# Abstract

This report initiates a thorough descriptive analysis of car insurance claims made in the USA. Fraud detection of claims made for car accidents is important as it reflects on an area’s community, that is if the population of that are law-obeying citizens. Further, false claims could impact a country’s economy. Identifying patterns and trends that could indicate fraudulent activities is of great interest to insurance companies.

The data considered for analysis were taken from the online platform Kaggle and contains various attributes of insurance claims. Each of these claims have various features attributed to them, including demographic information about the drivers, details of the vehicles insured, and specifics of the claims made. This project utilizes statistical methods and visualization techniques to glean an understanding of the data, thereby enabling identification of influential characteristics which can be used for fraud detection. We perform a combination of univariate, bivariate, and multivariate techniques for the same.

Moreover, multiple models are fitted for the data to fulfill our purpose of detecting car insurance fraud. Further, we explore the effectiveness of these models by comparing mean squared error. The results show that the model built with the gradient boosting method is the best model for fraud detection.

# Table of Contents

[Abstract 1](#_jq5noo5w6n3t)

[Table of Contents 1](#_1wscqc2q1cv9)

[1. Introduction 1](#_2uux8vjx8hrh)

[2. Description of the question 2](#_tleu40jjuj7b)

[3. Description of the dataset 2](#_x6lu4jz2p7b2)

[Data preprocessing 3](#_wj2hpp9h6knw)

[4. Important results of explorative analysis 3](#_dsxbx9gpwxj2)

[5. Important results of advanced analysis 4](#_uolv2nnfp1ll)

[5.1 Selected models and justification of choice 4](#_bkt8hrqkz5gh)

[5.2 Evaluation of Classification Models on Original Training Set 5](#_2ao02rpy0m29)

[5.3 Handling Class Imbalance and Evaluation of Models on Data Set after applying different sampling techniques 6](#_d19s31elwixm)

[5.4 Selection of best model 9](#_67t4rnipbzrc)

[5.5 Feature importance in best model 9](#_uvs3vfa4wrwt)

[6. Output of study 10](#_ot8t7924d8v3)

[7. Discussion and conclusions 10](#_qhk4s91wzmwd)

[Appendix 10](#_6oiee1280pg)

# 1. Introduction

The increasing frequency of car insurance claims necessitates thorough analysis of claims to ensure integrity and economic stability. Transportation has been a basic need of humans since the early days of humankind, and today almost every professional owns their own vehicle. Car insurance claims form a dynamic domain where patterns and trends are influenced by various features of the drivers, vehicles, and accident circumstances. Claims data showcase valuable insights of overall risk conditions and the integrity and efficiency of an insurance provider. This project engages in a descriptive analysis of the car insurance claims filed in the USA. The different attributes of claims that are included in the used dataset are explored and the features with an impact on the response variable “fraud” are identified. This is achieved using the Python and R programming language as well as the Jupyter Notebook and RStudio platforms. The findings of this project will be of importance to vehicle insurance payout firms, and any others with an interest in United States vehicle insurance policies.

# 2. Description of the question

This report aims to explore trends in the data and gain a thorough understanding of variables influencing car insurance fraud in the United States. It targets identifying variables that are highly associated with the variable “fraud”, as predicting fraud is the ultimate goal. Vehicle insurance companies that investigate claims can do so better based on certain aspects of the claims, and thus will be able to weed out fraudsters.

The main questions that this project aims to answer are:

•Which variables have a higher impact on accurately detecting car insurance and to what extent?

•What machine learning models are able to detect fraud in insurance claims on cars in the United States and which is the best model for the same?

# 3. Description of the dataset

The dataset considered contains information of 17998 claims on cars. Each observation (claim) has 25 attributes which may contribute to fraudulence. A brief description of them are as follows:

| **Variable Name** | **Description** |
| --- | --- |
| **claim\_number** | Claim Number |
| **age\_of\_driver** | Age of Driver |
| **gender** | Gender (M – Male, F-Female) |
| **marital\_status** | Marital Status (0-Unimarried,1-Married) |
| **safty\_rating** | Safety Rating |
| **annual\_income** | Annual Income |
| **high\_education\_ind** | Higher Education Indicator (0-No,1-Yes) |
| **address\_change\_ind** | Address Change Indicator (0-No,1-Yes) |
| **living\_status** | Living Status (Own,Rent) |
| **zip\_code** | ZIP Code |
| **claim\_date** | Claim Date |
| **claim\_day\_of\_week** | Claim Day of Week(Sunday , Monday, Tuesday, Wednesday, Thursday, Friday ,Saturday) |
| **accident\_site** | Accident Site(Local, Parking Lot, Highway) |
| **past\_num\_of\_claims** | Number of Past Claims |
| **witness\_present\_ind** | Witness Present Indicator (0-Not Present,1-Present) |
| **liab\_prct** | Liability Percentage |
| **Channel** | Channel(Phone, Broker, Online) |
| **policy\_report\_filed\_ind** | Police Report Filed Indicator(0-No,1-Yes) |
| **claim\_est\_payout** | Claim Estimated Payout |
| **age\_of\_vehicle** | Age of Vehicle |
| **vehicle\_category** | Vehicle Category (Compact, Large, Medium) |
| **vehicle\_price** | Vehicle Price |
| **vehicle\_color** | Vehicle Color(Black, Blue, Gray, Red, Silver, White, Other) |
| **vehicle\_weight** | Vehicle Weight |
| **fraud** | Fraud Indicator(0-No,1-Yes) |

## Data preprocessing

Prior to analyzing the data, it needed to be preprocessed to derive meaningful and clear interpretations. This was achieved by performing the following:

* The “claim\_date” variable was first converted to a datetime variable and then split into 3 different columns named “year”, “month” and “date”.
* Missing values were observed and handled by imputing 0 (for variables “claim\_est\_payout” and “witness\_present\_ind” ) and 2 for “marital\_status” signifying status unknown.
* No duplicates were observed and the “claim\_number” column was made into the index.
* Dataset was split into training (80%) and testing (20%) datasets.
* Variables were encoded using appropriate techniques (one-hot encoding, target encoding) prior to fitting models.

# 4. Important results of explorative analysis

* Descriptive analysis of the response variable, “fraud” helped understand it better. It is evident from the figure 4.1 of the distribution of the response variable, “fraud” that the majority of drivers’ claims have been identified as not fraud.
* When exploring driver characteristics it was evident that most drivers are male. Most drivers have had higher education while most of them are married. Noticeably, the highest number of claims had been made from those living in the zip code 15001 which is Aliquippa, Pennsylvania which could be due to the stronger laws in that area, and most drivers who filed for claims have changed their address.
* The age of fraudulent vehicles has a wider spread, suggesting a variety of ages, while the same of non-fraudulent vehicles) peaks around 3-5 years, indicating that older vehicles are more associated with fraud. This highlights potential differences in vehicle ages between fraudulent and non-fraudulent cases.
* The data revealed that women are more likely to commit fraud compared to men. Married individuals are more likely to commit fraud and those with higher education are also more likely to commit fraud.
* Local sites exhibit higher counts for both non-fraud and fraud cases. However, when examining highways and parking lots, a distinct pattern emerges: highways are more frequently associated with fraud activities, whereas parking lots are more commonly linked with non-fraud cases.
* By comparing p-values obtained from Chi-squared test to a 5% significance level, we identified 8 variables that are significantly associated with the 'fraud' variable. These variables were gender, marital status, higher education, address changed, living status, accident site, presence of witness and filing of policy report.
* The correlation analysis of the dataset reveals a significant positive relationship between annual income and age of the driver, with a correlation coefficient of 0.90. Thus, multicollinearity exists in this data. Pearson correlation was calculated for continuous variables while Cramer’s V was used for categorical variables.
* Principal component analysis and multiple correspondence analysis were attempted on the data with hopes of reducing dimensions. However, these did not yield favourable results.
* We applied factor analysis to the entire dataset (FAMD - Factor Analysis for Mixed Data); however, the scree plot showed that even with five dimensions combined, less than 2% of the total variation was captured. This minimal explanatory power indicated that factor analysis is not effective for our dataset. Consequently, further analysis using biplots and loading plots is unlikely to provide meaningful insights.
* Cluster analysis was also attempted but however it did not yield satisfactory results. Kmediods algorithm was applied on the original data and kmeans algorithm was applied after implementing fisher’s discriminant analysis on numerical variables and multiple correspondence analysis on categorical variables. The silhoutte scores of the clusters were close to 0 and thus the data was not clustered well. Therefore, this avenue was not further explored.

# 5. Important results of advanced analysis

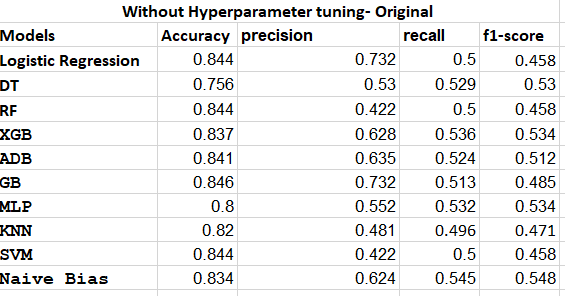
The data set was split into training (80% of original data) and test (20%) tests in the pre-processing stage and a few select models were fitted on the training data and tested on the test data.

## **5.1 Selected models and justification of choice**

1. **Logistic Regression**: A simple, interpretable technique that provides probabilities of class membership, makes it useful for understanding the likelihood of fraud incidence based on various features.
2. **Support Vector Machine (SVM)**: SVMs are able to find optimal hyperplanes to separate classes, potentially capturing intricate decision boundaries between claims that are fraudulent and not fraudulent.
3. **Decision Tree**: Decision trees are intuitive and capture non-linear relationships between features and the target variable, making them suitable for prediction tasks like detecting car insurance fraud. They can handle both numerical and categorical data.
4. **Random Forest**: Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and generalization and is robust to overfitting.
5. **XGBoost**: A powerful gradient boosting algorithm, it is known for its performance and scalability. Effectively handling imbalanced datasets and it captures complex interactions between features, making it well-suited for fraud detection where class distribution may be highly skewed.
6. **KNN**: KNN is simple and easy to understand, thereby well suited for initial exploration of the data and potentially capturing localized patterns relevant to fraud detection.
7. **Adaboost**: Adaboost is an ensemble learning method that combines multiple weak classifiers to create a strong classifier. Adapting to the complexity of the dataset, it potentially improves predictive performance when data is highly imbalanced.
8. **Gradient boost**: This method builds a series of decision trees sequentially, with each tree correcting errors made by previous ones.
9. **MLP (Multi-Layer Perceptron):** MLP is an artificial neural network capable of learning complex relationships between input features and the target variable. It's apt for predicting liver disease where there are non-linear relationships and interactions between multiple variables.
10. **Naive Bayes Classifier**: Naive Bayes is a probabilistic classifier which applies Bayes' theorem with a "naive" assumption of independence between features. It is computationally efficient and often effective for many classification tasks.

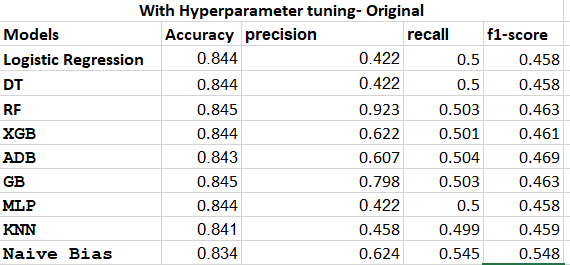
## **5.2 Evaluation of Classification Models on Original Training Set**

Upon completion of preprocessing, we fitted the above models to the data. The table below presents the performance metrics obtained from these models **without hyperparameter tuning** (using default parameters):



The **Gradient boosting** model emerged as the top performer with evaluation metrics, accuracy: 84.6%, precision: 73.2%, F1-score: 48.5% and recall 51.3%.

The table below presents the performance metrics obtained from these models with **hyperparameter tuning**:

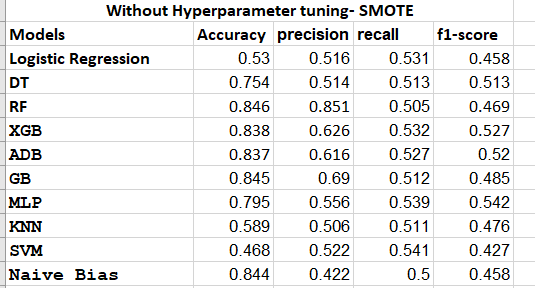


The **Random forest** model emerged as the top performer with evaluation metrics, accuracy: 84.5%, precision: 92.3%, F1-score: 46.3% and recall: 50.3%.

## **5.3 Handling Class Imbalance and Evaluation of Models on Data Set after applying different sampling techniques**

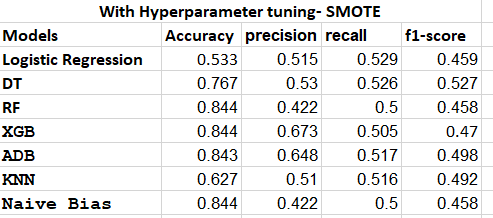
We observed a substantial class imbalance within our dataset, with approximately 84.32% of claims not being fraudulent within the studied population. To address this imbalance, we applied several techniques which were SMOTE (Synthetic Minority Over-sampling Technique), upsampling and downsampling. The same set of algorithms were then used to fit various models and these were evaluated using the previously-mentioned metrics.

**5.3.1 Smote dataset**

We fitted the above models to the data. The table below presents the performance metrics obtained from these models **without hyperparameter tuning** (using default parameters) on SMOTE data: 

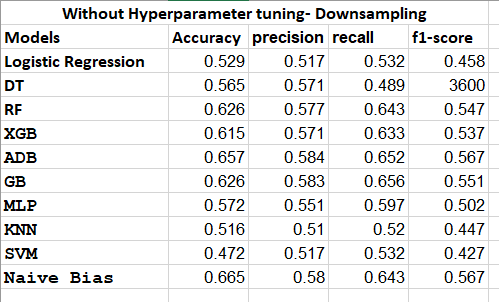
The **Random Forest** model emerged as the top performer with evaluation metrics, accuracy: 84.6%, precision: 85.1%, F1-score: 46.9% and recall: 50.5%.

The table below presents the performance metrics obtained from these models with **hyperparameter tuning** on SMOTE data:



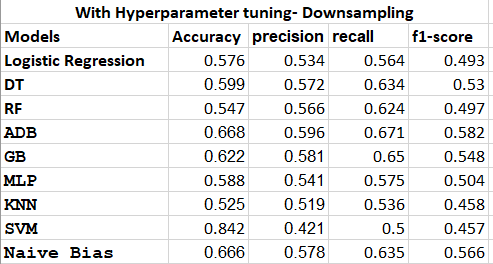
The **XGBoost** model emerged as the top performer with evaluation metrics, accuracy: 84.4%, precision: 67.3%, F1-score: 47% and recall: 50.5%.

**5.3.2 Downsampled dataset**

The table below presents the performance metrics obtained from these models **without hyperparameter tuning** (using default parameters) on downsampled data: 

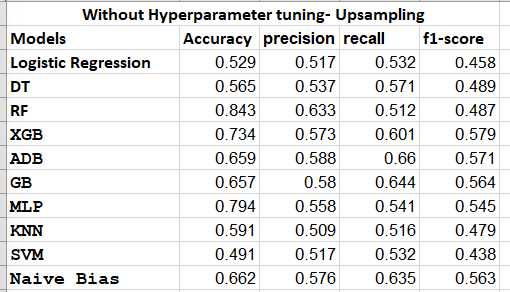
The **Naive Bayes** model emerged as the top performer with evaluation metrics, accuracy: 66.5%, precision: 58%, F1-score: 56.7% and recall: 64.3%.

Next, we evaluated the performance of several models on the downsampled dataset **with hyperparameter tuning**. The table below shows those results.



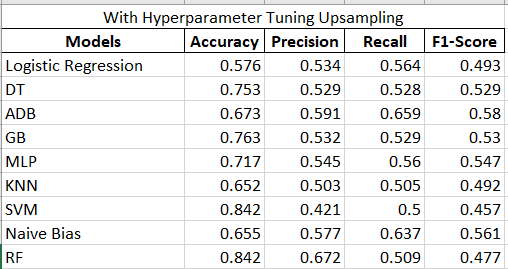
The **Ada boosting** model emerged as the top performer with evaluation metrics, accuracy: 66.8%, precision: 59.6%, F1-score: 58.2% and recall: 67.1%.

**5.3.3 Upsampled dataset**

We fitted the above models to the data. The table below presents the performance metrics obtained from these models **without hyperparameter tuning** (using default parameters) on upsampled data: 

The **Random forest** model emerged as the top performer with evaluation metrics, accuracy: 84.3%, precision: 63.3%, F1-score: 48.7% and recall: 51.2%.

The table below presents the performance metrics obtained from these models with **hyperparameter tuning**:



The **Random Forest** model emerged as the top performer with evaluation metrics, accuracy: 84.2%, F1-score:47.7% and recall:50.9%.

## 5.4 Selection of best model

We consider the model with highest accuracy, F1 Score, precision and recall value as our best model. The gradient boosting model fitted on original data without hyperparameter tuning has accuracy of 84.6% which is the highest overall. The random forest model fitted on original data without hyperparameter tuning has the second highest accuracy of 84.5%, with a higher precision, however both f1-score and recall were higher with the previous model. Therefore, the former was chosen as the best model.

## 5.5 Feature importance in best model

Among all the models evaluated, Gradient Boosting emerged

as the best performer when applied to the original data without

hyperparameter tuning. It achieved the highest overall accuracy,

F1 score, and recall. However, when we selected the top 10 and

30 variables using the feature importance plot and refitted the

Gradient Boosting model, we observed a significant reduction

in model performance. Interestingly, when we used the top

20 variables from the feature importance plot and refitted the

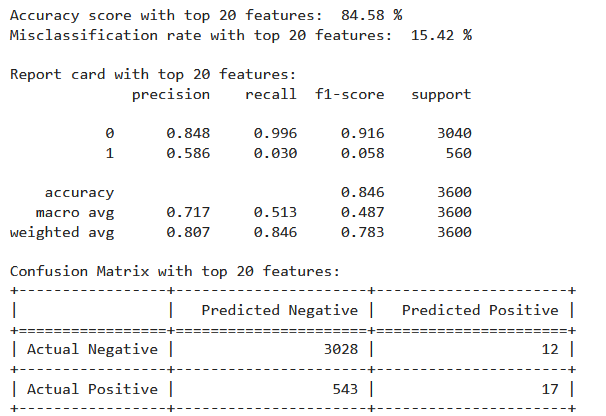
Gradient Boosting model, the performance was approximately

the same as that of the original model without feature selection.

This suggests that the top 20 variables are sufficient to maintain

the model's effectiveness, offering a more streamlined approach

without compromising performance.

****

# 6. Output of study

The main output of our study is the data product which is a website that interested parties can use for fraud detection of insurance claims in the United States.

# 7. Discussion and conclusions

Initial examination of the data revealed class imbalance in the response variable “fraud”. We addressed this issue by implementing SMOTE, upsampling and downsampling. The advanced analysis was conducted on the original dataset as well as the balanced datasets. Various models were constructed, with the optimal model being selected by comparing accuracy, f1-score, precision and recall among other evaluation metrics.

The gradient boosting model emerged as the top performer based on the following metrics:

Accuracy (84.6%): This high accuracy indicates that the model correctly predicts the outcome for almost all instances in the dataset.

F1-score (48.5%): The F1-score reflects a perfect balance between precision and recall for both classes. This implies that the model effectively identifies positive cases while minimizing false positives and false negatives.

Recall (51.3%): With a recall of 100%, the model captures all positive cases without missing any, demonstrating its robustness in identifying instances of interest.

# Appendix

Link to code: <https://drive.google.com/drive/folders/1sKCpWiiXKpd3nNjK1TQpaRcFzmpwEgKQ?usp=sharing>