INTERNATIONAL
HELLENIC
UNIVERSITY

Knowledge Discovery from Insurance Data

Katrilakas Georgios

SID: 3308180005

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2019

THESSALONIKI – GREECE

Knowledge Discovery from Insurance Data

Katrilakas Georgios

SID: 3308180005

Supervisor:		Prof. A. Papadopoulos
Supervising Members:	Committee	Assoc. Prof. Name Surname
		Assist. Prof. Name Surname

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2019

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in Data Science at the International Hellenic University. In this dissertation we are going to use a data set that includes insurance claims. We are going to use machine learning techniques that can classify our cases to fraud or non-fraud ones. The purpose is to train our algorithm in order to predict whether a new claim is fraud or not.

Katrilakas Georgios

Date

02/12/2019

Contents

1 Introduction

This dissertation is about the never-ending problem of insurance claims fraud. Upon the beginning of the history of insurance companies, people tried to receive greater refunds for their claims than the amounts that should receive. In the multiple sectors that insurance companies are involved, there are many ways for people to cheat. A company can sell medical, life, boat, vehicle coverage and for vehicles can split the damages into no-fault insurance, theft, crystal breakage and many more. In this dissertation, we are going to use a dataset of simple car accidents where a driver causes damage and his/her insurance company tries to compensate the loss and we will try to use the ability of the classification algorithms to separate true claims from false ones by using the attributes of a case.

At the beginning we are going to analyse some of the literature that exists about the topic where we can have an idea on how the problem was reached until now. Afterwards, we are going to use data mining techniques in order to end up to some conclusions about the information that lays hidden in the dataset. According to (Charu C. Aggarwal, "Data Mining") "data mining is the study of collecting, processing, analysing and gaining useful insights". As one can deduce the term can cover a broad range of data processing. By using data mining techniques and algorithms we can handle and provide useful deductions and results by using all this flood of data which come from the nowadays technology.

Analysts, in order to cope with the data, use so called "pipelines of processing". The pipeline is a methodology by which the data are collected, cleaned and transformed into specific types for handling. In this point, we need to mention that the greater amount of work is needed in the preprocessing phase, where the data need to be untangled and reconstructed in ways that can be processed. Further to the above, we are going to have an analysis of the data we have downloaded. This step is crucial because it gives the opportunity to the user to become familiar with the data that has to handle. The results are going to determine which of the attributes are useful to use in our models and which of them are not. Finally, after the analysis, we are going to apply some classification

methods using the attributes and we will try to train our algorithm in order to learn from the data and differentiate the fraud claims from the proper ones.

2 Literature Review

This dissertation is about knowledge discovery from insurance data. This work tries to provide solution to a problem that insurance companies care about most, fraud. In this dissertation we are going to use car insurance dataset which contains whether an accident is a fraud or not and we will try to build an algorithm that explores the data and tries to "understand" it by using the attributes provided.

2.1 Further Reading

In order to achieve the above target, we need to look through some bibliography and try to approach the problem and propose an enhanced solution to it. To begin with, we need to provide a simple definition of what vehicle insurance fraud is. As per the author of (Subudhi & Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud") "insurance fraud is when a person attempts to obtain economic refund by submitting false evidence of a car accident or by submitting damage caused from previous accidents or intently does not declare full information or even wrong about the enmeshed people". Additionally the authors of (Belhadji, Dionne, & Tarkhani, "A Model for the Detection of Insurance Fraud", 2000) claim that every one of the cases that are characterized as fraud, whether or not were brought in court, are called established frauds. As per the article (An expert system for detecting automobile insurance fraud using social network analysis), it is also vital to understand that fraud is not always planned. The individuals are trying to just seize the opportunity to claim more money for their property. So, from analysis by the author of (Šubelj, Furlan, & Bajec, "An expert system for detecting automobile insurance fraud using social network analysis", 2010), we can create a pattern of staged accidents which include late hours, non-urban areas, young ages. Upon now, a fraud is only suspected by the adjuster and it is on his own will and experience if the company would examine a case further. Therefore, as mentioned by the author of (Šubelj, Furlan, & Bajec, "An expert system for detecting automobile insurance fraud using social network analysis", 2010) only 20% of frauds are indeed examined, because still investigation is carried out by hand and not by using the help of a computer.

The indicators recorded by the authors of (Belhadji, Dionne, & Tarkhani, "A Model for the Detection of Insurance Fraud", 2000) are the ones which are significant in predicting the probability that the file is fraudulent. If the goal is detection, a large number of files must be sampled. Having large dataset has two shortcomings: 1) Reviewing a large data file is costly for the company 2) The re-examination of a broad range of files would entail some unfairness towards clients who are not cheaters but they will be closely investigated by the insurance company for fraud.

Basically, the writers mention that it depends on what threshold the company wants to examine for fraud. If the company needs to re-examine all the files that have probability of fraud over 90%, then the investigators will come across a small dataset to closely examine. The second advantage is that in 90% threshold the results would be more accurate. The authors, having examined these extreme cases (P =10% and P = 90%), saw that there is a trade-off between detection and accuracy: the higher the fraud probability threshold the greater the accuracy and the weaker the detection. The use depends on the company, if the company does not want to get too involved into fraud detection, it would choose a big threshold (like 90%). On the other hand, if a company is strict enough, they would rather choose a lower threshold of re-examining files. When a company decides to pursue an investigation, they check if the cost of settlement is lower than the cost of investigation, then there will not be any investigation, otherwise the investigation will be pursued. So in (Belhadji, Dionne, & Tarkhani, "A Model for the Detection of Insurance Fraud", 2000) they ended up to the following results: first there is a need to run the model and decide whether we will conduct an investigation and afterwards we will need to decide, if the file tends to be fraudulent based on our estimate, whether we will conduct a depth investigation and how possible will be to be successful.

In the beginning of (Belhadji, Dionne, & Tarkhani, "A Model for the Detection of Insurance Fraud", 2000) the authors try to establish a quick definition of insurance. They claim that insurance is an agreement between a company and an individual that aims to deteriorate economic loss in case of an accident or theft. In these cases, fraud is when the customer tries to obtain budgetary asset by using a staged accident or by claiming older damages.

At this point, the authors mention that two reasons make the fraud detection really difficult. The first point is that there is a lot missing information from the claim and second, there is a lack of experience of this kind of deceit as these cases are much less compared to the whole. In terms of machine learning, as per the bibliography, this uneven arrangement of the data tends to cause trouble to supervised classifiers. These algorithms are inclined to classify according to the major class disregarding the minority. This problem led the authors to propose a new data balancing tool which they called “ADASYN” (Adaptive Synthetic Approach for Imbalanced Learning).

After running some tests, the authors came to conclusion that the process of identifying fraud cases is pretty difficult given the fact of information lack and skewness of data. They proposed “ADASYN” to make matter easier and the article explained how the algorithm generates synthetic points and merges them in order to produce a balanced data set. Finally, the results showed that the use of this algorithm really improved the procedure.

Among plenty algorithms, Naive Bayes seems to be the most powerful one which can detect fraud more efficiently. However, we cannot use this specific technique for bunch of data which include smaller amount of fraud cases. According to the authors of (Robust fuzzy rule based technique to detect frauds in vehicle insurance) there are several types of fraud: The first contain the staged auto accidents, the second are about counterfeit air bags replacements, the third are wind shield replacement rip off and the last is towing scams. The fraud caused by some people lead to higher insurance fees which outrages other consumers

However, insurance companies need to protect themselves as per the authors by using some existing techniques like Bayesian Network, Decision Tree, Back propagation. Further to this, this article now proposes a new application, called Fuzzy, which tries to represent the data in different forms of knowledge in order to solve the problem and extract relationships that exist among the variables. Its strong advantage, according to the team that produced it is that Fuzzy logic is an extension to the classical methods as the conventional techniques do not deal with uncertainty and imprecision. It provides solutions to deal with the uncertain and imprecision environments. Also, fuzzy can be applied to rule-based systems because by using approximate reasoning methods it gives the advantage of handling uncertainty and inference methods which are robust. Finally, they (the authors) mention that in order to apply fuzzy rule base the user needs

to do two steps. One is to remove noise and “predict” missing values by using k means and two, use PCA to reduce data dimensionality.

The article (Artís, Ayuso, & Guillén, "Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims", 2002) shows how binary choice models work on fraud detection and uses algorithms for misclassification in the response variable. The authors also mention that detecting automobile fraud has become a very crucial problem for insurance companies that needs to be solved. They mention that when a company agrees with a client to cover his risk, the company needs to “calculate” that the insured person will not be honest and honor their contract. A theory, known as costly state verification, claims that the company can have information on whether a claim is fraudulent or not by “using” the cost of the claim. However, there is a high chance that some claims may be misclassified. In order to determine if an honest claim is really honest there is a theory developed by Weisberg and Derrig which uses a multiple linear regression model to select various features which can lead to a fraudulent claim. Other techniques that are suggested by the bibliography are: fuzzy set techniques, self-organizing neural net to transform claims characteristics to claim types and discrete choice method that uses previous knowledge and estimates the probability of fraud.

The aim of all these models is to provide tools to recognize fraud. As it understandable, the previous data regarding claims that companies have are imperfect though. In some sets only the indicators that lead to characterize a claim as fraudulent are available so we cannot have a general picture of “legal” claims. Thus, the observations are limited only to fraudulent claims. The authors of this paper suspect that only the honest claims hide inside some fraudulent claims and not the other way around. This phenomenon is known as omission. Thus, the authors of (Artís, Ayuso, & Guillén, "Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims", 2002) propose that given the dataset, the user should take as a fact that fraud exists, if the insured admits to it because legal prosecution is very rare situation.

The authors of (Nian, Zhang, Tayal, Coleman, & Li, "Auto insurance fraud detection using unsupervised spectral ranking for anomaly", 2016) have suggested the use of unsupervised learning for detecting fraud, because the procedure of obtaining labels is rather costly and many times infeasible. The article tries to explain why

spectral ranking of anomaly is a better, and a more relaxing SVM type, unsupervised learning method for data mining. They have used an auto insurance claim dataset in which they tried to show that using ranking instead of labels gives the problem a new type of solution which is not based on the usual outlier detection methods.

The purpose of using the SRA is to detect fraud via the method of interdependence relation. As per the authors, the main reason to use unsupervised learning instead of supervised is because the first method can discover patterns of fraud and this can lead to information for the case before it is really done due to patterns that exist in all these cases. This is way more efficient rather than having investigators who they just speak their mind and they may be also wrong. However, it is very challenging to define all the parameters to “discover” a pattern thus all the bibliography tries to solve the problem via supervised methods.

As per their bibliography, the authors of (Viaene, Dedene, & Derrig, "Auto claim fraud detection using Bayesian learning neural networks", 2005) notice that detection of fraudulent claims has become very important the last years as there is an increasing frequency for fake accidents. Nowadays the companies are able to store and organize their data electronically. These techniques create more necessity to handle all this data better and by using the appropriate tools the users can analyse and model formal relations between fraudulent claims and suspicious ones. This is the reason that machine learning is recruited to do all the exploration automatically. Statistical models, regression and linear analysis are widely used for prognosis. However, parts of them are rigid and limit the usefulness of the models. This is the reason why neural nets are used as they provide a more flexible way to do the analysis. A strong disadvantage of neural nets is that they are black boxes and humans cannot parametrize them as needed. Thus, as the authors mention, the experts need to grasp the patterns and the reasons that are hidden below the results of neural nets.

Insurance companies concern about two categories of claims. The first is the ones that refer to illegitimate frauds or build up frauds while the second one is a category where the person’s claims are magnified. The authors mention that we tend to see this phenomenon, more often, in claims that include injuries. As a matter of fact, research has shown that individuals tend to exaggerate about the damage done on their vehicle rather than stage an accident from scratch. Thus, there is a large field of literature that investigates the role of creating strategies of auditing on claims.

The strategies mentioned above try to minimize the total cost of a claim. This cost, as per the writers of the (Viaene, Dedene, & Derrig, "Auto claim fraud detection using Bayesian learning neural networks", 2005), is split into the cost of auditing and the amount paid on a specific claim. In fact, there are some problems that pop up when trying to create a system that detects a fraud. These problems are produced by the complexity of the attributes that the insurer needs to pay attention to. This literature examines less the deterrent role of auditing and more the detecting role, the success of which is counted by the smaller number of claims audited or larger number of fraud claims detected.

The theory of auditing simplifies the speculations of the nature and the conditions that a claim is characterized as fraud. One of these conditions may be the following: if a claim is large enough, it should have higher probability to be chosen for possible fraud. In the conclusion of the article, the authors came to deduction that the greater the claim the more probable to be examined and the authors came to a result of the research, that auditing is good for both deterrence and a measure of detecting fraud.

The (Pérez, Muguerza, Arbelaitz, Gurrutxaga, & Martín, "Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance", 2005) tries to present the result of using classification trees in fraud detection. It is important to mention that in the fraud detection problem, explanation of the results has a vital role. This means that we do not only need accuracy but we need to "know" the translation of the results. One major problem that occurs in real life is the need to face imbalanced data sets. That is to say, that if we have two classes where one of the two is rather bigger than the other one, we need to undersample our set. In order to achieve it we need to decrease the class that contains more data.

However, by using undersampling techniques we may have loss of information. So, in order to decrease the phenomenon, the authors suggested that one could use boosting or bagging algorithms. Nevertheless, these algorithms may not be very helpful in areas that need explanation. For this reason, the authors have established a new type of tree which creates sub samples that come from the original dataset and they adjoin every tree and info into one final tree. One problem that appears is that the claims that are detected as fraud in reality are less than the true number of true fraud claims, so companies have few examples of true fraud cases. In spite of this, the authors suggest

that companies would need a tool that creates profile of high-risk clients and would notify for further investigation.

In conclusion, the team managed to build one final tree which contains the best results of the sub samples. As noticed, the information produced was very good despite splitting the sample into several trees. They could compare their algorithm (CTC) to C4.5, CART, CAID and they came up with the suggesting that the difficulty in detecting fraud maybe is because the companies do not have such good data so further investigation could be done with different data sets and have better results.

In article (Lookman & Balasubramanian, "Survey of Insurance Fraud Detection Using Data Mining Techniques", 2013) the authors try to capture a definition of insurance policy. They define it as a deal made by an individual and a company in order the first to be reimbursed in case of a loss. There are several types of insurance fraud, but they used this paper to mainly focus on motor claims fraud where fraud may exist at the application stage or afterwards where the claimant raises the cost of fixing his property. The difficulty that the authors mention is that real data are not easily accessible due to personal data restrictions and of competition. This is why most researches are based on synthetic data which sometimes may be misleading.

Despite the problems, the authors claim that as data mining is evolving it will become easier for the companies to organize and audit their claims. Thus, they actuate experts to continue their work even with synthetic data sets in order to use it on real data in the future.

Following through the previous article, detection has lately become a very important subject for insurance companies. According to the bibliography, from the early years of existence of the insurance companies, underwriting was achieved by putting together claims and comparing them. However, this technique had a flaw from the part of the claimants. They could easily hide information or intently enlarge the amount of the damage. So, the only way for the insurers to be protected is to just compare similar case or to be based on the experience of a claim adjuster.

As per the article, another measure to protect the companies was emerged in US where there was formed a group of highly experienced claims adjusters who tried to develop internal procedures to deal with fraud attempts. The result of this attempt was to classify the claims based on the characteristics that proposed this group. Nowadays, the collection of data has become really huge, but the attempt to identify all the

fraudulent cases is worthwhile. For this reason, the article mentions some example of datasets and the methods that were used. At first, it mentions the Massachusetts data set in which regression, fuzzy clustering and unsupervised learning techniques were used. Afterwards, the authors mention Canadian data set worked with regression and probit networks and finally the Spanish data set with regression models.

All this effort led to the conclusion that we need to automate more the investigation procedure in order to reduce the time and the effort spent to define a fraudulent claim. Further to this, the article mentions the claims handling procedure as a two-step process. At first phase the claim passes a first screening and a cost estimation is proposed. The second phase is the split in two categories: if the case seems legit it moves on to be paid but if it raises suspicions then it may be referred to the more experienced claims handler.

At this point, the authors try to propose some techniques for this pattern recognition stage such as decision trees (C4.5), regression, k-nearest, multilayer perceptron, Least squares SVM, tree- augmented naive Bayes classification and they report some optimizations in some of them. Also, they try to compare the algorithms' result to the one of humans and try to see if these techniques can go deeper to investigating the claims.

Some difficulties that may occur are that the evaluation of the algorithms is critically affected by the shape of the sample. Thus, for all the above-mentioned algorithms the team needed to tune their hyperparameters.

Concluding, in their research they ended up that no matter what the target of sample set was, including all the attributes to the algorithm radically improves the performance. Furthermore, they saw that and simple algorithms have also a very good result in predictions and that too complex algorithms are not improving the result as much. From a practical and business view, the predictions made by more complex tools are not so useful if the user cannot define the "why". That is why they propose using a simple "white - box" algorithm rather than a complex "black - box" one which needs a way too difficult parametrization.

Another fact given by the authors of the article (Viaene, Ayuso, Guillen, Gheel, & Dedene, "Strategies for detecting fraudulent claims in the automobile insurance industry", 2011) is that fraud claims range from 5 to 10 percent of total number of incidents that are submitted. This means that all the clients who do not try to deceit their company are in way victims of this situation because fraud leads to increase of the

policy prices. Companies, in order to avoid these instances, they adopt various investigating strategies. Since, fraud is very well hidden, claims adjusters need to put some extra effort and resources in order to dig out the real cases. The main difficulty is to recognize the attributes that describe a deceit, because in most cases these are subjectively defined or based on the experience of the claims adjuster.

Nowadays that processes have become faster, there is no time to investigate extensively every claims' attributes. This lack of time led companies to develop and apply algorithms that calculate the measure of suspicion faster, given the facts. Thus, in the paper (Viaene, Ayuso, Guillen, Gheel, & Dedene, "Strategies for detecting fraudulent claims in the automobile insurance industry", 2011) the authors are going to prove that these algorithms, which in fact try to minimize the error rate to bring a result, are better to target minimising the cost of classification.

As a matter of fact, companies that use models to recognize deceit, mostly focus on diversion of characteristics, while few of them take under consideration cost. The research made for the paper showed that there are indeed profits when there is a screen which warns the user for possible fraud. Furthermore, when the adjusters have a more accurate prediction for the claims' costs or an average cost given the possible claim amount and auditing cost can lead to more profitable results. Finally, using better model is advised in order to lower the audit costs before decision making.

Another approach to the problem is given in the (Derrig, "Insurance Fraud", 2002), where the writers propose choosing refashioned claim attributes to calculate the misclassification error. In this model, they declare that the user can assign weights on each attribute to help the model decide. However, they explore alternative algorithms such as naive Bayes, neural nets, decision trees to see how better these are comparing with simple regression.

They begin with mentioning that the first approach of the problem should be the outlier detection and the proceed to implement the algorithms to company and prove that the results are useful. Further to the above, they admit that "discrete choice" models have a great impact on calculating the fraud percentage in each claim and also, are a vital way to evaluate the importance of attributes that are related to cheat.

The authors bring to attention that AUROC, "area under operating curve", is more efficient as an assessment measure because it shows immediately sensitivity and specificity of the model. It can be deduced that claims that have got attributes nearly

same to the ones that are characterized as fraud, maybe are also frauds. So, the writers turned to another research which tries to clear out and select the most suitable features. They deducted that the characteristics of such attributes have an infrequent and non-linear scoring model by which the cases are classified.

Concluding, the results proved that AUROC evaluation is better than PCC and ended up that sets which contain augmented with non-fraud indicator data produces better results than the ones which are filtered with only fraud-indicators.

3 Problem Definition

Until now we have seen in the literature review that a never-ending problem for insurance companies is fraud claims. The extra amount that a claimant tries to obtain from the companies has negative effects to the law-abiding people. In most cases, the consumers have to pay increased insurance fees because of these companies' loses. In this dissertation we tried to have a review of the problem into the literature and present some solutions that have been proposed before. After that, we used a dataset in order to apply some machine learning models in order to classify the claims into fraud or not. Our purpose is to try and find out what are the attributes which tend to characterize a claim like that.

As mentioned in the literature review, the claims adjusters spend most of their time trying to explore a case and dig out the deceit situations. With the above mentioned method we tried to create a method by which the claims adjusters are going to just have a preview of whether a case tends to be faulty or not. The dataset we used is about vehicle insurance claims only.

More specifically, we downloaded the data set from (https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4954928053318020/1058911316420443/167703932442645/latest.html) as a text file. The text file was transformed into a csv (comma separated value) file by using python in order to be executable with other tools if needed. Afterwards, we needed to transform the csv file into a data frame by using pandas library so as to handle the data easier as it contains 1000 rows and 41 columns.

3.1 EXPLORATORY DATA ANALYSIS

As the procedure of exploring the dataset goes on, we tried to understand it better. We tried to see how the claims are distributed and after we ran the algorithm we noticed that the non fraud claims were more than the fraud ones in the specific dataset.

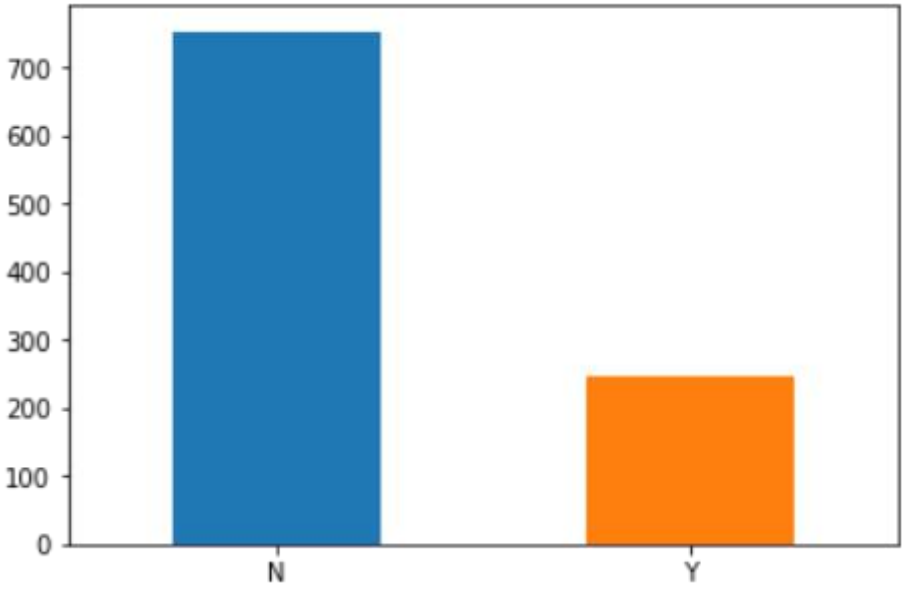

Table 1: Classes count

From the above diagram we can deduce that the target class is imbalanced. This makes our problem harder to be solved and we should try to avoid the most false positive results possible. Further down, we tried to split class into attributes. We tried to depict for each class the clients education. Below, there are the results given from this set.

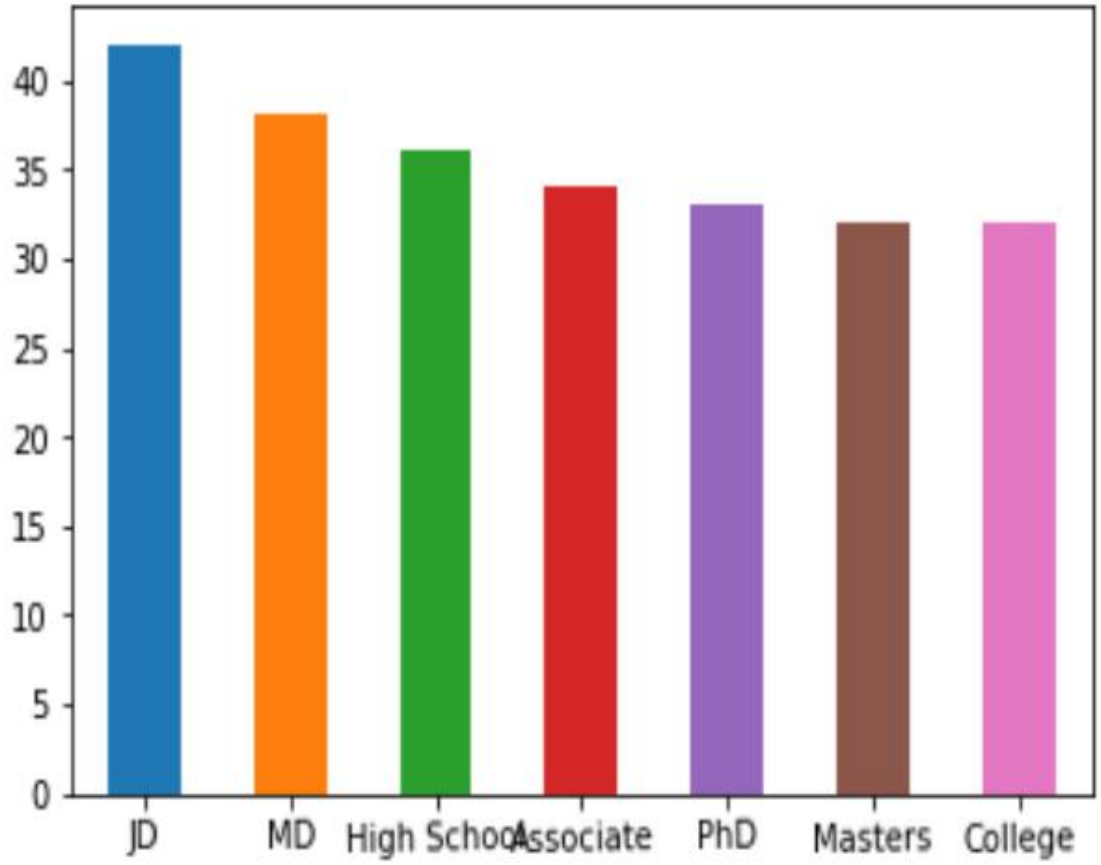

Table 2: Fraudulent Clients Education

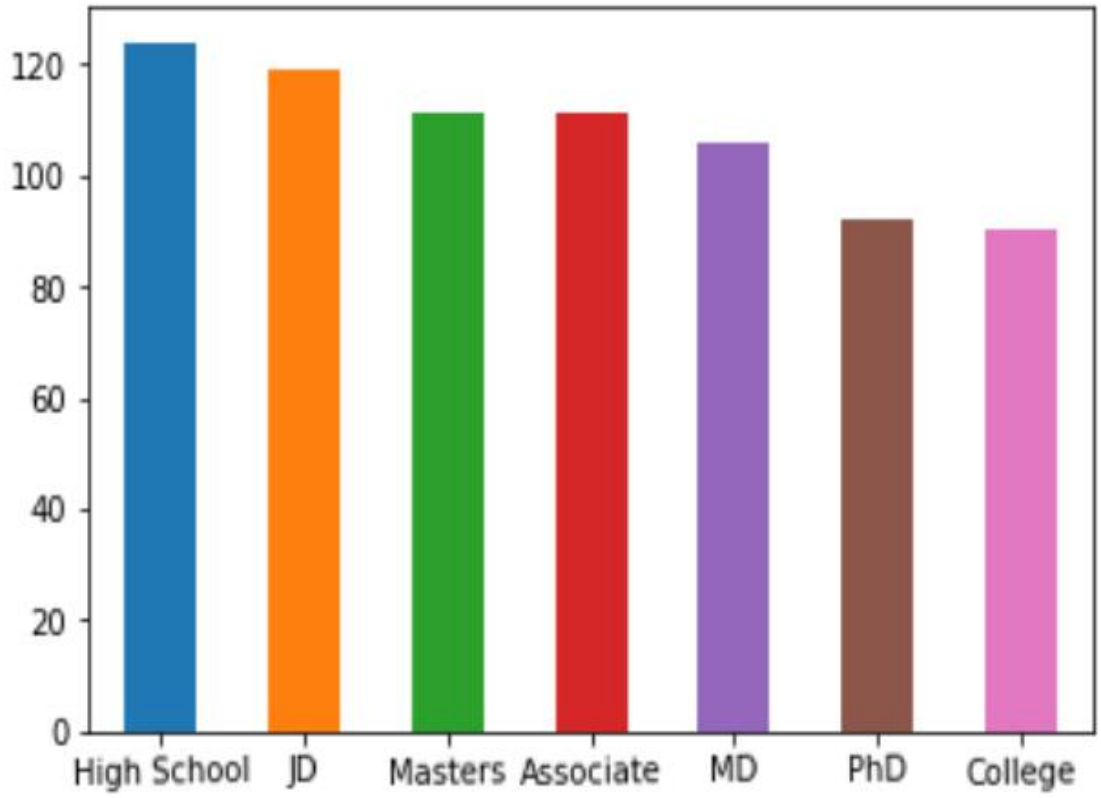

Table 3: Non-fraudulent Clients Education

Now we can proceed and eliminate some attributes that by literature seem to be non relevant to our measurements. The columns that we deleted were:

- ✓ Policy_number
- ✓ Policy_bind_date
- ✓ Insured_zip
- ✓ Incident_location
- ✓ Incident_date
- ✓ Incident_date_count

And we kept the rest of them in order to have our prediction model. The attributes kept were:

- ✧ months_as_customer
- ✧ age
- ✧ policy_state
- ✧ policy_csl
- ✧ policy_deductable
- ✧ policy_annual_premium
- ✧ umbrella_limit
- ✧ insured_sex
- ✧ insured_education_level
- ✧ insured_occupation
- ✧ insured_hobbies
- ✧ insured_relationship
- ✧ capital-gains

- capital-loss
- incident_type
- collision_type
- incident_severity
- authorities_contacted
- incident_state
- incident_city
- 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

3.2 PREPROCESSING

Now as the procedure continues, we start preprocessing the data in order to apply our prediction model. At first, we separated the target column from the rest of the set and we converted the Yes / No to binary as 1 and 0 accordingly. The next step is to change categorical data to numeric values. By changing the type we achieve homogeneity in our set and make easier to handle. In this way the dataset is ready to apply our models Further to this, we are going to examine the effect of every model by using precision, recall and f-1 measure criteria. More specifically:

- *Precision is defined as the percentage of reported positives that truly turn out to be positive*
- *Recall is defined as the percentage of ground-truth positives that have been reported as positives.*
- *F-1 measure is the harmonic mean between the precision and the recall f-1 = (2 * precision * recall) / (precision + recall). Provides better quantification than precision or recall. Because it is still dependant on threshold t, it is still not a complete representation of trade off between precision and recall.* ("Charu C. Aggarwal", Data Mining)

With a view to tune the hyper-parameters, we have used two types of libraries from sci-kit learn (https://scikit-learn.org/stable/) machine learning site. Tuning the hyper-parameters is a vital search we need to do before applying our model. We need to use the algorithms to search our parameters space to find the optimum score of which to use. In our set we applied randomized search and grid search. The main difference of the two is that grid search exhausts the parameters combinations whereas randomized search just tries to choose some parameters from the given space with specific distribution given by the user.

The next step is to split our set into a training set and a test set. The reason of doing this is because we need to train our models to known areas first and the apply the trained model to “unknown” data. In this way we can appraise how accurately our model responds to new information. Thus, by running the randomized search and grid search we had the following results:

Table 4: Randomized Search results

The mean accuracy of the **randomized search** model is: 0.881
The best parameters for the model are: {'penalty': 'l1', 'C': 7.833154072944154}

It is rather an unexpected result as grid search (Table 5) does an extensive research so the accuracy should have been bigger than the one of the randomized search (Table 4).

Table 5: Grid Search Results

The mean accuracy of the **grid search** model is: 0.872
The best parameters for the model are: {'C': 1.0, 'penalty': 'l1'}

The results returned using logistic regression for each case can be found in tables 6 and 7:

Table 6: Grid Search results

Class	**Precision**	**Recall**	**F1-score**
0	0.88	0.92	0.90
1	0.68	0.59	0.63
Avg/ total	0.83	0.84	0.84

Table 7: Randomized search results

Class	**Precision**	**Recall**	**F1-score**
0	0.87	0.89	0.88
1	0.60	0.54	0.57

Avg/ total	0.80	0.81	0.81

From our results we decide to use the grid search results as they perform more accurately in regression than the ones of randomized search.

3.3 MODELS

Proceeding our processing we tried to use a data pipeline in order to make our data more solid. In this way it was feasible to execute multiple models in one run and compare their results. The models used to classify our data set were naive Bayes, svm, regression, random forest and decision tree. After the pipeline ran, it showed that the best model to use to characterize a claim is Decision Tree with accuracy 0.854 as per the reference below. We can also observe that the optimal parameters that the algorithm chose were gini index as a criterion and max depth of the tree would be 5 in order to avoid overfitting. Moreover, it took the algorithm about 5 minutes to compare among the classifiers and decide the best. The parameters which were compared by the algorithm are shown below (Figure 1):

Figure 1: Models used

```
models = [      {"classifier": [GaussianNB()]},
                {"classifier": [svm.SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
                  decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
                  max_iter=-1, probability=False, random_state=None, shrinking=True,
                  tol=0.001, verbose=False)]},
#                {"classifier": [xgb.XGBClassifier(objective="binary:logistic", random_state=42)]},
                {"classifier": [LogisticRegression()],
                 "classifier__penalty": ['l2','l1'],
                 "classifier__C": np.logspace(0, 4, 10)
                 },
                {"classifier": [LogisticRegression()],
                 "classifier__penalty": ['l2'],
                 "classifier__C": np.logspace(0, 4, 10),
                 "classifier__solver":['newton-cg','saga','sag','liblinear']
                 },
                {"classifier": [RandomForestClassifier()],
                 "classifier__n_estimators": [10, 100, 1000],
                 "classifier__max_depth":[5,8,15,25,30,None],
                 "classifier__min_samples_leaf":[1,2,5,10,15,100],
                 "classifier__max_leaf_nodes": [2, 5,10]},
                {"classifier": [DecisionTreeClassifier()],
                 "classifier__splitter":['best', 'random'],
                 "classifier__max_depth":[5,8,15,25,30,None],
                 "classifier__min_samples_leaf":[1,2,5,10,15,100],
                 "classifier__max_leaf_nodes": [2, 5,10]},
                 {"classifier": [DecisionTreeClassifier(class_weight = 'balanced')],
                 "classifier__splitter":['best', 'random'],
                 "classifier__max_depth":[5,8,15,25,30,None],
                 "classifier__min_samples_leaf":[1,2,5,10,15,100],
                 "classifier__max_leaf_nodes": [2, 5,10]},]
```

Figure 2 Decision Tree training set results

```
The mean accuracy of the model is: 0.854
The best parameters for the model are: {'classifier': DecisionTreeClassifier(class_weight='balanced', criterion='gini', max_depth=5,
                       max_features=None, max_leaf_nodes=5,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=10, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best'), 'classifier__max_depth': 5, 'classifier__max_leaf_nodes': 5,
'classifier__min_samples_leaf': 10, 'classifier__splitter': 'best'}
[Parallel(n_jobs=-1)]: Done 4090 out of 4090 | elapsed:  4.8min finished
```

After having the above results (Figure 2), we used the Decision Tree algorithm in the training and test set in order to have results on whether the model works accurately or not. Thus, we have the below results using the "unknown" part of our dataset (test set):

Table 8: Test set results

Class	**Precision**	**Recall**	**F1-score**
0	0.95	0.81	0.87
1	0.57	0.85	0.68
Avg/ total	0.86	0.81	0.83

To conclude, one can see that the decision tree can dig out the non-fraud claims with precision of 0.95 whereas, it does not work good enough with the fraud ones. An obvious explanation of this is that the particular set, as mentioned in exploratory analysis, is way too imbalanced over the non-fraud claims. So the models cannot determine very accurately the faulty cases.

CONCLUSIONS

This thesis attempted to provide a prediction on whether a claim is a fraud based on an insurance claims data set. It would be rather useful for a claims coordinator to have a hint by the program that a claim maybe fake. The algorithms of classification we have used detect specific attributes of a claim and mark it as a potential problem. The second reason for trying is to conserve time for a claim-handler to other tasks of his job during the day, like payments. Thus, the above-mentioned reasons are vital not only for the

sustainability of the company but also for the clients. The companies would be assured that do not overpay damages and this could lead to better premiums that clients are paying to companies due to this anomaly.

As we moved on deeper into the dataset, an important result of the data analysis showed that the data sets rarely have balanced context. Companies rarely keep track of fraud claims and the main explanation is that they sparsely refuse to pay a claim even when it is fraud due to friendly consumer policy. As a result, in the specific data set we noticed that the instances that are depicted as non-fraud claims are way more than the characterized as fraud ones.

Afterwards, we ran randomized and grid search in order to find out which are the best parameters to use in order to have the best results. The findings showed that randomized search would be more accurate. In fact, we tested the parameters in logistic regression and, as per the results tables, it was proved that the grid search parameters would provide better accuracy.

The paper concludes by adducing that from the models we chose to use and test, the more efficient would be the Decision Tree algorithm. The results showed that, based on this imbalanced set, we would have accuracy of predicting 95% correctly a non-fraudulent case. However, we should have more cases for our algorithm to be trained efficiently for the fraudulent ones. Further research is suggested on using clustering techniques in order to group the attributes and create profiles of customers or attributes, or even apply some unsupervised and deep learning techniques which could be more adaptive based on the info provided in real time and not only in past data.

Bibliography

Abdallah, A., Maarof, M. A., & Zainal, A. (2016, April 13). Fraud detection system: A survey. Retrieved from https://www.sciencedirect.com/science/article/pii/S1084804516300571.

Artís, M., Ayuso, M., & Guillén, M. (2002, October 25). Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1539-6975.00022.

Aggarwal, C. C. (2016). Data mining: the textbook. Retrieved from https://www.amazon.com/Data-Mining-Textbook-Charu-Aggarwal/dp/3319141414.

Belhadji, B., Dionne, G., & Tarkhani, F. (2000, October 1). A Model for the Detection of Insurance Fraud. Retrieved from https://link.springer.com/article/10.1111/1468-0440.00080.

Derrig, R. A. (2002, October 25). Insurance Fraud. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1539-6975.00026.

Lookman, H., & Balasubramanian. (2013, September 3). Survey of Insurance Fraud Detection Using Data Mining Techniques. Retrieved from https://arxiv.org/abs/1309.0806.

Minority report in fraud detection: classification of skewed data. (n.d.). Retrieved from https://dl.acm.org/citation.cfm?id=1007738.

Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016, March 9). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. Retrieved from https://www.sciencedirect.com/science/article/pii/S2405918816300058.

Pérez, J. M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., & Martín, J. I. (2005, August 22). Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance. Retrieved from https://link.springer.com/chapter/10.1007/11551188_41.

Phua, Clifton, Lee, Vincent, Smith, Kate, … Ross. (2010, September 30). A Comprehensive Survey of Data Mining-based Fraud Detection Research. Retrieved from https://arxiv.org/abs/1009.6119.

SKlearn. (n.d.). Retrieved from https://scikit-learn.org/stable/.

Subudhi, S., & Panigrahi, S. (n.d.). Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud. Retrieved from https://ieeexplore.ieee.org/document/8588973.

Šubelj, L., Furlan, Š., & Bajec, M. (2010, August 5). An expert system for detecting automobile insurance fraud using social network analysis. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417410007712.

Supraja, k., & Saritha, S. J. (n.d.). Robust fuzzy rule based technique to detect frauds in vehicle insurance. Retrieved from https://ieeexplore.ieee.org/document/8390160.

Tennyson, S. (2002, October 25). Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1539-6975.00024.

Viaene, S., Dedene, G., & Derrig, R. A. (2005, May 10). Auto claim fraud detection using Bayesian learning neural networks. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417405000825.

Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (1970, January 1). A comparison of state-of-the-art classification techniques for expert automobile insurance fraud detection. Retrieved from https://www.semanticscholar.org/paper/A-comparison-of-state-of-the-art-classification-for-Viaene-Derrig/c3a54de403073b57c01d7bc6953e30c603f566ac.

Viaene, S., Ayuso, M., Guillen, M., Gheel, D. V., & Dedene, G. (2011, January 13). Strategies for detecting fraudulent claims in the automobile insurance industry. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0377221705006405.