**Predicting Auto Insurance Claim Fraud**

**Mary Donovan Martello**

**DSC 630 Predictive Analytics**

**Milestone 5**

**Executive Summary**

Automobile insurance companies must fight to detect and counter fraud in its insurance claims process. The goal of this project was to develop predictive classification models that will take as inputs features of automobile insurance claim transactions and classify the transactions as either fraudulent or non-fraudulent.

Many factors are involved in automobile insurance claims, including information on the policy, insured, and automobile, aspects of the damage incident, and elements of the claim filed. Including all these factors as inputs in the predictive model did result in a high degree of accuracy in predicting whether each claim record had been reported as fraudulent or not reported as fraudulent by the insurance company providing the claim records. However, in exploring the claim records before creating the predictive models, I determined that a smaller subset of the factors showed some distinction in the “fraud reported” cases but not in the “no fraud reported” cases. Therefore, I also built predictive models that only used as inputs the factors that showed distinction in the “fraud reported” cases but not in the “no fraud reported” cases. These predictive models with the smaller distinct inputs performed better than the models with all the claim factors included as inputs.

The conclusions from this project are that the predictive models using all the claim record factors were effective in predicting fraudulent claims at a good degree of accuracy, but the models that limited the claim record features to factors related to the number of witnesses to the incident, incident severity, hobbies of the insured, policy deductible, number of vehicles involved, and the length of time the claimant was an insured showed even higher performance results. The models built in the project will be valuable to automobile insurance companies for detecting fraudulent claims.

**Introduction/Background**

Fraudulent insurance claims contribute to between 5 and 10 percent of total claims and are costing insurance companies more than $40 billion annually (Moon, 2019), (Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claim Fraud, n.d.). Automobile insurance claims are susceptible to fraudulent claims and automobile insurance companies, and ultimately consumers, lose money due to fraudulent claims. Automobile insurance companies must endeavor to predict insurance claim transactions that are fraudulent claims.

Fraud detection in insurance has always been challenging because the target class is very imbalanced (i.e. the incidence of frauds is far less than the total number of claims) and each fraud is unique in its own way (Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claim Fraud, n.d.) (Dommalapati, 2019).

The goal of this project is to develop predictive classification models that will take as inputs features of automobile insurance claim transactions and classify the transactions as either fraudulent or non-fraudulent.

**Methods**

The analysis is following the CRISP-DM stages for data science projects including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The Business Understanding analysis is summarized in the Introduction/Background section above and deployment is beyond the scope of this project. The Data Understanding, Data Preparation, and Modeling, stages are discussed in this Methods section. Evaluation is subsequently discussed below.

**Data Understanding**

1,000 prior claim transaction records were assembled for use in developing a prediction model. Each record has a mix of 38 quantitative and categorical data features about the claim filed, including information on the policy, insured, and automobile, aspects of the damage incident, and elements of the claim filed.

The data set also has a feature that indicates whether fraud was reported on each observation (i.e., either Y or N). The Data Understanding analysis utilized this target feature designation to examine differences in features when the target feature was designated as fraud vs no-fraud. The Data Understanding analysis revealed that a few of the features did repeatedly show distinction or significance with respect to the fraud reported target feature.

Histograms of the input features overlayed with a normal curve colored by the fraud target feature showed some variation in the insured’s occupation, insured’s hobbies and auto year features. Total claim amounts for the fraud reported = Y cases showed higher claim amounts in the mean ($60,300 vs $50,230) and quartile descriptive statistics. Number of witnesses to the incident for the fraud reported = Y cases appeared higher in the histograms and descriptive statistics (mode of 2 vs 0 and mean of 1.6 vs 1.5). Point-biserial correlation is used for correlation between the binary fraud reported variable and continuous variables. None of the point-biserial correlations for the fraud\_reported feature are strong. However, the p-value for umbrella limit shows that it may be significant.

The biggest distinction between the two subsets appears to be in the incident severity feature. The fraud reported = Y cases have a significantly higher count of Major Damage than the no fraud reported cases: 68% of fraud reported cases claimed Major Damage but only 14% of fraud not reported cases claimed Major Damage.

The Data Understanding phase revealed that a few of the features did repeatedly show distinction or significance with respect to the fraud reported target feature, and the Data Understanding phase also showed some issues with the dataset. Some records are missing content on the collision type, existence of property damage, or whether a police report was filed. The most frequent value in the category was imputed to preserve the record in the dataset.

Further, fraud detection is challenging as the dataset often has a target class that is very imbalanced (i.e. the incidence of frauds is far less than the total number of claims). Data Understanding revealed that this dataset is imbalanced as the fraud reported records make up 25% of the records, which might be a sufficient proportion for the prediction.

Finally, the dataset has a large number of features in each transaction, and the fraud indications may be in different features in different records. Therefore, the number of features included in the model is evaluated in this project.

**Data Preparation**

After creating a new derived date duration feature and cleaning the data for an outlier, missing values, and messy features, the diverse quantitative features and the large quantity of features were addressed. Because there was a diverse range of numeric values in the quantitative variables, the quantitative variables were scaled. Because there was skew in some of the quantitative features, the quantitative features were log-transformed.

The large number of different features and highly correlated features were addressed by performing a Principal Component Analysis (PCA) on the quantitative features. The PCA model reduced the 18 quantitative features to 16 features. This suggests that the quantitative features are not highly correlated amongst each other.

Low variance among the quantitative variables was also tested by running a variance threshold analysis with a 50% threshold. The result did not reduce any features.

Feature selection also includes determining which predictive features will be used as inputs in the model. Identifying the key features in a dataset that results in accurate predictions is a goal of the predictive model. However, in fraud detection claim transactions, the fraud indication may be in varied features in different records. Therefore, the predictive models will be tested using all the predictive features as inputs. Nonetheless, the models will also be tested with key features identified by Data Understanding and by CART Regression and Random Forest models that identified the most important features with respect to the fraud target feature. The CART Regression and Random Forest models revealed the following features as important features: incident severity (major damage), insured hobbies (chess and cross-fit), policy deductible, witnesses, number of vehicles and months claimant was a customer before the claim.

**Modeling**

Various classification techniques have been used for insurance fraud predictions, including logistic regression, decision tree, Random Forest, k-nearest neighbor, k-means, DBSCAN and agglomerative clustering algorithms, Baysian learning multilayer perceptron neural network, support vector machine, naïve Bayes, tree-augmented naïve Bayes classification algorithms, and Gradient Boosting Model (Moon, 2019), (Nian, 2016), (Vadoodparast, 2015), (Sibal, 2017). Based on recent research, the models used for this analysis are: (1) logistic regression, (2) LASSO (least absolute shrinkage and selection operator) logistic regression, (3) Random Forests, and (4) support-vector machine (SVM).

The data has over 50% categorical features. Because many statistical models can handle only numerical attributes, the categorical features were converted to numerical dummy features through one-hot encoding.

Model Evaluation and Selection

**Phase One**

Model evaluation includes assessing the general performance of a predictive algorithm and evaluating how well the model will work in the real world. Thus, the first phase of modeling included evaluating the four classification algorithms with baseline models. The baseline models use the hold-out method to split data into training and test data and used all the input features in the models. The accuracy scores from the baseline models indicate that they are all viable models.

|  |  |
| --- | --- |
| **Phase One Baseline Models** | **Accuracy Percentage** |
| Logistic Regression | 80.61% |
| LASSO Logistic Regression | 80.30% |
| Support Vector Machine (SVM) | 82.12% |
| Random Forest | 79.00% |

**Phase Two**

The second phase of modeling determined if the baseline models would be improved by using subsets of the select features identified in the Data Understanding and feature selection analysis discussed above. Twelve subsets of features were tested on the four baseline models. Half of the subset baseline models showed improvement over the baseline models that had used all features as inputs in the model.

**Phase Three**

The third phase of modeling improved the baseline models by tuning the four models’ hyperparameters to find the best version of the model and by using k-fold cross-validation instead of the holdout method to split the data to address bias and variation issues with the holdout method. All the input features, instead of a subset of features, were used in the models during this third phase of modeling. To address the imbalance in the target variable, each model specified a class weighting configuration of the inverse of the target class distribution present in the dataset.

The performance of each of the four models in the third phase improved over the baseline models.

|  |  |
| --- | --- |
| **Phase Three CV Grid Models** | **Recall Score** |
| Logistic Regression | .7882 |
| LASSO Logistic Regression | .7882 |
| Support Vector Machine (SVM) | .8643 |
| Random Forest | .6668 |

**Phase Four**

The fourth and final phase of modeling was performed to test if the Phase Three models improved by using the best subset of select features identified in Phase Two of modeling instead of including all the input features. A grid search and k-fold cross-validation were again used to tune the four models’ hyperparameters and to find the best version of the model. Each model specified a class weighting configuration of the inverse of the target class distribution present in the dataset. The performance of each of the four models in the fourth phase of modeling improved over the Phase Three models.

|  |  |
| --- | --- |
| **Phase Four CV Grid Models** | **Recall Score** |
| Logistic Regression | .8244 |
| LASSO Logistic Regression | .8244 |
| Support Vector Machine (SVM) | .8244 |
| Random Forest | .7370 |

**Results**

The Phase Four classification models, which used the best identified subset of input features, a grid search and k-fold cross-validation, are the best performing models as they have the highest recall scores. A classification model is evaluated by recall scores when, as in fraud detection cases, there is a high cost associated with false negative predictions. Recall scores calculate how many of the actual positive fraud reported cases are labeled as true positives by the model.

The Logistic Regression, LASSO Logistic Regression, and Support Vector Machine (SVM) models had the best recall scores and in fact all had recall scores of .8244. These identical scores are not a mistake. In these Phase Four models, the feature inputs were limited to a select few features. When the input features were all the features in the Phase Three models, these recall scores were not identical. Also, the SVM model had much higher computing times.

82.44% of the Logistic Regression, LASSO Logistic Regression, and SVM models’ actual fraud records were labeled as true positives by the respective models, which is a favorable result for fraud detection cases because false negative predictions are costly to the insurance company.

**Discussion/Conclusion**

The Logistic Regression, LASSO Logistic Regression, and SVM models are viable, useful models for predicting fraud. The Random Forest model did not perform as well. Using hyperparameters and k-fold cross-validation improved the accuracy of the models. The models that used only a few select distinctive features performed better than models employing all the input features. The best subset of input features included: incident severity (Major Damage), insured hobbies (chess), and policy deductible. The Logistic Regression, LASSO Logistic Regression, and SVM models performed the same with the small number of input variable, but not will the full set of input features. This seems to indicate that models handle many features/noise differently but handle a small number of features the identically.

In conclusion, incident severity, the insured’s hobbies, and the policy deductible are distinctive in predicting fraudulent claims. The number of witnesses to the incident, number of vehicles involved, and the length of time the claimant was an insured also have strong distinction for predicting fraud. If the concern is that the fraud can be indicated by varied features in the many-factored claim transaction, then the models using all the features also produced decent scores and could help an insurance company identify claims to investigate further. Both versions of the Logistic Regression, LASSO Logistic Regression, and SVM models would all be useful for predicting fraudulent automobile insurance claims.

# References

*Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claim Fraud*. (n.d.). Retrieved from wipro: https://www.wipro.com/en-US/analytics/comparative-analysis-of-machine-learning-techniques-for-detectin/

Dommalapati, A. (2019, June 5). *Using a Data Pipeline to Predict Insurance Claim Fraud*. Retrieved from Noteworthy - The Journal Blog: https://blog.usejournal.com/using-a-data-pipeline-to-predict-insurance-claim-fraud-6f8e85367294

Moon, H. ,. (2019). A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. *Theoretical Economics Letters*, 1886-1900.

Nian, K. Z. (2016). Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly. *The Journal of Finance and Data Science*, 58-75.

Sibal, N. (2017, December 14). *How insurers can identify fraudulent claims before they are paid*. Retrieved from Property Casualty 360: https://www.propertycasualty360.com/2017/12/14/how-insurers-can-identify-fraudulent-claims-before-they-are-paid/?slreturn=20200824134326

Vadoodparast, M. H. (2015). Fraudulent Electronic Transaction Detection Using Dynamic KDA Model . *International Journal of Computer Science and Information Security*.

<https://www.kaggle.com/patilk1/fraudulentinsuranceclaim>