Fall 2022

Fraudulent Insurance Claims Detection Using Machine Learning

Arif Ismail Alrais
aia5557@rit.edu

Fraudulent Insurance Claims Detection Using Machine Learning

by

Arif Ismail Alrais

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

Fall 2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Arif Ismail Alrais**

Graduate Capstone Title**: Fraudulent Insurance Claims Detection Using Machine Learning**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee/Mentor

Acknowledgments

In The Name of Allah The Almighty God The Merciful The Compassionate.

I would like to express my thanks to those who helped me and pushed me into continuing my academic journey starting from my parents, my Department Head and his vice, ending with my co-workers with all their support and encouragement.

My mentor and advisor, Dr. Ehsan Warriach, deserve immeasurable amount of thanks and gratitude. Even though we had a rough start at the start of this project he have never failed to support me though it all knowing I don't deserve half of it from how lazy I am and how I tend to procrastinate a lot. A lot of thinks to my friend and colleague who helped me through out the courses and his advice through the project.

Finally, My appreciation and gratitude to Dr. Sanjay Modak for his directions and support toward not all me but all of my colleagues and students in the program in numerous occasions throughout the program and the Capstone Project.

Table of contents

Table of Figures

Abstract

As the different countries around the world evolve into a more economical-based and stimulating their economy is the goal. The main purpose of most of these countries is to fight off money launderers and fraudsters for better economic growth. A popular fraud topic in this regard is insurance fraud since it costs the companies and the public billions. Applying data analysis and machine learning are great ways used to address many problems regarding any automated system. To address this problem, first extensive research should be made to check out what has been applied and what the most promising solution using machine learning and data analytics is out there. After learning, then applying and building upon the findings of the research we propose a model that can flag these suspicious fraudulent claims for the insurance companies to help them out in saving money and time and helping them become more efficient in reacting to these fraudulent claims.

Keywords: Machine learning, prediction analysis, supervised learning, fraudulent detection, data visualization, data analysis.

Chapter 1 - Introduction

1.1. Background Information

As we live in a very materialistic world everyone is looking out to protect something they have or own in one way or another. Covid – 19 pandemic has proven difficult to many countries at the beginning of the vaccine revolution since every country is trying to protect their people. Many people were rushing to get the vaccine as insurance to protect themselves. That is the main point and idea behind insurance businesses. People are willing to pay money as a contingent against the unknown loss that they might face. In the U.S alone the insurance industry is valued at 1.28 trillion dollars and the U.S consumer market losses at least 80 billion to insurance fraud every year. That causes the insurance companies to increase the cost of their policies which puts them in a less competitive position against the competition. This in turn also increased the threshold of the minimal payment for a policy since they can afford to do so while everyone is raising prices

This paper aims to suggest the most accurate and simplest way that can be used to fight fraudulent claims. The main problem with detecting fraudulent activities is the massive number of claims that run through the companies systems. This problem can also be used as an advantage if the officials were to take into account that they hold a big enough database if they combined the database of the claims. Which can be used in order to develop better models to flag the suspicious claims

This paper will look into the different methods that have been used in solving similar problems to test out the best methods that have been used previously. Searching if examining these methods and trying to enhance and build a predictive model that could flag out the suspicious claims based on the researching and testing out the different models and comparing these models to come up with a simple enough time-efficient and accurate model that can flag out the suspicious claims without stressing the system it runs on.

1.2. Statement of the problem

The main purpose of this paper is to come up with a model to be used to find out if a certain insurance claim made is a fraud or not. The model will be designed after testing multiple algorithms to come up with the best model that can detect if a claim is fraudulent or not. This is aimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

1.3. Project Definition and Goals

Research Question 1: what are some of the possible approaches to designing a model?

Justification: many approaches can be taken, and many solutions can take place in designing a fraud detecting model. Each one of these solutions and models can be taken into account depending on the situation of the implementation of the model.

Research Question 2: comparison of the different models and any used models used in a similar environment?

Justification: To propose this solution to the desired customer or the end-user of this model a comparison must be available to justify why had this model has been picked over any other models

Research Question 3: The impact of this model results in the real world application if it will be taken by an institution taking into account the effect of the different results of the model

Justification: the main concern is how these results are translated into the real world. How will this affect the performance of a certain institution that will adopt this model as a solution in their operations

There are millions of companies in the insurance industry globally, and collect premiums totaling more than $1 trillion each year. Insurance fraud occurs when a person or organization submits a fraud insurance claim in an effort to collect money or benefits to which they are legally entitled. An insurance fraud is thought to have a total financial impact of over $40 billion so, detection of fraud is a hard task for the insurance sector.

The main goal is to come up with a model as a proposed solution for a possible insurance company for example to detect fraudulent claims made by individuals or entities. Coming up with a solution while answering the above questions is the ultimate goal to have an edge over any solutions that might be adopted by the customer.

1.4. Methodology

Overview:

For this project, a mixed research method will be used. The project will use some explanatory research methods along with experimental research methods and some qualitative and quantitative research methods to explain the findings and the result of the paper to explain the different results.

First, we will identify what is important in running the data according to the business that might use the solutions or the model. In this case, it will most likely be an insurance company that will take into account the financial aspect of each claim as a priority and the personal details will take no account in designing the model itself.

The paper will describe the data at hand and what are the different attributes and how is every attribute relevant to identify if this claim is fraud or not. What are the different types of data present at hand and is it possible to enhance and modify the existing data without tampering with the result of the end goal which is flagging out the suspicious claims. To do this it might be necessary to clean the data and remove some of the values or the attributes or even create new attributes and values by joining and using data integration methods to come up with new values.

After that, we will run the data through different kinds of algorithms and models to try and find out the best model or algorithms for these kinds of data. From the previous research of a lot of the work that has been done on the subject. Decision trees, support vector machines, neural networks, and logistic regression are the top candidate algorithms. If it is possible, running the data through more than one model before coming up with the conclusion that this claim is fraud or not. Of course, this depends highly on the type of business and how much time they are willing to give the model to come up with a decision for a single claim.

Finally evaluating the findings of each model and algorithms with a set of data that haven't been used in the training phase of the model to check how accurate is this model and find out if there is any overfitting or underfitting before the selection of the model or modifying it if it was possible to come up with a better and a more accurate model for the end-user or the customer.

This research is divided into following phases.

Tools and addons:

To draw the insights and related visualizations from the dataset, we employed a variety of tools and addons. In order to employ python commands, jupyter notebook from anaconda. Given that python primarily used to compare suggested solutions to multiple methods and record statistical results to be examined. We were able to plot and graphs using python in this study.

Python libraries Used:

In order to implement the model, we have used different python libraries that are listed here.

Table 1. Python Libraries

Library	**Function**	**Purpose**
Pandas	read_csv	To read the dataset
Matplotlib	pyplot	To plot certain plots and graphs
Seaborn	heatmap	To check correlation between features
warnings	filterwarnings	To ignore warnings
sklearn.model.selection	train_test_split	To breakdown dataset into training and testing part
collections	counter	To count pairwise element in testing and training parts
sklearn.ensemble	RandomForestClassifier	To use random forest classifier for model implementation
sklearn.linear_model	LogisticRegression	To use Logistic regression for model implementation
sklearn.neighbours	KNeighborsClassifier	To use Logistic regression for model implementation
Xgboost	XGBClassifier	To use XGBoost classifier for

		model implementation
sklearn.pipeline	pipeline	
sklearn.model_selection	GridSearchCV RandomizedSearchCv kfold	In order to tune the hyperparameter of the models.
sklearn.metrics	accuracy_score	To check the accuracy of the model
sklearn.utils	class_weight	To choose if the dataset is balance balance out the dataset

1.5. Limitations of study:

One significant flaw in the study was the lack of computational resources to carry out machine learning algorithm's training, one repetition requires a significant amount of time and computational resources. This approach took longer to train the model since we do not have computational resources and a good platform to conduct training and testing analysis, which limited how thoroughly we would complete the research.

Chapter 2 - Literature Review

- Fraud Detection System (FDS):

A fraud detection system is a system that tries to identify suspicious activities while they go through the main system (Aisha Abdallah, 2016). Previously this process has been done manually and going through a sample of real fraud data to detect and identify these activities. The operation has been time-consuming and prone to human error, misinterpretation and overlooking of some of the details. Hence the evolution of fraud detection systems to automate the process and remove the human element from the operation level of the system, but a lot of the data mining methods were missing previously, and they are much more enhanced and effective nowadays to come up with better results and findings for an effective fraud detection system.

- Supervised learning methods for credit card fraud detection:

Many financial institutions are looking to limit their losses in credit card fraudulent transactions. Since it's one of the most attractive research topics in the financial world many methods have been tried and tested in this domain from supervised learning to ensemble learning (Johannes Jurgovskya, 2018). Supervised learning has been found as the ideal way to detect fraud as it takes the datasets and uses its class attribute to find out and distinguish the two classes of the training data set (fraud, not fraud “genuine”). (E.W.T. Ngai, 2011) ran a systematic review of many journal articles and have found out that out of all the supervised methods running Decision trees, support vector machines, neural networks, and logistic regression have been found the widely used methods in coming up with a model for the fraud detection system.

- Experimental assessment of a similar data sets:

In an article by (Bart Baesens, 2021) the data set was divided into 30% and 70%, as in 70% of the data have been selected as a training set, and 30% of the data have been selected as the test set for a total of 31,763 records and 14 attributed. To experiment with their data set they have used many classification methods like logistic regression CART algorithms, decision tree, and many more algorithms. Many justifications have been given for each of these algorithms and why they have been taken into account. Logistic regression is very popular in establishing models for its speed and low cost of computation power, and that decision tree could give a better understanding of the decision process in understanding more about how the fraud was committed.

- Insight into credit card fraud types:

Like any type and number of criminal activities, credit card fraud has more than one type. (Siddhartha Bhattacharyya, 2011) have described and broken down credit card fraud into two types: application and behavior fraud. Describing application fraud is the act of getting new cards from the issuing credit card companies without using their own personal information or using fake information to obtain and issue new cards. Behavior fraud can be broken into four subtypes: mail theft, counterfeited cards, stolen/lost cards, and 'cardholder not present fraud. In which the information of the real holder of the credit card has been stolen or obtained in an illegal way. While this information is used to run 'cardholder not present' types of transactions like through the internet and phone. Another issue in credit card fraud is the cruciality of time in detecting the fraud because the faster the fraud is detected the greater the avoidable loss.

- Classification of the existing financial fraud detection systems:

(Jarrod West, 2016) have compared in his article many research papers about the fraud detection system and the different models that have been used in the detection of fraud and the accuracy of each model. Also previously in the same paper, a comprehensive discerption of each and every model has been conducted. Moreover, a comparison of the different types of fraud investigation types versus what type of methods have been used in the recent research and papers. For example, the most used method or algorithm for credit card fraud is the support vector machines, decision trees, hybrid methods, and artificial immune systems.

Bruns et al. (2019) suggested In an effort to learn Complex Event Processing rules to extract relevant information from large-scale data streams, an evolutionary algorithm-based model was proposed, taking heuristics into account when determining the process's ideal parameters. The empirical validation of this model utilized real-world transportation data, allowing assessment of the approach's benefits and limitations in the event that it were to be applied to real-world decision-making. Eweoya et al. (2019) employed decision trees to predict frauds in bank loan administration and subsequently lower losses from loan defaults for a fraud detection application in finance utilizing real-world data from a financial institution.

- Dealing with a large dataset:

In the following articles (Alejandro Correa Bahnsen, 2016) the author describes how he dealt with a huge number of records. The dataset at hand was 100+ million records with as many as 27 attributes while one of these attributes is the class attribute that labels out the fraud transaction. And the count of the fraud-labeled records was only 40,000 transactions. Which account for only 0.025% of the whole dataset. Running a training model on such a huge number of records would prove difficult for the running application. Not to mention very hard to describe the findings with such a small number of fraud within the whole dataset. The proposed solution was to cut down the data set to come up with only 236,735 transactions but with a higher fraud ratio of 1.5%. From this new dataset three datasets have been extracted. The training dataset with 50% of the dataset, the validation dataset with 25% of the dataset, and finally the testing dataset with 25% of the dataset. Mostly the validation dataset will be used to modify the model and tune and enhance the model to test it back again.

- Running hybrid approach for detecting credit card fraud:

Most of the papers and articles regarding the detection of fraud mostly use supervised techniques to detect fraudulent activities. The results are promising to keep using the supervised but a major concern is that the world changes and customer behavior mostly changes with it and the supervised techniques don't account for that difference. (Fabrizio Carcillo, 2021) have tried in their paper running hybrid approach to teaching the model fraud detection. Their findings were that the results are not that conclusive and convincing in any terms that would make someone implement this approach in the current days, at least not the classification methods. The clustering methods showed some promise, but the result needs additional work before knowing for sure if these results are real or only overfitting to the training set of the data they used. (Stijn Viaene, 2007) found also that using hybrid methods was the most effective in detecting insurance fraud.

- FDS in the real world and layers of controls used.

Understanding how the real world works in processing the transaction from the user and running it through the system is crucial in the making of a fraud detection system. (Andrea Dal Pozzolo G. B., 2018) describe the working system of one of their industrial partners. The system is divided into automated tools and finally, it goes to the investigators both combined to make up their fraud detection system. 1) Terminal: which is the basic security checks of the banks like the PIN code if the card is blocked or not, how many attempts have been made and is the transaction over the limit of the card or not. 2) Transaction-Blocking Rules: are 'if' statements that block the transactions that are clearly a fraud. Like if the customer attempts to run the card online on an untrusted website this layer will deny the transaction to avoid an instance of fraud. 3) Scoring Rules: which at the first glance looks like the previous layer but it's not. These layer statements are made by experts and investigators that run the transaction through many statements gaining a score and finally if, for example, it had a score over 0.9 then that system flags it out as a fraud. 4) Data-Driven Model: from the name it runs the transaction that has been authorized through the model with a classifier to train the model for a better detection of the fraud attempts and vice versa, if a transaction has been found out to be fraud then it runs through the system too with a classifier of fraud also to train the system 5) Investigators: they are responsible for the alerts of the scoring rules layer. They make up the statements and they can change the score of the statements. They determine if the correspondences of fraud are legit or false positive in this case.

- Requirement of process mining to enhance the system:

Process mining aims to analyze the process through the event logs. For this to take place and for process mining to work first every event needs to be connected to an activity for a particular case or instance. If possible, every event needs to be connected to a unique entity that generates these events. If the above criteria are met then process mining can be used for that process (Mieke Jans, 2011).

- Unbalanced sample problem in designing fraud detection systems

One of the main problems in designing a fraud detection system is having unbalanced datasets. To deal with this problem by using sampling techniques to rebalance the datasets and remove the noise between the different classes of the dataset (Andrea Dal Pozzolo O. C.-A., 2014). This solution also has its own setbacks and that introduces a new problem. Because these techniques don't take into account the effect of removing or adding instances into a specific class. It might also introduce under-sampling or oversampling of a certain class. For a better result using ensemble methods that combine the best of both worlds through balancing the samples and the classifier of both the minority and majority classes within the dataset (X. Liu, 2009).

- Ensemble learning approach

The ensembled learning approach seeks to combine different algorithms to reach better accuracy and better performance when compared to the individual algorithms (Javad Forough, 2021). The recent papers regarding fraud detection have shown that using the ensembles approach resulted in better performance and reduction in latency, better performance in the presence of noise, and better handling of the datasets in the case of class imbalance.

- Deep learning approach

Using a deep learning approach is rare in the case of a fraud detection system. (Eunji Kim, 2019) have tested out the deep learning approach in a competitive environment using a model called 'champion-challenger' model to test out a deep learning approach against one of the most popular and well-performing approaches: the ensemble learning approach. The results found out the deep learning approach was the winner. Using hundreds of neurons to deal with much more complex patterns shows better recall performance and a higher cost reduction rate compared to the ensemble approach.

- Fraud detection using Hidden Markov Model:

Hidden Markov Model (HMM) is a stochastic statistical model that observes a process that is influenced by another a certain outcome. (Abhinav Srivastava, 2008) took that model and applied it to the credit card transaction process to find out the fraudulent transaction. Applying such a model has been difficult but after applying it they found out that their model is 80 percent accurate, but one of the main points is that their system is very specific for the data they had and can't be easily adjusted to any system.

- Detection of Crimes Reason in Country:

It can be found in so many research papers that machine learning algorithms are also used in the detection of the overall crimes in the country and as well as the identification of the reasons behind these crimes. As an example, we can see that some research works related to the UK's crime reasons identification and analysis.

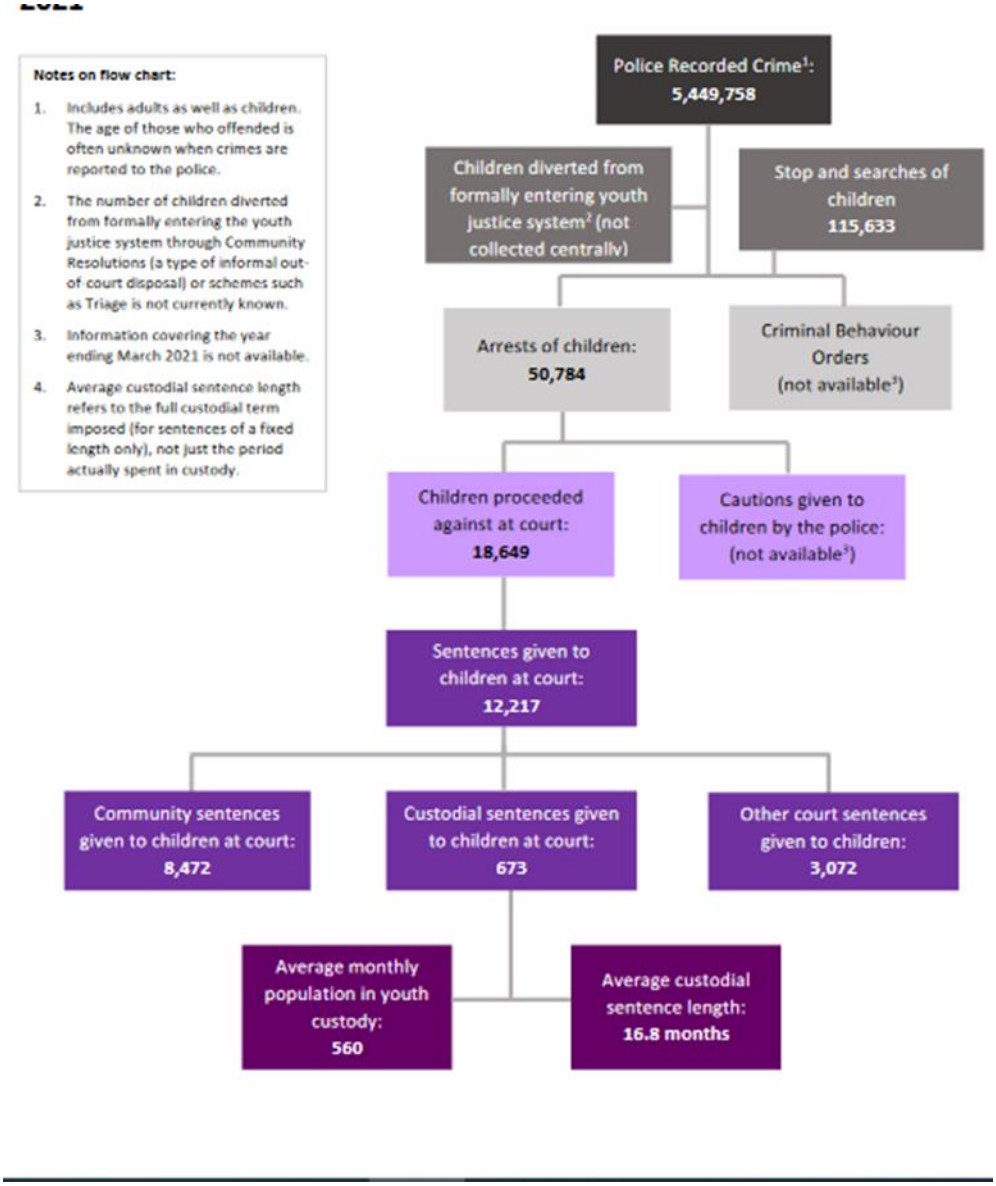


Figure 1 Research Outcomes For Crimes Detection

There have been studies found which have used crime data provided by police as primary data of the research and has been proved to be more authentic than any other self-collected primary data.

Crimes can be predicted by using different methods such as criminals' activities, they remain active and operate in their comfort zone. In one of the research projects in 2018, researchers used crimes reported by police easily available from the website of the UK police department. They have used various visualization techniques along with machine learning algorithms to predict crime distribution over an area. In this research they have tried to find out patterns through which a crime can be predicted and prevented (Hitesh Kumar Reddy Toppi, Bhavna, & Ginika, 2018).

- Machine learning and data science in healthcare:

For healthcare, machine learning has so many applications. It is a complete industry, and we can see the most common and famous use of this particular in the era of COVID 19 detection and prevention detection mechanisms.

Similarly, there are so many studies which work on particular diseases like Malaria, Heart failure and so on for its complete deep learning and deep analysis. Wiens, J., & Shenoy, E. S. (2018) and Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020) and so many other researched highlighted the applications as well as the usage of this and explain that machine learning algorithms are used for analyzing the healthcare data analytics. Once machines are trained, there is no need to apply again and again the same tests for models.

- Heart Failure Reasons detection using Machine Learning:

Detection of heart failure reasons is also one of the applications of machine learning and data science. We have found many relevant papers to support this research. The purpose of this study is to use big data analytics to identify the causes of heart failure. Many studies have been conducted recently that show big data analytics has assisted the improvement of health. Heart disease is becoming more and more fatal, day by day. With the passing of each year and the availability of enormous amounts of data, there have been studies where researchers have used data mining techniques to identify, anticipate the cause, and find a remedy. In one study, researchers hypothesized that data mining is used because of the abundance of data and that data might be advantageous if turned into valuable information. (Niraj & R, Predictive Analysis on Heart Disease Using Different Machine Learning Techniques, 2019).

They found data mining to be effective when analyzing a huge amount of data. These predicting techniques are beneficial for practitioners in analyzing. They used a dataset of heart patients available on UCI machine learning datasets. They used several machine learning algorithms like K-Nearest Neighbor, Random Forest, SVM, Decision Trees, Adaptive boosting and Logistic Regression. Their result found after applying these algorithms on data may help physicians to estimate the risk of heart failure in different age groups.

- Machine Learning applications in Retail:

Machine learning and data sciences has applications in almost every field. Similarly we can see in the study of (Akhare, Rakhi, et al, 2021 and Silva, E. S., Hassani, H., & Madsen, D. Ø. (2019)) as well as the real life examples of use of data analytics tools and technologies in retail industry to manage the records of daily sales and purchases as well as the dashboards/tool and technologies to predict the sales and profit with the passage of time

For example we can see the applications in Retail like the statuses of Recovery, Inventory etc. and the products which are going through complete life cycle on daily basis.

- Detection of Cyber Bullying:

Another very interesting case study could be found in the research which is related to the detection of cyber bullying using the machine learning algorithms. Authors (Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020) and Hani, J., Mohamed, N., Ahmed, M., Emad, Z., Amer, E., & Ammar, M. (2019)) has used various techniques to extract the data from the social media site like Twitter and then perform complete cycle which includes the following phases: Dataset Origin Identification, Labeling the data, Develop the features of Input, Learning the Model, Class Weighting and then Evaluation. The authors has run the algorithms for different set of repetitions to analyze the accuracy of the model.

Key Takeaways

- Many ways can be used to achieve the same result and in Data Analysis, but it's all about efficiency
- it's possible to divide your dataset into 3 subset Training – Validation - Testing
- Ensemble learning methods & deep learning yielded the most result in the past
- To achieve a specific result for a non-changeable variable it is better to use supervised learning methods

Chapter 3 - Project description

3.1. Overview:

Dataset discovery, data preprocessing, data exploration, testing and modeling, and outcomes are some of the stages that this study will go through. The chosen dataset will be created for usage in the project's various modeling and testing phases. The initial stage of the project will be the preprocessing phases of data cleaning and normalization, which involve dealing with redundant and missing data and making sure the data that is present in the dataset adheres to integrity. The data will be visualized using a variety of tools to produce a wide range of visuals that will help illustrate the relationships between the different features as part of a thorough analysis of the information. Our proposed solution will be validated using a range of techniques, including testing and modeling, incorporating model comparison and several estimations for a range of characteristics, such as accuracy, computational expense, and processing time.

The dataset contains fellows-specific features.

Feature Name	Description	Type
Month	Months in which the accident occured	Object
WeekOfMonth	The week in the month the accident occured	Object
DayOfWeek	Are these the days of the week the accident occured on?	Object
Make	The car model manufacturers	Object
AccidentArea	Classifies area of accident rural or urban	Object
DayOfWeekClaimed	Contains the day of the week the claim was filed	Object
MonthClaimed	Contains the day of the month the claim was filed	Object
WeekOfMonthClaimed	contains weeks in the month that the claimed in field	int64
Sex	Gender of making individual claim	Object
MaritalStatus	Marital status of individual claim	Object
Age	Ages of individual making claim	int64
PoliceReportFiled	Indicates whether a police report was filed for the accident	Object

Fault	Categorization of who was deemed at fault.	object
BasePolicy	Type of insurance coverage	Object
NumberOfCars Year	Number of car involved in accidents	Object
AddressChange_Claim	Time from claim was filed to when the person moved.	Object
NumberOfSuppliments	Not sure what supplement is insurance	Object
WitnessPresent	Indicate whether a witness is present	Object
AgentType	Classify the agent who is handling the claim	Object
AgeOfPolicyHolder	Each value is a range of ages	Object
VehicleCategory	Categorization of vehicle	Object
PolicyType	1. Type of insurance 2. Category of vehicle	Object
VehiclePrice	Ranges for the value of vehicle	Object
Deductible	The ductile amount	int64
AgeOfVehicle	Age of vehicle at the time of accident	int64
PastNumberOfClaims	Previous number of claims	Object
Days_Policy_Claim	Number of days the purchased and claimed was filed	Object
RepNumber	Rep number	int64
Days_Policy_Accident	The number of days between the accident occured	Object

DriverRating	Driver rating	Object
FraudFound_P	Indicate whether a claim is fraudulent	int64
PolicyNumber	The masked policy number	int64

3.2. Data Collection

The data is taken originally from kaggle.com which is an open-source website that has many data sets, but after some search it turns out that the dataset have been published but oracle. Even after knowing that we failed to reach the exact release link or origin of the dataset, but we found some related links from oracle that contained the data and it described more about who collected that data. The dataset is collected by Angoss Knowledge Seeker software from January 1994 to December 1996. The data has 33 attributes and ultimately the classification of whether this is considered fraud or not as a class attribute. It contains 15,420 records of policy claims and 33 features. Some variables may be utilized to simply identify the individuals or entities that produced or received the claim since they are not relevant for running the models and algorithms, or they may be changed through data integration.

Links:

- **https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud detection?select=fraud_oracle.csv**
- **https://github.com/AnalyticsandDataOracleUserCommunity**
- **https://blogs.oracle.com/machinelearning/post/a-two-step-process-for-detecting-fraud-using-oracle-machine-learning**

Chapter 4 - Project Analysis

4.1. Dataset overview:

The first important step is to collect and collect data. After formulating the business problem, it is important to understand the data sources. The data collected in this phase is raw data because it is collected maybe from different means and systems, so it is not organized as such in this phase. The set of data we collected from Kaggle is called binary data and ultimately classifies whether it is considered a scam or not. It has 15,420 insurance records and 33 functions. Features included in the following are:

'Month', 'WeekOfMonth', 'DayOfWeek', 'Make', 'AccidentArea', 'DayOfWeekClaimed', 'MonthClaimed', 'WeekOfMonthClaimed', 'Sex', 'MaritalStatus', 'Age', 'Fault', 'PolicyType', 'VehicleCategory', 'VehiclePrice', 'FraudFound_P', 'PolicyNumber', 'RepNumber', 'Deductible', 'DriverRating', 'Days_Policy_Accident', 'Days_Policy_Claim', 'PastNumberOfClaims', 'AgeOfVehicle', 'AgeOfPolicyHolder', 'PoliceReportFiled', 'WitnessPresent', 'AgentType', 'NumberOfSuppliments', 'AddressChange_Claim', 'NumberOfCars', 'Year', 'BasePolicy'.

4.2. Data Pre-processing:

Data pre-processing in machine learning can be an important step that can make a difference in improving the quality of information to facilitate the extraction of meaningful knowledge from the information. Data preprocessing in machine learning refers to the method of preparing (cleaning and organizing) raw data to make it suitable for building and training machine learning models. In simple terms, machine learning data processing can be an information mining technique that transforms rough information into justifiable and lucid organization. After collecting the raw data, it is time to organize it so that it can be used for further processing.

4.3. Data Cleaning:

It is important that the data set is free of defects that could prevent testing or, more seriously, lead to insufficient analysis. These deficiencies or problems caused by redundant records, missing values, or loss of dimension must be effectively resolved. So, in this step bad data will be removed, and missing data will be added. The information we currently have is a comprehensive general information from which we need to remove unnecessary information and perhaps add the missing information.

```
Month                 0
WeekOfMonth           0
DayOfWeek             0
Make                  0
AccidentArea          0
DayOfWeekClaimed      0
MonthClaimed          0
WeekOfMonthClaimed    0
Sex                   0
MaritalStatus         0
Age                   0
Fault                 0
PolicyType            0
VehicleCategory       0
VehiclePrice          0
FraudFound_P          0
PolicyNumber          0
RepNumber             0
Deductible            0
DriverRating          0
Days_Policy_Accident  0
Days_Policy_Claim     0
PastNumberOfClaims    0
AgeOfVehicle          0
AgeOfPolicyHolder     0
PoliceReportFiled     0
WitnessPresent        0
AgentType             0
NumberOfSuppliments   0
AddressChange_Claim   0
NumberOfCars          0
Year                  0
BasePolicy            0
```

Figure 2. Sum Of Null Values In The Dataset

As we can see, there are no missing values in the data set, so handling null values is a very important step in the data cleaning process. Because it helps to deal with the problems that appear during the subsequent procedures. There are several ways to handle null values and missing values. We can remove all records from the data set or impute missing values using mean, median, or regression methods.

4.4. Exploratory Data Analysis:

The data contains a lot of information that needs to be discovered first in order to better understand and investigate the information and by visualizing the data we can get a better sense and information about the data**.**

```
Data columns (total 33 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Month                 15420 non-null  object
 1   WeekOfMonth           15420 non-null  int64
 2   DayOfWeek             15420 non-null  object
 3   Make                  15420 non-null  object
 4   AccidentArea          15420 non-null  object
 5   DayOfWeekClaimed      15420 non-null  object
 6   MonthClaimed          15420 non-null  object
 7   WeekOfMonthClaimed    15420 non-null  int64
 8   Sex                   15420 non-null  object
 9   MaritalStatus         15420 non-null  object
 10  Age                   15420 non-null  int64
 11  Fault                 15420 non-null  object
 12  PolicyType            15420 non-null  object
 13  VehicleCategory       15420 non-null  object
 14  VehiclePrice          15420 non-null  object
 15  FraudFound_P          15420 non-null  int64
 16  PolicyNumber          15420 non-null  int64
 17  RepNumber             15420 non-null  int64
 18  Deductible            15420 non-null  int64
 19  DriverRating          15420 non-null  int64
 20  Days_Policy_Accident  15420 non-null  object
 21  Days_Policy_Claim     15420 non-null  object
 22  PastNumberOfClaims    15420 non-null  object
 23  AgeOfVehicle          15420 non-null  object
 24  AgeOfPolicyHolder     15420 non-null  object
 25  PoliceReportFiled     15420 non-null  object
 26  WitnessPresent        15420 non-null  object
 27  AgentType             15420 non-null  object
 28  NumberOfSuppliments   15420 non-null  object
 29  AddressChange_Claim   15420 non-null  object
 30  NumberOfCars          15420 non-null  object
 31  Year                  15420 non-null  int64
 32  BasePolicy            15420 non-null  object
dtypes: int64(9), object(24)
```

Figure 3. Information About The Data

The insurance data is made up with 33 features out of 33 features only 9 are numerical features and remaining all are categorical features.

4.5. Data Transformation:

In this phase data will be organized or managed so that it will be helpful to achieve the required goal.

	count	mean	std	min	25%	50%	75%	max
WeekOfMonth	15420.000000	2.788586	1.287585	1.000000	2.000000	3.000000	4.000000	5.000000
WeekOfMonthClaimed	15420.000000	2.693969	1.259115	1.000000	2.000000	3.000000	4.000000	5.000000
Age	15420.000000	39.855707	13.492377	0.000000	31.000000	38.000000	48.000000	80.000000
FraudFound_P	15420.000000	0.059857	0.237230	0.000000	0.000000	0.000000	0.000000	1.000000
PolicyNumber	15420.000000	7710.500000	4451.514911	1.000000	3855.750000	7710.500000	11565.250000	15420.000000
RepNumber	15420.000000	8.483268	4.599948	1.000000	5.000000	8.000000	12.000000	16.000000
Deductible	15420.000000	407.704280	43.950998	300.000000	400.000000	400.000000	400.000000	700.000000
DriverRating	15420.000000	2.487808	1.119453	1.000000	1.000000	2.000000	3.000000	4.000000
Year	15420.000000	1994.866472	0.803313	1994.000000	1994.000000	1995.000000	1996.000000	1996.000000

Figure 4. Descriptive Statistics Summary Of The Dataset.

Figure 4 shows the entire summary of the descriptive statistics of the data set. We can clearly see that in the data set we have 15420 sets of data that the weekofmonth shows the week of the month when the accident occurred. The average number of weeks that occurred in a month is two and the maximum number of weeks in a month is five, similarly with the weekofmonth claimed contains weeks in the month that the claimed in field the mean of the weekofmonth is two and the max is 5. Age is the ages of individuals that make claims the average age of individual is 40 while the max age of 80. Column FraudoundP indicating whether the claim was fraudulent or not i.e. 1 or 0.

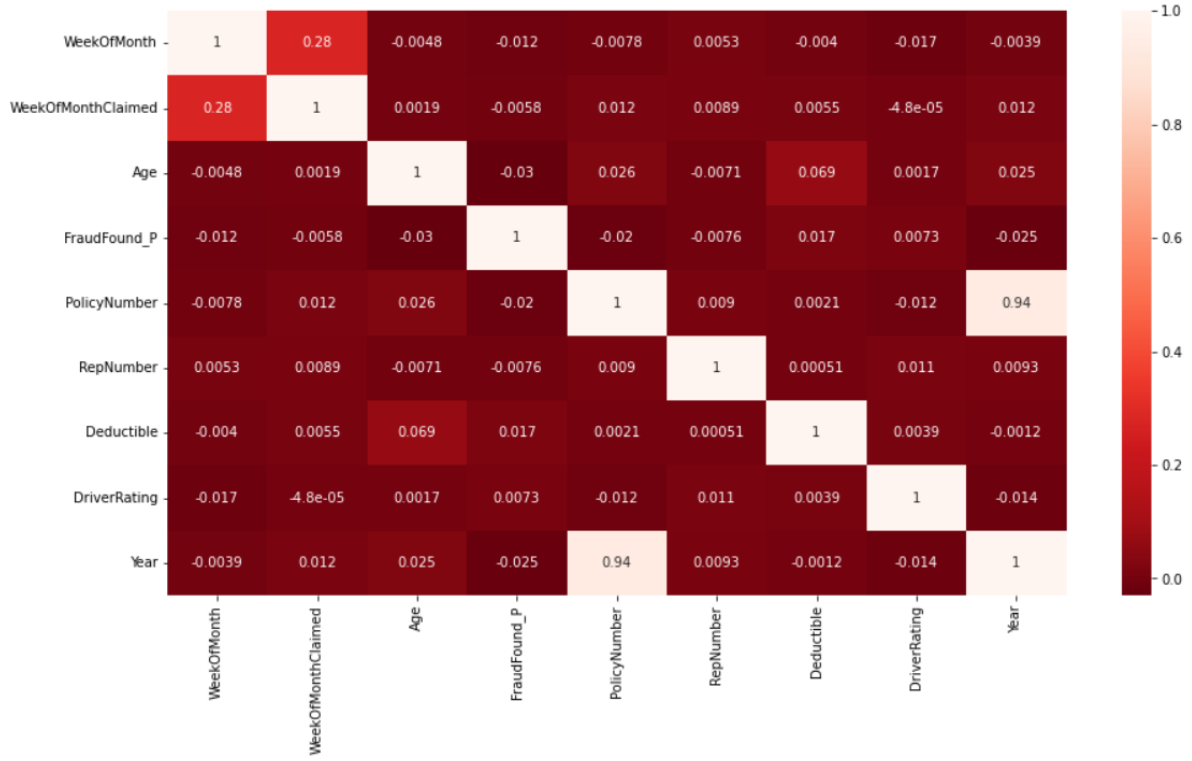


Figure 5. Relation Between Independent And Dependent Variables.

We need to predict the output using supervised machine learning techniques. sometimes when we have small amount of data, we can easily work with it, but as the amount of data increases, it becomes difficult to find predictors or variables. In this situation, using all the data can often be detrimental, which affects not only the accuracy of the model, but also the computational resources when we use all the data. This is where the concept of correlation comes in as we explore the relationship between dependents and independents features, then select the features that are important for prediction. As shown in Figure 5, we can see the relationship between each feature and how they correlate with each other.

4.6. Data Exploration:

Using Pandas data frames, the required data is extracted from the initial loaded data. Data is completely examined and discovers information and at the end concludes it to generate the report.

The data is displayed using line graphs for the user to analyze using matplotlib and seaborn. For simpler comparison, all of the line graphs are displayed in one graph at the end. Prior to being combined into one large line graph for analysis and comparison with the same date, the data is first displayed on distinct graphs to demonstrate trends in the various dataset values obtained using data visualization.

	Month	WeekOfMonth	DayOfWeek	Make	AccidentArea	DayOfWeekClaimed	MonthClaimed	WeekOfMonthClaimed	Sex	MaritalStatus	Age	Fault
0	Dec	5	Wednesday	Honda	Urban	Tuesday	Jan	1	Female	Single	21	Policy Holder
1	Jan	3	Wednesday	Honda	Urban	Monday	Jan	4	Male	Single	34	Policy Holder
2	Oct	5	Friday	Honda	Urban	Thursday	Nov	2	Male	Married	47	Policy Holder
3	Jun	2	Saturday	Toyota	Rural	Friday	Jul	1	Male	Married	65	Third Party
4	Jan	5	Monday	Honda	Urban	Tuesday	Feb	2	Female	Single	27	Third Party
5	Oct	4	Friday	Honda	Urban	Wednesday	Nov	1	Male	Single	20	Third Party
6	Feb	1	Saturday	Honda	Urban	Monday	Feb	3	Male	Married	36	Third Party
7	Nov	1	Friday	Honda	Urban	Tuesday	Mar	4	Male	Single	0	Policy Holder
8	Dec	4	Saturday	Honda	Urban	Wednesday	Dec	5	Male	Single	30	Policy Holder
9	Apr	3	Tuesday	Ford	Urban	Wednesday	Apr	3	Male	Married	42	Policy Holder

Figure 6. Exploration Of Data To Check Categorical Data

In order to build a predictive model by using machine learning it is important to have all the input and output variables in numeric format not in categorical format as we know machines only understand numeric values, so we have to convert all the categorical variables into numeric to fit and access the model.

	Month	WeekOfMonth	DayOfWeek	Make	AccidentArea	DayOfWeekClaimed	MonthClaimed	WeekOfMonthClaimed	Sex	MaritalStatus	Age	Fault	Policy
15410	9	3	3	3	1	6	10	4	0	2	31	1	
15411	9	4	5	6	0	7	10	5	1	1	42	1	
15412	9	4	5	13	1	7	10	4	0	2	28	0	
15413	9	4	4	9	1	2	10	4	1	1	40	0	
15414	9	4	0	2	1	2	10	4	1	2	58	1	
15415	9	4	0	17	1	6	10	5	1	1	35	0	
15416	9	5	4	13	1	1	3	1	1	1	30	0	
15417	9	5	4	17	0	1	3	1	1	2	24	0	
15418	2	1	1	17	1	5	3	2	0	1	34	1	
15419	2	2	6	17	1	5	3	3	1	2	21	0	

Figure 7. Conversion Of Categorical Values Into Numeric

As in figure 7 we can see that all the categorical variables have been converted into numeric values in order to build a classification model.

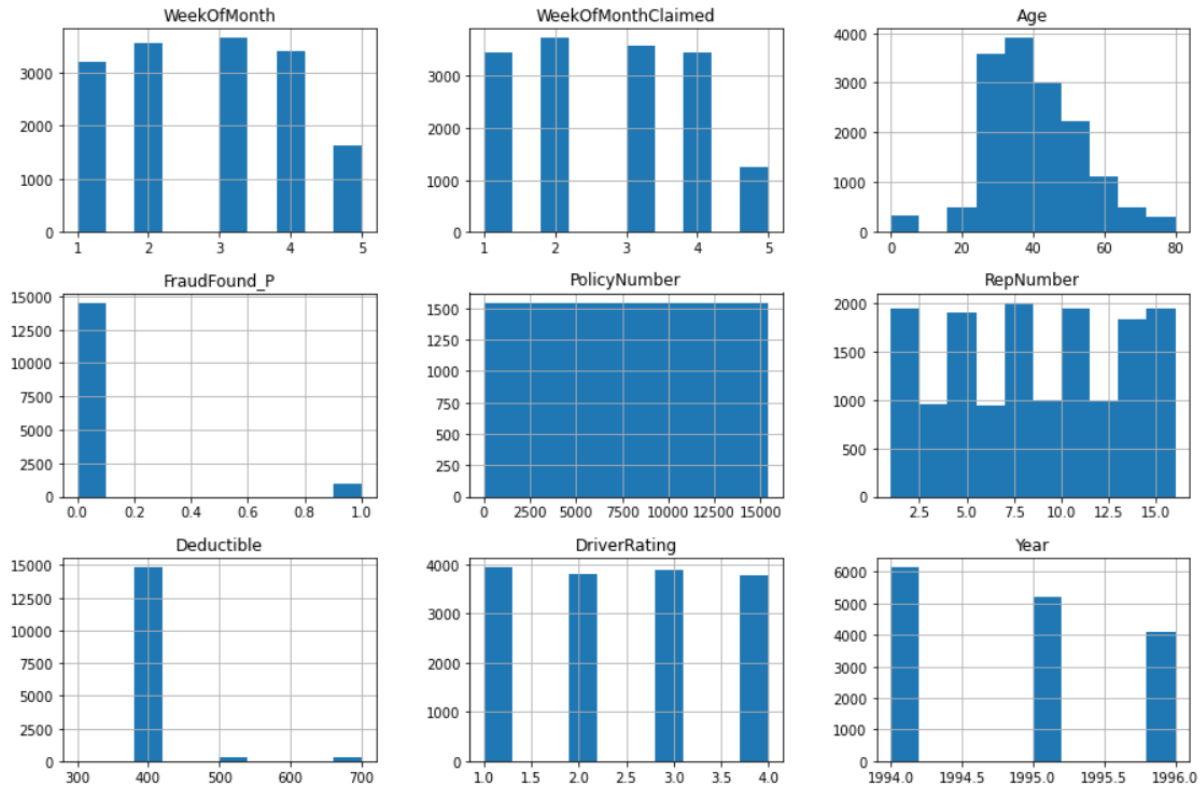


Figure 8. Visualization Of Categorical Columns

Figure 8 showing the pairwise relationships between the categorical features that are present in the dataset.

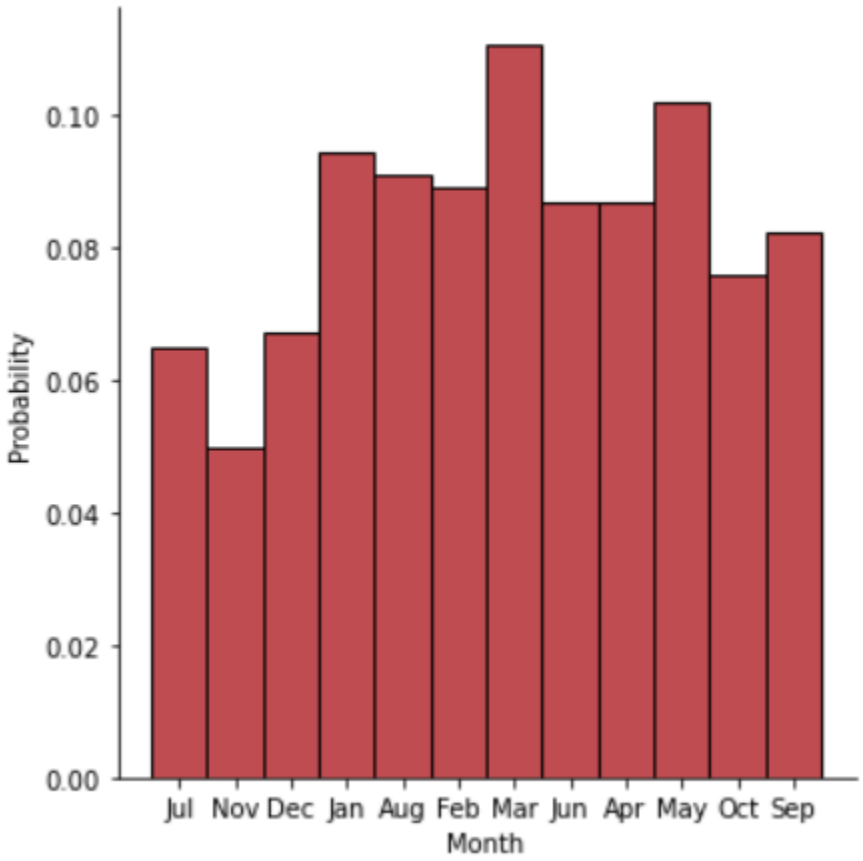


Figure 9. Fraudulent Cases Probability

In figure 9 we can clearly see that Amongst fraudulent cases months of march and may have higher probability.

Figure 10. Distribution Of Fraudulent Claims.

The above figure 10. indicates whether the claim was fraudulent (1) or not (0) so we can clearly see that 94% are fair and only 6% are fraudulent claims.

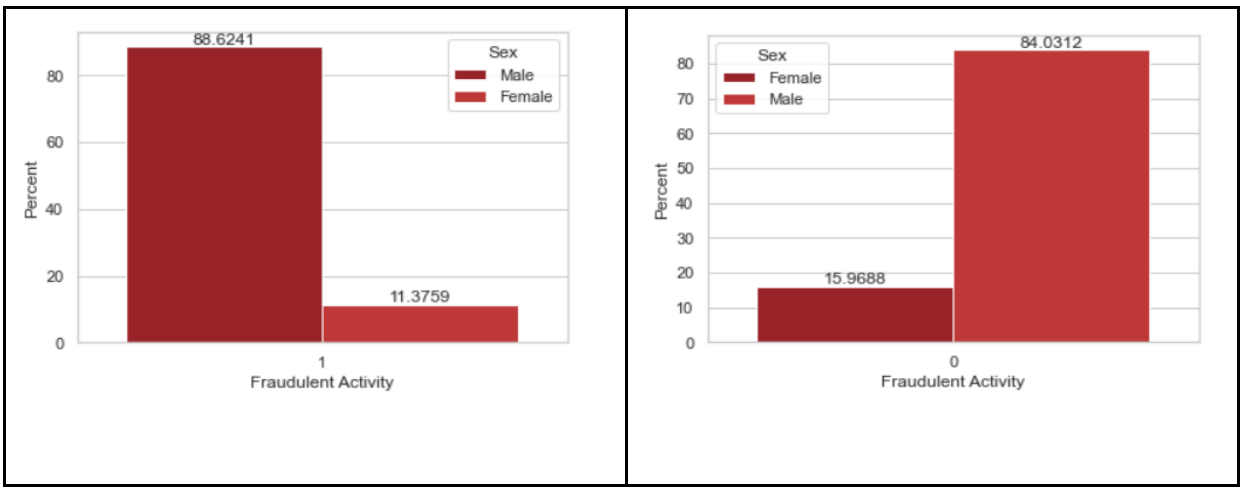


Figure 11. Gender Involved In Fraudulent Activity

Figure 11 shows the fraudulent claims according to gender, and it is clear that while men make up 84.03% of the total claims for legitimate transactions, they make up 88.62% of the fraudulent ones. Males are more prone than females to file false claims.

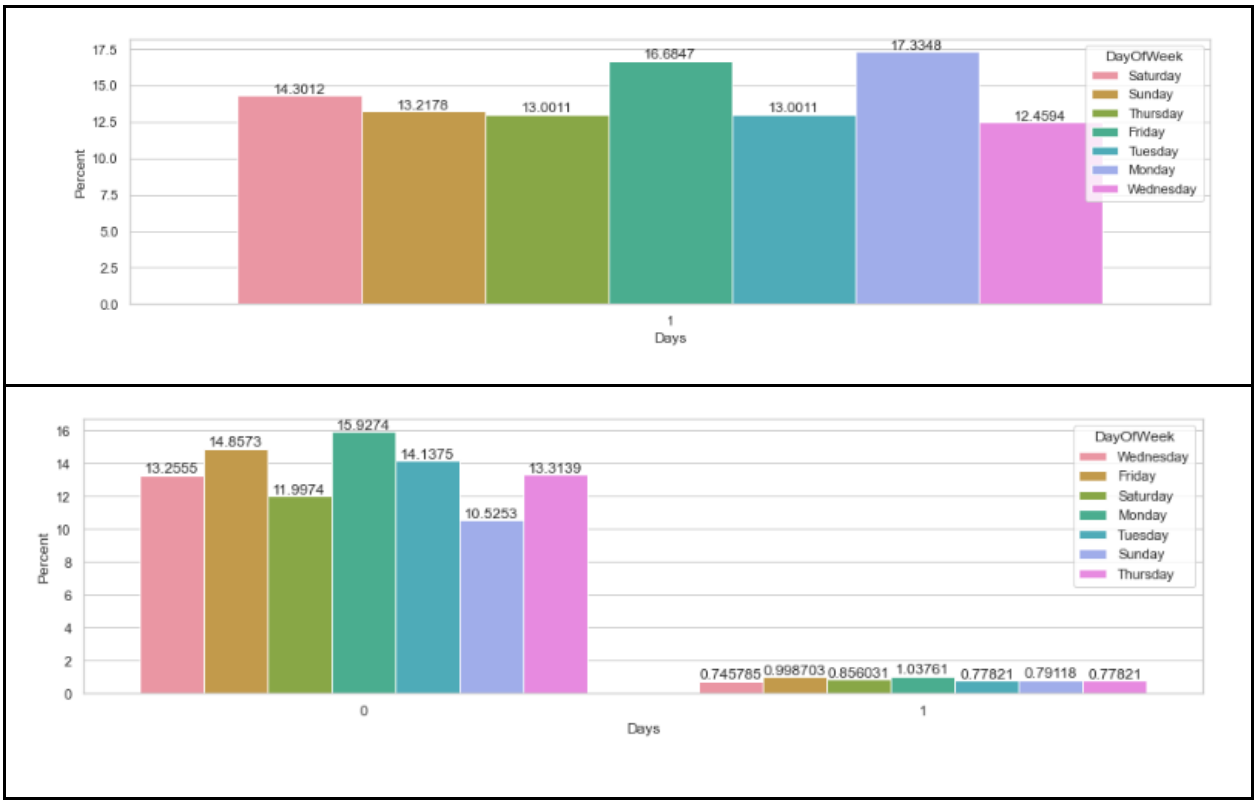


Figure 12. Fraudulent Activity On Days Of Week

The fraudulent activity depicted in figure 12 above is based on weeks and days. The accompanying image reveals that Monday and Friday have the highest percentage of fraudulent actions. Similar to this, Monday and Friday also have the most claims.

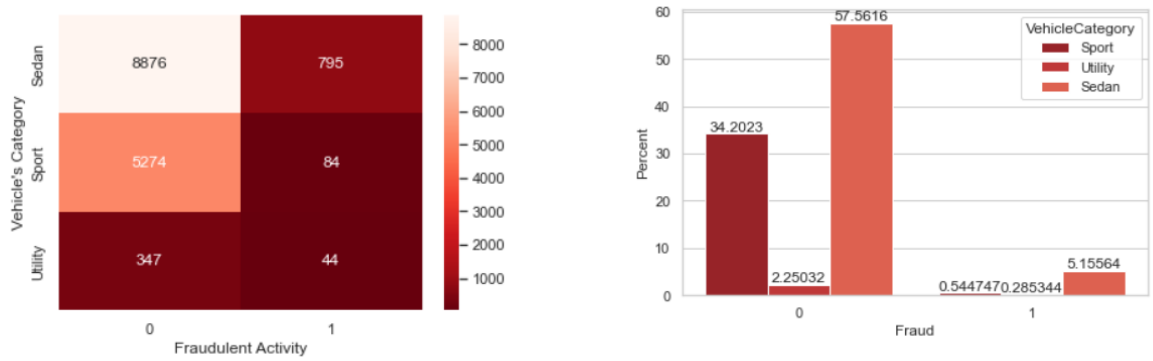


Figure 13. Car Categories With Respect To Fraudulent Claims.

The fact that sports cars are less likely to be used in fraud is one item that stands out. Sedans are the primary source of claims. They are also the most motivated, though. Sports cars are just 0.02% likely to be involved in false claims, whereas utility cars are generally 0.11% likely to have fraudulent claims. Utility vehicles have higher expectations of being involved in fraudulent operations, according to the correlation matrix above.

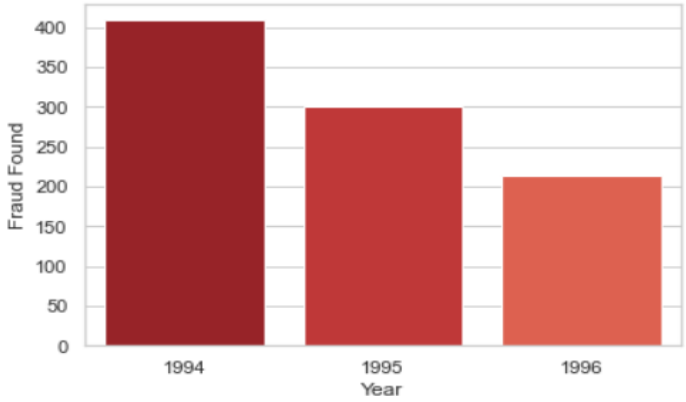


Figure 14. Year-Wise Fraudulent Claims

Inferred from the following graph is that the majority of fraud was committed in 1994. We can observe the qualities of the data and the type of correlation that exists between them. It is the quickest technique to determine whether the features match the output. However, when working on machine learning projects, we typically overlook the two components known as data and mathematics. We know that ML is a data-driven methodology, and our ML model will only deliver outcomes that are as excellent or as awful as the data we feed it.

4.7. Build & training model:

Once a pattern has been identified, a suitable model needs to be constructed. A model is created by studying, practicing, and then using it. The model will be put to use and produce fraud detection. We went through these 6 processes to create and train the model:

- Contextualize machine learning in your organization
- Explore the data and choose the type of algorithm
- Prepare and clean the dataset
- Split the prepared dataset and perform cross validation
- Perform machine learning optimization
- Deploy the model

The generic flow of machine learning data model is presented below:

An integrated research approach will be applied for this project. To explain the findings and conclusions of the paper and the various results, the project will employ some explanatory research methods in addition to experimental research methods, some qualitative research methods, and some quantitative research methods.

We will start by determining what is crucial for managing the data in accordance with the business that may use the model or the solutions. In this situation, an insurance provider will probably prioritize the financial aspects of each claim while giving no consideration to personal information when developing the model.

The report will detail the available data, the many qualities, and how each attribute relates to determining whether or not this claim is fraudulent. What are the different forms of data that are available, and can the existing data be improved or changed without affecting the outcome of the end goal, which is identifying dubious claims? To do this, data cleaning is necessary. Some values or characteristics may need to be removed, or new values may need to be created by merging existing ones and using data integration techniques. Like how we presented all of the ideas we discovered from data analysis in the exploratory analysis.

Once the data have been cleaned and thoroughly understood, following exploratory data analysis We also dealt with issues related to class inequality. Class imbalance issues dominate classification problems. It shows that the dependent or target class frequency is substantially out of balance, with one class happening far more frequently than the other. The target is therefore biased or skewed in our dataset. Because of our imbalanced target values, which are 94% fair and 6% fraudulent claims, we also have a problem with class imbalance. says that 94% of the claims are false, while only 6% are legitimate. Since this will affect the accuracy, we must first address the issue of the unbalanced class. To determine if the claim is false or not, we used a variety of supervised machine learning methods.

Random forest:

In essence, Random Forest is used for both classification and regression issues. It is an ensemble classification method and a supervised machine learning classifier. The more trees there are, the more precise the outcome would be. The RF machine learning method is simple, easy to use, and capable of achieving outstanding outcomes in most cases without the need for hyper-tuning. Over-fitting is one of the decision tree algorithm's main problems. The decision tree appears to have remembered the data. Random Forest is utilized to prevent this and is an illustration of ensemble learning in action. The use of several repetitions of one or more algorithms is referred to as "ensemble learning." A "random forest" is a collection of decision trees.

Decision tree:

The decision tree approach is also included in the supervised learning category. Regression and classification problems can be solved with DT. However, it is used in this work to overcome categorization problems. DT breaks down the input into ever-smaller bits in attempt to solve the problem, which results in the prediction of a target value (diagnosis). A decision tree (DT) consists of decision nodes and leaf nodes, each of which is linked to a class label and traits that are shown on the interior node of the tree. Though DT is pretty simply many algorithms drive from its roots one of these algorithms is called XGBoost. Which is an incredibly quick machine learning technique that uses tree-based models to try to get the best accuracy possible by making the best use of available computing power, Extreme Gradient Boosting, often known as XGBoost, becomes the obvious option.

Logistic regression:

It is a condensed version of "linear regression," a potent tool for visualizing data. The likelihood of an illness or other health concern as a result of a plausible cause is ascertained using logistic regression. The link between independent factors (X), also known as exposures or predictors, and a binary dependent (target) variable (Y), also known as the outcome or response variable, is examined using both basic and multivariate logistic regressions. It is often applied to forecast changes in the dependent variable that will be binary or multiclass.

KNN:

K-Nearest Neighbor is one of the most straightforward machine learning algorithms, based on the supervised learning approach (KNN). The model is saved, and when a new data point is given, the model searches for similarities between the data points. K-nearest neighbors are identified by calculating the distance between each data point with respect to the new data point, and based on the similarity, it produces the output. This method is also referred to as lazy learning. This indicates that new data can be reliably and quickly categorized using the K-NN approach.

4.8 Classifier Score:

The 80:20 test-train-split package in Python and its machine learning tools were used to compute classifier scores for several models to check against our suggested fraudulent detection once the unbalanced dataset was fixed:

Machine learning model	Train Accuracy	Test Accuracy
Logistic Regression	0.747790	0.754268
Random Forest	1.000000	0.997758
KNearest Neighbor	0.935590	0.897569
XGBoost	0.840224	0.840317

Table 1 Accuracy Comparison

The accuracy function in sklearn uses the following formula by default:

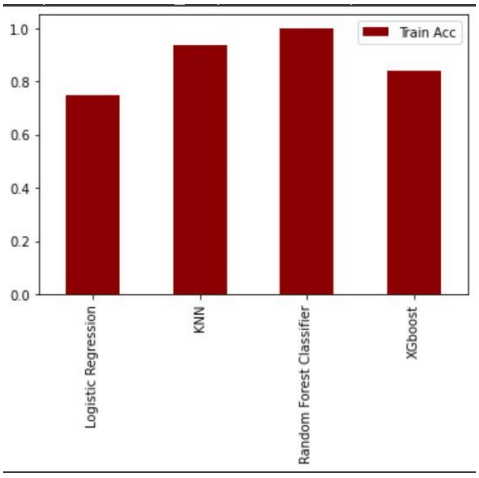
$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

	Train Acc
Logistic Regression	0.747790
KNN	0.935590
Random Forest Classifier	1.000000
XGboost	0.840224

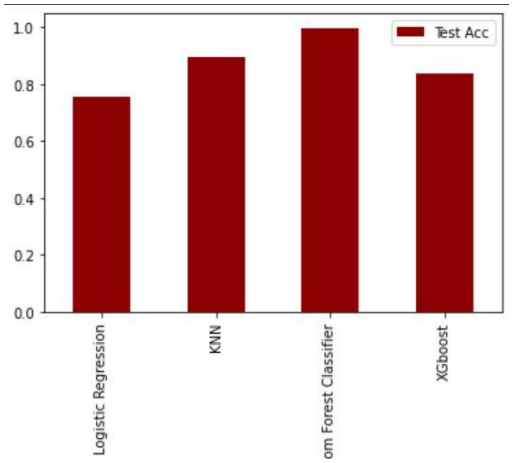
Train Accuracy Table

	Test Acc
Logistic Regression	0.754268
KNN	0.897569
Random Forest Classifier	0.997758
XGboost	0.840317

Test Accuracy Table



Train Accuracy Plot



Test Accuracy Plot

Figure 15. Train/Test-Sets Comparison

we can see in the above figure KNN , XGboost and Random Forest is working exponentially well on the dataset and achieving 93.5590% , 84.0224% and 100.0000% on train data respectively but on the other hand logistic regression is not working well here we implemented hyperparameter tuning as well in order to improve the model accuracy, but Accuracy of this model didn't much improved by tuning. So, we can say that Logistic Regression is not a reliable model for this dataset. But performance of other models is comparatively better than Logistic Regression Model.

4.9 Classification Report:

The categorization report includes a number of metrics that are crucial for assessing any model. Accuracy, precision, recall, and F1 are the included measures.Accuracy: it is the ratio of correct predictions against total observations.

- Precision: is the proportion of correctly made positive predictions to all positively observed data.
- Recall: The ratio of correctly predicted positive observations to all of the observations in a class, sometimes referred to as sensitivity.
- F1: is the average of the recall and precision scores.

The scikit-learn packages come within default parameters. The default parameters have resulted in undesired results, and so the tunning for the models have been done through tunning the hyperparameter.

Accuracy of this model did not improve much in Logistic regression by hyperparameter tuning either. So, we can say that Logistic Regression is not a reliable model for this dataset. While other algorithms like KNN, XGboost and random forest work pretty well on the dataset and have given 96%, 89% and 100% respectively. We can say that these algorithms can be used to achieve accurate results with new and huge data. By fine tuning the model the accuracy of KNN has increased from 89% to 96%. XGBoost accuracy has increased from 84% to 89%. For the other two models the Logistic Regression did not improve after the fine tuning and the Random Forest was performing exceptionally well and most likely its overfitting on the default value of the model. After fine tuning the model to get a more realistic result it got 98.5% as its best result.

The other matrices of the models are shown in the following table:

Table 2: Classification report of Models.

Metrics	**Logistic Regression**	**KNN**	**Random Forest**	**XGBoost**
Tuned Accuracy	75.03%	96.28%	98.5 %	89.0%
Precision	0.78	0.97	0.98	0.89
Recall	0.75	0.96	0.98	0.89
F1	0.76	0.96	0.96	0.88

Chapter 5- Conclusion

5.1. Conclusion:

As the different countries around the world evolve into a more economical-based one, stimulating their economy is the goal. To fight these fraudsters and money launderers was quite a complex task before the era of machine learning but thanks to machine learning and AI we are able to fight these kinds of attacks. The proposed solution can be used in insurance companies to find out if a certain insurance claim made is a fraud or not. The model was designed after testing multiple algorithms to come up with the best model that will detect if a claim is fraudulent or not. This is aimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

5.2. Recommendations:

The dataset which we used for building this insurance predictive model was taken from kaggle and it was the data between 1994 to 1996. It will be good for us if we collect the new dataset of the past 2 to 5 years To determine whether the suggested solution would perform well when compared to other datasets that might serve as imitations, testing random combinations or a predetermined set of parameters is advised or to test the model to a similar type of dataset. surroundings that are different from the environment created after data cleansing. Reducing the number of characteristics is advised to cut further computational costs. The study was done by the Machine learning supervised techniques, which are used to build insurance claims predictive models. We have used Random Forest, KNN, logistic regression and XGBoost, amongst all these four algorithms KNN and Random Forest performed exponentially well on the dataset.

5.3. Future Work:

In order to compare the effectiveness of machine learning and deep learning methodologies, future research should focus on attempting to use an advanced or recently obtained dataset. Additionally, it is advised to utilize a different dataset in light of the fact that the one being used is unbalanced. Additional evaluation should be done to determine feature relevance across various datasets that may or may not have similar characteristics in order to develop a much more universal method to feature selection and focus. Because this research has been done by using all features in the future, we will do the feature selection to measure the variance between the total and selected features.

Bibliography

1. Abhinav Srivastava, A. K. (2008). Credit Card Fraud Detection Using. *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, 37 - 48.
2. Aisha Abdallah, M. A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 90-113.
3. Alejandro Correa Bahnsen, D. A. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems With Applications*, 134-142.
4. Andrea Dal Pozzolo, G. B. (2018). Credit Card Fraud Detection: A Realistic Modeling. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, vol 29, NO 8.
5. Andrea Dal Pozzolo, O. C.-A. (2014). Learned lessons in credit card fraud detection from a practitioner. *Expert systems with applications* .
6. Bart Baesens, S. H. (2021). Data engineering for fraud detection . *Decision Support Systems* .
7. E.W.T. Ngai, Y. H. (2011). The application of data mining techniques in financial fraud detection: A classification. *Decision Support Systems*, 559-569.
8. Eunji Kim, J. L.-k.-a.-i. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid. *Expert Systems With Applications*, 214 - 224.
9. Fabrizio Carcillo, Y.-A. L. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 317-331.
10. Jarrod West, M. B. (2016). Intelligent financial fraud detection: A comprehensive review. *ScienceDirect*, 47-66.
11. Javad Forough, S. M. (2021). Ensemble of deep sequential models for credit card fraud detection. *Applied Soft Computing Journal*.
12. Johannes Jurgovskya, M. G.-E.-G. (2018). Sequence classification for credit-card fraud detection. *Expert Systems With Applications*, 234-245.

13. Mieke Jans, ,. J. (2011). A business process mining application for internal transaction fraud mitigation. *Exert Systems with Applications* .
14. Siddhartha Bhattacharyya, S. J. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 602-613.
15. X. Liu, J. W. (2009). Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, 39* , 539 - 550.
16. Severino, M.K. and Peng, Y., 2021. Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 5, p.100074.
17. Hitesh Kumar Reddy Toppi, R., Bhavna, S., & Ginika, M. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. International Conference on Computational Intelligence and Data Science (ICCIDS 2018) (pp. 696–705). Elsevier
18. Callahan, Alison, and Nigam H. Shah. "Machine learning in healthcare." Key Advances in Clinical Informatics. Academic Press, 2017. 279-291.
19. Wiens, Jenna, and Erica S. Shenoy. "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology." Clinical Infectious Diseases 66.1 (2018): 149-153.
20. Niraj, K., & R, K. (2019). Predictive Analysis on Heart Disease Using Different Machine Learning Techniques. International Journal of Computer Sciences and Engineering, 97-101.
21. Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. International Journal of Forecasting, 36(4), 1420-1438.
22. Donepudi, P. K. (2018). AI and machine learning in retail pharmacy: systematic review of related literature. ABC journal of advanced research, 7(2), 109-112.
23. Hani, J., Mohamed, N., Ahmed, M., Emad, Z., Amer, E., & Ammar, M. (2019). Social media cyberbullying detection using machine learning. International Journal of Advanced Computer Science and Applications, 10(5).

24. Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710.
25. Stijn Viaene, M. A. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research, Volume 176, Issue 1, Pages 565-583.