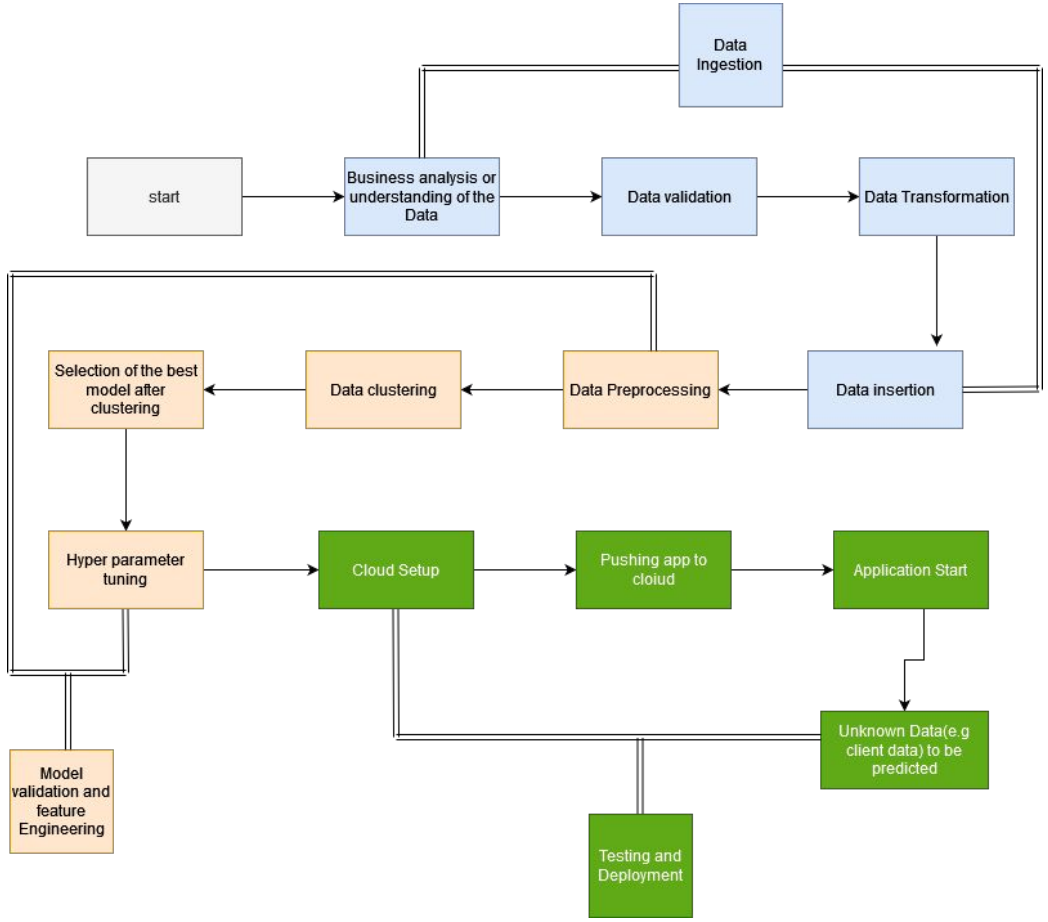
Credit Card Default Prediction System

Objective:

The Credit card default prediction system is the binary classification model which helps us to predict the defaulters(customers who haven’t pay their bills) based on the analysis of the data set from the client.

Architecture



Data Validation

In this step, we perform different sets of validation on the given dataset,

1 Name validation – Validating the data based on the pattern based.

After validating the data checking the length of the date in file as well as length of time in fil

2 Number of the columns – Validating the number of the columns in the file if the validation error occur the file will be not accepted.

3 The Datatype of the column – Validating the datatype of the column if there is no error then inserting the data into Database.

Data TransformationTaking the validating data and applying the data transformation process. Lets say if the data is in some categorical format or in any other language format then transforming the data into categorical transformation, date-time transformation, language transformation.

Data insertion

Database Creation and connection - Create a database with the given name .
If the database has already been created, open a connection to the database. The database will be in Cassandra.

Model Validation and feature engineering

Data Preprocessing

- Check for null values in the columns. If present, impute the null values using the categorical imputer.
- Scale the numeric values using the standard scaler.
- Check for correlation.

Data 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.
The idea behind clustering is to implement different algorithms

Model selection and hyperparameter tuning

After clustering process has been done selecting the one or two best model with the hyperparameter tuning approach in gridsearch ,random search or optuna.

Calculating the different scores for the best fitted model using the evalution metrices approach and graph visualization.