STUDIA UNIVERSITATIS BABEȘ-BOLYAI OECONOMICA
VOLUME 64, ISSUE 2, 2019, pp. 53-71
DOI: 10.2478/subboec-2019-0009

IDENTIFYING KEY FRAUD INDICATORS IN THE AUTOMOBILE INSURANCE INDUSTRY USING SQL SERVER ANALYSIS SERVICES

BOTOND BENEDEK[*]
Babeş-Bolyai University, Romania

EDE LÁSZLÓ
Babeş-Bolyai University, Romania

Abstract.
Customer segmentation represents a true challenge in the automobile insurance industry, as datasets are large, multidimensional, unbalanced and it also requires a unique price determination based on the risk profile of the customer. Furthermore, the price determination of an insurance policy or the validity of the compensation claim, in most cases must be an instant decision. Therefore, the purpose of this research is to identify an easily usable data mining tool that is capable to identify key automobile insurance fraud indicators, facilitating the segmentation. In addition, the methods used by the tool, should be based primarily on numerical and categorical variables, as there is no well-functioning text mining tool for Central Eastern European languages. Hence, we decided on the SQL Server Analysis Services (SSAS) tool and to compare the performance of the decision tree, neural network and Naïve Bayes methods. The results suggest that decision tree and neural network are more suitable than Naïve Bayes, however the best conclusion can be drawn if we use the decision tree and neural network together.

JEL classification: C49, C88, G22, K42;

Keywords: automobile insurance, insurance fraud, fraud indicators, data mining.

1. Introduction

Nowadays the existence of a company may depend on the quick and accurate information gathering, on the analysis, on flexible development and on the innovation. More and more senior executives realize that the Internet, and the electronic data storage can be put at the service of the company. However, the data is not useful on its own, exists a demand by the companies for that information that

[*]Corresponding author. Address: Faculty of Economics and Business Administration, Babeş-Bolyai University, 58-60 Teodor Mihali Street, FSEGA Campus, Office 450, 400591 Cluj-Napoca, Romania, Tel: 0040 743 506 142, E-mail: botond.benedek@econ.ubbcluj.ro

can be obtained from the data, and are adaptable to their needs. This creates a new demand, a need for a tool, which is capable to analyze raw data for information gathering purposes (Bodon, 2010). This is especially true in marketing, where due to the transformation of the conventional segmentation policy, we have to find methods, which can help us to efficiently follow today's dynamically changing preferences. The solution can be the application of data mining algorithms, as data mining means searching those connections and global patterns in large databases, which are hidden behind the great mass of data. These connections can provide valuable information about the database and its objects, and if the database is a true reflection of reality, then also about the real world (Holsheimer and Siebes, 1994).

In the case of mandatory car insurance contracts, customer segmentation is a major challenge, as several factors influence the extent of potential compensation, and thus the determination of the optimal insurance premium. Such factors include, for example, the driving experience of the person who is driving the insured vehicle, the technical characteristics of the automobile or the potential car repair service where the repair takes place. Therefore, as a first step, the customers should be differentiated (more exactly the insurance premium paid by the customers) based on the probability of causing an accident and the amount of the possible claim for compensation due to the accident, must be defined. In Romania, this differentiation depends on the age of the insured person and on the engine capacity of the insured vehicle. Although the system is currently being used, there is a growing need for other factors which could be used in the determination of the car insurance premium. One of the factors is the probability of insurance fraud.

The studies conducted recently in Australia, United States and China reflected a growth trend of costs caused by the car insurance fraud. For instance, in 2014, the British Insurers Association (BIA) investigated the increase in the number of claims for false damages, resulting that it was 18% more than in the previous year (Insurance Fraud Bureau, 2015). The fraud proportion in the car insurance sector in Romania is at least 15%, compared to the maximum of 5-10% European average, declared Thomas Brinkmann, the Friss Country Manager of Greece and Cyprus at the International Insurance and Reinsurance Forum in 2017. In absolute numbers, this means that the level of fraud reaches annually 1.7-1.8 billion RON or approximately 400 million Euros. However, at the moment, the ability to control risks and detect frauds by the insurance companies is relatively undeveloped (Abdallah, et al., 2016). In addition, losses caused by car insurance fraudsters are not just a growth in temporary losses. They also have a serious effect on the development of the insurance industry and on the determination of insurance premiums. Due to higher compensation paid by insurers, insurance premiums are also higher (Wang and Xu, 2018). For this reason, the fast and reliable customer segmentation is essential because with its help, the higher insurance premiums will be paid by the potential fraudsters.

The paper is organized as follows: In the next section we provide a description of the background literature. Section 3 describes the data and presents the tools/technologies used in this study. Section 4 describes the results of the research, while in the last section we discuss the findings and draw the conclusions of the study.

2. Literature review

Based on Kotler and Armstrong (2010) market segmentation is no other than a strategy. A strategy which based on the separation of the whole market into segments of consumers. Consumers with different characteristics or behavior and with different needs which require different marketing policies from the companies (Kotler and Armstrong, 2010). Based on Liu et al (2019), market segmentation "can help firms know more about preferences and needs of consumers and tailor different policies for targeted segments in order to improve consumer satisfaction and increase revenue” (Liu et al., 2019 pp: 3). Nevertheless, market segmentation is a continuously developing process. In accordance with Wedel and Kamakura (2012) the development of market segmentation theories is affected by two main factors: the availability of marketing data and the advances in analytical techniques. A detailed review of various segmentation methods, approaches and solution applicable in case of different consumers can be found in (Wedel and Kamakura, 2012; Huerta-Munoz et al. 2017).

Based on Green, (1977); Wind (1978) or Wedel and Kamakura (2012) are two main type of segmentation methods: the priori and the post-hoc approach. In accordance with priori approaches companies identify the number and the characteristics of segments in advance based on their prior knowledge about the consumers. This knowledge based on indicators such as geographic areas, demographic characteristics (age, sex, etc) or purchase amounts (Frank and Strain, 1972; Green, 1977; Han et al., 2014). The other approach, the post-hoc recommends the execution of the segmentation just after the analysis of market data. In previous studies a lot of different segmentation method have been proposed in a post-hoc approaches. For example Dowling and Midgley (1988), Tsafarakis et al., (2008) or Balakrishnan et al., (2011) used clustering algorithms, Han et al., (2014) used category management, Fan and Zhang (2009) applied classification and regression trees, Kiang et al., (2006) used self-organizing maps, or Liu et al., (2010) used a multi-objective evolutionary algorithms for data analysis before the market segmentation.

Following the post-hoc approaches, with the help of a data mining tool, which identify the relationship between different attribute, we try to identify the fraudster’s segment.

Related work in car insurance fraud detection

One of the first and most cited methods was proposed by Phua et al. in 2004. They suggested the combination of stacking and bagging classifiers in order to detecting the fraud in car insurance. Initially the stacked ensemble selects the best classifier method from a group of base learner methods. Later, the bagging technique is used on the chosen classifier in order to analysis the oversampled dataset (Phua et al., 2004). Another approach recommended by Pathak et al. in 2005 recommends the application of fuzzy logic for detecting the fraudulent insurance claims, from a huge dataset (Pathak et al., 2005). Pinquet et al. in 2007 developed a statistical bivariate probit method to identify the illegitimate claims from a Spanish car insurance dataset (Pinquet et al., 2007). In 2008 Bermúdez et al. recommended a Bayesian dichotomous logit method, for detecting fraudulent insurance claims in a real car insurance dataset from Spain (Bermúdez et al., 2008). Šubelj et al, in 2011 proposed an expert system based on Iterative Assessment Algorithm. The method could detect the collaboration of automobile insurance fraudsters. Contrary with other solutions, the proposed method uses networks for data representation and because of this needs only unlabeled data for processing. (Šubelj et al., 2011). In 2011 Xu et al proposed a neural network combined with a random rough subspace method in order to identify the insurance fraud in automobile industry. As a first step the model segments the dataset into several subspaces with the help of rough set data space reduction method. After this, the neural network classifier was trained by using all the subspaces. Finally, the results of all neural network classifiers trained on the subspaces are combined using ensemble strategies (Xu et al., 2011). Tao et al. in 2012 proposed a fuzzy support vector machine for detecting the fraud in the automobile insurance (Tao et al. in 2012). In 2015 Sundarkumar and Ravi developed a detection method which is able to remove outliers from the dataset. By applying this method, the imbalance effect – typically presents in car insurance datasets – can be reduced. They used two unsupervised techniques in tandem, for resolving the skewed distribution problem. Namely the two techniques are the k-Reverse Nearest Neighborhood and the One Class Support Vector Machine (Sundarkumar and Ravi, 2015). Nian et al. in 2016 proposed the use of the spectral ranking anomaly concept for fraud detection in automobile insurance (Nian et al., 2016). Later in 2016 Hassan and Abraham in order to boost the performance of existing classifiers, recommend the use of the partitioning-under-sample method on the majority class in case of the imbalanced datasets. The results show, that decision tree based algorithms performs better than others and because of this, a decision tree based model was choose to compare the various partitioning-under-sampling approaches. The empirical results show that the proposed novel model have a better performance than previously suggested approaches (Hassan & Abraham, 2016). Li et al. in 2018 build up a principle component analysis based random forest then combine with a potential nearest neighbor method and tested on 12 datasets selected from various fields. One of the datasets was a real-word car insurance dataset. The experimental results illustrate that the recommended method has a higher classification accuracy and lower variance as the standard random forest, oblique decision tree ensemble or the rotation forest methods (Li, et al., 2018). Finally, Wang and Xu suggested a deep learning model which uses Latent Dirichlet allocation-based text analytics. Based on the experimental results on a real-world automobile insurance dataset, the suggested text analytics-based framework has a better performance than the traditional one. Furthermore, the authors highlight that the performance of the proposed framework is better, than the performance of the widely used random forest or support vector machine (Wang & Xu, 2018).

3. Methodology

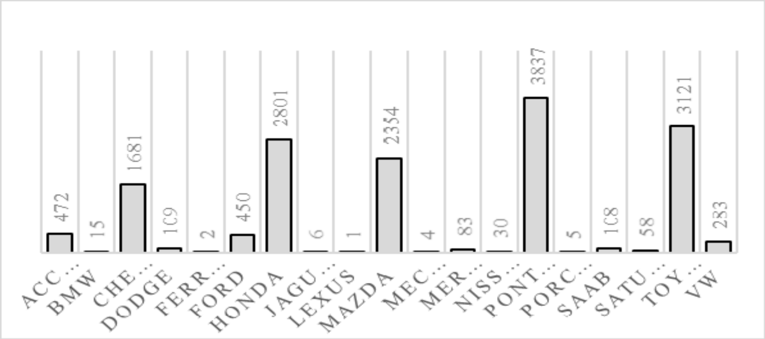
To build up the data mining models, used to analyze the data from the automobile insurance dataset and to identify the relations between different properties of the insurance policy holders in order to determine groups who are more susceptible to make a fraud, we used the *SQL Server Analysis Service.* It is an analytical processing and data mining tool in SQL Server, used to develop business intelligence, data warehousing and data mining applications. In this case the multidimensional database was used to organize data and express relations between attributes. There are two reasons we choose the SSAS. The first one is the price of the tool. We think that the current price of this tool makes it accessible for all the insurance companies and broker agencies. The second reason we choose this tool is the ease of use. This makes the tool available for the general users and not requires a software engineering background.

Data description

The dataset used for data mining was provided by Angoss Knowledge Seeker software with 15420 cases of car insurance policies from the United States of America. The dataset also contains the information if a case was a fraud or not. The data set contains 11338 cases logged between 1994 and 1995 and 4082 cases logged in 1996, however we won't analyze the time of these accidents. The main focus is to determine the relations between a fraudulent case and the characteristics of the policy holder. The car characteristics that was involved in the accident also could be important. The fraudulent cases represent 6% of the total cases. A detailed presentation of the attributes from the original dataset can be found in the appendix. However, in the next part we present the distribution of the most important indicators.

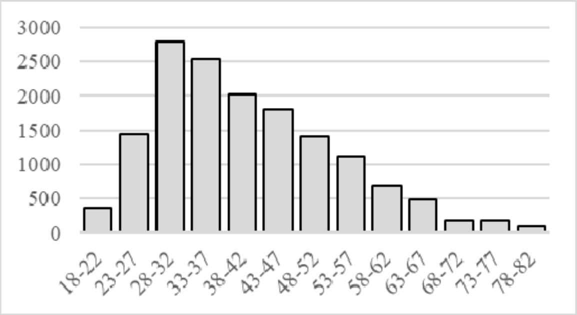
Figure 1. presents the statistical representation of the manufacturers. We can observe that Honda, Mazda, Pontiac and Toyota represent the majority of cases.

Figure 1. Manufacturers distribution



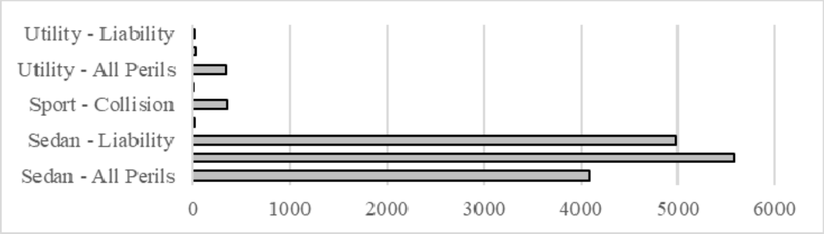
Further analyzing the dataset, we can see that the majority of cases happens in the urban area with 13822 cases. Figure 2 shows how the age of the policy holders are distributed.

Figure 2. Age distribution of policy holders



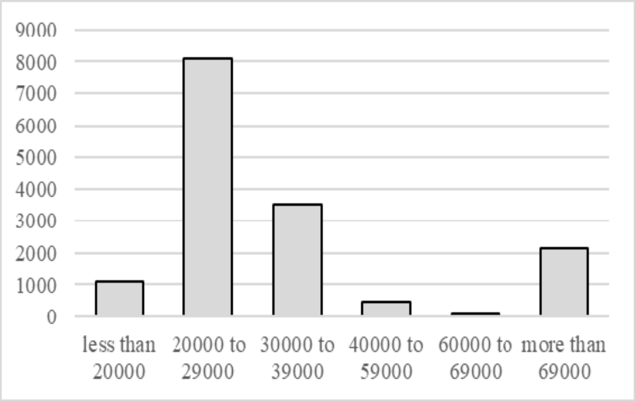
The next attribute we want to analyze is the marital status. We would also like to know if this attribute has any effect on fraud probability. The numbers are the following: 10625 married, 4684 single, 76 divorced and 35 widow policy holders. Another attribute that could have huge impact on fraud is the policy type. Figure 3 presents the distribution on policy types.

Figure 3. Distribution of policy types



Finally, the category and the price of the vehicle are important. The dataset contains 9671 sedan, 5358 sport and 391 utility type vehicles. The vehicle price distribution is presenting on figure 4.

Figure 4. Vehicle prices distribution



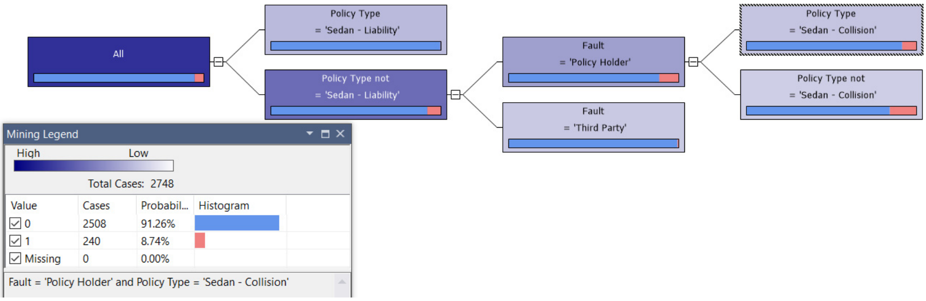
4. The applied data mining algorithms

In this section we present the methods that we use for data mining, on the real word automobile insurance fraud dataset. The data mining methods we used are the following: decision trees, naïve Bayes and neural networks. They are general and widely used models however their accuracy can vary based on the structure of data. Based on our literature review, these models could have the best performance in case of automobile insurance.

Before applying the above-mentioned methods, we needed to define some structure in SSAS which contains the input variables. In our first approach we used a structure which contains 24 from the previously presented 33 attributes. This mining structure was named as fraud characteristics mining structure. The PolicyNumber, Deductible, Days_Policy_Accident, Days_Policy_Claim, AgentType, NumberOfSuppliments, AddressChange_Claim, Year, BasePolicy attributes were not analyzed, because in our case, they are unique (ex. PolicyNumber), irrelevant (ex. Year), or not commonly used (ex. AgentType) attributes.

Based on figure 5 and figure 6, we can observe that by using decision trees, the two most important fraud indicators are Fault and Policy Type.

Figure 5. Distribution of Probability in case Fault = 'Policy Holder' and Policy Type = 'Sedan-Collision'

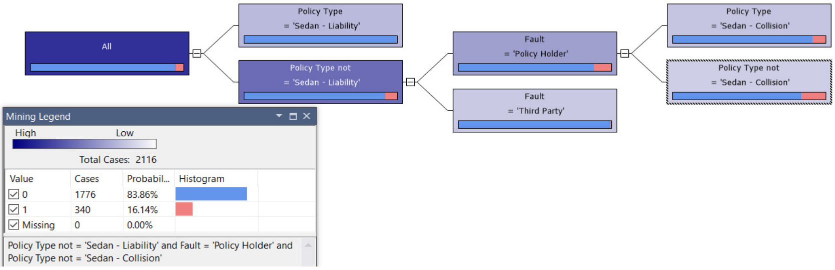


Total Cases: 2748

Value	Cases	Probabil...	Histogram
☑ 0	2508	91.26%	
☑ 1	240	8.74%	
☑ Missing	0	0.00%	

Fault = 'Policy Holder' and Policy Type = 'Sedan - Collision'

Figure 6. Distribution of probability in case Policy Type is not 'Sedan Collision' and 'Sedan Liability' and Fault = 'Policy Holder'



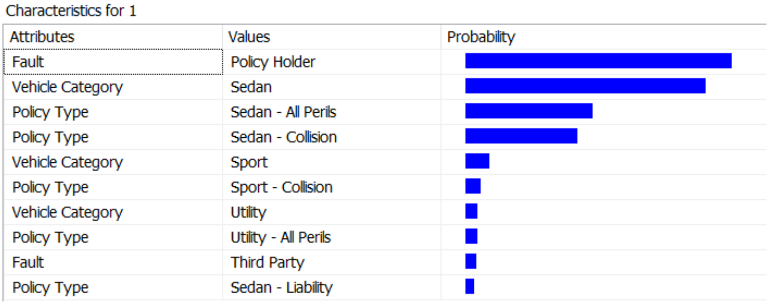
Total Cases: 2116

Value	Cases	Probabil...	Histogram
☑ 0	1776	83.86%	
☑ 1	340	16.14%	
☑ Missing	0	0.00%	

Policy Type not = 'Sedan - Liability' and Fault = 'Policy Holder' and Policy Type not = 'Sedan - Collision'

Using the Naïve Bayes algorithm on the fraud characteristics mining structure, we get three important fraud indicators: Fault = ‘Policy Holder’, Vehicle Category = ‘Sedan’, Policy Type = ‘Sedan – All Perils’ or ‘Sedan - Collision’. On the figure 7 we have these indicators ordered by their impact on the probability of fraud.

Figure 7. Key indicators and their values ordered by their impact on the probability of fraud

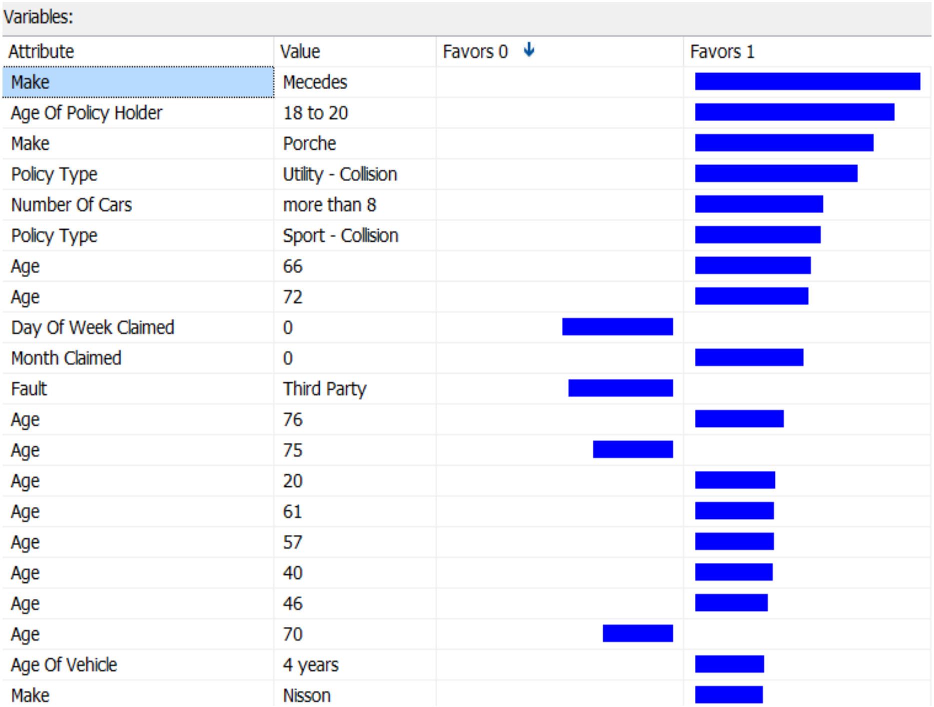


Characteristics for 1

Attributes	Values	Probability
Fault	Policy Holder	
Vehicle Category	Sedan	
Policy Type	Sedan - All Perils	
Policy Type	Sedan - Collision	
Vehicle Category	Sport	
Policy Type	Sport - Collision	
Vehicle Category	Utility	
Policy Type	Utility - All Perils	
Fault	Third Party	
Policy Type	Sedan - Liability	

By applying the Neural Networks algorithm, we received a list of input parameters (see figure 8) and their values ordered by their impact on the probability of fraud. On the figure can be observed that the highest probability of fraud was determined in case of the Mercedes vehicles. The same conclusion cannot be drawn in the case of the Decision Trees and Naïve Bayes algorithms, as the number of observations which have Mercedes as manufacturer in our dataset are quite small, compared to the other vehicle manufacturers. So, the results gained by using the Neural Networks algorithm are really important, however the big picture - the overall number of policies recorded and the share of Mercedes from it - couldn't be forgotten. Finally, on the figure is also visible the disadvantage of this algorithm, especially that it does not group these attributes as the other two methods.

Figure 8. Neural Network results on fraud characteristics mining structure



Variables:

Attribute	Value	Favors 0 ↓	Favors 1
Make	Mecedes		■
Age Of Policy Holder	18 to 20		■
Make	Porche		■
Policy Type	Utility - Collision		■
Number Of Cars	more than 8		■
Policy Type	Sport - Collision		■
Age	66		■
Age	72		■
Day Of Week Claimed	0	■	
Month Claimed	0		■
Fault	Third Party	■	
Age	76		■
Age	75	■	
Age	20		■
Age	61		■
Age	57		■
Age	40		■
Age	46		■
Age	70	■	
Age Of Vehicle	4 years		■
Make	Nisson		■

After the initial result, obtained from the mining structure with 24 attributes, we went into more details. To further analyze the data and find more key fraud indicators, we divided the attributes in four different structures (as recommended by Weisberg, 1998 or Belhadji, 2000) which are the following:

- car characteristics: Make, VehicleCategory, VehiclePrice, AgeOfVehicle
- accident characteristics: Month, WeekOfMonth, DayOfWeek, AccidentArea, Fault, PoliceReportFiled, WitnessPresent
- driver characteristics: Sex, MaritalStatus, Age, DriverRating, PastNumberOfClaims

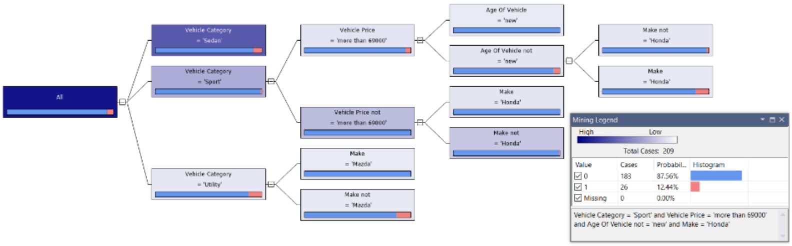
- claimant characteristics: DayOfWeekClaimed, MonthClaimed, WeekOfMonthClaimed, PolicyType, RepNumber, AgeOfPolicyHolder, NumberOfCars

Using these mining structures, we applied the same three algorithms.

Car Characteristics Mining Structure

By applying the decision tree algorithm on the car characteristics structure, we got a more detailed result of how the vehicle category and the manufacturer of the vehicle influence the fraud probability. On figure 9 can be observed that Vehicle Category 'Sedan' and 'Utility' are representing the majority of frauds. In the case of 'Utility' vehicles when the manufacturer is not 'Mazda' the probability of fraud is 13.90% almost twice compared to the base probability. In case of 'Sport' vehicles if the manufacturer is Honda and the vehicle is considered 'new' the probability of fraud is 12.44%.

Figure 9. Decision Tree applied on Car Characteristics mining structure


All
Vehicle Category = 'Sedan'
Vehicle Category = 'Sport'
Vehicle Category = 'Utility'
Vehicle Price = 'more than 69000'
Vehicle Price not = 'more than 69000'
Make = 'Mazda'
Make not = 'Mazda'
Age Of Vehicle = 'new'
Age Of Vehicle not = 'new'
Make = 'Honda'
Make not = 'Honda'
Make not = 'Honda'
Make = 'Honda'
Mining Legend
High
Low

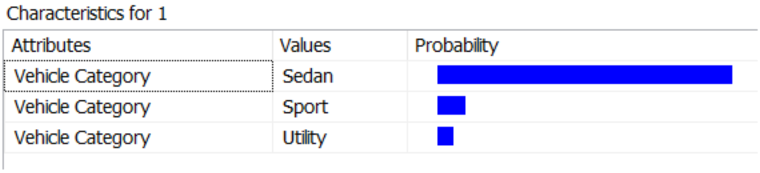
Total Cases: 209

Value	Cases	Probabil...	Histogram
☑ 0	183	87.56%	
☑ 1	26	12.44%	
☑ Missing	0	0.00%	

Vehicle Category = 'Sport' and Vehicle Price = 'more than 69000' and Age Of Vehicle not = 'new' and Make = 'Honda'

Using the Naïve Bayes algorithm, we only receive the distribution of the vehicle categories ordered by their impact on the probability (see figure 10).

Figure 10. Naive Bayes applied on Car Characteristics

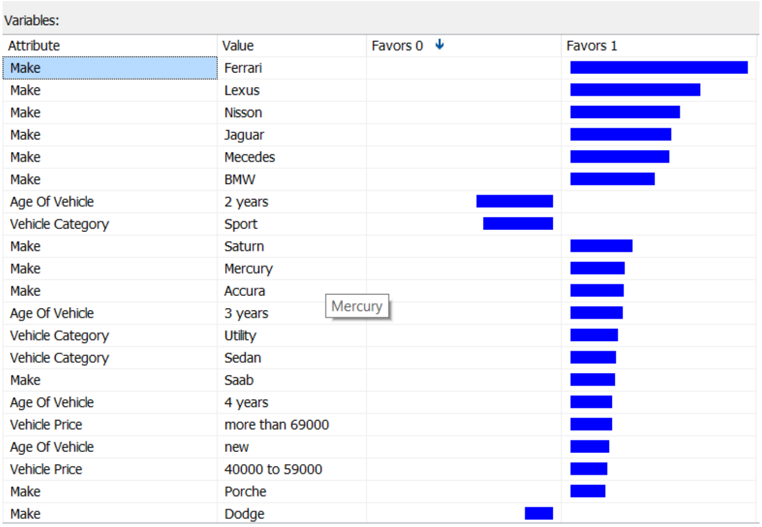


Characteristics for 1

Attributes	Values	Probability
Vehicle Category	Sedan	
Vehicle Category	Sport	
Vehicle Category	Utility	

The Neural Network results are showing us that more luxurious cars have a higher impact on the probability of fraud (see figure 11). The other key indicators of fraud are Age of Vehicle, Vehicle Category and Vehicle Price. For example, if the price of the vehicle is higher than 69000$ there is a higher chance of fraud.

Figure 11. Neural Network applied on Car Characteristics



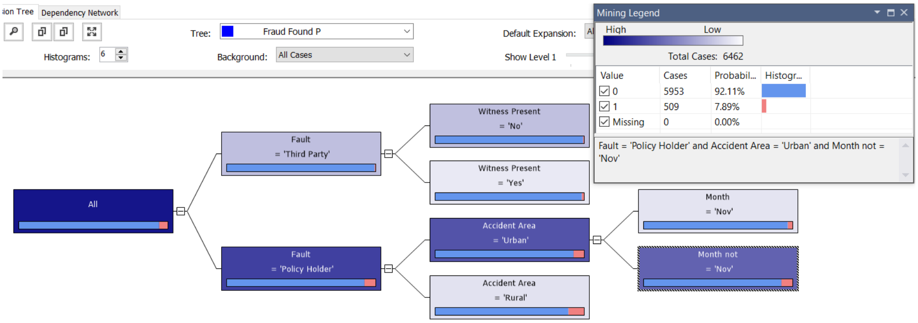
Variables:

Attribute	Value	Favors 0	Favors 1
Make	Ferrari		■
Make	Lexus		■
Make	Nisson		■
Make	Jaguar		■
Make	Mecedes		■
Make	BMW		■
Age Of Vehicle	2 years	■	
Vehicle Category	Sport	■	
Make	Saturn		■
Make	Mercury		■
Make	Accura		■
Age Of Vehicle	3 years		■
Vehicle Category	Utility		■
Vehicle Category	Sedan		■
Make	Saab		■
Age Of Vehicle	4 years		■
Vehicle Price	more than 69000		■
Age Of Vehicle	new		■
Vehicle Price	40000 to 59000		■
Make	Porche		■
Make	Dodge	■	

Accident Characteristics Mining Structure

When applying the Decision Tree on the accident characteristics' structure we receive a segmentation based on Fault, Witness Present and Accident Area (see figure 12). The most important indicator is the Fault. If the value of this attribute is 'Policy Holder' there is a higher fraud probability. Further going down on this path, can be observed, that fraud probability in case of 'Rural' accident area is higher than 'Urban', 11.05% compared to 7.54%.

Figure 12. Decision Tree applied on accident characteristics structure



Tree: Fraud Found P

Histograms: 6

Background: All Cases

Default Expansion: Al

Show Level 1

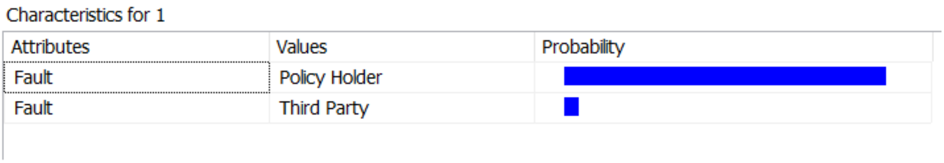
Total Cases: 6462

Value	Cases	Probabil...	Histogr...
☑ 0	5953	92.11%	
☑ 1	509	7.89%	
☑ Missing	0	0.00%	

Fault = 'Policy Holder' and Accident Area = 'Urban' and Month not = 'Nov'

On figure 13 can be obtained that Naïve Bayes identified one indicator (Fault) as important attribute which influences the fraud probability.

Figure 13. Naive Bayes result on Accident Characteristics

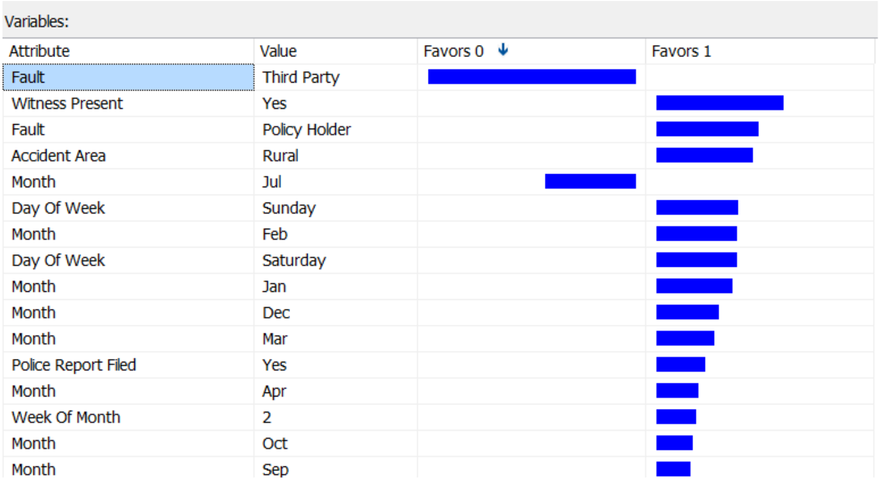


Characteristics for 1

Attributes	Values	Probability
Fault	Policy Holder	■
Fault	Third Party	■

Using the Neural Networks algorithm, of the figure 14 can be observed that if the Fault is ‘Third Party’ or Month of the accident is July, the likelihood of fraud decreases.

Figure 14. Neural Network results on accident characteristics



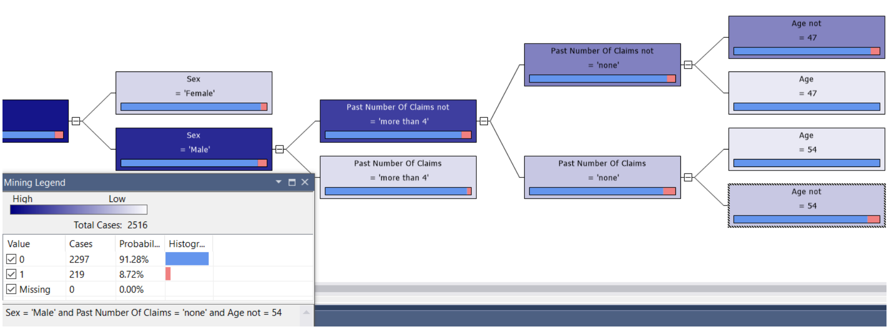
Variables:

Attribute	Value	Favors 0	Favors 1
Fault	Third Party	■	
Witness Present	Yes		■
Fault	Policy Holder		■
Accident Area	Rural		■
Month	Jul	■	
Day Of Week	Sunday		■
Month	Feb		■
Day Of Week	Saturday		■
Month	Jan		■
Month	Dec		■
Month	Mar		■
Police Report Filed	Yes		■
Month	Apr		■
Week Of Month	2		■
Month	Oct		■
Month	Sep		■

Driver Characteristics Structure

As shown of figure 15, the Decision Tree algorithm has identified 3 key indicators, Sex, Past Number of Claims and Age, which influence the fraud probability. For example, can be observed that if the driver is ‘Male’, the Past Number of Claims is less than 4, and the age differs form 54, the probability of fraud is 8.72%.

Figure 15. Decision Tree results on Driver Characteristics



Sex = 'Female'
Sex = 'Male'
Past Number Of Claims not = 'more than 4'
Past Number Of Claims = 'more than 4'
Past Number Of Claims not = 'none'
Past Number Of Claims = 'none'
Age not = 47
Age = 47
Age = 54
Age not = 54
Mining Legend
High
Low

Total Cases: 2516

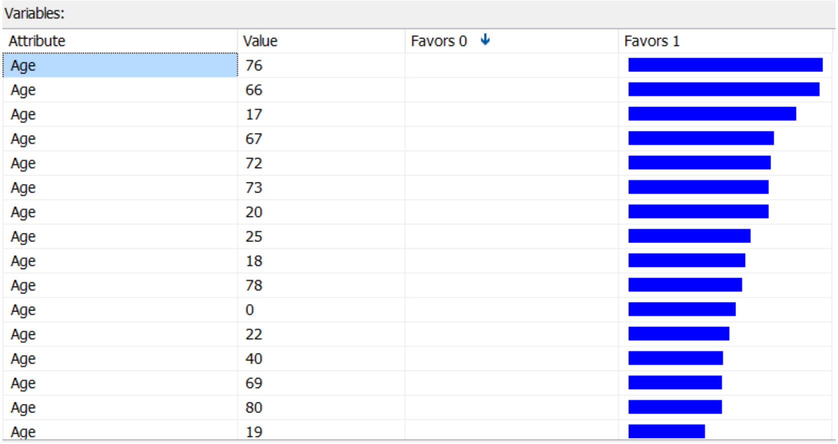
Value	Cases	Probabil...	Histogr...
☑ 0	2297	91.28%	■
☑ 1	219	8.72%	■
☑ Missing	0	0.00%	

Sex = 'Male' and Past Number Of Claims = 'none' and Age not = 54

Using the driver characteristics structure, the Naïve Bayes was not able to identify any impactful indicators.

Based on the Neural Network algorithm (see figure 16) can be stated that, if the age of the driver is between 16 and 25 or between 60 and 80, the probability that a fraud will happen is higher.

Figure 16. Neural Network results on Driver Characteristics structure



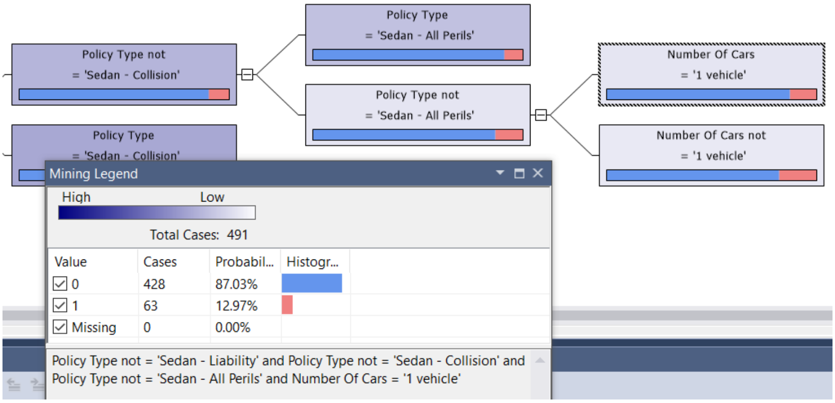
Variables:

Attribute	Value	Favors 0	Favors 1
Age	76		
Age	66		
Age	17		
Age	67		
Age	72		
Age	73		
Age	20		
Age	25		
Age	18		
Age	78		
Age	0		
Age	22		
Age	40		
Age	69		
Age	80		
Age	19		

Claimant Characteristics Mining Structure

By using the Decision Tree on the claimant characteristics mining structure, we obtain a tree (see figure 17) that is based on the Policy Type and Number of Cars input parameters and their values. We can identify them as the key indicators of fraud in this case.

Figure 17. Decision Trees applied on Claimant Characteristics structure


Policy Type not = 'Sedan - Collision'
Policy Type = 'Sedan - Collision'
Policy Type = 'Sedan - All Perils'
Policy Type not = 'Sedan - All Perils'
Number Of Cars = '1 vehicle'
Number Of Cars not = '1 vehicle'
Mining Legend
High
Low

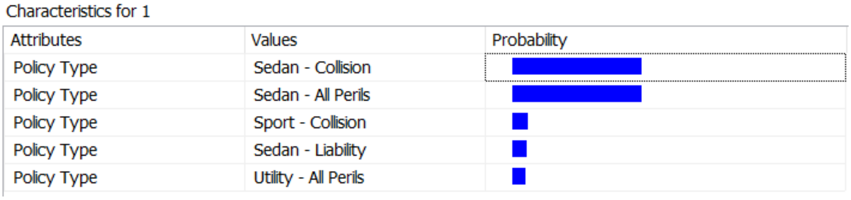
Total Cases: 491

Value	Cases	Probabil...	Histogr...
☑ 0	428	87.03%	🟦
☑ 1	63	12.97%	🟥
☑ Missing	0	0.00%	

Policy Type not = 'Sedan - Liability' and Policy Type not = 'Sedan - Collision' and Policy Type not = 'Sedan - All Perils' and Number Of Cars = '1 vehicle'

The Naïve Bayes algorithm (see figure 18) identified Policy Type, as the most important indicator and ordered their values based on their impact on the probability of fraud.

Figure 18. Naive Bayes applied on Claimant Characteristics

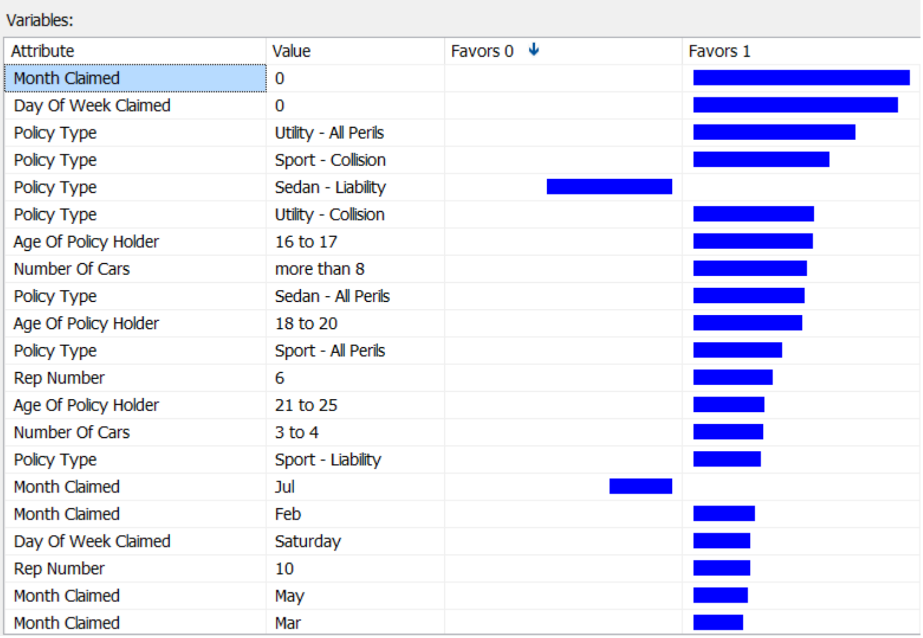


Characteristics for 1

Attributes	Values	Probability
Policy Type	Sedan - Collision	
Policy Type	Sedan - All Perils	
Policy Type	Sport - Collision	
Policy Type	Sedan - Liability	
Policy Type	Utility - All Perils	

Finally, the results of the Neural Network on the accident characteristics mining structure indicate an interesting correlation between the month the accident takes place and between the month the claims take place. More exactly, if the accident happened in the month of July and the month claimed is also July, these values favor that a fraud will not happen. Another interesting result is that Rep Number 6 and 10 appears also as a value that favors fraud. The most important indicators in these results are Policy Type, Age of Policy Holder, Number of Cars, Month Claimed and Rep Number.

Figure 19. Neural Network results in case of Claimant Characteristics



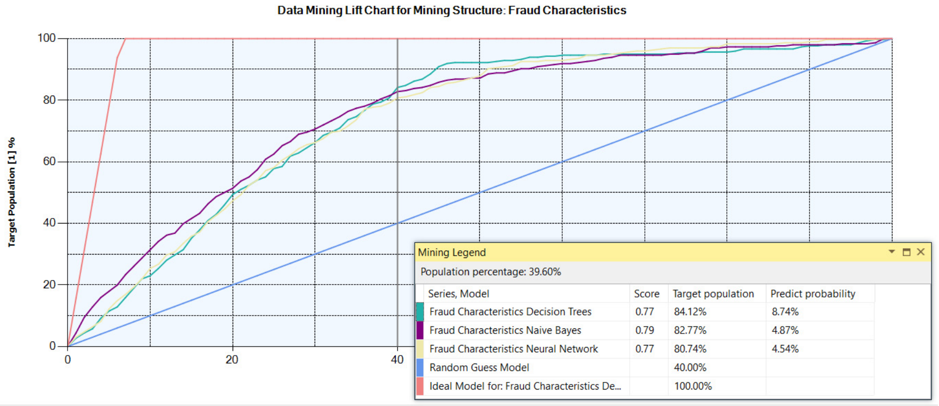
Variables:

Attribute	Value	Favors 0	Favors 1
Month Claimed	0		
Day Of Week Claimed	0		
Policy Type	Utility - All Perils		
Policy Type	Sport - Collision		
Policy Type	Sedan - Liability		
Policy Type	Utility - Collision		
Age Of Policy Holder	16 to 17		
Number Of Cars	more than 8		
Policy Type	Sedan - All Perils		
Age Of Policy Holder	18 to 20		
Policy Type	Sport - All Perils		
Rep Number	6		
Age Of Policy Holder	21 to 25		
Number Of Cars	3 to 4		
Policy Type	Sport - Liability		
Month Claimed	Jul		
Month Claimed	Feb		
Day Of Week Claimed	Saturday		
Rep Number	10		
Month Claimed	May		
Month Claimed	Mar		

Comparison of the applied algorithms

After applying the decision tree, naïve Bayes and neural network methods on the five mining structure, we compare the performance of this algorithms. To do this, we used a built-in function from the SSAS. This function – known as the Lift Chart – compares the performance of each used algorithms. In the next five figures can be observed the performance of each algorithms on each mining structure.

Figure 20. Performance of the decision tree, naïve Bayes and neural network algorithms on the initial mining structure

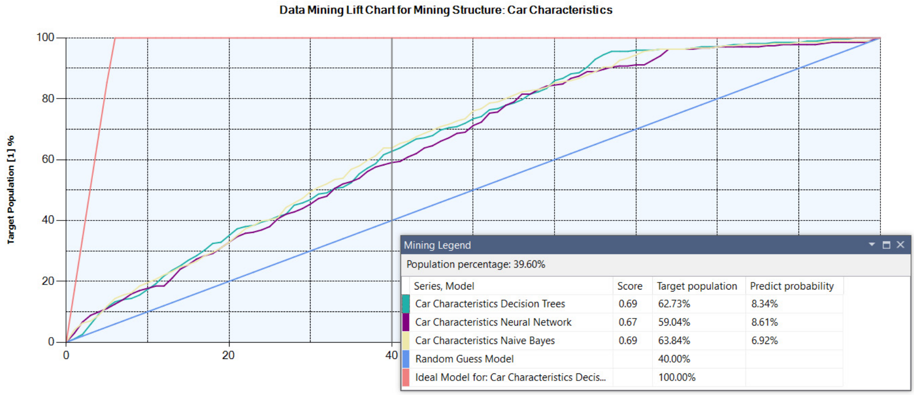


Mining Legend

Population percentage: 39.60%

Series, Model	Score	Target population	Predict probability
🟩 Fraud Characteristics Decision Trees	0.77	84.12%	8.74%
🟪 Fraud Characteristics Naive Bayes	0.79	82.77%	4.87%
🟨 Fraud Characteristics Neural Network	0.77	80.74%	4.54%
🟦 Random Guess Model		40.00%	
🟥 Ideal Model for: Fraud Characteristics De...		100.00%	

Figure 21. Performance of the decision tree, naïve Bayes and neural network algorithms on the car characteristics mining structure

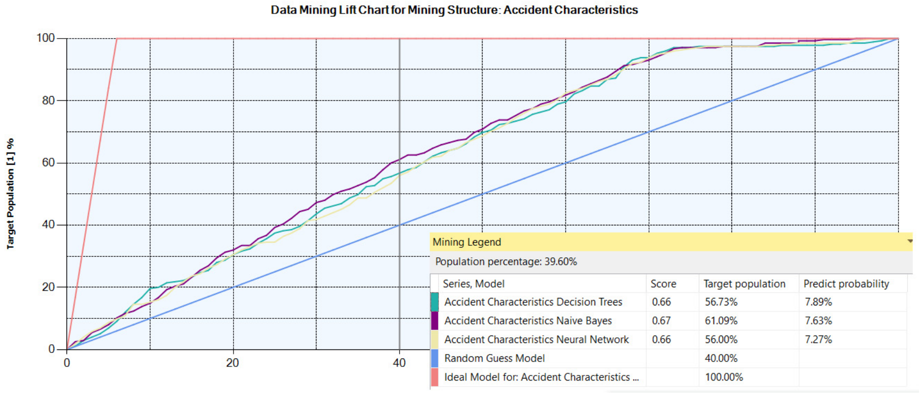


Mining Legend

Population percentage: 39.60%

Series, Model	Score	Target population	Predict probability
🟩 Car Characteristics Decision Trees	0.69	62.73%	8.34%
🟪 Car Characteristics Neural Network	0.67	59.04%	8.61%
🟨 Car Characteristics Naive Bayes	0.69	63.84%	6.92%
🟦 Random Guess Model		40.00%	
🟥 Ideal Model for: Car Characteristics Decis...		100.00%	

Figure 22. Performance of the decision tree, naïve Bayes and neural network algorithms on the accident characteristics mining structure

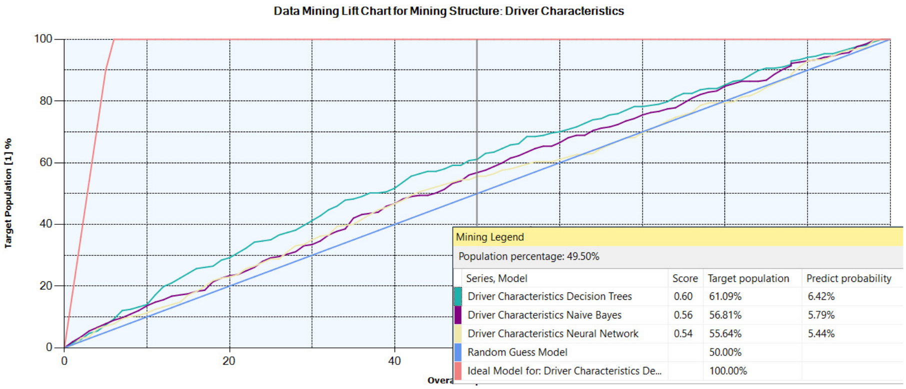


Mining Legend

Population percentage: 39.60%

Series, Model	Score	Target population	Predict probability
🟩 Accident Characteristics Decision Trees	0.66	56.73%	7.89%
🟪 Accident Characteristics Naive Bayes	0.67	61.09%	7.63%
🟨 Accident Characteristics Neural Network	0.66	56.00%	7.27%
🟦 Random Guess Model		40.00%	
🟥 Ideal Model for: Accident Characteristics ...		100.00%	

Figure 23. Performance of the decision tree, naïve Bayes and neural network algorithms on the driver characteristics mining structure

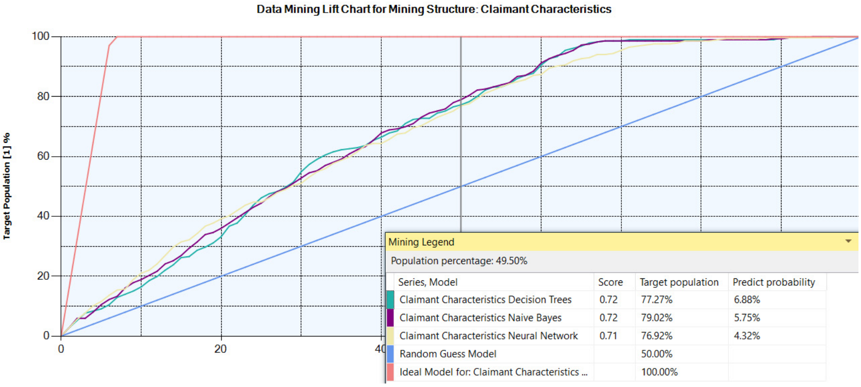


Mining Legend

Population percentage: 49.50%

Series, Model	Score	Target population	Predict probability
Driver Characteristics Decision Trees	0.60	61.09%	6.42%
Driver Characteristics Naive Bayes	0.56	56.81%	5.79%
Driver Characteristics Neural Network	0.54	55.64%	5.44%
Random Guess Model		50.00%	
Ideal Model for: Driver Characteristics De...		100.00%	

Figure 24. Performance of the decision tree, naïve Bayes and neural network algorithms on the claimant characteristics mining structure



Mining Legend

Population percentage: 49.50%

Series, Model	Score	Target population	Predict probability
Claimant Characteristics Decision Trees	0.72	77.27%	6.88%
Claimant Characteristics Naive Bayes	0.72	79.02%	5.75%
Claimant Characteristics Neural Network	0.71	76.92%	4.32%
Random Guess Model		50.00%	
Ideal Model for: Claimant Characteristics ...		100.00%	

Overall, it can be said, that in our case the decision tree algorithm has a better performance than other algorithms.

5. Discussion and conclusion

By using data mining methods for identifying key fraud indicators on a real-world automobile insurance dataset, we obtained connections between our input attributes (such as age, gender, car value, etc) and insurance fraud probability. With the use of SQL Server Analysis Services we took advantage of already defined and widely recognized algorithms like decision trees, naïve Bayes and neural networks. Another advantage of using this visual tool is, the ease of use. It is not necessary to have a software engineering background; however, the understanding of the algorithms could be an advantage in the interpretation of the results. So, the first important statement of this study, could be the following: the SSAS could be an accessible tool for all the insurance companies from Romania.

In order to identify key fraud indicators, we analyzed the fraud characteristics mining structure with the decision tree, naïve bayes and neural network algorithms. We identified Policy Type, Fault and Vehicle Category as key indicators. The results of the neural network gave us a better insight on our dataset and point out the need of a deeper analysis. Therefore, we created four separate mining structures, where we focused on the characteristics of the driver, accident, claimant and car, separately. By analyzing the characteristics of an accident, we identified two more key indicators: accident area and month of the accident. In case of the driver characteristics mining structure the tree most important indicators where: sex, past number of claims and age. The result obtained from the car characteristics mining structure showed that the key indicators are vehicle category, make, age of vehicle and vehicle price.

While examining the performance of models the results suggest that decision tree and neural network are more suitable than naïve Bayes. These results are in accordance with the statement of Hassan and Abraham, that decision tree-based algorithms perform better than others. However, based on our opinion the best conclusion can be drawn if we combine the decision tree and neural network results.

Limitations and further research

One of the main limitations of obtaining more fraud indicators were the small number of cases that we could analyze in this database. The data mining algorithms are designed to deal with millions (or hundred millions) of observations. They need huge amount of data for training purpose. By using 24 input attributes, the research would require a bigger number of cases.

Another limitation of the research is caused by the unbalanced type of the dataset. The related literature recommends some workarounds, but unfortunately most of them can't be implemented in SSAS. We think that this limitation could be resolve be using another (more complex of insurance specific) analytical tool.

Finally, the dataset is "old" and comes from the American market so it's less relevant in case of Romania, but as far as we know doesn't exist a Romanian dataset.

Acknowledgements

This research was supported by the Ministry of Human Resources of Hungary through NTP-NFTÖ-18 program.

References

Abdallah A., Maarof M.A., Zainal A. (2016) Fraud detection system: A survey, Journal of Network and Computer Applications, 68, 90-113.

Balakrishnan P., Kumar S., Han P. (2011) Dual objective segmentation to improve targetability: An evolutionary algorithm approach, Decision Sciences, 42(4), 831-857.

Bermúdez L., Pérez J.M., Ayuso M., Gómez E., Vázquez F.J. (2008) A Bayesian dichotomous model with asymmetric link for fraud in insurance, Insurance: Mathematics and Economics, 42(2), 779-786.

Bodon F., (2010) Adatbányászati algoritmusok, [Online] Available at: *www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf* [Accessed 06 01 2019].

Dowling G.R., Midgley, D.F. (1988) Identifying the coarse and fine structures of market segments, Decision Sciences, 19(4), 830-847.

Fan B., Zhang P. (2009) Spatially enabled customer segmentation using a data classification method with uncertain predicates, Decision Support Systems, 47(4), 343-353.

Frank R.E., Strain C.E., (1972) A segmentation research design using consumer panel data, Journal of Marketing Research, 385-390.

Han S., Ye Y., Fu X., Chen Z. (2014) Category role aided market segmentation approach to convenience store chain category management, Decision Support Systems, 57 296-308.

Green P.E., (1977) A new approach to market segmentation, Business Horizons, 20(1), 61-73.

Hassan A.K.I., Abraham A. (2016) Modeling insurance fraud detection using imbalanced data classification, Cham, Springer, 117-127.

Holsheimer M., Siebess A. (1996) Data mining: The search for knowledge in databases, Amsterdam: Centrum voor Wiskunde en Informatica.

Huerta-Munoz D.L., Rios-Mercado R.Z., Ruiz R. (2017) An iterated greedy heuristic for a market segmentation problem with multiple attributes, European Journal of Operational Research, 261(1), 75-87.

Kiang M.Y., Hu M.Y., Fisher D.M. (2006) An extended self-organizing map network for market segmentation - a telecommunication example, Decision Support Systems, 42(1), 36-47.

Kotler P., Armstrong G. (2010) Principles of marketing, Pearson Education.

Insurance Fraud Bureau, 2015. Cutting corners to get cheaper motor insurance backfiring on thousands of motorists warns the ABI. [Interactiv] Available at: *https://www.insurancefraudbureau.org/media-centre/news/2015/cutting-corners-to-get-cheaper-motor-insurance-backfiring-on-thousands-of-motorists-warns-the-abi/* [Accesat 01 09 2018].

Li Y., Yan C., Liu W., Li, M. (2018) A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification, Applied Soft Computing, Volumul 70, 1000-1009.

Liu J., Liao X., Huang W., Liao X. (2019). Market segmentation: A multiple criteria approach combining preference analysis and segmentation decision, Omega, 83, 1-31

Liu Y., Ram S., Lusch R.F., Brusco M. (2010) Multicriterion market segmentation: a new model, implementation, and evaluation, Marketing Science, 29(5), 880-894.

Nian K., Zhang H., Tayal A., Coleman T., Li, Y. (2016) Auto insurance fraud detection using unsupervised spectral ranking for anomaly, The Journal of Finance and Data Science, 2(1), 58-75.

Pathak J., Vidyarthi N., Summers S.L. (2005) A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims, Managerial Auditing Journal, 20(6), 632-644.

Phua C., Alahakoon D., Lee, V. (2004) Minority report in fraud detection: classification of skewed data, Acm sigkdd explorations newsletter, 6(1), 50-59.

Pinquet J., Ayuso M., Guillén M. (2007) Selection bias and auditing policies for insurance claims, Journal of Risk and Insurance, 74(2), 425-440.

Šubelj L., Furlan Š., Bajec M., (2011) An expert system for detecting automobile insurance fraud using social network analysis, Expert Systems with Applications, 38(1), 1039-1052.

Sundarkumar G.G., Ravi V. (2015) A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, Engineering Applications of Artificial Intelligence, Volumul 37, 368-377.

Tao H., Zhixin L., Xiaodong S. (2012) Insurance fraud identification research based on fuzzy support vector machine with dual membership. s.l., IEEE, 457-460.

Tsafarakis S., Grigoroudis E., Matsatsinis N. (2008) Targeting the undecided customer, In Proceedings of the 37th EMAC Conference.

Wang Y., Xu W (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud, Decision Support Systems, Volumul 105, 87-95.

Wedel M., Kamakura W.A. (2012) Market segmentation: Conceptual and methodological foundations, volume 8. Springer Science and Business Media.

Wind Y. (1978) Issues and advances in segmentation research, Journal of marketing research, 317-337.

Xu W., Wang S., Zhang D., Yang, B. (2011) Random rough subspace based neural network ensemble for insurance fraud detection. s.l., IEEE, 1276-1280.

APPENDIX

Appendix 1: Attributes of the dataset

No	Attributes name	Description
1	Month	The month in which the accident took place
2	WeekOfMonth	The week of the month
3	DayOfWeek	The day of the week
4	Make	The car manufacturer
5	AccidentArea	Accident aria, rural or urban
6	DayOfWeekClaimed	The claim day of the week
7	MonthClaimed	The month of the claim
8	WeekOfMonthClaimed	The week of the claim
9	Sex	Gender, male or female
10	MaritalStatus	Marital Status
11	Age	Age of policy holder
12	Fault	Policy holder or third party
13	PolicyType	Type of the policy
14	VehicleCategory	Sedan, sport or utility
15	VehiclePrice	Price of the vehicle with 6 categories represented in dollars
16	FraudFound_P	Fraud
17	PolicyNumber	Unique identification number of the policy
18	RepNumber	Id of the person who process the claim
19	Deductible	Amount to be deducted before claim disbursement
20	DriverRating	Driving experience with 4 categories
21	Days_Policy_Accident	Days left in policy when accident happened
22	Days_Policy_Claim	Days left in policy when claim was filed
23	PastNumberOfClaims	Past number of claims
24	AgeOfVehicle	Vehicle's age with 8 categories
25	AgeOfPolicyHolder	Policy holder's age with 9 categories
26	PoliceReportFiled	Yes or no
27	WitnessPresent	Yes or no
28	AgentType	Internal or external
29	NumberOfSuppliments	Number of supplements
30	AddressChange_Claim	No of times change of address requested
31	NumberOfCars	Number of cars
32	Year	1994, 1995 and 1996
33	BasePolicy	All perils, collision or liability