Insurance Claim-Fraud Detection

Let’s begin with some simple facts of insurance that we heard from the past and we are still hearing, some of them are,

* It is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness or death in return for payment of a specified premium.
* A thing which provides protection against a possible eventuality.
* It is a form of risk management, primarily used to hedge against the risk of uncertain loss.
* “It is contract, represented by a policy in which an individual or entity receives financial protection or reimbursement against losses from an insurance company.”

Insurance is the next need of a life of a human being as tragedies such as illness, injury or permanent disability; even death can leave you and your family tremendous emotional stress, and emotional grief. With Insurance in place, you and your family’s financial stress will be reduced, and you can focus on recovery and rebuilding your lives.

There are different types of insurance, like.....Life insurance, health insurance, motor insurance, travel insurance, property insurance, mobile insurance and many more. So many companies provide life insurance in the early stages while signing the offer letter and accepting the terms and conditions provided by the organisation, even they provide medical insurances to provide medical safety. So, now days it is very tough to claim insurance even after paying all the premium on time, sometimes customers make some fraud and many times from companies end it happened, this problem is growing big day by day.

IRDA(Insurance Regulatory and Development and Authority) is the main body who regulates all the insurance activities in India. It’s mission, “to protect the interest of policy holders, to regulate, promote and ensure orderly growth of the insurance industry.

**Fraud Scenarios**:

A few scenarios of insurance fraud:

* Producing forged documents
* Non-disclosure of critical information
* Buying of policies in the name of dead person or a person with a terminal illness
* Stating false reasons for claims
* Misappropriating assets
* Manipulating pre-policy health check-up records
* staged accidents and fake disability claim

**IRDA Fraud Policy**:

According to the Insurance Regulatory and Development Authority (IRDA), every insurance company is required to set up a Fraud Monitoring Framework. The framework shall include measures to protect, prevent, detect and mitigate the risk of fraud from policyholders/claimants, intermediaries and employees of the insurance companies.

**Anti-Fraud Policies**:

Insurers are expected to adopt a holistic approach to adequately identify, measure, control and monitor fraud risk and accordingly lay down appropriate risk management policies and procedures.  The Insurance company Board of Directors are mandated by the IRDA to review their respective Anti-Fraud Policies on an annual basis, and at such other intervals as it may be considered necessary. Such policies need to provide a comprehensive guideline on fraud monitoring procedures, identification of potential avenues of fraud, guidelines to cooperate and coordinate with State and Claw enforcement agencies for identifying the act of fraud as well as the perpetrators.

These policies also guide in building a framework that will allow them to exchange information with other insurance companies with regard to sharing intelligence on the occurrence of incidents and scenarios of such frauds so that these can be red-flagged within the insurance ecosystem.

Role of Machine Learning in predicting Insurance Fraud**:**

Artificial intelligence and machine learning plays an important role in banking industry, as insurance claim is part of BFSI industry and while claiming the insurance, number of fraud cases are increasing day by day, so it is very important for the country to use machine learning models to predict the fraud while analyzing large datasets. The popular form of machine learning applied to the insurance industry is called deep anomaly detection. Anomaly detection works by analyzing normal, genuine claims made by the customer and forming a model of what a typical claim looks like.

Analysis of Problem Statement:

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, dataset is provided which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not?

“The problem statement is, by using the different data provided, we have to predict, the insurance claim is Fraudulent or not?”

(<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile_insurance_fraud.csv>)

Above dataset consist of details for Auto Insurance Industry in which we have different details of customer age, policy number, premium amount, incident date, incident city, injury claim, vehicle claim, property claim, auto model , auto year and many more, a total of 1000 rows and 40 columns. While dataset is a combination of string and integer data types with no missing values, as if there are any missing values in the dataset, it affect the final accuracy of model, which results in poor prediction.

To begin with data analysis, we imported required libraries as matplotlib.pyplot and seaborn which gives us graphical representation and relationships of different columns present in the dataset. Some important observations from the analysis of the data are:

* Number of fraud cases are most in 'OH' state and least number of fraud cases are in 'IL' state. 'IL' is the state where number of no-fraudulent is high and rate of no fraudulent cases is low in 'IN' state.
* For the Male category, in both the cases i.e., fraudulent and no-Fraudulent, is highest.
* Most of the fraudulent cases are noticed for the occupation as 'exec-managerial'. Most of the no-fraudulent cases are noticed for the occupation as 'machine-op-inspect'.
* Fraudulent case rate is high for the category sub-category of other\_relative.no-fraudulent case is high for sub-category of own-child.
* Rate of fraud cases is high for single vehicle collision. Rate of no-fraud cases is high for Multi-Vehicle-Collision.
* Fraudulent and no-fraudulent rate is high for the case where the number of vehicles involved is 1.
* Fraud cases are high for Mercedes company. No-fraudulent cases are high for Nissan company.
* Highest number of fraud cases are there in the year 2007 and lowest in the year 1998.Highest number of no-fraud cases are there in the year 1999 and lowest in the year 2004.
* There is linear relationship in most of the cases for the two columns, as; amount for property claim is increasing, total amount claim also increasing and vice-versa.
* In the dataset, more number of people are in the age-range (30 to 40).
* policy deductable amount is INR 1000-1100 for 350 users. Policy deductable amount is INR 400-600 for 330 users. Policy deductable amount is INR 1800-2000 for 320 users.
* Most of the users are from ‘OH’ state and ‘Columbus’ city.
* Number of vehicles is 1 for 580 users, 2 for 20 users, 3 for 350 users and 4 for 50 users.
* Main focus on statistical summary as it indicates us about mean, median, maximum value, minimum value, and standard deviation for each column individually. And we get insight about missing values, outliers , skewness for individual column.
* Highest density of total\_claim\_amount is in the range of INR 50,000 to 70,000.
* Highest density injury claim amount is in the range between INR 5000 to 8000.
* Highest density property claim amount is in the range between INR 5000 to 7000.
* Highest density for vehicle claim amount is in the range between INR 35000 to 57000.
* Plotting the heat-map to check the correlation of columns with each other.
* For most of the columns, outliers are there but they are near to the whiskers, so we kept them for further operations.
* As the difference between number of fraudulent and no-fraudulent cases is very high, so there is class imbalance problem.

“We did EDA of the dataset by using different plots as countplot, heatmap, boxplot, histogram and to check data distribution we used distplot”.

# Preprocessing of Data:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

Here, in the dataset maximum columns are of string data type, as EDA were already done, we converted the string data type into integer data type as Machine Learning Models can only work with numeric data. So for that, we used Label Encoder to encode the required data and then by using Standard-Scaler operation, data scaling were done to get great model Accuracy. As class imbalance problem is there in the dataset, we use oversampling operation, as a result class imbalance problem is solved.

# Building Machine Learning Models:

Following Models and libraries were imported:

* from sklearn.linear\_model import LogisticRegression
* from sklearn.metrics import accuracy\_score
* from sklearn.metrics import confusion\_matrix,classification\_report
* from sklearn.tree import DecisionTreeClassifier
* from sklearn.ensemble import RandomForestClassifier
* from sklearn.model\_selection import cross\_val\_score
* from sklearn.model\_selection import GridSearchCV
* from sklearn.svm import SVC
* from sklearn.naive\_bayes import GaussianNB
* from sklearn.model\_selection import train\_test\_split

Model Accuracies for different machine learning models are as follows:

* Logistic Regression – 77.43%
* DecisionTreeClassifier – 77.65%
* SupportVectorClassifier – 84%
* GaussianNb – 72.34%
* RandomForestClassifier – 84.95%

Always take a note of one thing, we are getting above model accuracies due to overfitting or underfitting of the data which in future will affect our model performance.So, to overcome this problem, always check the cross validation score as it help us to choose perfect model for prediction purpose. After applying cross validation, we came to know that SupportVectorClassifier model performing best among other models as the difference between accuracy and cross val score is less for SVC.

Here, while building any model for prediction, always go for Hyper parameter tuning as it will increase our model accuracies by using the best parameters(C and kernel)

After applying hyper parameter tuning with the best parameters, we got the final accuracy of the model as, 87.38%.

Importing joblib to dump the model and then to load the SVC model as follows,

joblib.dump(SVC\_M,"Insurance\_Claim.pkl")

Final\_Model=joblib.load("Insurance\_claim.pkl").

# Conclusion:

“Model is built which help to predict the claim of auto insurance as fraudulent or no-fraudulent.”

An article by:

Name - Mr. Akshay Junankar

email – [akkijunankar@gmail.com](mailto:akkijunankar@gmail.com)

Institute – Datatrained Education

Batch - 1828

reference links for help:

<https://www.financialexpress.com/>

<https://www.irdai.gov.in/>