Effective Fraud Detection in Insurance Claims Through Machine Learning

MAMIDALA SAI KUMAR1, NUKAMREDDY SRINADHREDDY2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ Insurance fraud is an intentional illegal conduct undertaken with the goal of profit. This is currently the most pressing issue for many insurance companies throughout the world. In the majority of cases, the primary issue has been identified as one or more holes in the investigation of false claims.

As a result, there has been an increase in the desire to adopt computer solutions to prevent fraud activities, providing clients with not only a dependable and stable environment, but also dramatically reduced fraud claims.

We demonstrated our findings by automating the examination of insurance claims utilising a range of data methodologies, with the detection of erroneous claims performed automatically using Data Analytics and Machine Learning techniques.

The algorithm might also be able to develop heuristics for fraud warning indications. Because it improves both company reputation and consumer satisfaction, this technique benefits the whole insurance industry.

1.INTRODUCTION

The insurance sector is now embracing efficient fraud control. While some people pay premiums, others defraud businesses in order to receive compensation. Hard insurance fraud and soft insurance fraud are the two main types of fraud.

Hard insurance fraud is described as the deliberate fabrication of an accident. Soft insurance fraud occurs when a person files a legitimate insurance claim but falsifies a portion of it. Both types of fraud have serious repercussions that can include increased insurance costs for everyone as well as criminal charges.

To avoid fraud and preserve the integrity of the insurance system, it is crucial for insurance companies to fully investigate any suspicious claims. Customer satisfaction will increase if a company has a good fraud detection and prevention management system. Loss adjustment costs will go down as a result of the higher satisfaction. There are now numerous methods for identifying fraud claims.

The most popular technique is data analysis using specific instructions. Therefore, they require in-depth investigations that take a lot of time and deal with various fields of knowledge. Overcome the entire issue by using machine learning techniques.

Automating the fraud detection process and minimising the time and resources needed for investigations are both possible with machine learning techniques. These methods can analyse large amounts of data from numerous sources and spot trends that point to fraud.

2.LITERATURE SURVEY

2.1 A fraud detection system looks for suspicious activity as it is being processed by the main system (Aisha Abdallah, 2016).

Previously, this process involved manually detecting and identifying these activities by looking through a sample of actual fraud data. The process has taken a lot of time and has been prone to misunderstandings, human error, and overlooking certain details. Thus, fraud detection systems have evolved to automate the process and eliminate the human element from the system's operational level. However, many data mining techniques were lacking in earlier iterations, and they are now much more advanced and efficient to produce better results and findings for an effective fraud detection system.

2.2 Experimental evaluation of related data sets: Bart Baesens, S. H. (2021). Data engineering for fraud detection . Decision Support Systems .

In an article by (Bart Baesens, 2021), the data set was split into 30% and 70%, meaning that 30% of the data were chosen as the test set and 70% of the data were chosen as a training set, for a total of 31,763 records and 14 attributed. They have experimented with a variety of classification techniques on their data set, including decision trees, logistic regression, CART algorithms, and many others. Each of these algorithms has a variety of justifications that explain why it was used.That decision tree could provide a better understanding of the decision process in order to understand more about how the fraud was committed. Logistic

regression is very popular in the establishment of models due to its speed and low cost of computation power.

2.3 Classification of the current financial fraud detection systems: (Jarrod West, 2016)

Compared numerous research studies on the fraud detection system, the various models that have been used in the detection of fraud, and the efficacy of each model in his article. A thorough analysis of each and every model has previously been done in the same paper. In addition, a comparison of the various fraud investigation types with the techniques applied in recent studies and papers. For instance, support vector machines, decision trees, hybrid methods, and artificial immune systems are the most popular methods and algorithms for credit card fraud.

2.4 Crocker, K. J., and S. Tennyson," Insurance Fraud and Optimal Claims Settlement Strategies: An Empirical Investigation of Liability Insurance Settlements" The Journal of Law and Economics, 45(2), april 2010.

The study and creation of a system that can learn from data constitute the fundamental idea behind machine learning. It serves as the foundation for teaching computers to behave more intelligently. Depending on how each technique operates, there are four main categories. They are reinforcement learning, unsupervised, semi-supervised, and supervised. Supervised learning occurs when the correct class of training data is known; otherwise, unsupervised learning occurs

3.PROPOSED SYSTEM

To detect fake insurance claims, we used Random Forest and Lightgbm. Because of their ability to handle big datasets and complicated feature interactions, the Random forest and Lightgbm algorithms were chosen. The results demonstrated that both the Lightgbm and Random forest models were quite accurate at detecting bogus claims. These findings show that sophisticated machine learning algorithms might significantly improve the detection of bogus insurance claims, potentially saving insurance firms millions of dollars in damages. In the future, it may be possible to combine these algorithms with other approaches to improve the precision and efficiency of the findings.

3.1 IMPLEMENTATION

3.1.1 LightGBM Algorithm:

LightGBM is a gradient boosting framework built on decision trees that improves model performance while using less memory. It employs two cutting-edge methods: All GBDT (Gradient Boosting Decision Tree) frameworks use gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which overcomes the drawbacks of histogram-based algorithm. The characteristics of the LightGBM Algorithm are formed by the two GOSS and EFB techniques that are described below. Together, they enable the model to function effectively and give it an advantage over other GBDT frameworks. One Side Sampling Method for LightGBM Based on Gradients: Different data instances play a variety of roles in the information gain calculation. The under-trained instances, which have larger gradients, will contribute more to the information gain.

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that is designed to be fast, efficient, and scalable for handling large-scale machine learning tasks. It was developed by Microsoft and is widely used in both academia and industry for various applications such as classification, regression, ranking, and anomaly detection.

Here's a step-by-step explanation of how LightGBM works:

Gradient Boosting: LightGBM belongs to the family of gradient boosting algorithms. Gradient boosting is an ensemble learning method where weak learners, typically decision trees, are combined to create a strong learner. It works in an iterative manner, building each new tree to correct the mistakes made by the previous trees.

Decision Trees: LightGBM uses decision trees as base learners. A decision tree is a flowchart-like structure where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome or prediction. Decision trees are constructed in a top-down manner by recursively partitioning the data based on the selected features.

Gradient-based Learning: LightGBM uses gradient-based learning to optimize the model's performance. During the training process, LightGBM calculates the gradients (partial derivatives) of a loss function with respect to the predicted values. These gradients represent

the direction and magnitude of the error, allowing the algorithm to update the model's parameters in a way that minimizes the loss.

Leaf-wise Tree Growth: Unlike traditional decision tree algorithms that grow trees in a level-wise manner, LightGBM grows trees leaf-wise. This means that it chooses the leaf with the maximum delta loss (improvement in the loss function) to grow in each iteration. By growing trees leaf-wise, LightGBM can achieve a higher growth rate and reduce the number of levels in the trees, leading to faster training times.

The diagram below illustrates the process of leaf-wise tree growth in LightGBM, which differs from other boosting algorithms that grow trees level-wise. In leaf-wise growth, the algorithm selects the leaf with the maximum delta loss to expand. This approach results in lower loss compared to level-wise growth since the leaf is fixed. However, it's important to note that leaf-wise growth may increase the complexity of the model and potentially lead to overfitting, particularly in small datasets

3.1.2 Random forest:

Random Forest is a powerful ensemble learning algorithm that combines the predictions of multiple decision trees to make more accurate and robust predictions. It is widely used in various fields, including machine learning and data analysis.

Ensemble Learning: Random Forest is an example of ensemble learning, where multiple models are combined to improve predictive performance. In this case, the individual models are decision trees.

Decision Trees: A decision tree is a tree-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents an outcome or prediction. Decision trees are built by recursively partitioning the data based on the values of the features, with the goal of minimizing impurity or maximizing information gain at each step.

Bootstrapping: Random Forest uses a technique called bootstrapping to create multiple subsets of the original training data. Bootstrapping involves randomly sampling the training data with replacement, resulting in multiple datasets of the same size as the original but with slight variations.

Random Feature Selection: For each decision tree in the Random Forest, a random subset of features is selected at each split point. This helps to introduce randomness and diversity into the ensemble, reducing the correlation between the trees.

Tree Construction: With the bootstrapped datasets and random feature subsets, each decision tree in the Random Forest is constructed independently. The trees are grown by recursively partitioning the data based on the selected features and their respective split points, until a stopping criterion is met (e.g., reaching a maximum depth or minimum number of samples in a leaf node).

Voting or Averaging: Once all the decision trees are constructed, predictions are made by each tree individually. For regression problems, the predictions of all the trees are typically averaged to obtain the final prediction. For classification problems, each tree's prediction is considered as a vote, and the class with the majority of votes is selected as the final prediction.

Out-of-Bag (OOB) Error: During the bootstrapping process, some samples may not be selected in a particular bootstrap sample, known as out-of-bag (OOB) samples. These OOB samples can be used to estimate the model's performance without the need for cross-validation or a separate validation set. The OOB error is calculated by aggregating the predictions of the trees on their corresponding OOB samples.

Advantages of Random Forest: Random Forest offers several advantages. Firstly, it reduces overfitting by averaging or voting over multiple decision trees. Secondly, it handles a large number of features effectively by selecting a random subset at each split. Thirdly, it provides an estimate of feature importance, which can be useful for feature selection. Lastly, it can handle both regression and classification problems.

Parameters: Random Forest has several parameters that can be tuned to optimize its performance, such as the number of trees in the ensemble, the maximum depth of the trees, and the number of features to consider at each split.

Overall, Random Forest is a versatile and powerful algorithm that combines the strength of multiple decision trees to make accurate predictions while mitigating the weaknesses of individual trees

4.RESULTS AND DISCUSSION

Enter the age: 45

Enter the policy number: 655656

Enter the insured sex: male

Enter the insured education_level: md

Enter the insured occupation: craft-repair

Enter the insured relationship: husband

Enter the indident type: single-vehicle-collision

Enter the collision type: side-collision

Enter the incident_severity: major-damage

Enter the authorities_contacted: police

Enter the property damage: yes

Enter the police_report_available: yes

OUTPUT:

Fraud insurance

4.1 Dataset and features:

- The dataset consists of 266964 values.
- The features in dataset are:

Months_as_customer , age,, policy_number, policy_bind_date, policy_state, policy_csl, policy_deductable, policy_annual_premium, umbrella_limit, insured_zip, insured_sex , insured_education_level, insured_occupation, Insured_hobbies, insured_relationship, capital-gains, capital-loss, incident_date, incident_type, collision_type, incident_severity, authorities_contacted, incident_state, incident_city, incident_location, incident_hour_of_the_day, number_of_vehicles_involve, 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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	months_a	age	policy_nu	policy_bir	policy_sta	policy_csl	policy_de	policy_an	umbrella_	insured_z	insured_s	insured_e	insured_o	insured_h	insured_r	capital-ga	capital-lo	incident_	incident_t	collision_	incident_s
2	328	48	521585	########	OH	250/500	1000	1406.91	0	466132	MALE	MD	craft-repa	sleeping	husband	53300	0	########	Single Vel	Side Colli	Major Dar
3	228	42	342868	########	IN	250/500	2000	1197.22	5000000	468176	MALE	MD	machine-(	reading	other-rela	0	0	########	Vehicle Th	?	Minor Dar
4	134	29	687698	9/6/2000	OH	100/300	2000	1413.14	5000000	430632	FEMALE	PhD	sales	board-gan	own-child	35100	0	########	Multi-veh	Rear Colli	Minor Dar
5	256	41	227811	########	IL	250/500	2000	1415.74	6000000	608117	FEMALE	PhD	armed-for	board-gan	unmarriec	48900	-62400	########	Single Vel	Front Coll	Major Dar
6	228	44	367455	6/6/2014	IL	500/1000	1000	1583.91	6000000	610706	MALE	Associate	sales	board-gan	unmarriec	66000	-46000	########	Vehicle Th	?	Minor Dar
7	256	39	104594	########	OH	250/500	1000	1351.1	0	478456	FEMALE	PhD	tech-supp	bungie-ju	unmarriec	0	0	1/2/2015	Multi-veh	Rear Colli	Major Dar
8	137	34	413978	6/4/2000	IN	250/500	1000	1333.35	0	441716	MALE	PhD	prof-speci	board-gan	husband	0	-77000	########	Multi-veh	Front Coll	Minor Dar
9	165	37	429027	2/3/1990	IL	100/300	1000	1137.03	0	603195	MALE	Associate	tech-supp	base-jum	unmarriec	0	0	########	Multi-veh	Front Coll	Total Loss
10	27	33	485665	2/5/1997	IL	100/300	500	1442.99	0	601734	FEMALE	PhD	other-sen	golf	own-child	0	0	########	Single Vel	Front Coll	Total Loss
11	212	42	636550	########	IL	100/300	500	1315.68	0	600983	MALE	PhD	priv-hous	camping	wife	0	-39300	1/5/2015	Single Vel	Rear Colli	Total Loss
12	235	42	543610	########	OH	100/300	500	1253.12	4000000	462283	FEMALE	Masters	exec-man	dancing	other-rela	38400	0	1/6/2015	Single Vel	Front Coll	Total Loss

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	incident_	authoritie	incident_	incident_	incident_	incident_	number_c	property_	bodily_inj	witnesses	police_re	total_clair	injury_cla	property_	vehicle_cl	auto_mak	auto_mod	auto_year	fraud_rep	_c39
2	Major Dan	Police	SC	Columbus	9935 4th D	5	1	YES	1	2	YES	71610	6510	13020	52080	Saab	92x	2004	Y	
3	Minor Dar	Police	VA	Riverwoo	6608 MLK	8	1	?	0	0	?	5070	780	780	3510	Mercedes	E400	2007	Y	
4	Minor Dar	Police	NY	Columbus	7121 Franc	7	3	NO	2	3	NO	34650	7700	3850	23100	Dodge	RAM	2007	N	
5	Major Dan	Police	OH	Arlington	6956 Mapl	5	1	?	1	2	NO	63400	6340	6340	50720	Chevrolet	Tahoe	2014	Y	
6	Minor Dar	None	NY	Arlington	3041 3rd A	20	1	NO	0	1	NO	6500	1300	650	4550	Accura	RSX	2009	N	
7	Major Dan	Fire	SC	Arlington	8973 Wash	19	3	NO	0	2	NO	64100	6410	6410	51280	Saab	95	2003	Y	
8	Minor Dar	Police	NY	Springfiel	5846 Weav	0	3	?	0	0	?	78650	21450	7150	50050	Nissan	Pathfinde	2012	N	
9	Total Loss	Police	VA	Columbus	3525 3rd H	23	3	?	2	2	YES	51590	9380	9380	32830	Audi	A5	2015	N	
10	Total Loss	Police	WV	Arlington	4872 Rock	21	1	NO	1	1	YES	27700	2770	2770	22160	Toyota	Camry	2012	N	
11	Total Loss	Other	NC	Hillsdale	3066 Franc	14	1	NO	2	1	?	42300	4700	4700	32900	Saab	92x	1996	N	
12	Total Loss	Police	NY	Northben	1558 1st R	22	1	YES	2	2	?	87010	7910	15820	63280	Ford	F150	2002	N	
13	Major Dan	Fire	SC	Springfiel	5971 5th H	21	3	YES	1	2	YES	114920	17680	17680	79560	Audi	A3	2006	N	

5.CONCLUSIUON

As the world develops towards a more economically based society, the goal is to stimulate each nation's economy. Fighting these fraudsters and money launderers was a difficult chore prior to the era of machine learning. However, machine learning and artificial intelligence have enabled us to combat these types of attacks. The proposed technique can be utilised in insurance companies to identify whether a certain insurance claim is fraudulent or not. The model was developed after experimenting with numerous algorithms to determine which one was most efficient in assessing whether a claim was true or untrue. This is a presentation to insurance companies to create a model that is more tailored to their needs for their own systems.

REFERENCES

1. Crocker, K. J., and S. Tennyson," Insurance Fraud and Optimal Claims Settlement Strategies: An Empirical Investigation of Liability Insurance Settlements" The Journal of Law and Economics, 45(2), april 2010

2. Clifton Phuna damminda, Alahakoon, and Vincent phua" Minority Report in Fraud Detection: Classification of Skewed Data". Sigkdd Explorations, Volume – 6, Issue – 1,sep 2011

3. Chen, Y.; Wang, X. Research on medical insurance fraud early warning model based on data mining. Comput. Knowl. Technol. 2016, 12, 1–4.

4. Jarrod West, M. B. (2016). Intelligent financial fraud detection: A comprehensive review. ScienceDirect, 47-66. 11. Javad Forough, S. M.

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

6. Bart Baesens, S. H. (2021). Data engineering for fraud detection . Decision Support Systems .

7. Jarrod West, M. B. (2016). Intelligent financial fraud detection: A comprehensive review. ScienceDirect, 47-66.

8. Alejandro Correa Bahnsen, D. A. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems With Applications, 134-142.

9. Javad Forough, S. M. (2021). Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing Journal.

10. Abhinav Srivastava, A. K. (2008). Credit Card Fraud Detection Using. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, 37 - 48.

11. E. Belhadji, G. Dionne and F. Tarkhani, A Model for the Detection of Insurance Fraud Geneva Papers on Risk and Insurance Theory, vol. 25, pp. 517-538, may 2012.

12. K. J. Crocker and S. Tennyson, "Insurance Fraud and Optimal Claims Settlement Strategies: An Empirical Investigation of Liability Insurance Settlements", *The Journal of Law and Economics*, vol. 45, no. 2, april 2010.

13. Kajia Muller, The Identification of Insurance Fraud-an Empirical Analysis Working papers on Risk Management and Insurance, no. 137, June 2013

AUTHOR PROFILES