**Problem Statement**

To build a classification methodology to determine whether a customer is placing a fraudulent insurance claim.

**Architecture**



**Data Description**

The client will send data in multiple sets of files in batches at a given location. The data has been extracted from the census bureau.

The data contains the following attributes:

**Features:**

1. **months\_as\_customer:** It denotes the number of months for which the customer is associated with the insurance company.
2. **age**: continuous. It denotes the age of the person.
3. **policy\_number**: The policy number.
4. **policy\_bind\_date**: Start date of the policy.
5. **policy\_state**: The state where the policy is registered.
6. **policy\_csl**-combined single limits. How much of the bodily injury will be covered from the total damage.

<https://www.berkshireinsuranceservices.com/arecombinedsinglelimitsbetter>

1. **policy\_deductable**: The amount paid out of pocket by the policy-holder before an insurance provider will pay any expenses.
2. **policy\_annual\_premium**: The yearly premium for the policy.
3. **umbrella\_limit**: An umbrella insurance policy is extra liability insurance coverage that goes beyond the limits of the insured's homeowners, auto or watercraft insurance. It provides an additional layer of security to those who are at risk of being sued for damages to other people's property or injuries caused to others in an accident.
4. **insured\_zip**: The zip code where the policy is registered.
5. **insured\_sex**: It denotes the person's gender.
6. **insured\_education\_level:** The highest educational qualification of the policy-holder.
7. **insured\_occupation:** Theoccupation of the policy-holder.
8. **insured\_hobbies**: Thehobbies of the policy-holder.
9. **insured\_relationship:** Dependents on the policy-holder.
10. **capital-gain**: It denotes the monitory gains by the person.
11. **capital-loss:** It denotes the monitory loss by the person.
12. **incident\_date:** The date when the incident happened.
13. **incident\_type:** The type of the incident.
14. **collision\_type:** The type of collision that took place.
15. **incident\_severity:** The severity of the incident.
16. **authorities\_contacted:** Which authority was contacted.
17. **incident\_state:** The state in which the incident took place.
18. **incident\_city:** The city in which the incident took place.
19. **incident\_location:** The street in which the incident took place.
20. **incident\_hour\_of\_the\_day:** The time of the day when the incident took place.
21. **property\_damage:** If any property damage was done.
22. **bodily\_injuries:** Number of bodily injuries.
23. **Witnesses:** Number of witnesses present.
24. **police\_report\_available:** Is the police report available.
25. **total\_claim\_amount:** Total amount claimed by the customer.
26. **injury\_claim:** Amount claimed for injury
27. **property\_claim:** Amount claimed for property damage.
28. **vehicle\_claim:** Amount claimed for vehicle damage.
29. **auto\_make:** The manufacturer of the vehicle
30. **auto\_model:** The model of the vehicle.
31. **auto\_year:** The year of manufacture of the vehicle.

**Target Label:**

Whether the claim is fraudulent or not.

1. fraud\_reported: Y or N

Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

**Data Validation**

In this step, we perform different sets of validation on the given set of training files.

1. Name Validation- We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the name, we check for the length of date in the file name as well as the length of time in the file name.
2. Number of Columns - We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file,.
3. Name of Columns - The name of the columns is validated and should be the same as given in the schema file.
4. The datatype of columns - The datatype of columns is given in the schema file. It is validated when we insert the files into Database.
5. Null values in columns - If any of the columns in a file have all the values as NULL or missing.

**Model Training**

1) **Data Preprocessing**

1. Drop the columns not required for prediction.
2. For this dataset, the null values were replaced with ‘?’ in the client data. Those ‘?’ have been replaced with NaN values.
3. Check for null values in the columns. If present, impute the null values using the categorical imputer.
4. Replace and encode the categorical values with numeric values.
5. Scale the numeric values using the standard scaler.

2) **Clustering -** KMeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using "KneeLocator" function. The idea behind clustering is to implement different algorithms

The Kmeans model is trained over preprocessed data, and the model is saved for further use in prediction.

3) **Model Selection –** After the clusters have been created, we find the best model for each cluster. We are using two algorithms, “SVM” and "XGBoost". For each cluster, both the algorithms are passed with the best parameters derived from GridSearch. We calculate the AUC scores for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.