Predicting the “Fraud in auto insurance claims”
&
Pattern extraction

Background & Problem statement:

A major general insurance company has a business problem with significant number of claims being reported are fraudulent in nature and it is leading to leakages. So, the Insurer decided to predict the fraudulent ones before even processing the claims to allocate costs appropriately, to keep the thorough investigation process in place and to design proper action plan for the claims etc.

Insurance fraud refers to any claim with the intent to obtain an improper payment from an insurer. Motor and health insurance are the two prominent segments that have seen a spurt in fraud. Frauds can be classified from source and/or nature point of view.

Sources can be policyholder, intermediary and/or internal with the latter two being more critical from internal control framework point of view. Frauds can be classified into nature wise, for example, application, inflation, identity, fabrication, staged/contrived/induced accidents etc.

Fraud affects the lives of innocent people as well as the insurance industry and thus it may be of interest for the health of the Insurance Industry and Society. In fact, Insurers report certain classified cases to Regulator and Law enforcement agencies like Police, Crime Bureaus and others as mandated by the Regulators/Government and required by Law. With the advent of organised gangs and/or collusion, the problem has become more complex and sophisticated and the frauds have been difficult to detect and to prove, if detected.

The framework of prediction of fraud and pattern extraction wil be useful for the insurance companies, regulatory body, intelligence department etc.

Prediction at the time of processing claims will reduce costs and minimize losses.

The intelligence arising out of ever improving prediction algorithms will help retrofitting in terms of improvement of the underwriting process, exercise of good selection of policyholders based on identified profile attributes, strengthening of internal risk management mechanisms and finally, a clear guidance and communication to employees and other stakeholders involved.

At the Industry level, the shared aggregate information helps build appropriate intelligence and resilience while paving the way for collective effort for prevention as well as minimizing losses and to match the efforts of perpetrators.

At the Regulator and Law enforcement level, the intelligence arising out of prediction will help revamp the Regulations/Laws and plan not only enforcement but Industry based initiatives/systems for resilience and to share information for consumption of the Industry and the Society.

Prediction at the time of processing claims will reduce costs and minimize losses for the insurance company. Hence, prediction of fraud plays very important role in auto insurance claims. The company wants to understand the hidden patterns in the data which lead to construction of investigation process as well as claim settlement decision.

Beyond building a model to predict fraud you will have to identify the patterns for fraud which will help in-turn to the company to take action accordingly to initiate the investigations on claim clases to identify if fraud exists and also to handle the fraudulent cases while settling the claims.

You are expected to create an analytical and modelling framework to predict the fraud in auto insurance claims based on the demographic, policy, claim, and vehilce related features provided in the datasets and also generate the top 20 patterns for fraud on target attribute using the decision tree algorithms only, while answering other questions too cited below.

Data set details and Attribute Description:

1. **Demographics Data :** These files consist of the demographic data of each customer, like CustomerID, Country, InsuredAge, InsuredGender, InsuredEducationLevel, etc .
2. **Policy Information :** These files consist of the customer auto insurance policy information, connected to the claim with the insurance company, like, CustomerID, InsurancePolicyNumber, DateOfPolicyCoverage, InsurancePolicyState, UmbrellaLimit, etc.
3. **Data of Claim :** These files consist of the details about the insurance claim, that the customer applied for, like DateOfIncident, TypeOfIncident, AmountOfInjuryclaim, AmountOfPropertyClaim, etc
4. **Data of Vehicle :** These files consist of the details about the Vehicle, connected to the policy.
5. **Fraud Data :** This Train.csv contains the Fraud information details, like CustomerID, ReportedFraud.

The datasets are provided as cited below for the analysis:

1. Demographics Data :

- **"Train_Demographics.csv" & "Test_Demographics.csv"**
- These files consist of the demographic data of each customer, like CustomerID, Country, InsuredAge, InsuredGender, InsuredEducationLevel, etc .

2. Policy Information :

- "Train_Policy.csv" & "Test_Policy.csv"
- These files consist of the customer auto insurance policy information, connected to the claim with the insurance company, like, CustomerID, InsurancePolicyNumber, DateOfPolicyCoverage, InsurancePolicyState, UmbrellaLimit, etc.

3. Data of Claim :

- "Train_Claim.csv" & "Test_Claim.csv"
- These files consist of the details about the insurance claim, that the customer applied for, like DateOfIncident, TypeOfIncident, AmountOfInjuryclaim, AmountOfPropertyClaim, etc

4. Data of Vehicle:

- "Train_Vehicle.csv" & "Test_Vehicle.csv"
- These files consist of the details about the Vehicle, connected to the policy.

5. Fraud Data:

- "Train.csv" & "Test.csv"
- This Train.csv table contains the Fraud information details, like CustomerID, ReportedFraud.
- **This Test.csv contains only CustomerID (Not the target attribute, which is to be predicted)**

6. Attributes Details : "AttributeInformation.docx"

- **This has the details of attributes for the datasets cited above (1 to 5)**

Note: For analysis, consolidate/aggregate all the datasets cited above (1) to (5)

Note: Missing values are denoted as “”,”NA” in general in the datasets. Please go through the document "AttributeInformation.docx" thoroughly to address different aspects in the data.

Feature Engineering

Study the problem carefully and identify the scope of creating new variables

Data Visualizations

Choose suitable visualization

Model Building

- What is the performance metric you choose and explain the reason for the choice
- Build 5 different machine learning models and experiment with at least 3 parameters for hyper-parameter tuning
- Feature engineering
- Consolidate the performance of all the models in a data frame and write your comments
- Error Metrics: Consider “F1 statistic” for “Y” level of Target attribute as error metric and tune the model accordingly.
- Consider appropriate evaluation metric for deciding the top 20 patterns for fraud on target attribute.

Project Deliverables and Deadlines

Capstone Schedule with Dates (12th April 2023 – 2nd May 2023)		
Date	**Event/Deliverable**	**Type of file to submit through SLMS and timings**
1st May 2023	Visualizations and Hypothesis Testing results	Zip file with code in ipynb and HTML formats only

13th May 2023	Model Building results	Zip file with code in ipynb and HTML formats only
14th May 2023	PPT	PPT in PPTX format as per the guidelines
15th , 16th,17 May 2023	In-person 30 minutes viva	

Submission

- Submissions uploaded in SLMS only will be considered for the evaluation.
- Please submit a Zip folder that includes all outputs in jupyter notebook (.ipynb) and HTML format only.
- Zip folder name should follow the format as mentioned below

  < _XXXXXX_Project_Enrollid.zip >

Viva Metrics

Code Quality	Code Execution
	Code Structure and Completeness
	Commenting in Code
Presentation Skills	Quality of Presentation Material
	Overall, Clarity of Thought and Flow of Presentation
Coding Skills	Python
Modelling Skills	Data Handling
	Models
	Model Validation
Conceptual Skills	Statistics
	ML
	AI
	Data Visualizations
	Big Data and Data Engineering