Roadmap

Steps-

1).Making of the model for prediction

1). Doing EDA on the csv file, making it into a proper dataset

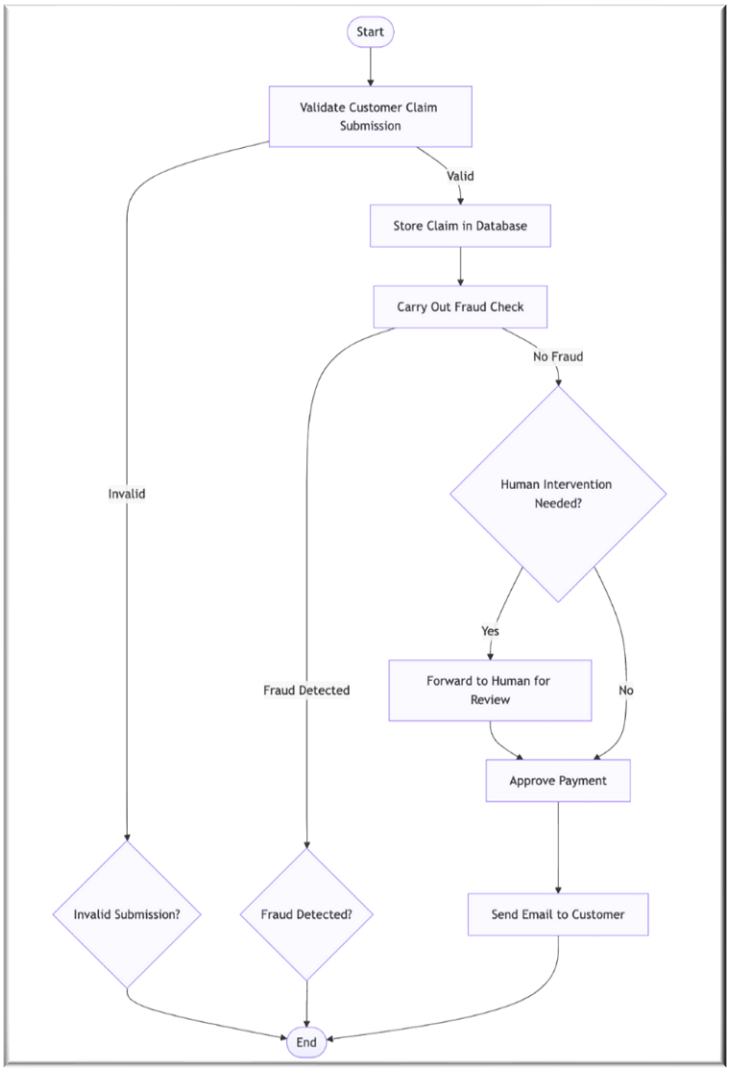
2). Find the important/required fields for modelling.

3). Making multiple models for comparison and bagging for better prediction.

4). Combining unsupervised anomaly detection with supervised random forest/decision tree algorithms

2). Extracting data –

- Data to be extracted form pdfs.
- Extract the data according to the required columns that we will have idea of from doing the first step.
- Pdf text can be read through various libraries like PyPDF2or pdfquery and stored in the list manner
- Now iterate through the lists, defining a dictionary and using regex library to extract data in front of the required columns and store it in a key value pair which can sent to the model for prediction.
- If any key has no value, it will lead to invalid submission.



Basic if-else statement to be used to follow the flow.

A confidence value of 0.8 should be enough for the profile to pass without a human interference.

A profile from 0.5 to 0.8 will be sent to human intervention.

Profile below 0.5 will be rejected.

Streamlit will be used for the webapp and a .pkl file of the trained model to be used for the prediction.

Making summary –

BERT or LLaMA to be used to make a digestable summary of the fraud data and the prediction ouput.

Adding to the database –

Will be using a tabular database MySQL to add the verified data from our python file.