

**Dineshchandar Ravichandran**

dravich@g.clemson.edu

**Archana Lalji Saroj**

[asaroj@g.clemson.edu](mailto:asaroj@g.clemson.edu)

**Prashanth Reddy Kadire**

[pkadire@g.clemson.edu](mailto:pkadire@g.clemson.edu)

**CPSC-6300**

**Insurance Fraud Detection**

**Checkpoint 1**

Contents

[Introduction & Problem statement 3](#_Toc115180715)

[Motivation & Goals 3](#_Toc115180716)

[Data Exploration 3](#_Toc115180717)

[Data Summary 3](#_Toc115180718)

[What is the Unit of Analysis? 4](#_Toc115180719)

[How many observations in total are in the dataset? 4](#_Toc115180720)

[How many unique observations are in the dataset? 5](#_Toc115180721)

[What time period is covered? 5](#_Toc115180722)

[Data cleaning: 5](#_Toc115180723)

[Description of outcome with an appropriate visualization technique 6](#_Toc115180724)

[Project Approach: 11](#_Toc115180725)

# Introduction & Problem statement

The insurance industry consists of over 7,000 companies that collect over $1 trillion in premiums yearly. The massive size of the industry contributes significantly to the cost of insurance fraud by providing more opportunities and bigger incentives for committing illegal activities. The total cost of insurance fraud (non-health insurance) is estimated to be more than $40 billion annually. That means Insurance Fraud costs the average U.S. family between $400 and $700 per year in the form of increased premiums.[[1]](#footnote-2) Insurance fraud also steals at least $308.6 billion yearly from American consumers.[[2]](#footnote-3)

For this project, we will focus on Automobile-insurance fraud where 25%-33% of insurance claims have an element of fraud.[[3]](#footnote-4) Automobile claim fraud and buildup added $5.6 billion-$7.7 billion in excess payments to auto-injury claims paid in the U.S. in 2012. 21 % of bodily injury (B.I.) claims and 18 % of personal injury protection (PIP) claims closed with payment had the appearance of fraud and/or buildup. Buildup involves inflating otherwise legitimate claims.[[4]](#footnote-5)

The government and other organizations are responding to this by investing in technology to detect fraudulent claims, 21% of insurance institutes plan to invest in A.I. (Artificial Intelligence) in the next two years.[[5]](#footnote-6)

# Motivation & Goals

We will construct a supervised machine learning model-based Fraud Detection system to predict the chances of a fraudulent insurance claim. This system aims to analyze the insurance claim data set containing 38 distinctive features and generate a classification logic to identify genuine and fraudulent claims. By constructing such a fraud detection system, we can alert the insurance institutes and allow them to take necessary action against fraudulent claims.

# Data Exploration

## Data Summary

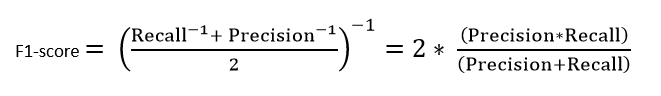
We collected the insurance claims dataset from Kaggle "insurance\_claims\_data."

This dataset comprises 1000 total data points. The dataset is imbalanced since there are a total of 247 fraud claims and 753 genuine claims, according to the preliminary data exploration.

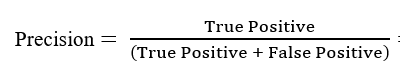
### What is the Unit of Analysis?

For this project, the accuracy, Precision, and F1 Score will be calculated on each model. These measures are considered as our unit of analysis for selecting the best model.

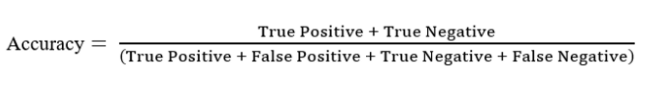
* **F1 score**: this is the harmonic mean of precision and recall and gives a better measure of the incorrectly classified cases than the accuracy matrix.



* **Precision:** It is implied as the measure of the correctly identified positive cases from all the predicted positive cases. Thus, it is useful when the costs of False Positives are high.



* **Accuracy:** One of the more obvious metrics is the measure of all the correctly identified cases. It is most used when all the classes are equally important.

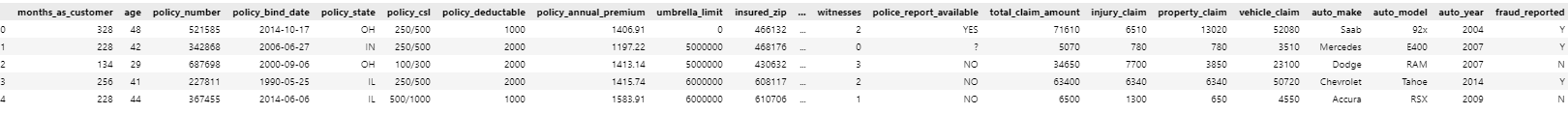


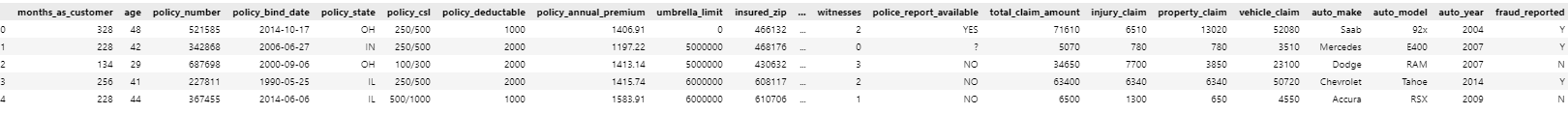
### How many observations in total are in the dataset?

We have a data set containing 1000 insurance claims and 38 features pertaining to it. In addition, we also have a column stating which of these insurance claims are fraudulent and which are genuine as illustrated below:

Graphical user interface, text, application, Word

Description automatically generated





### How many unique observations are in the dataset?

In our dataset, every claim is unique, though we do not have a unique identifier like claim I.D. the policy number for all the claims are unique.

### What time period is covered?

Our dataset contains claims with incidents occurring from 1st Jan 2015 to 1st March 2015.

## Data cleaning:

Three columns had '?' Missing value:

1. Collision Type
2. Property Damage
3. Police Report Available

The missing values are imputed using KNN imputer. The categorical features have been transformed into numerical features using one hot encoding and label encoding.

## Description of outcome with an appropriate visualization technique

#### Description of the dataset:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| months\_as\_customer | int64 | No.of months customer has been a customer to the insurance organization. |
| age | int64 | Age of the customer. |
| policy\_number | int64 | Policy no.of the customer's insurance. |
| policy\_bind\_date | object | The moment when the insurance coverage goes into force, it's date and time specific. |
| policy\_state | object | State in which the policy was procured. |
| policy\_csl | object | Combined single limits are a provision of an insurance policy limiting coverage for all components of a claim to a single dollar amount. A combined single limit policy has a maximum dollar amount that covers any combination of injuries or property damage in an incident. |
| policy\_deductable | int64 | The amount customers have to pay for covered services before their insurance plan starts to pay. |
| policy\_annual\_premium | float64 | Annual payment for the policy. |
| umbrella\_limit | int64 | Extra insurance that provides protection beyond existing limits and coverages of other policies. Umbrella insurance can cover injuries, property damage, certain lawsuits, and personal liability situations. |
| insured\_zip | int64 |  |
| insured\_sex | object | Male/ Female. |
| insured\_education\_level | object | Education level of the customer. |
| insured\_occupation | object | Occupation of the customer. |
| insured\_hobbies | object | Customers' hobbies. |
| insured\_relationship | object | Relationship of the person involved in the incedent to the actual insurance holder. |
| capital-gains | int64 | Increase in a capital asset's value. |
| capital-loss | int64 | Loss in a capital asset's value. |
| incident\_date | object | Date of incident. |
| incident\_type | object | Type of incident (Single Vehicle Collision, Vehicle Theft, Multi-vehicle Collision, and Parked Car). |
| collision\_type | object | Type of collision (Side Collision, Rear Collision, and Front Collision). |
| incident\_severity | object | The severity of the incident classified as per damage (Major Damage, Minor Damage, Total Loss, and Trivial Damage). |
| authorities\_contacted | object | Type of authorities contacted after the incident. |
| incident\_state | object | U.S. state where the incident occurred. |
| incident\_city | object | City in which the incident occurred. |
| incident\_location | object | Address of the incident. |
| incident\_hour\_of\_the\_day | int64 | Time of incendent in 24hrs. |
| number\_of\_vehicles\_involved | int64 | No.of vehicles involved with the incident. |
| property\_damage | object | If 'YES,' then the insurance carrier helps pay to repair the damage customer caused to other involved parties. |
| bodily\_injuries | int64 | No.of bodily injuries. |
| witnesses | int64 | No.of witnesses to the incident. |
| police\_report\_available | object | If a police report exists of the incident. |
| total\_claim\_amount | int64 | The total amount claimed by the customer fot the incedent. |
| injury\_claim | int64 | The portion of the total claim requested for the injury claim. |
| property\_claim | int64 | The portion of the total claim requested to pay for property damages. |
| vehicle\_claim | int64 | The portion of the total claim requested for vehicle damage. |
| auto\_make | object | Manufacturer of the customer's vehicle that was involved in the incident. |
| auto\_model | object | The specific automobile model of the customer's that was involved in the incident. |
| auto\_year | int64 | Model year |
| fraud\_reported | object | Determining whether a claim is fraudulent or not. |

#### Annual policy premium distribution

Chart, histogram

Description automatically generated

The above graph gives us the distribution of the annual premium paid by the customers for their insurance, where the bulk of the population pays between 1000-1500.

#### Age distribution of policyholders

Chart, histogram

Description automatically generated

The above graph gives us the distribution of the age of the customers.

*Genuine claims VS Fraudulent claims comparisons*

*Chart, pie chart

Description automatically generated*

As illustrated in the above graph, only 24.70% of the total claims are fraudulent.

*Incident state-wise fraudulent claims*

*Chart

Description automatically generated*

As illustrated above, most fraud-reported incidents occurred in N.Y. and S.C.

*Age-wise fraudulent claims*

Chart, bar chart

Description automatically generated

As illustrated above, most fraud claims are committed by customers aged between 19-23, and hence is an essential feature of our detection model.

*Incident type and damage relation to fraudulent claims*

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Based on the above graphs, we can determine that most of the fraudulent claims tend to be collisions with reported major damage.

# Project Approach:

Based on the above analysis, we will solve this binary classification problem using different Machine Learning algorithms such as Regression, SVM, Random Forest, Adaboost classifier, and XGboost classifier. According to the performance metrics, we will choose the best classifier.

To start with we will be selecting the following models:

#### Logistic Regression from sklearn.linear\_model:

Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable. The target or dependent variable's nature is binary, meaning there would be only two possible classes: 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X.

#### RandomForestClassifier from sklearn.ensemble:

Random forest algorithm builds decision trees on data samples, obtains predictions from each one, and then uses voting to determine the best option. Because it averages the outcomes, the ensemble method is superior to a single decision tree in that it lessens over-fitting.

#### XGBClassifier from XGBoost:

Extreme Gradient Boosting is abbreviated as XGBoost. The "eXtreme" part of XGBoost's name refers to speed-improving features such as parallel processing and cache awareness that make it around ten times faster than conventional gradient boosting. A special split-finding method and integrated regularization that lessens over-fitting are also included in XGBoost. XGBoost is a quicker, more accurate version of gradient boosting.

1. Federal Bureau of Investigation insurance reports and publication <https://www.fbi.gov/stats-services/publications/insurance-fraud> [↑](#footnote-ref-2)
2. Coalition Against Insurance Fraud is working to update this figure in 2022. [↑](#footnote-ref-3)
3. http://www.insurancejunction.co.za/insurance-fraud/ [↑](#footnote-ref-4)
4. https://www.michigan.gov/difs/consumers/fraud/insurance-fraud-statistics [↑](#footnote-ref-5)
5. https://insurancefraud.org/fraud-stats/ [↑](#footnote-ref-6)