Detecting Fraud Vehicle Insurance Claims using Machine Learning Technique
Kusuma¹, Brundhashree. U. R², Famida³, Nagamani. K. N⁴, Nivedha. S⁵
Assistant Professor³
Department of Computer Science and Engineering
The National institute of Engineering, Mysuru, India

Abstract:
The fraud vehicle insurance claim is a global problem. Detecting them manually is the traditional approach which is less accurate and time consuming. Machine Learning techniques along with Supervised Learning method provide an effective way to detect fake vehicle insurance claims. The accuracy of the result can be calculated and performance can be measured by constructing confusion matrix.

Keywords: Machine learning (ML), Supervised Learning (SL), Naïve Bayes (NB).

I. INTRODUCTION
An insurance claim is a formal request to the insurance industry asking insurance company for the claim amount based on the terms and conditions of insurance policy. The fraudulent claims in vehicle insurance are mostly done by faking accidents[1]. Traditionally, the false insurance claims are detected manually. Typically there are two types of insurance fraud; hard fraud and soft fraud[2]. Hard fraud is deliberately fabricating an insurance claim, and adding fake details to a legitimate claim is categorized as soft fraud. Fake claims in the insurance domain make the insurer to take more time for recovering. When the fraud claims increase, overall cost of insurance is increased to settle the financial health. So the fraudulent claims are a serious problem to the insurance companies and to the government. This paper proposes a solution for the problem of detecting fraud vehicle insurance claims by using Machine Learning techniques and Naïve Bayes algorithm. Usage of Naïve Bayes algorithm increases the percentage of right prediction[3]. This allows the insurance branch in-charger to classify the claims as fraud or genuine.

II. EXISTING WORK
In traditional approach, the fraud vehicle insurance claim detection is consigned to insurance claim agents or branch in-chargers for support and maintenance. The claim agent uses the facts related to vehicle insurance claim and gather information to sort the claims and wait for the investigation report. Using the gathered information and report, the claim agent predicts whether the claim is fraud or genuine. Figure 1 shows the existing method of fraud vehicle insurance claim detection system. According to this system, the branch in-charger review the claims based on the facts discovered around indicators. Unlike the existing system, claims are assigned to an investigator irrespective of the score received. The raw data from investigation report is transformed to parameters. The collected insurance claim data is split into three different segments –training data, testing data and cross-validation data. The algorithm is trained on the training data set and then tested on testing data set and finally the correctness is examined by cross validating the result with cross-validation data set. Based on the training data, the fraud vehicle insurance claim detection system will classify the new claim as fraud or genuine. The claims classified as fraud are not eligible to claim the insurance amount. This paper proposes a solution using Supervised Learning technique to detect fake
vehicle insurance claims. It contains two target classes namely fraud and genuine.

The raw data used in this model includes,

- Claim occurrence date
- Claim report date
- Claim open date
- Policy effective date
- Part market cost
- Policy premium
- Count of customer communication
- Claim document submitted

The raw data is transformed into a data set and given as an input to Naïve Bayes classification algorithm. The data set is transformed to the following parameters

- Difference between claim occurrence date and claim report date
- Difference between claim report date and claim open date
- Difference between policy effective and claim occurrence date.
- Are claim document submitted?
- Part cost difference
- Credit rating
- Policy premium
- Count of customer communication

By using these parameters, Naïve Bayes algorithm classifies the claim as fraud or genuine.

The workflow of classification is given below,

Step 1: Data collection - Define the problem of fake claims and assemble the necessary data attributes to detect fraudulent claims.
Step 2: Data preparation - Clean the collected data (remove duplicates, correct errors, deal with missing values, etc…), randomize and visualize data for better performance.
Step 3: Split data - Separate the data into training and testing data set to train and test the machine learning system.
Step 4: Choose algorithm - Choose the Naïve Bayes algorithm to predict the vehicle insurance claim as fraud or genuine.
Step 5: Train the algorithm - Train the Naïve Bayes algorithm with training data set.
Step 6: Evaluate the algorithm - Test the algorithm with testing data set.
Step 7: Parameter tuning - Tune Naïve Bayes algorithm for new parameter dataset for improved performance.

The confusion matrix is used to measure the performance of algorithm. The four basic terms of confusion matrix are given below,

**True Positive (TP):** The prediction and actual claim is fraud.
**True Negative (TN):** The claim is predicted as fraud but it is a genuine claim.
**False Positive (FP):** The claim is predicted as genuine but it is a fraud claim.
**False Negative (FN):** The prediction and actual claim is genuine.

The confusion matrix estimates the percentage of correctly labeled test data.

**IV. SYSTEM IMPLEMENTATION**

The branch in-charger identifies and collects the traditional dataset to detect fraud vehicle insurance claims. The collected data set is segregated into training and testing dataset and stored in database. The system is trained by using training data set, and the testing data set parameters are fed as input to the Naïve Bayes algorithm. The result produced by the algorithm is cross-validated. The accuracy of the Naïve Bayes algorithm is given by the percentage of correctly predicted claims.

**WORKING STEPS OF NAÏVE BAYES ALGORITHM**

Step1: Scan the dataset
Step2: Calculate the probability of each parameter values.
Step3: Apply the formulae

\[ P(\text{av/sv}) = \frac{\text{nc}+(\text{m*p})}{(\text{n}+\text{m})} \]

where,
- \( \text{av} \rightarrow \text{attribute value} \)
- \( \text{sv} \rightarrow \text{subjective value} \)
- \( \text{m} \rightarrow \text{Number of parameters} \)
- \( \text{p} \rightarrow \text{Probability of each parameter values} \)
- \( \text{m} \rightarrow \text{Number of training examples which have a = av} \)
- \( \text{a} \rightarrow \text{Parameter values in training data set.} \)
- \( \text{nc} \rightarrow \text{Number of training examples which have a = av and v = sv} \)
- \( \text{v} \rightarrow \text{Result in training data set.} \)
- \( p \rightarrow \text{a priori estimate for P(\text{av/sv})} \)

Step4: Multiply the probabilities with p.
Step5: Compare and classify the parameter values against target class.
Consider the sample example Table 1 contains the training data set which has both parameters and result. The system is trained using training data set and new claim records are tested against it.

There are 3 parameters P1, P2, P3, therefore \( m = 3 \).

The Possible outcomes are Fraud and Genuine, therefore \( p = \frac{1}{2} = 0.5 \)

<table>
<thead>
<tr>
<th>Data</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>10</td>
<td>20</td>
<td>Fraud</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>11</td>
<td>20</td>
<td>Fraud</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>11</td>
<td>21</td>
<td>Genuine</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>10</td>
<td>21</td>
<td>Fraud</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>12</td>
<td>12</td>
<td>Genuine</td>
</tr>
</tbody>
</table>
Let new claim record parameters be P1-7, P2-10, P3-20

Apply the formula on each parameter

\[ P = \frac{nc + (m*p)}{(n+m)} \]

<table>
<thead>
<tr>
<th>Table 2. The calculated probability of each parameter.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fraud</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>n = 2, nc = 2, m = 3, p = 0.5</td>
</tr>
<tr>
<td>p = [2+(3*0.5)]/(2+3)</td>
</tr>
<tr>
<td>p = 0.7</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>n = 2, nc = 2, m = 3, p = 0.5</td>
</tr>
<tr>
<td>p = [2+(3*0.5)]/(2+3)</td>
</tr>
<tr>
<td>p = 0.7</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>n = 2, nc = 2, m = 3, p = 0.5</td>
</tr>
<tr>
<td>p = [2+(3*0.5)]/(2+3)</td>
</tr>
<tr>
<td>p = 0.7</td>
</tr>
</tbody>
</table>

Fraud = 0.7 * 0.7 * 0.7 * 0.5 = 0.1715
Genuine = 0.3 * 0.3 * 0.3 * 0.5 = 0.0135

Since Fraud > Genuine
So this claim is classified as Fraud.

Using Naïve Bayes classification algorithm the quality of prediction is improved.

V. CONCLUSION AND FUTURE WORK

The proposed solution uses Naïve Bayes classification algorithm to detect fraud vehicle insurance claim. For every claim, the parameters are prepared from raw data of investigation report and given as an input to the Naïve Bayes algorithm. Based on the trained data set the algorithm classify the claim to one of the target class i.e fraud or genuine. It is observed that the Machine learning techniques increases the accuracy and improve the performance of the system. The future work is smoothing of the algorithm. If a claim has parameter values which are not part of training data set, then the algorithm not be able to predict resulting in zero probability. The deduction system can be improved by adding smoothing techniques which helps in solving this problem.

VI. REFERENCES

[1]. Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud. DOI: 10.1109/ICDSBA.2018.00104
[4]. policybazaar.com/motor-insurance/?pb_source=organic
[5]. blogs.sas.com/content/brightdata/2017/03/03/insurance-fraud-operational-processes-automated-controls