Project :

Insurance claim fraud detection :

---------------------------------------------------------------------------------------------------------

Task : Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, you are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

So , this is the dataset where I have to predict if an insurance claimed by the customer is fraudulent or not .

Data Analysis :

1. So , first after opening the dataset in the jupyter notebook , I have given proper column names to each column .
2. After getting a complete dataset , I have checked the basic requirements of the dataset like is any null value present in the dataset , any duplicated value present in the dataset and the column information .

* The dataset does not contain any null value .
* The dataset does not contain any duplicate value .
* The dataset contains both numerical and categorical columns .

1. After getting these basic information , I move towers the ‘fraud reported’ column where if a insurance claim fraud or not is given .

I have seen that the dataset is imbalanced. The values of fraud claims is very less ( 246 ) than the values of not fraud claims ( 753 ).

Hence , I have made the dataset balanced by keeping only 300 not fraud claim values so that I can get correct information in performing EDA and my model also can perform well on the dataset .

EDA :

Here , in this dataset , I think it is important to analyse each column with the fraud reported column so that I can get enough information . In this type of dataset , a small mistake can lead to a big loss .

The columns are :

1. months\_as\_customer: Number of months of patronage
2. age: the length of time a customer has lived or a thing has existed
3. policy\_number: It is a unique id given to the customer, to track the subscription status and other details of customer
4. policy\_bind\_date:date which document that is given to customer after we accept your proposal for insurance
5. policy\_state: This identifies who is the insured, what risks or property are covered, the policy limits, and the policy period
6. policy\_csl: is basically Combined Single Limit
7. policy\_deductable: the amount of money that a customer is responsible for paying toward an insured loss
8. policy\_annual\_premium: This means the amount of Regular Premium payable by the Policyholder in a Policy Year
9. umbrella\_limit: This means extra insurance that provides protection beyond existing limits and coverages of other policies
10. insured\_zip: It is the zip code where the insurance was made
11. insured\_sex: This refres to either of the two main categories (male and female) into which customer are divided on the basis of their reproductive functions
12. insured\_education\_level: This refers to the Level of education of the customer
13. insured\_occupation: This refers Occupation of the customer
14. insured\_hobbies: This refers to an activity done regularly by customer in his/her leisure time for pleasure.
15. insured\_relationship: This whether customer is: single; or. married; or. in a de facto relationship (that is, living together but not married); or. in a civil partnership
16. capital-gains: This refers to profit accrued due to insurance premium
17. capital-loss: This refers to the losses incurred due to insurance claims
18. incident\_date: This refers to the date which claims where made by customers
19. incident\_type: This refers to the type of claim/vehicle damage made by customer
20. collision\_type: This refers to the area of damage on the vehicle
21. incident\_severity: This refers to the extent/level of damage
22. authorities\_contacted: This refers to the government agencies that were contacted after damage
23. incident\_state: This refers to the state at which the accident happened
24. incident\_city: This refers to the city at which the accident happened
25. 1ncident\_location: This refers to the location at which the accident happened
26. incident\_hour\_of\_the\_day: The period of the day which accident took place
27. number\_of\_vehicles\_involved: This refers to number of vehicles involved the accident
28. property\_damage: This refers to whether property was damaged or not
29. bodily\_injuries: This refers to injuries sustained
30. witnesses: This refers to the number of witnesses involved
31. police\_report\_available: This refers to whether the report on damage was documented or not
32. total\_claim\_amount: This refers to the financial implications involved in claims
33. injury\_claim: This refers to physical injuries sustained
34. property\_claim: This refers to property damages during incident
35. vehicle\_claim: This refers to property damages during incident
36. auto\_make: This refers to the make of the vehicle
37. auto\_model: This refers to the model of the vehicle
38. auto\_year: This refers to the year which the vehicle was manufactured
39. \_c39:
40. fraud\_reported
41. So , first I have gone for the numerical columns .
    1. I have seen that the months of customer is not effecting that much in fraud report . But it is little bit high for those claims which are fraud .
    2. The factors ‘Age’ , ‘Policy deductable’ and ‘Policy annual premium’ is also not effecting in fraud report .
    3. The factor ‘Umbrella limit’ is effecting the fraud report . The claims with No fraud has less overall umbrella limit than fraud claims.
    4. The ‘capital gains’ and ‘capital loss’ is also not effecting that much .
    5. The total claim amount is effecting here . The claims with no fraud has less total claim amount.
    6. Similarly Injury claim , property claim and vehicle claim also has little bit of effect . The claims with no fraud has less amount of above factors .

Now I have checked the categorical columns :

* + 1. First I have check the relation of fraud reported with policy state . Policy state has an effect here .

the relation is :

The claims with no fraud has IL policy state most . then OH . But the claims with fraud has least IL policy state .

* + 1. Injured occupation is also has an effect here . The claim with no fraud has injure occupation ‘prof speciality’ . But it’s ‘exec merginal’ for the claims with fraud .
    2. In the case of incident type , the claims with no fraud has the incident type multi vehicle collision the most but in the case of the fraud claims , the single vehicle collision is the most.
    3. Collision type is not effecting that much here .
    4. In case of incident severity , the claim with no fraud has minor damages the most and then total loss , but in case of fraud claims , the incident severity is major damage the most.
    5. In case of authorities connected , the people whose claims are not fraud has connect with police the most but on the other hand , people whose claims are fraud has connected others the most.
    6. ‘ The number of vehicles involved ‘ is not effecting that much here .
    7. In case of property damage , the people with no fraud claims , there property has not damaged but , the people who has fraud claims , in the case of them , the amount of property damage is not known.
    8. In case of bodily injured , the people with no fraud claims , the bodily injured is 0 most of the time, the people who has fraud claims , in the case of them , the bodily injured is 2 most of the time .
    9. The witnesses is not effecting that much here .
    10. In the case of police reported available , the people with no fraud claims , most of the time they do not have police report and then Yes . But , the people who has fraud claims , in the case of them , the police report availability is not known and then No.
    11. For vehicle information , in case of ‘auto make’ , the claims with no fraud has the auto make Suburu the most , then Nissan and then Chevrolet . But in the case of fraud claims , the auto make is Ford the most , then Mercedes and then Audi .

So, after the analysis of all the columns , I have selected some most important columns . The columns are :

Customer Information :

1. months\_as\_customer

2. policy\_state

3. policy\_annual\_premium

4. capital-gains

5. capital-loss

Incident\_Information :

1. incident\_type

2. incident\_severity

3. authorities\_contacted

4. police\_report\_available

5. witnesses

6. total\_claim\_amount

7.collision\_type

Vehicle information :

1. auto\_make

2. auto\_model

Others :

1. insured\_occupation

2. injury\_claim

3. vehicle\_claim

4. property\_damage

5. bodily\_injuries

Preprocessing :

* + - 1. In the preprocessing phase , first I have replaced fraud\_reported values Y and N with numerical number . 1 for Y and 0 for N .
      2. Then I have split the data into X and y .
      3. Then I have used OneHotEncoder as column transformer to change the categorical columns . I have used OneHotEncoder , so that tree models can perform better in this dataset . I will use tree models to do Machine learning here .

Building Machine Learning models :

To perform this dataset , I will chose the tree base algorithms . Because I think tree based algorithms can perform correctly in this dataset by separating the conditions .

More over , along with the accuracy\_score , I have checked the Confusion metrix also to check what is the false positive value ( Actually fraud but Model predicting not fraud ) for each model . I have select false positive so that company can not face more loss . If model reject the deserved insurance claim, then next time we will get even more believable customer .

Using Decision Tree Classifier :

Here , After doing the GridSearchCV and also checking the best random state , I have traind DecisionTree Model . The accuracy is 82 % which is good . More over the false positive value is 20 and the true negative value is 8 , that means model is good .

Using Random Forest Classifier :

* Before Using Random Forest , I have use GridSearchCV to have best parameters and select the best random state .
* After Using Random Forest also , I have got the same result . Same for false positive and true negative also .

Using Gradient Boosting Classifier :

* Before UsingGradient Boosting, I have use GridSearchCV to have best parameters and select the best random state .
* After using Gradient Boosting also , I have got the same accuracy as 82 % but , here , the value of false positive is 17 which is less than DecisionTree Classifier and Random Forest Classifier .

I have used Voting Classifier also , but Gradient Boosting was giving the best result over here .

Hence I have selected Gradient Boosting Classifeir here . It is also profitable here because it learns from mistake .

Then I have made a Predictive model which is required .

----------------------------------------------- x ----------------------------------------------