Insurance Fraud Detection

Requirement: To build an automated Insurance Fraud Detection system where, client/firm can upload daily/weekly claim details and system will predict insurance claim as Fraud or not.

The client will provide data in multiple sets of files in batches at a specified location. The data, extracted from the census bureau, includes the following attributes:

**Features:**

1. **months\_as\_customer**: Number of months the customer is associated with the insurance company.
2. **age**: Continuous. Denotes the age of the person.
3. **policy\_number**: The policy number.
4. **policy\_bind\_date**: Start date of the policy.
5. **policy\_state**: State where the policy is registered.
6. **policy\_csl**: Combined single limits. Coverage for bodily injury from total damage. [More info](https://www.berkshireinsuranceservices.com/arecombinedsinglelimitsbetter)
7. **policy\_deductable**: Amount paid out of pocket by the policyholder before the insurance provider pays expenses.
8. **policy\_annual\_premium**: Yearly premium for the policy.
9. **umbrella\_limit**: Extra liability insurance coverage beyond the insured's other insurance limits.
10. **insured\_zip**: Zip code where the policy is registered.
11. **insured\_sex**: Gender of the person.
12. **insured\_education\_level**: Highest educational qualification of the policyholder.
13. **insured\_occupation**: Occupation of the policyholder.
14. **insured\_hobbies**: Hobbies of the policyholder.
15. **insured\_relationship**: Dependents on the policyholder.
16. **capital\_gain**: Monetary gains by the person.
17. **capital\_loss**: Monetary loss by the person.
18. **incident\_date**: Date of the incident.
19. **incident\_type**: Type of the incident.
20. **collision\_type**: Type of collision that occurred.
21. **incident\_severity**: Severity of the incident.
22. **authorities\_contacted**: Authority that was contacted.
23. **incident\_state**: State where the incident occurred.
24. **incident\_city**: City where the incident occurred.
25. **incident\_location**: Street where the incident occurred.
26. **incident\_hour\_of\_the\_day**: Time of day when the incident occurred.
27. **property\_damage**: Whether any property damage occurred.
28. **bodily\_injuries**: Number of bodily injuries.
29. **witnesses**: Number of witnesses present.
30. **police\_report\_available**: Availability of the police report.
31. **total\_claim\_amount**: Total amount claimed by the customer.
32. **injury\_claim**: Amount claimed for injury.
33. **property\_claim**: Amount claimed for property damage
34. **vehicle\_claim**: Amount claimed for vehicle damage.
35. **auto\_make**: Manufacturer of the vehicle.
36. **auto\_model**: Model of the vehicle.
37. **auto\_year**: Year of manufacture of the vehicle.
38. **fraud\_reported**: Whether the claim is fraudulent (Y or N). Target Variable

Raw Data Validation

Raw data received from client is validated for the following basic criterias.

* 1. Name of File Validation
  2. Number of Columns
  3. Null Values in columns

Based on validation move good/proper files to separate folder for further analysis and bad files to another folder.

Data Preprocessing

After raw validation we perform data preprocessing which includes EDA, feature engineering and we build a custom pipeline to perform this. We save model so that we can predict the test/predict data

Model Building

Based on EDA we found Xg-Boost to be the best model, so we will be using same for the model development. We save model so that we can use it later for predictions

For Prediction Data, we perform Raw Data Validation, Data Preprocessing and Prediction using saved model