables so that they may be used in the training data and test data in the machine learning model.

The classification model I used was the k nearest neighbor model. I ran several scenarios varying the number of k clusters and varying the proportion of training data and test data. The results were unexpected in that the highest accuracy rates resulted when only False/0 results were predicted. Several of the models had 75% accuracy, but in these models, there were zeros for the predicted True/1 values. At a little less accuracy, the knn model predicted both 0 and 1 values. For example, with a k = 5 and a proportion of training data and test data of 65/35, the accuracy was 71.06017% and the confusion matrix was:

```
#            test_claims
#knn.roundup   0   1
#           0 238  76
#           1  25  10
```

Insights.

I was hoping that the analysis would show certain variables were strongly indicative of predicting fraudulent auto insurance claims. For example, perhaps the bad actors overwhelmingly purchased insurance with low deductibles because they intended on making fraudulent claims. Another example might be that 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. The histograms for incident_severity, total_claim_amount and weeks_bf_incident showed higher values for fraud_reported = Y versus fraud_reported = N even though the fraud_reported = Y subset has only about 1/3 of the number of observations compared to the fraud_reported = N observations.

The counts for incident_severity also shows this discrepancy. incident_severity for fraud_reported = Y has 50% higher "Major Damage" counts than fraud_reported = N even though it only has 1/3 of the number of observations. The counts for insured_hobbies showed a distinction in chess and cross-fit hobbies.

Just as the histograms showed some significance for incident_severity and total_claim_amount, the correlation analysis showed significance for incident_severity and total_claim_amount. These two variables plus umbrella_limit have p-values that are 6% or less. Partial correlation showed that the combination of these variables increases the correlation results.

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. The accuracy of the logistic regression model and the k nearest neighbor classification model shows similar accuracy results. These accuracy results of 75% - 80% are good but are they good enough to predict fraudulent auto insurance claims?

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

Implications.

Auto insurance companies could use the logistic regression and k nearest neighbor models to evaluate auto insurance claims submitted to them and get a good idea if they should be flagged as fraudulent. However, the accuracy rates are not high enough to completely rely upon. Further investigation would be needed.

Limitations.

Other models may improve the predictions on this data set. Learning and utilizing other predictive models is a further step to improve the predictions.

Three variables have significant numbers of values of "?" instead of actual responses (18%, 34% and 36%). In order to rely on these variables, valid responses would need to be investigated.

References

Correlation Test Between Two Variables in R. (n.d.). Retrieved from Statistical tools for high-throughput data analysis: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

Croll, M. (2018, February 22). *Insurance Fraud and What Regulators and Insurers Are Doing About It*. Retrieved from ValuePenguin: https://www.valuepenguin.com/2018/02/auto-home-insurance-fraud-and-what-regulators-and-insurers-are-doing-about-it

Field, A. M. (2012). *Discovering Statistics Using R.* Los Angeles: Sage Publications Ltd.

Klopper, J. H. (n.d.). *Biserial correlation*. Retrieved from RPubs by RStudio: https://rpubs.com/juanhklopper/biserial_correlation

stats.stackexchange.com. (n.d.). Retrieved from P-value for point biserial correlation in R: https://stats.stackexchange.com/questions/226157/p-value-for-point-biserial-correlation-in-r