Theme Chosen : Strengthening Cybersecurity - Threat Detection

FRAUD DETECTION IN VEHICLE INSURANCE CLAIM

SOCIETE GENERALE HACKATHON

Team Members:

1. Shreya Rajesh Patil (Final Year B.Tech.)
2. Shivani Digambar Patil (Final Year B.Tech.)

Brief Abstract:

The insurance industries consist of more than a thousand companies worldwide, and collect more than one trillions of dollars in premiums each year. Vehicle insurance fraud is the most prominent type of insurance fraud, which can be done by fake accident claims. Fraud in insurance is an unethical activity performed systematically to get some financial gain. A robust fraud detection and prevention management system can lead to heightened customer satisfaction and reduced loss adjustment expenses. Therefore, using machine learning algorithms we construct an automated fraud detection application framework in this study by accurately identifying fraud claims in a shorter amount of time is the goal and can lead to heightened customer satisfaction. However, traditional techniques often require intricate and time-consuming investigations, necessitating expertise across diverse domains. To overcome these challenges, the utilization of machine learning techniques has emerged as a solution. By incorporating machine learning methods, the insurance industry can streamline and enhance the process of detecting fraudulent claims. This approach minimizes the need for labor-intensive investigations and enables faster and more accurate identification of suspicious activities. As a result, the industry can better allocate resources, improve customer satisfaction, and ultimately combat insurance claim fraud more effectively.

Methodology proposed for achieving the end-result (output):

The basic principle of machine learning is that, study and construction of a system that can learn from data. So here we propose a machine learning based automated framework employed with Random Forest Classifier, Support Vector Classifier (SVC) and K-Nearest Neighbour(KNN) to classify claims. We will also compare the performance of algorithms with other algorithms to obtain most accurate results whether the vehicle fraud insurance claim is fraud or genuine. Once we are done with model building, and find the optimum model with highest accuracy among above mentioned models, we can create a web application to make this project user friendly by creating GUI with the help of HTML, CSS and other web development tools.

Resources required:

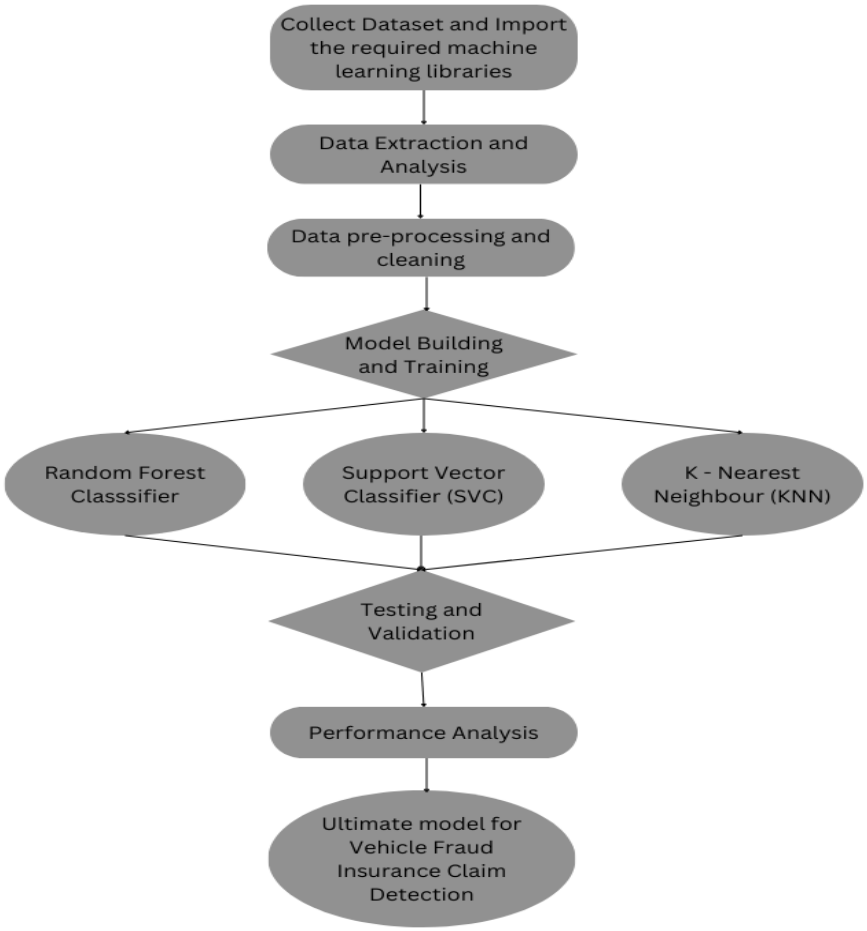
Softwares like: Jupyter notebook, Visual Studio and Dataset (from kaggle).

Project Objective:

To build a classification methodology using machine learning techniques to determine whether a customer is placing a fraudulent insurance claim by using historical data.
Building the models and finding the best suitable model for this application.
Comparative analysis of Machine Learning algorithms:

1. Random forest
2. Support Vector Classifier (SVC)
3. KNN (K Nearest Neighbour)

Flow diagram:



Implementation Plan:

- ➢ Importing the dataset, cleaning the dataset and handling the missing values using **pandas**.
- ➢ Visualizing the dataset (**Graphs and Heatmap**) with the help of **seaborn** and **matplotlib.**
- ➢ **Data preprocessing**:
  Used different approaches for both categorical and numerical data inputs.
  Found the most unique attributes in each column of the dataset, and dropped the column with highest unique values as they don't possess any trend and are not beneficial for our model to learn anything from it.
  a. Numerical data:
     - Used seaborn library for plotting the heatmap, and finding the correlation between data inputs, and dropped the highly correlated attributes as they can affect the data accuracy.
  b. Categorical data:
     - Done the one hot encoding for categorical data to convert it into numerical data for better interpretation.
     - Creating binary encodes for object type variables using **numpy**.
  c. Handling outliers:
     - Visualize the outliers in the dataset(numerical dataset), and get rid of it using scaling techniques.

- ➢ Standardizing the data.
- ➢ Building the machine learning models: Random Forest Classifier, Support Vector Classifier (SVC) and K - Nearest Neighbor (KNN)
- ➢ Analyzing the models (**confusion matrix, precision, recall, f1 score, accuracy score**).
- ➢ The model that gives more correct fraud prediction with more accuracy will be chosen as the final model for this fraud detection in vehicle insurance claim purposes.

Results:

COMPARING MACHINE LEARNING MODELS

↓ Evaluation Parameters	Random Forest Classifier	SVC (Support Vector Classifier)	KNN (K - Nearest Neighbour)
Accuracy	83.2%	76.8%	76.8%
False Positive (FP) Values (from Confusion Matrix)	33	58	58
Prediction Speed	Moderate	Fast	Depends
Average Prediction Accuracy	Higher	Lower	Lower
Training speed	Fast	Fast	Fast (Excluding Feature extraction)

Thus, from the above table we can conclude that the Random Forest Classifier would be the most optimized algorithm to detect the fraud for vehicle insurance claim detection. Because, it has the highest accuracy among the all three algorithms and also the value for False Positive parameter is less for this algorithm which indicates it is the most precise algorithm.
Also the average prediction accuracy for the random forest classifier is higher.

Importance of FP parameter: it indicates the number for the cases where the fraud is actual happened but the model is not being able to predict that fraud. (So, such number should be minimum)

Future enhancement:

Fraud detection system should be able to process the custom dataset uploaded at the front-end phase of the system, with visual graphs and remarks, and detect fraud automatically.
A front-end model can be built which would take the input from the user side and would then predict the output and classify it as Yes or No.
Also we can explore more by using the deep learning techniques, such as **convolutional neural networks (CNNs) and recurrent neural networks (RNNs),** for feature extraction and fraud detection. These approaches would help to capture complex patterns in image, text, and time-series data.
Enhancement on the model transparency and interpretability can be done to build trust among users and stakeholders. XAI techniques can help provide explanations for model predictions and highlight important features contributing to fraud detection.

Web applications can be built using a streamlit library in python.
There we will be creating 3 python files, namely:

1. app.py
   This will be the actual execution page, where we import other two file functions, show_predict_page() and show_explore_page(). This allows us to make a user interface tabs for prediction of fraud and exploring the dataset(data Visualization)
2. Predict.py
   This will be the file where our prediction model is loaded using joblib and the dataset is loaded using pickel. This same file will consist of input tabs that we ask from users. Input tabs can be in the form of selectbox, slider, textInputs, integerInputs depending upon data columns.
   The page will also have a predict button which will display the fraud or genuine insurance claims according to the conditions satisfied by the user inputs.
3. Explore.py
   All the visualization graphs, plots would be plotted here.