Vehicle Insurance Fraud Detection using Decision Tree-Based algorithms

Szymon Skiba

Abstract

Fraud is described as criminal activities for the economic and personal benefits [1]. Although it is not a new issue, rise of big data technologies and enlarging car market [2] allowed to collect large amount of insurance cases but it also become harder to detect. This paper presents vehicle insurance fraud detection method using Decision Tree-Based algorithms. A brief description is provided for both data used and methodology. We analyze and interpret the classifier predictions.

Keywords: Supervised learning, Decision trees, Data analysis, Fraud, Vehicle Insurance.

Introduction

Fraud detection and prevention has been a challenging task for many years. Insurance companies are especially targeted by fraudulent activities loosing up to 5% of revenue according to 2022 ACFE report[3]. It causes costs even for people not related to crime by generating judicial proceedings, causing job layoffs, damage to public and private property [4]. Unfortunately the exponential rise of the amount of data made it harder to detect it. Internal control team members should need to look at every transaction that takes place, but, unfortunately this issue can no longer be manually performed, requiring the use of data analysis tools and programs thus other effect of insurance frauds is development of IT systems and data analysis.

Although on the analytical market, offers many specialized tools capable to support enhance the antifraud activity, many managers and teams are not taking advantage of them. Data analytics (tools are currently in use in the organizations, but there is much lower adoption of more sophisticated tools, most people use Microsoft Excel, database tools such as MS Access or MS SQL Server[6] as presented on Figure 1. While these tools are important, they only focus on the grouping, matching, ordering and filtering of data and require substantial amount of human resources and time.

Forensic data	Percent
Spreadsheet tools such as Microsoft Excel	65%
Database tools such as Microsoft Access or Microsoft SQL Server	43%
Continuous monitoring tools, which may include governance risk and compliance (GRC) tools (SAP, SAI Global, Oracle)	29%
Text analytics tools or keyword searching	26%
Forensic analytics software (ACL, iDEA)	26%
Social media/web monitoring tools	21%
Visualization and reporting tools (Tableau, Spotfire, QlikView)	12%
Statistical analysis and data-mining packages (SPSS, SAS, R, Stata)	11%
Big data technologies (Hadoop, Map Reduce)	2%
Voice searching and analysis (Nexidia, NICE)	2%

Figure 1. Forensic data analytics tools in organizations

However classification techniques have proved to be very effective in fraud detection [5] and therefore it can used to analyze crime data and it is used more frequently. A unique contribution of this paper is presenting methodology and comparing the utility of Decision Trees, Random Forest [7], XGBoost [8], and AdaBoost [10], as tools for vehicle insurance fraud detection with Machine Learning. It also proves that classification Decision Tree-Based algorithms can be used to detect vehicle insurance fraud with high accuracy.

Methodology

A. Technology

Programing language used while conducting research is Python in version 3.9 . The most popular language used for data science according to 2022 Stack Overflow's "Developer Survey". Platform on which the experiments were conducted is "Jupyter Notebook". Hardware used is presented in Figure 2. While performing data analysis, building and evaluating models, python packages were used, the list goes as follows: pandas, numpy, matplotlib, seaborn, time, pandassql, imblearn, scipy, tabulate, xgboost, xgboost, hyperopt, sklearn, category_encoders.

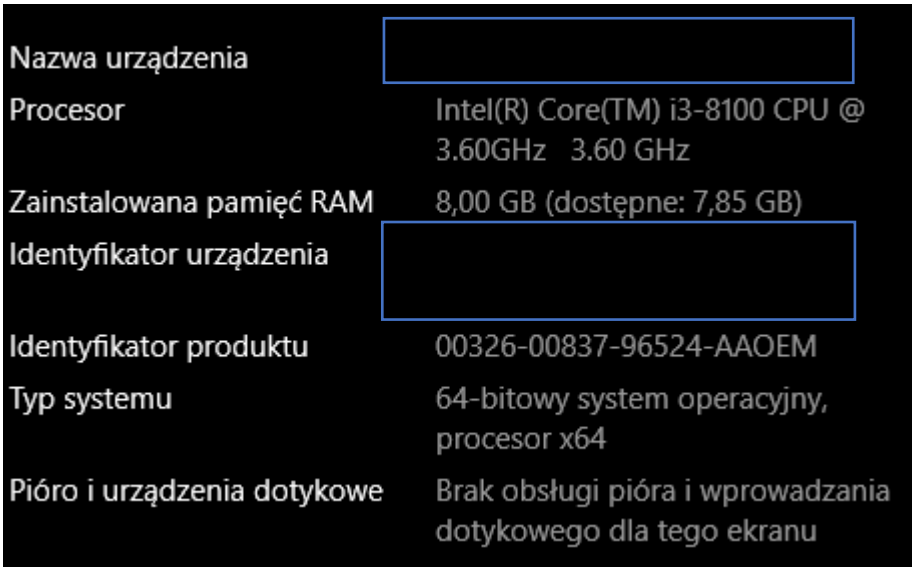

Figure 2. Hardware

B. Data

Data used for building and evaluating models comes from oracle machine learning database and is available on kaggle.com [14] after creating free account and is downloadable in form of .csv file. First the data had to be prepared thus steps to evaluate and modify had been taken i.e. process of data exploration:

1. Initial investigation and reworking the data.

   The data is in form of data matrix that contains 15420 rows where each row represents specific vehicle insurance case in years 1994-1996, and 33 columns that represent attributes of each case. The list of attributes, what they represent and data their type:

1. **Month** - object
   - contains 3 letter abbreviations for the months of the year
2. **WeekOfMonth** - int64
   - provides the week in the month when the accident occurred
3. **DayOfWeek** - object
   -contains days of the week the accident occurred on

4. **Make** - object
   - contains a name of car manufacturer
5. **AccidentArea** - object
   - classifies area for accident as "Urban" or "Rural"
6. **DayOfWeekClaimed** - object
   -contains the day of the week the claim was filed
7. **MonthClaimed** - object
   -contains 3 letter abbreviations for the months of the year
8. **WeekOfMonthClaimed** - int64
   - contains weeks in the month that the claimed in filed
9. **Sex** - object
   - gender of individual making claim
10. **MaritalStatus** - object
   - marital status of individual making claim
11. **Age** - int64
   -ages of individual making claim
12. **Fault** - object
   -categorization of who was deemed at fault
13. **PolicyType** - object
   -contains two pieces of information
14. **VehicleCategory** - object
   -contains the categorization of the vehicle (see PolicyType)
15. **VehiclePrice** - object
   -contains ranges for the value of the vehicle
16. **FraudFound_P** - int64
   -indicates whether the claim was fraudulent (1) or not (0)
17. **PolicyNumber** - int64
   - the masked policy number, appears to be the same as row number minus 1
18. **RepNumber** - int64
   - rep number is integer from 1 - 16
19. **Deductible** - int64
   - the deductible amount
20. **DriverRating** - int64
   - driver ratin in the scale: 1, 2, 3, 4
21. **Days_Policy_Accident** - object
   - number of days between when the policy was purchased and the accident occurred
22. **Days_Policy_Claim** - object
   - number of days that pass between the policy was purchased and the claim was filed
23. **PastNumberOfClaims** - object
   -previous number of claims filed by policy holder (or claimant?)
24. **AgeOfVehicle** - object
   -represents age of vehicle at time of the accident
25. **AgeOfPolicyHolder** - object
   - each value is a range of ages
26. **PoliceReportFiled** - object
   -indicates whether a police report was filed for the accident
27. **WitnessPresent** - object
   -indicted whether a witness was present
28. **AgentType** - object
   -this classifies an agent who is handling the claim as internal vs external
29. **NumberOfSuppliments** - object
   -unknown after initial investigation
30. **AddressChange_Claim** - object
   -time from claim was filled to when person moved (i.e. filed an address change)
31. **NumberOfCars** - object
   -number of cars involved in accident OR number of cars covered under policy
32. **Year** - int64
   - year accident occurred
33. **BasePolicy** - object
   -type of insurance coverage (see PolicyType)

After loading data stored in .csv file to pandas DataFrame function connected to this package allowed to extract unique values. In result it is possible to notice that many of attributes are given as range e.g. `VehiclePrice`. Another visible issues are some attributes are symbolic or categorical e.g `Martial Status`; `Age` contains values 0; `PolicyNumber`, does not represent any actual data it is only number of row plus 1. `FraudFound_P` represents if the specific case was fraud or not, this attribute in further described steps is used for training classification models since it is the only decision attribute(attribute describing result of classification).

During investigation of the feature 'Age' being set to 0, it was established that there are 7241 rows out 15419, which is roughly 46.96% of the data, who's 'Age' does not correspond to the age range for 'AgeOfPolicyHolder'. A discrepancy this prevalent feels unlikely to be a typo. A somewhat reasonable assumption, is that the individual driving at the time of the accident was not the policy holder, but another individual. Therefore the 0 was replaced with the mean value of the interval. Another step in reworking data set was also removing any duplicate rows.

2. Further Investigation.

The aim of this part was to gather some insight into the relationship between observations and the desired predicted feature, ‘FraudFound_P’. Visual representation and metrics were used to uncover correlations.

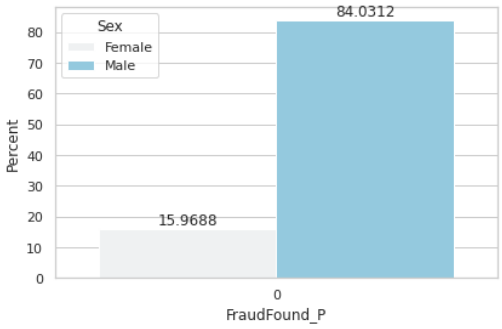

Figure 3. non-fraudulent transactions male to female

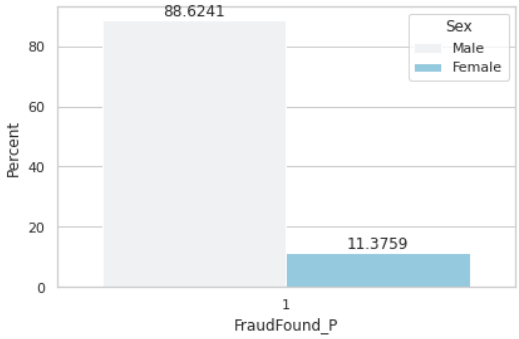

Figure 4. fraudulent transactions male to female

Column	Fraud cases percent	Non fraud cases
RepNumber	0.008095	0.002574
DriverRating	0.012922	0.004735
Age	0.013923	0.012332
Month	0.016755	0.004877
DayOfWeek	0.019506	0.019066
MonthClaimed	0.019883	0.006034
WeekOfMonth	0.058367	0.053399
WeekOfMonthClaimed	0.072729	0.067216
Make	0.079108	0.082115
DayOfWeekClaimed	0.096135	0.097706
Year	0.106360	0.064331
PastNumberOfClaims	0.128623	0.093488
AgeOfVehicle	0.132199	0.143240
AgeOfPolicyHolder	0.132668	0.129357
VehiclePrice	0.163007	0.195053
NumberOfSuppliments	0.179684	0.145157
PolicyType	0.182505	0.157858
BasePolicy	0.255064	0.053475
MaritalStatus	0.326613	0.325026
AddressChange_Claim	0.388396	0.409006
NumberOfCars	0.403495	0.407886
Days_Policy_Accident	0.436324	0.441237
Deductible	0.452058	0.476389
VehicleCategory	0.457765	0.295333
AccidentArea	0.503325	0.564281
Sex	0.546227	0.481258
Days_Policy_Claim	0.568908	0.573289
Fault	0.650416	0.301945
PoliceReportFiled	0.682592	0.666913
AgentType	0.700978	0.683985
WitnessPresent	0.702510	0.698912

Table 1.Standard deviation of each attribute

Column	Difference
WitnessPresent	0.002544
Days_Policy_Claim	0.003051
MaritalStatus	0.003428
Days_Policy_Accident	0.006578
NumberOfCars	0.007758
DriverRating	0.010947
PoliceReportFiled	0.011087
RepNumber	0.011534
AgentType	0.012016
WeekOfMonth	0.013914
WeekOfMonthClaimed	0.015602
DayOfWeekClaimed	0.016741
Age	0.019521
DayOfWeek	0.020216
Month	0.023727
AgeOfVehicle	0.024890
MonthClaimed	0.028345
AgeOfPolicyHolder	0.029036
Make	0.035431
Deductible	0.036020
AddressChange_Claim	0.038698
AccidentArea	0.043102
Sex	0.045940
Year	0.047701
NumberOfSuppliments	0.057872
VehiclePrice	0.068592
PastNumberOfClaims	0.090515
PolicyType	0.191769
BasePolicy	0.214045
Fault	0.246406
VehicleCategory	0.249084

Table 2. Max difference

Column	Difference
BasePolicy	-0.304057
PolicyType	-0.302539
VehicleCategory	-0.272817
Fault	-0.246406
VehiclePrice	-0.072162
PastNumberOfClaims	-0.060295
Sex	-0.045940
AccidentArea	-0.043102
NumberOfSuppliments	-0.042044
AddressChange_Claim	-0.037329
AgeOfVehicle	-0.037231
Deductible	-0.037062
Year	-0.036201
MonthClaimed	-0.035634
Month	-0.029840
AgeOfPolicyHolder	-0.029142
Make	-0.020643
DayOfWeek	-0.020375
DayOfWeekClaimed	-0.017552
WeekOfMonthClaimed	-0.016922
DriverRating	-0.015526
RepNumber	-0.013235
WeekOfMonth	-0.013147
Age	-0.012028
AgentType	-0.012016
PoliceReportFiled	-0.011087
Days_Policy_Accident	-0.008809
NumberOfCars	-0.007967
Days_Policy_Claim	-0.005060
MaritalStatus	-0.002685
WitnessPresent	-0.002544

Table 3.Min difference

Figure 3 and Figure 4 show that amongst the total claims males contribute 84.03% for non-fraudulent transactions and contribute 88.62% of fraudulent transactions. This kind of methodology was applied to other attributes. I measured percentage contribution of each unique value in attribute and contribution of fraudulent cases likewise. Then

	PolicyType	FraudFound_P	Total Accidents	Percentage by PolicyType	Percentage by Total
0	Sedan - All Perils	411.0	4086.00	10.059	2.666
1	Sedan - Collision	384.0	5584.00	6.877	2.490
2	Sedan - Liability	36.0	4987.00	0.722	0.233
3	Sport - All Perils	0.0	22.00	0.000	0.000
4	Sport - Collision	48.0	348.00	13.793	0.311
5	Sport - Liability	0.0	1.00	0.000	0.000
6	Utility - All Perils	41.0	340.00	12.059	0.266
7	Utility - Collision	3.0	30.00	10.000	0.019
8	Utility - Liability	0.0	21.00	0.000	0.000
9	923.0	15419.0	53.51	5.985	NaN

Table 4. Policy Type - Fraud Found

standard deviation was measured for each attribute and maximal and minimal amplitude for unique values in each attribute. The results are presented in tables 1, 2,3.

This comparison helps to differentiate attributes with bigger variability and influence on fraud. It helps to establish that RepNumber, Driver Rating, Age, Policy Number have low influence on fraud, therefore they were removed.

I also made comparison between fraud-found and policy type, fraud-found and martial fraud-found and months of the year when accident occurred to further investigate attributes correlation to fraud. Table 4 presents that Sport-Collision type of police is most likely to become a fraud. Table 5 showcases that most scams are committed in march and may.

	Month	FraudFound_P	Total Accidents	Percentage by Month	Percentage by Total
0	Apr	80	1280	6.25	0.519
1	Aug	84	1127	7.453	0.545
2	Dec	62	1285	4.825	0.402
3	Feb	82	1266	6.477	0.532
4	Jan	87	1411	6.166	0.564
5	Jul	60	1256	4.777	0.389
6	Jun	80	1321	6.056	0.519
7	Mar	102	1360	7.5	0.662
8	May	94	1367	6.876	0.61
9	Nov	46	1201	3.83	0.298
10	Oct	70	1305	5.364	0.454
11	Sep	76	1240	6.129	0.493

Table 5. Month - Fraud Found

3. Preparing data for classification

In this step data was prepare for classification algorithms that require data to be in numerical form. Category attributes were transformed to numerical data which is presented in Figure 5. Figure 6 presents ordinal-category attributes which were given values corensponding to their order e.g January -> 1, February ->2… .

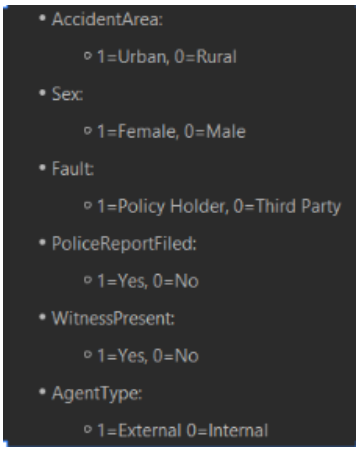

Figure 5.

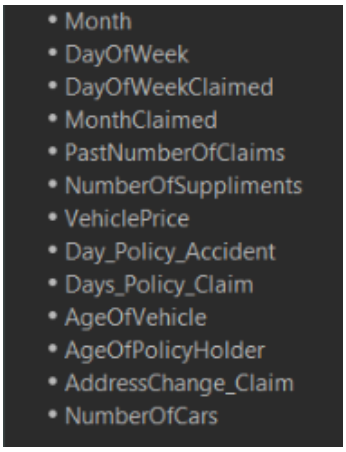

Figure 6.

C. Modeling

For modeling this data set, models from sklearn package were used: DecissionTreeClassifier(), RandomForestClassifier(), balancedRandomForestClassifier(), AdaBoostClassifier(), XGBClassifier(). The models were trained on randomly selected objects from data set. The size of training data set was 30% of original data set. Following the models were tested.

Results

model	run_time	avg_accy	avg_accy_std	avg_recall	avg_recall_std	avg_precision	avg_precision_std
DecisionTreeClassifier	0.01	0.892173	0.012398	0.194686	0.089629	0.085280	0.025956
RandomForestClassifier	0.18	0.939991	0.000869	0.000000	0.000000	0.059922	0.000755
BalancedRandomForestClassifier	0.26	0.635104	0.015657	0.897585	0.062797	0.123768	0.012108
AdaBoostClassifier	0.13	0.938521	0.002345	0.011413	0.018932	0.067050	0.013411
XGBClassifier	0.16	0.932727	0.005997	0.064915	0.044474	0.079075	0.024542

model	avg_f1	avg_f1_std	avg_matthew_corcoef	avg_matthew_corcoef_std	avg_roc_auc	avg_roc_auc_std
DecisionTreeClassifier	0.176183	0.072827	0.120509	0.078292	0.565655	0.044956
RandomForestClassifier	0.000000	0.000000	-0.000429	0.002309	0.499954	0.000248
BalancedRandomForestClassifier	0.227811	0.016293	0.248888	0.031623	0.757981	0.032719
AdaBoostClassifier	0.021342	0.035401	0.035003	0.076434	0.504511	0.009724
XGBClassifier	0.101730	0.068342	0.101305	0.084933	0.526478	0.022461

Figure 7. Results

Figure 7 presents statistical measures of each tested model. Results contain statistical measures like accuracy, precision and recall and also statistical measures made for binary classficators: f1_score, Matthews correlation coefficient. All models scored high on accuracy except BalancedRandomForestClassifier, but it scored the highest measures in every other category. Best accuracy is subscribed to RandomForestClassifier.

Discussion

In cause of fraud detection accuracy is the most important factor[15]. Comparing the different methods and metrics we see that the Decision Tree Classifier and XGBClassifier were the most successful, judged by the average F1 score and Matthew Correlation Coefficient from the 10-fold cross-validation. Of the four models tested the DecisionTreeClassifier and XGBClassifier are the top candidates. Other models scored very well as well. When it comes to BalancedRandomForestClassifier it is not really accurate but has high score in precision and recall wich indicates that it is good in identifying true positives (cases classified as fraud are fraud with high probability), although this may be good in some cases, in insurance industry it is preferred to classify as fraud as many cases as possible with high accuracy and then pass results for review to team of experts [15]. All in all decision trees score high in all metrics with low standard deviation wich indicates their effectiveness in identifying fraud. BalancedRandomForestClassifier is good in certain use cases do it almost perfect scores in precision and recall.

Limitations of this study is the data set itself. It contains only data from short period of time. Fraud cases in this period might be correlated to socio-econmical circumstances of this period therefore performed classfication might not apply more wider and be subjected to overfitting, process in wich models are only correct witch ceratain data and can not be effective with new data.

Other limitation comes from nature of classifiers. Models like presented in this study can only detect fraud cases based on already detected fraudulent cases. Therefore they will never found any new correlation between data and fraud. Therefore it will never eliminated need for expert evaluation.

Conclussions

In recent years, the interest in data analysis has increased, leading to growth of the industry. As a consequence, companies are investing in new technologies to improve their revenue. In insurance companies submitted insurance cases vertification is a crucial step for both processes and is currently largely dependent on human experts and simple filtering methods. This work aims at the prediction of fraud from objective analytical tests that are available with decision tree based algorithms. Performed case study and measurmnet of statistical metrics of prepared models showcases that classification can be really effective in detecting fraud. Although, it has limitations and will never replace human experts, it can be very effective tool and reduce time required to review submissions as well as expenses of reviewing them. Decision Tree-Based algorithms are effective in detecting fraud.

Refrences

[1] https://dictionary.law.com/default.aspx?selected=785
last visited: 20.01.2023

[2] https://ourworldindata.org/grapher/motor-vehicle-ownership-per-1000-inhabitants
last visited: 20.01.2023

[3]Association of Certified Fraud Examiners Occupational Fraud 2022.

[4] https://insurancefraud.org/fraud-stats/
last visited: 20.01.2023

[5] Chen, R., Chiu, M., Huang, Y., Chen, L.: Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines. In: IDEAL2004, 800--806(2004).

[6] EY – Global Forensic Data Analytics Survey 2018.

[7] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[8] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining—KDD ’16 (2016).

[9] Hancock, J.T., Khoshgoftaar, T.M. Gradient Boosted Decision Tree Algorithms for Medicare Fraud Detection. SN COMPUT. SCI. 2, 268 (2021).

[10] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol. 31, pp. 6638–6648. Curran Associates, Inc.; 2018.

[11] Adrian Bănărescu,Detecting and Preventing Fraud with Data Analytics,Procedia Economics and Finance,Volume 32,2015,Pages 1827-1836,ISSN 2212-5671

[12] Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal Of Engineering & Technology, 41(1), 33–40.

[13] FERNANDO, E. N. R. MACHINE LEARNING APPROACHES ON MOTOR INSURANCE FRAUD DETECTION. 2022. PhD Thesis.

[14] https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection?resource=download
last visited: 20.01.2023