INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 371-374

PREDICTION OF FRAUD IN HEALTH INSURANCE USING CAT BOOST AND LIGHT GBM

P. PooraniArul[1], Mrs. M. Saranya[2]

[1]B. Sc, Department of Computer Science Sri Kaliswari College, Sivakasi, Tamil Nadu, India.

[2]M.Sc., M.Phil., Department of Computer Science Sri Kaliswari College, Sivakasi, Tamil Nadu, India.

ABSTRACT

Health insurance is a crucial choice when it comes to ensuring a secure future. The insurance industry plays a significant role in fostering sustainable economic development in every nation. Fraud including health insurance claims, is a prevalent issue within the financial sector. This study focuses on analysing health insurance claims to detect and anticipate potential fraud. To achieve this, feature selection improve the performance of the model and eliminating irrelevant feature. The comparison between CatBoost and Light GBM showed that the CatBoost model is better than the Light GBM. Based on the result of accuracy, precision, recall and F1-score values on both models.

Keywords- Health Insurance, Cat Boost, Light GBM

1. INTRODUCTION

Fraudulent actions involving health insurance claims provide a serious problem for insurance providers, resulting in large losses in terms of money and damaged reputation [1]. The integrity of the insurance sector depends on the detection and prevention of fraud. Machine learning algorithms have become extremely effective tools for detecting fraud, giving insurers the ability to spot suspicious patterns and successfully stop fraudulent activity [2]. This research focuses on applying LightGBM and Catboost, two well-known machine learning algorithms, to predicting fraud in health insurance claims. The gradient-boosting frameworks LightGBM and Catboost are works at spotting bogus health insurance claims. Our plan is to utilize an extensive dataset that comprises past claims data, which includes vital details like age, gender, medical history, and other significant information. Our goal is to create reliable fraud prediction models by using this dataset to train the LightGBM and Catboost algorithms. The dataset will go through feature selection, which is used to make the process more accurate in order to accomplish this goal. After that, use this processed dataset to train the LightGBM and Catboost models, and assess their performance using suitable evaluation measures, including precision, recall, and F1-score. This project aims to enhance insurance firms' fraud detection systems, improve accuracy, and reduce financial losses by understanding the advantages and disadvantages of machine learning algorithm. It aims to contribute to the understanding of fraud prediction in health insurance claims, promoting operational effectiveness, cost reduction, and customer satisfaction [3].

2. PROPOSED ALGORITHM

a. CatBoost

Cat Boost, which stands for Category Boosting, is a well-known machine learning method that excels at processing a wide range of data types. It excels at integrating text, category, and numerical variables for thorough analysis, especially in datasets with these three types of features. CatBoost's capacity to perform well with very little data is one of its advantages. This feature ensures consistent model performance even in situations where data availability is limited, making it a useful tool. One of CatBoost's unique features is that its architecture makes use of balanced trees, which are similar to mirrors. The algorithm's stability and speed are enhanced by this balanced tree structure, which makes it easy to handle complicated datasets. Additionally, CatBoost uses symmetric trees, which divide data according to the same requirements. This symmetrical strategy preserves consistency throughout the decision-making process, which improves the interpretability of the model and helps achieve optimal performance. CatBoost also uses another important approach called target-based encoding. By figuring out the target variable's mean for every category, this method creates a link between categorized features and the target variable. A categorical value is then used in place of the mean value, adding useful information to the dataset and improving prediction accuracy. CatBoost's versatility as a machine learning algorithm is attributed to its strong tree structure, symmetrical splitting strategy, creative encoding methods, and ability to handle a wide range of input formats [4].

B. Light GBM –

Light Gradient Boosting Machine, or LightGBM for short, is a state-of-the-art machine learning technique that is highly effective and performant. Gradient-based one-side sampling, a method that greatly increases training speed and efficiency, is one of its primary features. Using a gradient-based selection process, this strategy focuses on samples

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 371-374

that make the largest contributions to the total loss function for each iteration. Furthermore, when building trees, LightGBM uses a leaf-wise growth technique, splitting nodes to minimize loss. This method produces more effective trees with fewer nodes, which improves model performance and computational speed even further. One additional noteworthy benefit of LightGBM is that it requires very little memory to function. LightGBM delivers performance that is not compromised by handling huge datasets by optimizing memory usage during training and inference. Because of this, it is especially well-suited for applications that deal with large datasets or memory constraints. Moreover, one-hot encoding—a preprocessing step frequently necessary for managing categorical features in conventional machine learning algorithms—is eliminated by LightGBM. Alternatively, LightGBM may handle categorical features directly, saving time and computational resources during preprocessing. With the help of this feature, categorical data can be more easily included in the model, and the training

pipeline is made simpler [5].

3. EXPERIMENT AND RESULT

Data collection is the initial step in any data analysis process, involving the acquisition of relevant data from various sources. Kaggle, a renowned data science platform, serves as a valuable resource for obtaining datasets spanning diverse domains. These datasets, including those from Kaggle, serve as the foundation for subsequent analysis tasks. Following data collection, data exploration is conducted to gain a comprehensive understanding of the dataset's characteristics and patterns. This phase encompasses various techniques, including descriptive statistics, where measures such as mean, median, standard deviation, minimum, and maximum are calculated to summarize the dataset's central tendencies and dispersion [6]. Additionally, data visualization techniques are employed to visually represent the data distribution and relationships between variables. This includes generating scatterplots, lineplots, boxplots, barplots, countplots, and pointplots, each offering unique insights into the dataset [7]. Moreover, as part of data exploration, missing value analysis is performed to identify any missing or null values within the dataset. This analysis is crucial for ensuring data integrity and reliability throughout the subsequent analysis stages, as appropriate strategies can be employed to handle missing data effectively. Overall, data collection and exploration lay the groundwork for informed decision-making and actionable insights in the data analysis process. Figure 1 shows the reject claims and accept claims.

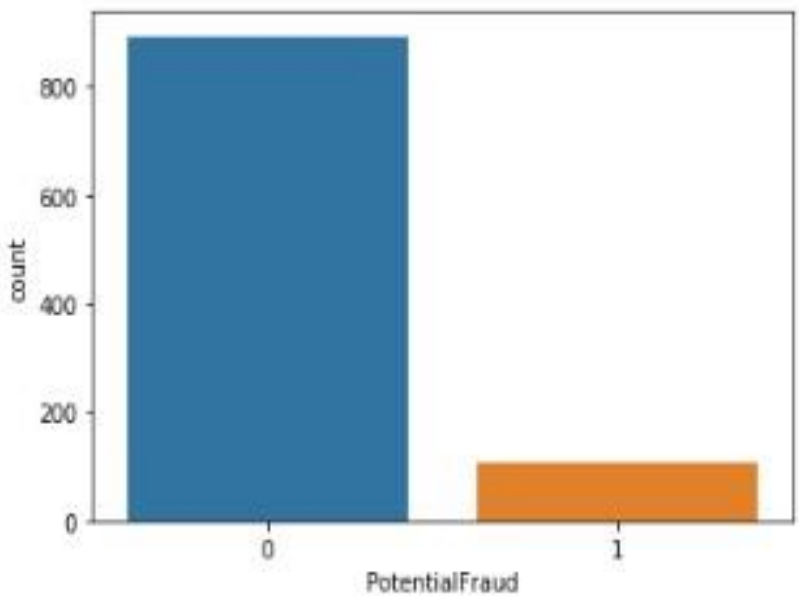

Fig. 1. Data Analysis

Data preprocessing is a critical phase in the data analysis pipeline, where raw data is transformed and prepared for further analysis [8]. One common preprocessing technique is binary encoding, which converts categorical variables into binary representations, facilitating machine learning algorithms' understanding of categorical data. Feature selection is another crucial aspect of data preprocessing aimed at identifying the most relevant and informative features for model training. Various methods are employed for this purpose, including correlation analysis, which examines the linear relationship between features [9]; variance thresholding to filter out low-variance features; and statistical tests like ChiSquare and Anova to assess feature significance. Figure 2 shows the feature selection, that select from the dataset.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 371-374

```
Feature Selection
_________________
Variance Threshold
------------------
Original Features :
Index(['Age', 'Gender', 'BMI', 'Children', 'Smoker', 'Region', 'Charges',
       'Race', 'ChronicCond_Alzheimer', 'ChronicCond_Heartfailure',
       'ChronicCond_KidneyDisease', 'ChronicCond_Cancer',
       'ChronicCond_ObstrPulmonary', 'ChronicCond_Depression',
       'ChronicCond_Diabetes', 'ChronicCond_IschemicHeart',
       'ChronicCond_Osteoporasis', 'ChronicCond_rheumatoidarthritis',
       'ChronicCond_stroke', 'IPAnnualReimbursementAmt',
       'IPAnnualDeductibleAmt', 'OPAnnualReimbursementAmt',
       'OPAnnualDeductibleAmt'],
      dtype='object')
Selected Features :
Index(['Age', 'BMI', 'Children', 'Region', 'Charges', 'Race',
       'IPAnnualReimbursementAmt', 'IPAnnualDeductibleAmt',
       'OPAnnualReimbursementAmt', 'OPAnnualDeductibleAmt'],
      dtype='object')
```

Fig. 2. Feature Selection

Additionally, information gain measures the importance of features based on their contribution to the predictive power of the model. Once data preprocessing and feature selection are completed, the next step is model evaluation, which assesses the performance of the trained model. Evaluation metrics include the Confusion Matrix, which provides insights into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as well as accuracy, precision, recall, and F1 score. Table 1 shows that the precision, recall, and F1-score values of Cat Boost and Light GBM. Accuracy measures the overall correctness of predictions, while precision quantifies the quality of positive predictions. Recall assesses the model's ability to correctly identify positive instances, while the F1-score provides a balanced measure of a model's overall performance, considering both precision and recall [10]. These evaluation metrics collectively offer valuable insights into the model's effectiveness and help guide further optimization efforts. Table 1 shows that the accuracy values of Cat Boost are greater than Light GBM.

Table -1 Comparison between Cat Boost and Light GBM

	Cat Boost	Light GBM
Precision	**0.89**	**0.88**
Recall	**1.00**	**0.99**
F1-score	**0.94**	**0.93**

Table -2 Experiment Result

Accuracy	Training Set	Testing Set
CatBoost	**0.8986**	**0.885**
LightGBM	**0.8986**	**0.88**

4. CONCLUSION

Cat Boost and Light GBM algorithms have the sophisticated ability to handle various types of data and optimize model performance, which makes them useful inidentifying health insurance fraud. By improving fraud detection methods, these algorithms increase risk management, lower costs, and maintain system integrity. The integration of machine learning techniques has the potential to significantly combat fraudulent conduct as healthcare advances.

5. REFERENCE

[1] https://www.sciencedirect.com/science/article/pii/S1877050923017775#:~:text=There%20are%20nine%20independent%20variables,better%20than%20the%20Logistic%20model.

[2] "Fraud Detection Algorithms | Fraud Detection using Machine Learning" https://intellipaat.com/blog/fraud-detection-machine-learning-algorithms/

[3] "Risks | Free Full-Text | Fraud Detection in Healthcare Insurance Claims Using Machine Learning" https://www.mdpi.com/2227-9091/11/9/160

[4] "Cat Boost in Machine Learning - GeeksforGeeks" https://www.geeksforgeeks.org/catboost-ml/amp/

[5] “Light GBM – Wikipedia” https://en.m.wikipedia.org/wiki/LightGBM

[6] “Data Exploration in Python with Examples | by Shreya Singh | Medium” https://medium.com/@jscvcds/data-exploration-in-python-with-examples-30a5324472aa

[7] "Python Big Data Exploration & Visualization: A Comprehensive Guide | Data And Beyond" https://medium.com/data-and-beyond/how-to-visualize-and-explore-big-data-using-python-2c4cd0d8dae4

[8] "How to Preprocess Data in Python | Built In" https://builtin.com/machine-learning/how-to-preprocess-data-python

[9] "Feature selection Techniques|python code - Shiksha Online" https://www.shiksha.com/online-courses/articles/feature-selection-techniques-python-code/

[10] ‘’12 Essential Evaluation Metrics for Evaluating ML Models” https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

ABOUT THE AUTHOR

P. Pooraniarul,

B. Sc, Sri Kaliswari College, Sivakasi, Tamil Nadu.