Auto Insurance Fraud Claim Detection

Hung Cong Tran

January 12, 2023

Overview

1. Motivation and Problem identification
2. Data
3. Modeling results and analysis

Motivation and Problem identification

Fraudulent auto insurance accident claims are very popular in India nowaday.

Insurance companies are paying very large amounts of of wrong insurance payout and this hurts their profits significantly.

Also, insurance companies have to raise their insurance premiums and this also hurts their current customers.

Therefore, there is a big demand for a good prediction model to detect fraudulent vehicle accident claims from insurance companies.

Motivation and Problem identification

In this project, I will create a classification model that help predict whether or not vehicle insurance accident claims in India are fraudulent and reduce the amount of wrong insurance payout for insurance companies.

Data

The built model is 28836 auto insurance claims with features in four main categories as follow:

Insurance Claim Information (e.g. Type Of Incident, Type Of Collission, Severity Of Incident, Authorities Contacted, etc)

Customer Demographics (e.g. Insured Age, Insured ZipCode, Insured Gender, Insured Education Level, etc)

Insurance Policy (e.g. Insurance Policy Number, Customer Loyalty Period, Date Of Policy Coverage, Insurance Policy State, etc)

Vehicle Information (e.g. Vehicle Make, Vehicle Model, Vehicle Year of Model, , etc)

Modeling results and analysis (Model selection)

I try different classification models on the training data set and use cross-validation to compute the average AUC score. Here is the result:

Models	AUC scores
KNeighborsClassifier	0.90
Logistic regression	0.79
SVM (with polynomial kernel)	0.91
SVM (with radial basis function kernel)	0.91
Decision Tree	0.79
Random Forrest	0.91
AdaBoost with random forests	0.91
GradientBoostingClassifier	0.88
XGBoost	0.91

Modeling results and analysis (Model selection)

It seems that SVM, Random forest, Adaboost with random forests, and XGBoost models all perform well. For the simplicity and interpretability I choose the Random forest model to proceed.

After running the Grid Search to find the best random forest model, I chose the one with on 560 estimators.

Modeling results and analysis (Model selection)

The chosen model achieves the 0.92 AUC scores. We also achieved the accuracy 92% and the precision, recall, and F1-scores as follows:

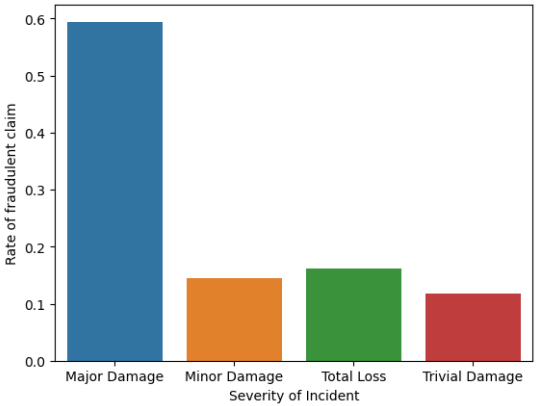
Report for the performance on the test set				
	Precision	Recall	F1-score	Support
Not Fraud	0.92	0.98	0.95	3953
Fraud	0.94	0.77	0.85	1477

Modeling results and analysis (Feature importance)

The five most important features that came up in the modeling as follows:

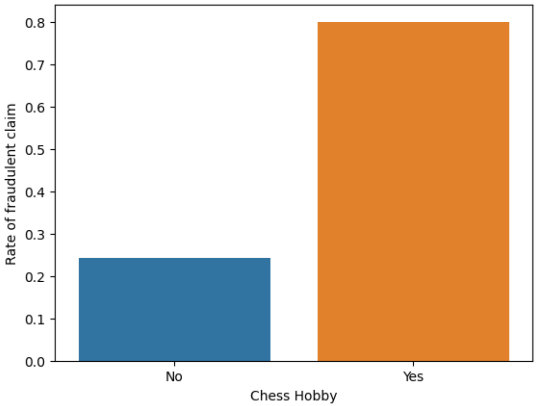
Feature Importance		
Rank	Feature	Contribution
1	Severity Of Incident	0.091433
2	Insured Hobbies chess	0.041979
3	Amount Of Property Claim	0.041606
4	Amount Of Injury Claim	0.039432
5	Amount Of Vehicle Damage	0.038610

Modeling results and analysis (Contribution of Incident severity)



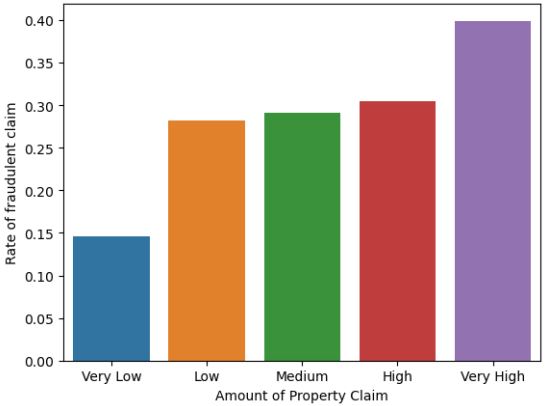
We can observe that there is a significantly higher rate of fraud among claims with major damage severity.

Modeling results and analysis (Contribution of Chess hobby)



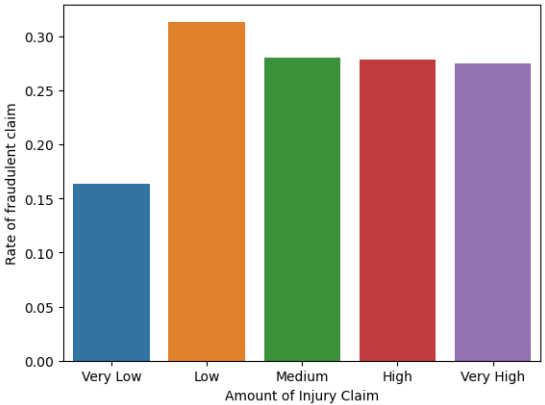
We can observe that there is a significantly higher rate of fraud among claims from people who like playing chess.

Modeling results and analysis (Contribution of amount of property claims)



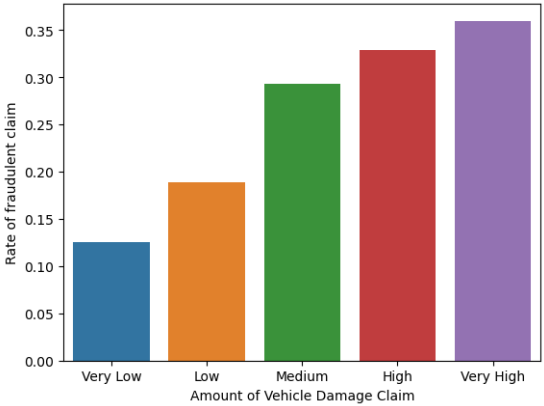
We can observe that the claims with very low amount of property claims are less likely as fradulent as the ones with very high amount of property claim.

Modeling results and analysis (Contribution of amount of injury claims)



We can observe that a claims with very low amount of injury claim has a small rate of fraud.

Modeling results and analysis (Contribution of amount of vehicle damage claims)



We can observe that the rate of fraudulent claims is positive correlated to the amount of vehicle damage claim.

The End