**ARTIFICIAL INTELLIGENCE IN THE INSURANCE INDUSTRY**

**BY**

**OMOYENI OGUNDIPE**

**OLAYEMI SHITTU**

**DEPARTMENT OF MATHEMATICS AND STATISTICS**

**AUSTIN PEAY STATE UNIVERSITY**

**GRADUATE TEACHING ASSISTANSHIP RESEARCH PROJECT**

**DR SAHI.K. RAMANJIT**

**DECEMBER 30TH 2022**

**ABSTRACT**

As AI becomes more deeply integrated in the world, insurance companies must position themselves to respond to the changing business landscape. Insurance executives must understand the factors that will contribute to this change and how AI will reshape claims, distribution, fraud detection, underwriting and pricing. With this understanding, they can start to build the necessary skills and talent that will be needed to embrace the emerging technologies, and create the culture and perspective needed to be successful players in the insurance industry of the future.

In this paper, we will look at how Artificial Intelligence can impact the insurance industry with core focus on Fraud detection. We will use AI to build classification models for fraud detection and compare the results.

**TABLE OF CONTENT**

ABSTRACT………………………………………………………………………….2

TABLE OF CONTENT ………………………………….………………………….3

INTRODUCTION……………………………………………………………………4

FRAUD DETECTION…………………….…………………………………………5

Traditional ways of detecting fraud by Insurers…………………………....5

How AI is Transforming Insurance Fraud Detection……………………....8

METHODOLOGY……………….…….……………………………………………10

DATA ANALYSIS………….…….……………..………………………………….11

Data Preprocessing…….………….……..………………………………….11

Modeling……………….………….……..………………………………….14

RESULTS…….…………………………………..………………………………….15

Model comparison…….………….……..………………………………….19

CONCLUSION………………………...…………………………………………….21

FURTHER RESEARCH……………………………………………….…………….21

REFERENCES……………………………………………………………………….22

ACKNOWLEDGEMENT ……………………………………………………..…….24

**INTRODUCTION**

With the power of the cloud and today’s rapidly changing technologies, forward-thinking insurers can leverage Artificial Intelligence to drive faster and more personalized customer experiences, increasing satisfaction with claimants and generating significant efficiencies in insurance underwriting.

The insurance industry has only begun its foray into AI, and companies are already experimenting new ways to incorporate AI into their day-to-day operations in anticipation of further technological development. The vision is that by 2030, AI could make more impact in the services they provide, and this evolution will shift insurance from its current state of “detect and repair” to “predict and prevent,” transforming every aspect of the industry in the process.

AI and its related technologies will have a seismic impact on all aspects of the insurance industry, from distribution, underwriting and pricing to claims. Some key areas include:

* Claim processing
* Client Journey
* Fraud Detection
* Customer retention and churn prediction
* Operation process optimization
* Risk Assessment

As we move forward in this research, we will streamline our focus to Fraud detection and try to understand better how Artificial Intelligence can be fully optimized in this area and how it can positively impact the insurance industry as a whole.

**FRAUD DETECTION**

Insurance fraud is an intentional deception against or by an insurance company for financial benefits. Claimants, policyholders, third-party claimants, or professionals providing services to claimants could commit fraud at different stages during insurance. Insurance agents and company employees may also engage in insurance fraud. Some of the common frauds perpetuated include "padding" (inflating claims), misrepresenting facts on an insurance application, submitting claims for injuries or damage that never occurred, and staging accidents.

Insurance fraud has been a Multi-Billion Dollar Fraud Problem in the insurance industry for decades now. A 2022 study by The Coalition Against Insurance Fraud (CAIF) showed that insurance fraud can cost U.S. consumers $308.6B yearly. That amount includes estimates of annual fraud costs across several liability areas, including Life Insurance, Property insurance, Workers Compensation, and Auto insurance.

**Traditional ways of detecting fraud by Insurers**

Historically, some of the ways insurance companies detect frauds are through:

1. **Analysis of claim history:** Insurers usually do a deep dive into the insurance claims of individuals, carrying out scrutiny to find patterns, frequency through historical claims. They do all sorts of data analysis to get information from the data they have and if your claim doesn't match the typical pattern, they’ll notice.
2. **Checklist of "Suspicious Loss Indicators":** Insurance agents look for "suspicious loss indicators" when they suspect a claim may be false. For example, looking to determine when and how a home fire started.

The National Insurance Crime Bureau (NICB) has developed a list of 23 “Suspicious Loss Indicators” that can signal that a claim may be fake, bogus or a rip-off.

Some of the suspicious loss indicators insurance agents look for:

* A claimant who's totally calm and unflustered after submitting a large claim
* A claimant who submits handwritten receipts for repairs on a covered item
* A claimant who adds or increases homeowners or auto insurance coverage shortly before submitting a claim
* A fire-damage claim for a home or auto where the fire started immediately after a family argument, or shortly after family members left the home/car
* A medical claim submitted by an employee whose job is ending

1. **Using Private investigators** to ensure claims aren’t false and leveraging social media for investigations.
2. Looking for Evidence of Personal Injury.
3. **Using sophisticated computer systems to ensure legitimate Billing:** Billing is one way people use to defraud their insurance. Often, they'll work with their auto repair, or health center as the case may be, to pad the bill to cover things like the deductible. In situations like this, insurance companies' computer systems can pull up claims where repairs appear inflated, or don't match with information provided about the claim.
4. **Handing Cases to Special Investigation Unit:** Many insurers have special investigation units, and they usually consist of well-trained individuals who have experiences working as detectives, police officers, medical personnel, etc. They're able and well trained to perform a variety of tests and checks to detect fraud. For example, they can:

* Conduct burn-pattern analyses and computer simulations on cars and homes damaged by fire to determine if the fire was intentionally set or accidental.
* Determine if a claimant's injuries match a reported accident.
* Investigate damaged vehicles to see if the resulting dents and scratches are from the reported accident. Also, use rust analysis and wear patterns to see if your car's damage is indeed from an old accident or not.
* Conduct financial reviews on claimants. Auto or homeowners claims from those who are behind on car or mortgage payments are immediately flagged as potentially fraudulent.

1. **Evaluate Prospective Employees' Credit Histories:** Claimants are not the only ones committing insurance fraud. This also happens with insurance agents, hence employees often check their credit before being hired. Insurance agents may commit fraud by "stealing" a customer's insurance premiums. In this common scam, an agent can take money for insurance and keep it without ever underwriting the policy for you. Insurance companies try to prevent such fraud by doing a credit check on all potential employees. Applications from people with credit or financial problems are flagged as most likely to commit fraud.
2. **Reporting claims in an online anti-fraud information system:** Thousands of insurance companies, self-insured entities and third-party administrators report all their claims to ISO Claim Search, an anti-fraud information system. The system was created by Insurance Services Office, Inc., and covers auto, property and liability claims. Cross-checking a new claim against all of those in this database (1 billion-plus) is one of the easiest ways for insurers to catch fraud.

**How AI is Transforming Insurance Fraud Detection**

Insurance fraud is a very big problem in the United States. And the people who commit it are increasingly creative, outsmarting some of the traditional checks and balances put in place, causing a large amount of insurance frauds to go undetected thereby constantly increasing the costs of insurance premiums.

It is becoming increasingly challenging for financial institutions and insurance companies to provide superior customer services, comply with regulatory requirements, and manage fraud, all at scale and in a cost-efficient manner. Traditional approaches are reactive, manual in nature, and ineffective, prompting institutions to consider leveraging Artificial Intelligence.

Some of the ways AI can be used to detect and prevent fraud in insurance are:

1. **Predictive analytics:** The first line of defense against insurance fraud is predictive analytics for early detection and fraud prevention. Based on historical behavior, predictive analytics can gain insights to an insured’s potential fraud risk, thereby taking the traditional system of “detect and repair” to “predict and prevent,” saving billions of dollars for insurance companies in the process. One predictive algorithm that is worth looking further into for this case is the logistic regression.
2. **Using NLP to analyze historical data:** In addition to processing mountains of information around the clock, Natural Language Processing can analyze historical data of past fraudulent claims and learn how to classify claims into groups of fraudulent or non-fraudulent claims by evaluating recorded conversations and other types of textual data, such as emails. Without using AI to detect claim fraud, this would be ineffective if not impossible to replicate with human labor alone. By tracking the historical trends in a person's claim history, the algorithms understand an individual's claim history and whether a particular claim seems normal or suspicious.
3. **Data mining:** Data mining can help third-party payers like insurance companies to extract useful knowledge from thousands of claims and identify a smaller subset of the claims or claimants for further assessment and scrutiny for fraud.

Combining automated methods and statistical knowledge led to a newly emerging interdisciplinary branch of science that is named Knowledge Discovery from Databases (KDD). Data mining is the core of the KDD process. In the domain of health care fraud and abuse detection, supervised data mining involves methods that use samples of previously known fraudulent and non-fraudulent records. These two groups of records are used to construct models, which allow us to assign new observations to one of the two groups of records. Supervised methods require confidence in the correct categorization of the records. Furthermore, they are useful in detecting previously known patterns of fraud and abuse. Some of the algorithms that are currently being used for fraud detection include clustering for outlier detection which can detect anomalous non-compliant fraudulent claims, decision tree, neural networks, Support Vector Machine (SVM).

1. **Real-time notifications:** With an artificial intelligence system that works around the clock and continuously monitors the habits and behavior of claims and policyholders, algorithms can easily flag potentially fraudulent activity and provide real-time alerts to the business when complaints require further investigation. The earlier insurers can be alerted to potentially fraudulent activity, the better protected they are against paying the requested amount and corresponding loss.

AI is a useful tool that improves a firm’s resource efficiency and saves insurers millions of dollars each year. With better early fraud risk detection, NLP to analyze historical claims data, advanced data mining, and real-time alerts, insurance companies can leverage AI to better manage fraud and resulting losses.

**METHODOLOGY**

We made use of historical insurance claim data including normal and fraudulent ones, to investigate the normal/fraud behavior features. Based on ML techniques, and using these features, we checked if a claim is fraudulent or not. After this, a comparative study was performed to decide which Machine Learning classifier performed best.

We used 12 Machine Learning algorithms to build models for fraud detection.

The algorithms are:

* Support Vector Classifier
* KNN
* Decision Tree Classifier
* Random Forest Classifier
* Ada Boost Classifier
* Gradient Boosting Classifier
* Stochastic Gradient Boosting (SGB)
* XgBoost
* Cat Boost Classifier
* Extra Trees Classifier
* LGBM Classifier
* Voting Classifier

**DATA ANALYSIS**

The data used for this study is an open data gotten from [Kaggle](https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data) and has been extracted from insurance claim settlement. The dataset contains 1000 rows and 40 columns.

The Data Analysis and modeling are subdivided into two steps;

* **Data Preprocessing**: In this step we will prepare our data for modeling by treating null values, carrying out feature engineering and scaling the data for modeling
* **Modeling**: After preparing the data, we will go ahead to build models for fraud prediction

**Data Preprocessing**

***Treating Null Values***

Chart

Description automatically generated with medium confidenceAfter carrying out exploratory data analysis we saw that some columns had ‘?’ and ‘missing values’ (Figure 1) which we treated using mode fill.

Figure 1: Visualization of missing values

***Checking for multicollinearity***

Using a heatmap to visualize the correlation between numerical columns in the dataset

Chart, treemap chart

Description automatically generated

Figure2:Correlation between numerical variables

We see from (Figure 2) that there is multicollinearity in the numerical data, so we removed some of the numerical columns that are not necessary for prediction and to remove the problem of multicollinearity. After doing this, we were left with 27 columns.

Chart, bar chart, waterfall chart

Description automatically generatedUsing a heatmap to check for multicollinearity again.

Figure3:Correlation between numerical variables

From the plot in (Figure 3), we can see that there is still multicollinearity in the data and there is a high correlation between **age** and **months\_as\_customer**, and between **total\_clam\_amount, injury\_claim, property\_claim,** and **vehicle\_claim.**

As total claim is the sum of all others, we will drop the total claim column and the age column. After doing that, we are left with 25 columns as seen below.

Data columns (total 25 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 months\_as\_customer 1000 non-null int64

1 policy\_csl 1000 non-null object

2 policy\_deductable 1000 non-null int64

3 policy\_annual\_premium 1000 non-null float64

4 umbrella\_limit 1000 non-null int64

5 insured\_sex 1000 non-null object

6 insured\_education\_level 1000 non-null object

7 insured\_occupation 1000 non-null object

8 insured\_relationship 1000 non-null object

9 capital-gains 1000 non-null int64

10 capital-loss 1000 non-null int64

11 incident\_type 1000 non-null object

12 collision\_type 1000 non-null object

13 incident\_severity 1000 non-null object

14 authorities\_contacted 1000 non-null object

15 incident\_hour\_of\_the\_day 1000 non-null int64

16 number\_of\_vehicles\_involved 1000 non-null int64

17 property\_damage 1000 non-null object

18 bodily\_injuries 1000 non-null int64

19 witnesses 1000 non-null int64

20 police\_report\_available 1000 non-null object

21 injury\_claim 1000 non-null int64

22 property\_claim 1000 non-null int64

23 vehicle\_claim 1000 non-null int64

24 fraud\_reported 1000 non-null object

***Encoding Categorical Variables***

Moving on, we went ahead to encode the categorical variables, by turning categorical variables into numerical variables. The columns and variables that were encoded were:

policy\_csl: ['250/500' '100/300' '500/1000']

insured\_sex: ['MALE' 'FEMALE']

insured\_education\_level:['MD' 'PhD' 'Associate' 'Masters' 'High School' 'College' 'JD']

insured\_occupation:

['craft-repair' 'machine-op-inspct' 'sales' 'armed-forces' 'tech-support'

'prof-specialty' 'other-service' 'priv-house-serv' 'exec-managerial'

'protective-serv' 'transport-moving' 'handlers-cleaners' 'adm-clerical'

'farming-fishing']

insured\_relationship: ['husband' 'other-relative' 'own-child' 'unmarried' 'wife' 'not-in-family']

incident\_type:['Single Vehicle Collision' 'Vehicle Theft' 'Multi-vehicle Collision' 'Parked Car']

collision\_type: ['Side Collision' 'Rear Collision' 'Front Collision']

incident\_severity: ['Major Damage' 'Minor Damage' 'Total Loss' 'Trivial Damage']

authorities\_contacted: ['Police' 'None' 'Fire' 'Other' 'Ambulance']

property\_damage: ['YES' 'NO']

police\_report\_available: ['YES' 'NO']

***Detecting Outliers and scaling the data***

Calendar

Description automatically generatedWe used a box plot to detect outliers as seen in figure 4 and scaled the numerical columns to treat those outliers so that every numerical column can have the same weight.

Figure3: Boxplot showing outliers

**Modeling**

Now that our data is ready for modeling, we used 12 machine learning algorithms to build models for fraud prediction and compared the models with each other to see which performs best.

We made use of the scikit-learn library in python as well as xgboost, catboost and lighgbm libraries to build the models and used GridSearch for hyperparameter tuning.

**RESULTS**

After training and testing our models, we printed the test accuracy, the confusion matrix and classification report which consist of the f1 score, and the precision and recall score of each model. These are the metrics we would be using to evaluate the performance of our models.

1. **Support Vector Classifier**

Training accuracy of Support Vector Classifier is : 0.8453333333333334

Test accuracy of Support Vector Classifier is : 0.72

[[180 0]

[ 70 0]]

precision recall f1-score support

N 0.72 1.00 0.84 180

Y 0.00 0.00 0.00 70

accuracy 0.72 250

macro avg 0.36 0.50 0.42 250

weighted avg 0.52 0.72 0.60 250

1. **KNN**

Training accuracy of KNN is : 0.7666666666666667

Test accuracy of KNN is : 0.72

[[180 0]

[ 70 0]]

precision recall f1-score support

N 0.72 1.00 0.84 180

Y 0.00 0.00 0.00 70

accuracy 0.72 250

macro avg 0.36 0.50 0.42 250

weighted avg 0.52 0.72 0.60 250

1. **Decision Tree Classifier**

Training accuracy of Decision Tree is : 0.8106666666666666

Test accuracy of Decision Tree is : 0.752

[[138 42]

[ 20 50]]

precision recall f1-score support

N 0.87 0.77 0.82 180

Y 0.54 0.71 0.62 70

accuracy 0.75 250

macro avg 0.71 0.74 0.72 250

weighted avg 0.78 0.75 0.76 250

1. **Random Forest Classifier**

Training accuracy of Random Forest is : 0.9546666666666667

Test accuracy of Random Forest is : 0.732

[[180 0]

[ 67 3]]

precision recall f1-score support

N 0.73 1.00 0.84 180

Y 1.00 0.04 0.08 70

accuracy 0.73 250

macro avg 0.86 0.52 0.46 250

weighted avg 0.80 0.73 0.63 250

1. **Ada Boost Classifier**

Training accuracy of Ada Boost is : 0.8106666666666666

Test accuracy of Ada Boost is : 0.812

[[154 26]

[ 21 49]]

precision recall f1-score support

N 0.88 0.86 0.87 180

Y 0.65 0.70 0.68 70

accuracy 0.81 250

macro avg 0.77 0.78 0.77 250

weighted avg 0.82 0.81 0.81 250

1. **Gradient Boosting Classifier**

Training Accuracy of Gradient Boosting Classifier is 0.9546666666666667

Test Accuracy of Gradient Boosting Classifier is 0.288

Confusion Matrix :-

[[ 2 178]

[ 0 70]]

Classification Report :-

precision recall f1-score support

N 1.00 0.01 0.02 180

Y 0.28 1.00 0.44 70

accuracy 0.29 250

macro avg 0.64 0.51 0.23 250

weighted avg 0.80 0.29 0.14 250

1. **Stochastic Gradient Boosting (SGB)**

Training Accuracy of Stochastic Gradient Boosting is 0.9466666666666667

Test Accuracy of Stochastic Gradient Boosting is 0.308

Confusion Matrix :-

[[ 9 171]

[ 2 68]]

Classification Report :-

precision recall f1-score support

N 0.82 0.05 0.09 180

Y 0.28 0.97 0.44 70

accuracy 0.31 250

macro avg 0.55 0.51 0.27 250

weighted avg 0.67 0.31 0.19 250

1. **XgBoost Classifier**

Training accuracy of XgBoost is : 0.8106666666666666

Test accuracy of XgBoost is : 0.812

[[154 26]

[ 21 49]]

precision recall f1-score support

N 0.88 0.86 0.87 180

Y 0.65 0.70 0.68 70

accuracy 0.81 250

macro avg 0.77 0.78 0.77 250

weighted avg 0.82 0.81 0.81 250

1. **Cat Boost Classifier**

Training Accuracy of Cat Boost Classifier is 0.9066666666666666

Test Accuracy of Cat Boost Classifier is 0.64

Confusion Matrix :-

[[132 48]

[ 42 28]]

Classification Report :-

precision recall f1-score support

N 0.76 0.73 0.75 180

Y 0.37 0.40 0.38 70

accuracy 0.64 250

macro avg 0.56 0.57 0.56 250

weighted avg 0.65 0.64 0.64 250

1. **Extra Trees Classifier**

Training Accuracy of Extra Trees Classifier is 1.0

Test Accuracy of Extra Trees Classifier is 0.724

Confusion Matrix :-

[[167 13]

[ 56 14]]

Classification Report :-

precision recall f1-score support

N 0.75 0.93 0.83 180

Y 0.52 0.20 0.29 70

accuracy 0.72 250

macro avg 0.63 0.56 0.56 250

weighted avg 0.68 0.72 0.68 250

1. **LGBM Classifier**

Training Accuracy of LGBM Classifier is 1.0

Test Accuracy of LGBM Classifier is 0.652

[[144 36]

[ 51 19]]

precision recall f1-score support

N 0.74 0.80 0.77 180

Y 0.35 0.27 0.30 70

accuracy 0.65 250

macro avg 0.54 0.54 0.54 250

weighted avg 0.63 0.65 0.64 250

1. **Voting Classifier**

Training accuracy of Voting Classifier is : 0.9373333333333334

Test accuracy of Voting Classifier is : 0.772

[[158 22]

[ 35 35]]

precision recall f1-score support

N 0.82 0.88 0.85 180

Y 0.61 0.50 0.55 70

accuracy 0.77 250

macro avg 0.72 0.69 0.70 250

weighted avg 0.76 0.77 0.76 250

**Model Comparison**

Accuracy is the proportion of correct predictions generated by the model. Hence, using the Test accuracy score, we are going to graph a table and plot in descending order from highest score to the lowest test accuracy score.

|  |  |  |
| --- | --- | --- |
|  | **Model** | **Score** |
| 4 | Ada Boost | 0.812 |
| 10 | XgBoost | 0.812 |
| 11 | Voting Classifier | 0.772 |
| 2 | Decision Tree | 0.752 |
| 3 | Random Forest | 0.732  0,724 |
| 8 | Extra trees | 0.724 |
| 0 | SVC | 0.720 |
| 1 | KNN | 0.720 |
| 9 | LGBM | 0.652 |
| 7 | Cat Boost | 0.640 |
| 6 | SGB | 0.308 |
| 5 | Gradient Boost | 0.288 |

Chart, bar chart

Description automatically generated

Figure4: Bar chart of models test accuracy scores

From the table above and the bar chart in figure 4, it is evident that the best performing model is Ada Boost and XgBoost model which give a test accuracy score of 0.812 and similar f1 scores of 0.87.

**CONCLUSION**

Rapid advances in technologies is already leading to disruptive changes in the insurance industry. The winning Insurance companies will be those that use new technologies to create innovative products, harness cognitive learning insights from new data sources, streamline processes, lower costs, build automated powerful fraud detection systems and exceed customer expectations.

With the massive amount of financial transactions taking place every day, the volume of fraudulent transactions is also increasing. It also means that the old techniques are no more useful. The good news is that there are sophisticated algorithms that can take advantage of the massive datasets for training and solving these problems.

**FUTURE RESEARCH**

Moving on in this research, we would like to investigate how AI can be used to transform other areas in the insurance industry like we highlighted above. For example, how AI can be used to predict and prevent customer churn.

**REFERENCE**

Accenture. (2022, September 5). *Transforming claims and underwriting with AI*. https://www.accenture.com/us-en/insightsnew/insurance/ai-transforming-claims-underwriting

Accern, T. (2022, September 12). *How AI is Transforming Insurance Claims Fraud Detection*. Accern. https://accern.com/blog/how-ai-is-transforming-insurance-claims-fraud-detection/

*Background on: Insurance fraud | III*. (n.d.). Retrieved October 25, 2022, from https://www.iii.org/article/background-on-insurance-fraud

Balasubramanian, R., Libarikian, A., & McElhaney, D. (2021, July 1). *Insurance 2030—The impact of AI on the future of insurance*. McKinsey & Company. https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2014). Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. *Global Journal of Health Science*, *7*(1). https://doi.org/10.5539/gjhs.v7n1p194

McManus, M. R. (2022, April 29). *10 Ways Insurance Agents Spot Fraudulent Claims*. HowStuffWorks. https://money.howstuffworks.com/personal-finance/auto-insurance/10-ways-insurance-adjusters-spot-fraudulent-claims.htm

Uzialko, A. (2022, September 19). *Artificial Insurance? How Machine Learning Is Transforming Underwriting*. Business News Daily. <https://www.businessnewsdaily.com/10203-artificial-intelligence-insurance-industry.html>

Yadav, N. (2022, March 4). *Insurance Fraud Detection (Using 12 Models)*. Kaggle. https://www.kaggle.com/code/niteshyadav3103/insurance-fraud-detection-using-12-models

**ACKNOWLEDGEMENTS**

Our sincere gratitude goes to our GTA supervisor Dr, R. Sahi for her guidance, tutoring and support to making this research achievable. Many thanks to the Department of Mathematics and Statistics for giving us the opportunity to carry out this research and enriching us with the necessary tools and skills that contributed to the completion of this research.