

**Insurance Claim Fraud**

Submitted by:

Mini Sharma

**About the project**

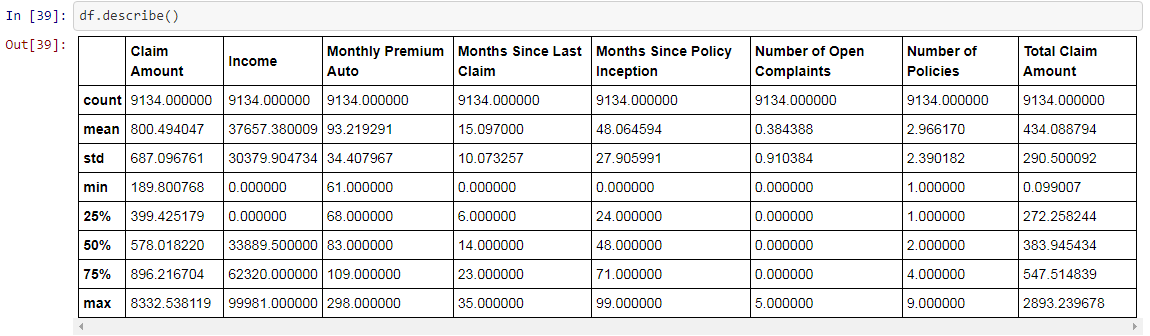
Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. However we can detect the fraud using machine learning. These are the factors from which we can detect a fraud.

* Claim Amount
* Response
* Coverage
* Education
* Effective To Date
* EmploymentStatus
* Gender
* Income
* Location Code
* Marital Status
* Monthly Premium Auto
* Months Since Last Claim
* Months Since Policy Inception
* Number of Open Complaints
* Number of Policies
* Policy Type
* Policy
* Claim Reason
* Sales Channel
* Total Claim Amount
* Vehicle Class
* Vehicle Size

# Analytical Problem Framing

## Mathematical/ Analytical Modeling of the Problem

Lets view some basic statistics about the data like the percentile, mean , maximum, minimum etc.



* + The average income is 37657.38. The average month since last claim is 15. Average month since policy inception is 48.



#### Lets See co-relation between the Columns

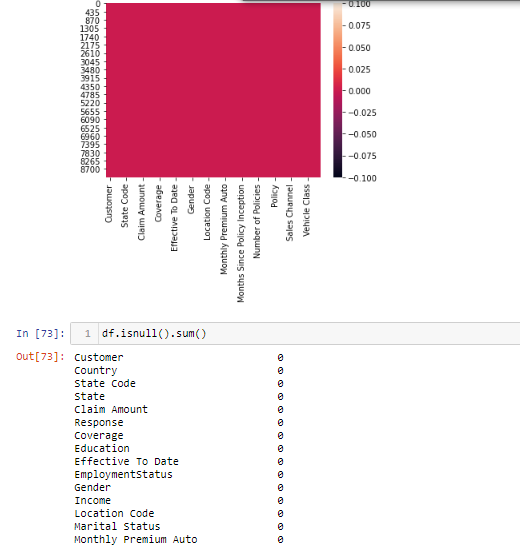
#### 

## Data Sources and their formats

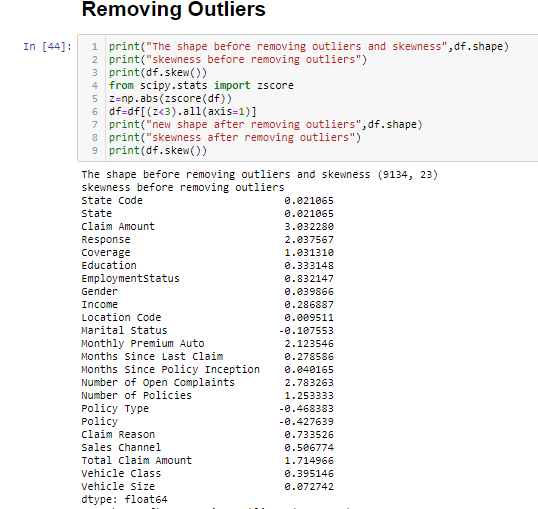
## The data is in comma separated value format and worked on it accordingly.

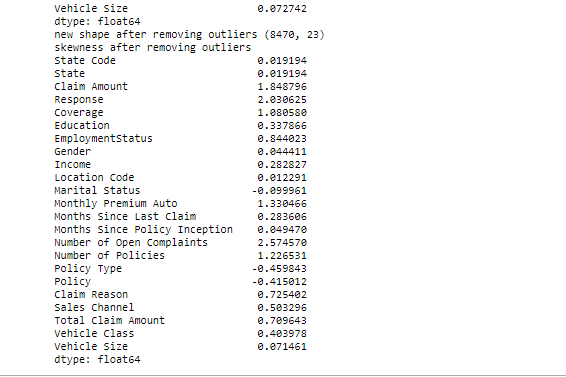
## Data Preprocessing Done

* + Lets check the shape and see count of the number of empty values in each column.



* + And we can see dataset contains no null values.



* + Data set contains many outliers so by using zscore we will remove outliers from the data set.
  + 

## Hardware and Software Requirements and Tools Used

We will use here Jupyter notebook to make your Prediction Model.

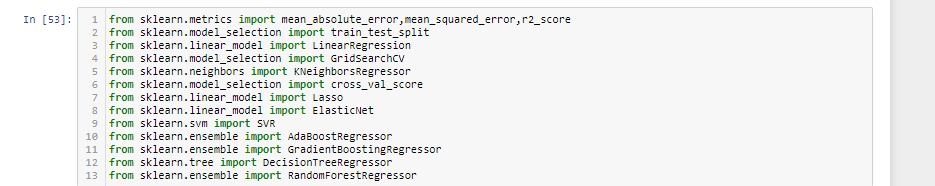
# Model/s Development and Evaluation

## Identification of possible problem-solving approaches (methods)

## 

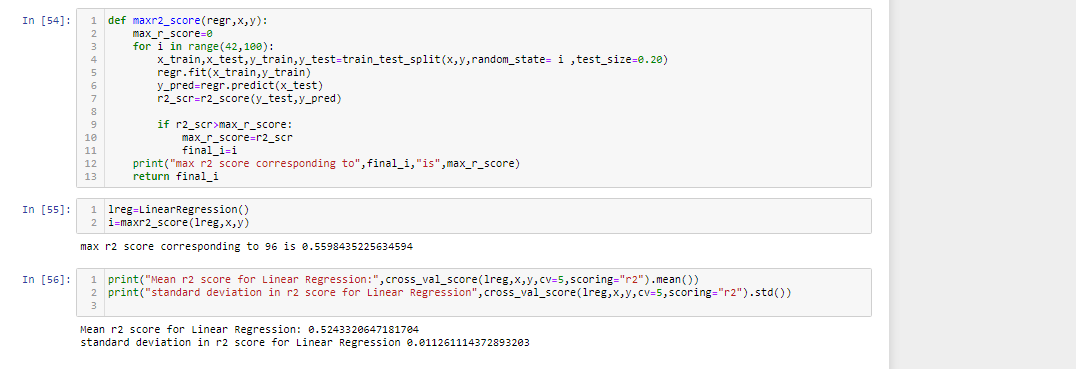
* + Now we will split the data set into input and output variable. As you can see above x is your input variable and y (label) is your target out variable.

## Testing of Identified Approaches (Algorithms)

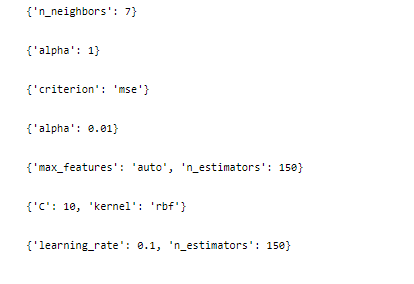


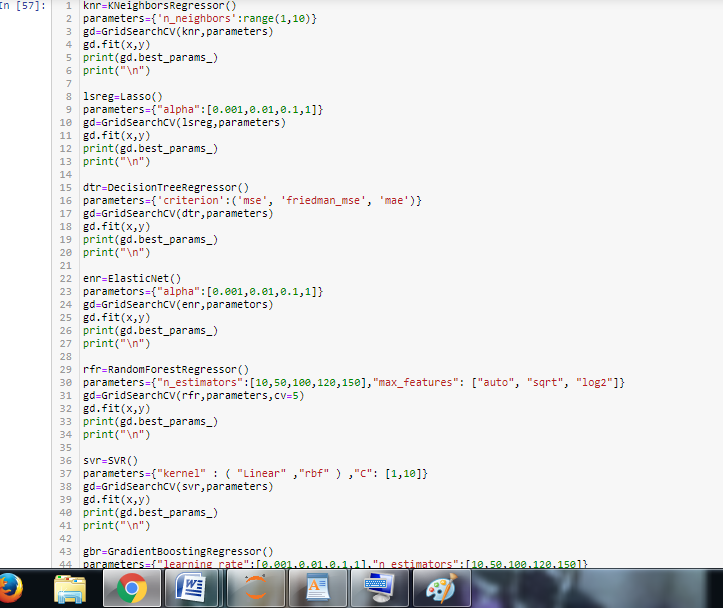
1. **Linear Regression**

* In Logistic Regression, we wish to model a dependent variable(y) in terms of one or more independent variables(x). It is a method for regression. This algorithm is used for the dependent variable that is Continous. Y is modeled using a function that gives output in linear relationship.

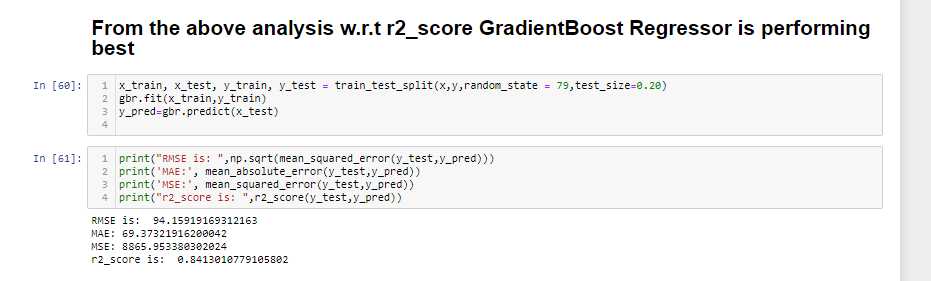


Further we will use grid search cv to find the best parameter for linear regression and other algorithms.





#### 



#### K Neighbors Classifier-

This is a supervised, non-parametric learning algorithm which classifies a given point based on its neighbours. The choice of the ‘k’ becomes very crucial since the data point is assigned to the class of the nearest ‘k’ neighbors. Once we get to know such ‘k’ nearest data points, the test data is assigned a label by taking the

majority vote from the class labels of the ‘k’ nearest data points

Further we will use grid search cv to find the best parameter for K Neighbors Classifier.

## Visualizations

