Executive Summary

Would it be better to save 276 million THB by implementing a fraud analysis and detection?

Nowadays, there are more than 4,500 fraudulent cases for an insurance company per year. Each of them causes different expenses and could lead to lumpsum amount of wasted cost up to 282 million THB annually.

This report aims to present the business impact of implementing a machine learning-based fraud detection system within the organization. It evaluates the potential benefits and considerations of adopting models such as Random Forest Classifier, Gradient Boosting Machines, Logistic Regression, and XGBoost.

Introduction to Fraud Detection

As the number of fraudulent cases have been growing up during the recent years, fraud analysis and detection systems have been significantly essential for identifying and preventing fraudulent transactions, thereby protecting the company's financial interests and maintaining customer trust.

Data Description

The data encompasses transactional records with various attributes, which are crucial for the models to identify patterns indicative of fraudulent activity. The data which is used for these machine learning algorithms belongs to policy holders and contains 100,000 rows and features consist of:

1. Age;
2. Employment Status;
3. Credit Score;
4. Policy Type;
5. Premium Amount;
6. Policy Duration;
7. Time Since Policy Inception;
8. Claim Type;
9. Claim Amount;
10. and Number of Past Claims.

As an example shown in the table below.

Age	Employment Status	Credit Score	Policy Type	Premium Amount	Policy Duration	Time Since Policy Inception	Claim Type	Claim Amount	Number of Past Claims	Fraudulent
62	Retired	454	Auto	848.65	10	48	Theft	19122.59	4	0
65	Employed	843	Health	1005.4	5	313	Illness	14812.86	0	0
82	Unemployed	446	Auto	443.81	8	128	Other	16249.89	2	0
85	Retired	379	Home	870.5	12	51	Natural Disaster	6741.75	4	0
85	Unemployed	763	Health	284.88	29	252	Accident	90184.39	5	1

The data have been manipulated by different method such as:

1. cleaning;
2. under sampling which helps reduce the imbalance of dataset to reduce the data bias, and there are around 9,800 rows data left which number of fraudulent cases and non-fraudulent cases are equal;
3. feature correlations which observe which x variables have strong correlates with others;
4. hypothesis testing such as t-Test, and Chi-Square to observe which x variables has a strong effect on y variables – fraudulent;
5. feature encoding which transforms text data into categories numerical data and is trainable in machine learning models;
6. feature scaling both standardization and normalization which allow the models to train faster and more efficiently. Moreover, by doing this could help us compare which kinds of scaling and models work best;
7. K-Folds cross validation to observe whether the model overfit with any particular dataset and measure of robustness, we did 3-5 folds for each model and the scores are about 0.6 on average which indicate that the model is robust;
8. Besides, do hyperparameter tuning to see if which parameters would provide the best parameters, and we will use that parameters for the best result.

Data Analysis Approach

First, I explored the data by plotting the box plot in figure 1 and found that every claim which is above 20,000 THB is fraudulent. By this, we could save up to 272,619,279.81 THB annually.

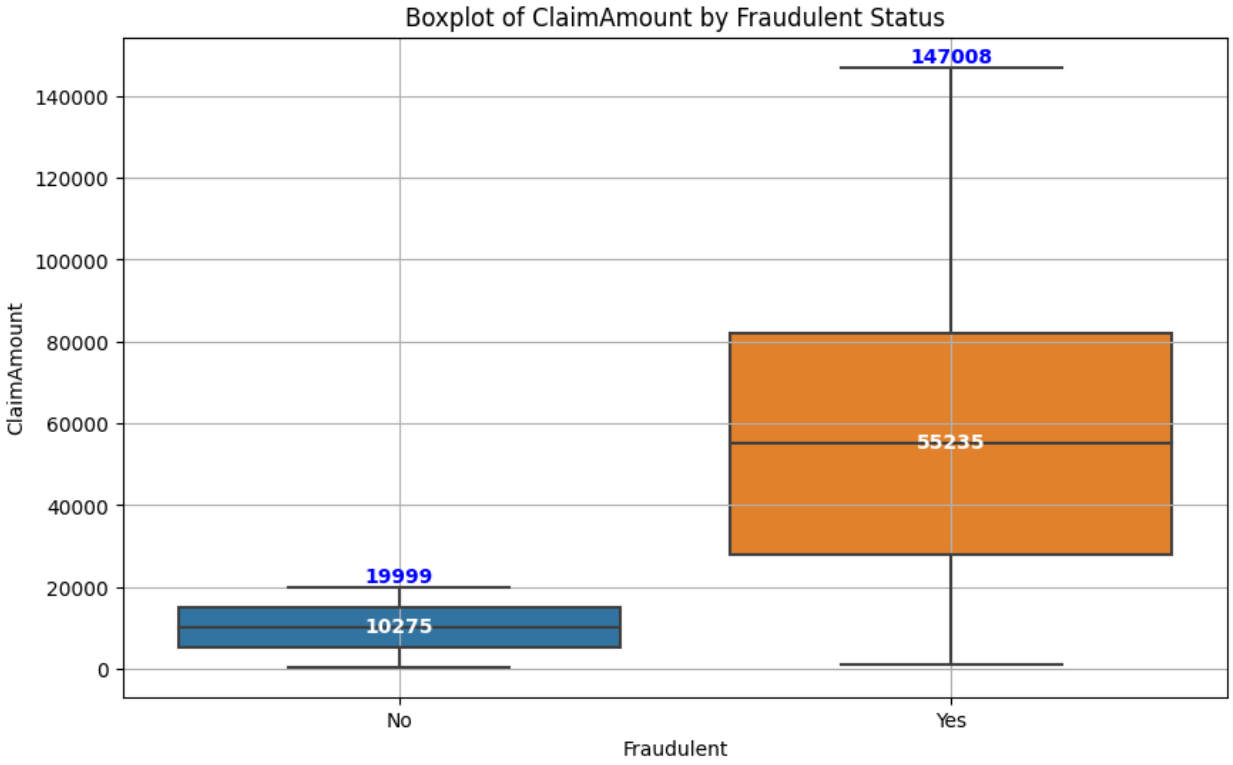


Figure 1

After that, I investigated the claims from 20,000 THB and below, but have number of claims from 5 and above as shown in figure 2. The data revealed there are 285 fraud cases which cost 3,000,748.57 THB. On the other hand, this is also the amount the company could save.

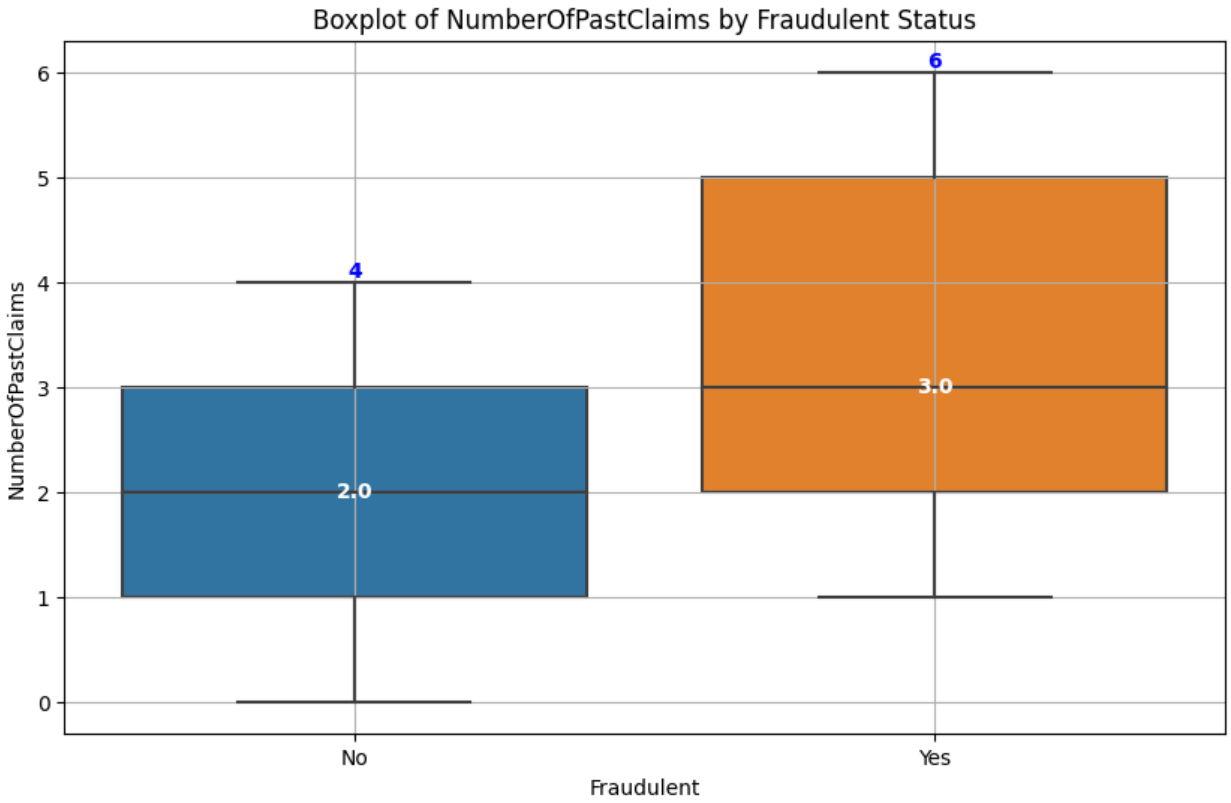


Figure 2

Modeling Approach

Once the company knows that 1) the claims above 20,000THB are fraudulent, and 2) the claims under 20,000 THB with number of claims from 5 and above are fraudulent. It could select the rest of data (the data with claims under 20,000 THB and number of claims under 5) as the data for modelling and take this dataset into data preparation process. The process includes:

1. cleaning;
2. under sampling;
3. feature correlations;
4. hypothesis testing;
5. feature scaling;
6. 10-Folds cross validation;
7. and hyperparameter tuning.

After the data preparation process has been done, the selected models which consist of 1) Random Forest Classifier, 2) Gradient Boosting Machine, 3) Logistic Regression, and 4) XGBoost and are known for their predictive capabilities in classification tasks are implemented.

Model Robustness

As I did 10-Folds cross validation, and standard deviation of score for each model is around 0.01 and very close to 0 while the mean score is around 0.6. It shows that the models are robust and not overfit.

Model Interpretability

This section discusses the interpretability of each model, highlighting how it contributes to understanding and trust in the fraud detection process.

For the performance,

1. Precision shows us the proportion of number of cases in which the model correctly predicts the positive class divided by total number of predicts the positive class (True Positive / Total Predicted Positive).

   For example, the models predicted 837 cases as positive and every prediction is correct. So, the precision is 837/837 = 1.

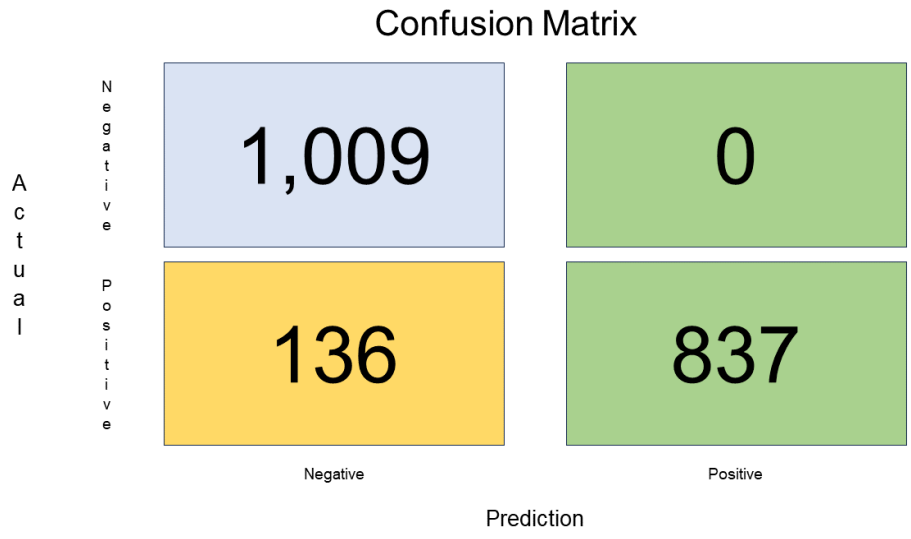
   It could indicate that from total positive predictions, how many predictions are correct.

2. Recall shows us the proportion of number of cases in which the model correctly predicts the positive class divided by number of actual positive class (True Positive / Total Actual Positive).

   For example, the models predicted 837 cases as positive. However, there are 136+837 = 973 actual positive cases. So, the recall is 837 / 973 = 0.86.

   It could indicate that from total fraud cases, how many fraud cases the model can detect.

3. F1 Score shows us the harmonic mean calculated by (2 x Precision x Recell) / (Precision + Recell). The higher the score, the better the performance.



Business Impact

After I implemented the models and got many results, and the model evaluation could take place. I evaluate by providing the impact for each categories for each model with different threshold in table. The business impact in each features differently as follows:

1. True Negative (Predict not fraud, and correct), the company could still get the policy premium of 1,052 THB on average;
2. True Positive (Predict fraud, and correct), the company could protect its extra compensation from fraudulent cases of 10,982 THB on average.
3. False Positive (Predict fraud, and wrong) the company could still get the policy premium but might offer 50% discount as present. The premium would be discounted to 525 THB each on average.
4. False Negative (Predict not fraud, and wrong) the company would have to spend its extra compensation from fraudulent cases of -10,982 THB on average.

By this, we could get the highest model generate value, and would choose the 1st rank which is the "XG Boost" model with threshold of 0.55. The model could generate value for 355,117 THB

Model	Threshold	True Negative	True Positive	False Positive	False Negative	Model Generate Value (THB)
Random Forest	0.55	76	74	39	53	331,049
Random Forest	0.5	53	93	20	76	252,950
Random Forest	0.45	39	107	6	90	230,872
Gradient Boosting	0.55	96	48	65	33	299,847
Gradient Boosting	0.6	129	2	111	0	215,947
Gradient Boosting	0.65	129	0	113	0	195,033
Logistic Regression	0.55	123	11	102	6	237,856
Logistic Regression	0.5	66	78	35	63	252,537
Logistic Regression	0.45	8	113	0	121	(79,440)
XG Boost	0.55	78	74	39	51	355,117
XG Boost	0.5	56	89	24	73	247,224
XG Boost	0.45	45	98	15	84	208,963

As you can observe from the table, the model could detect 110 fraudulent cases or up to 1,210,000 THB, calculated by average fraudulent claim amount 11,000 THB.

Therefore, by implementing fraud data analysis and detection, the company could save by following:

1. From fraudulent cases which have claim amount over 20,000 THB each for 272 million THB
2. From fraudulent cases which have claim amount under 20,000 THB and number of claims from 5 and above for 3 million THB
3. From the machine learning model 0.3 million THB

By doing this, the company could save up to 276 million THB.

Recommendations

As there has been an increasing in number of fraud cases in recent years and it could cost the insurance company up to 282 million THB annually. I would recommend doing fraud data analysis and detection to save the most proportion, 97%, of this amount – up to 276 million THB a year. For the implementation, I would recommend using the Random Forest algorithm with threshold 0.6.

However, I recognize that this is your decision and would welcome your views on this before we begin the implementation.

Disclaimer

The dataset used for the analytics in this report is a mockup version – not from the real business practice. The report is only for showcase purpose, please do not use the report for any commercial purposes or references.