Campus Placement Predictor

Objective:

The Placement of students is one of the most important objective of an educational institution. Reputation and yearly admissions of an institution invariably depend on the placements it provides it students with. That is why all the institutions, arduously, strive to strengthen their placement department so as to improve their institution on a whole. Any assistance in this particular area will have a positive impact on an institution's ability to place its students. This will always be helpful to both the students, as well as the institution.

The main goal is to predict whether the student will be recruited in campus placements or not based on the available factors in the dataset.

Data Sharing Agreement :

> Sample file name (train.csv)

> Number of Columns

> Column names

> Column data type

Architecture
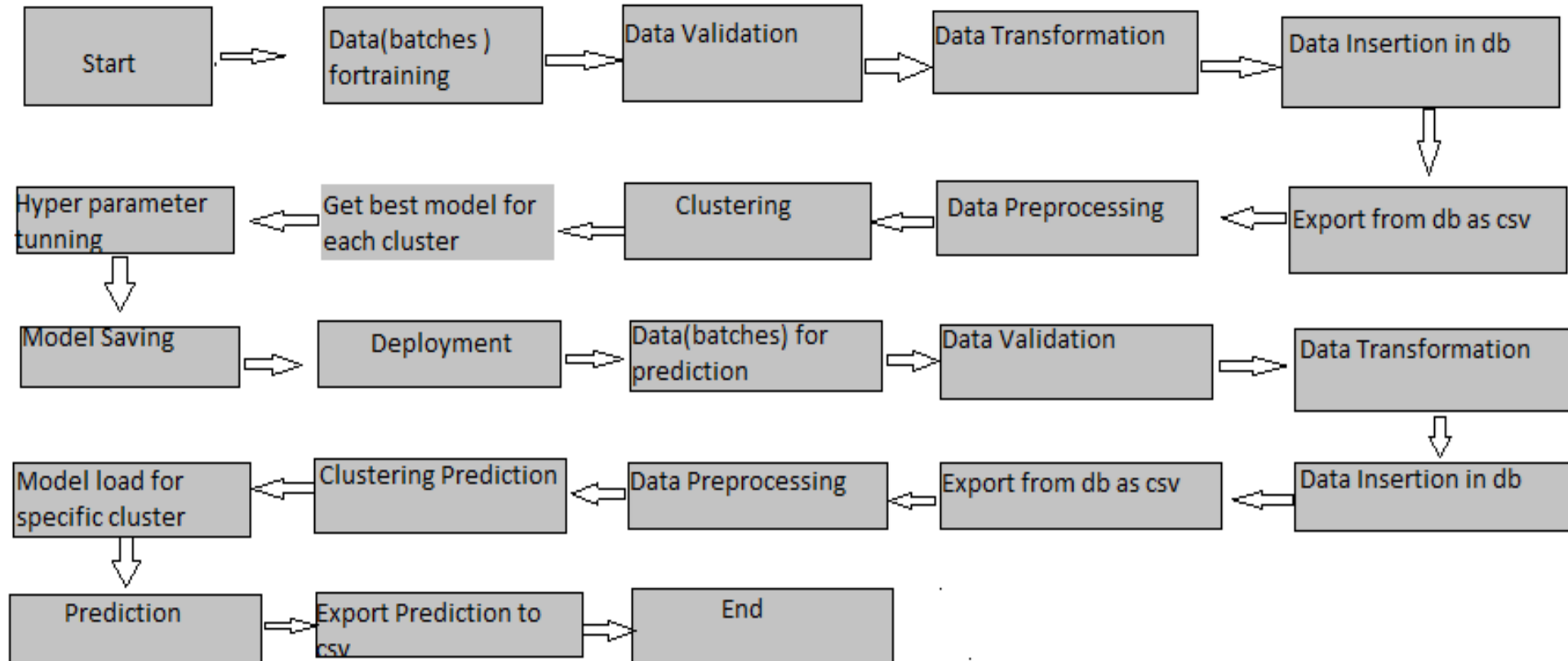
Data Validation and Data Transformation :

➢ Name Validation - Validation of files name as per the DSA. We have created a regex pattern for validation. After it checks for date format and time format if these requirements are satisfied, we move such files to "Good_Data_Folder" else "Bad_Data_Folder."

➢ Number of Columns – Validation of number of columns present in the files, and if it doesn't match then the file is moved to "Bad_Data_Folder."

➢ Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad_Data_Folder".

➢ Data type of columns - The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad_Data_Folder".

➢ Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad_Data_Folder".

Data Insertion in Database:

- ➢ Table creation :- Table name "placement_predictor" is created in the database for inserting the files. If the table is already present then new files are inserted in the

  same table.

- ➢ Insertion of files in the table - All the files in the "Good_Data_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table

Model Training:

- ➢ Data Export from Db :

  The accumulated data from db is exported in csv format for model training

- ➢ Data Preprocessing

  - ▪ Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc.

  - ▪ Check for null values in the columns. If present impute the null values.

  - ▪ Encode the categorical values with numeric values.

  - ▪ Perform Standard Scalar to scale down the values.

- Model Selection –

- We found the best model for the dataset. We checked four algorithms out of which Gradient Boosting was selected. We will calculate the r2 scores and mean absolute error for models and select the model with the best score.

Prediction:

➤ The testing files are shared in the batches and we perform the same Validation operations ,data transformation and data insertion on them.

➤ The accumulated data from db is exported in csv format for prediction

➤ We perform data pre-processing techniques on it.

➤ After pre-processing the data, respective model is loaded and is used to predict the data.

➤ Once the prediction is done, the predictions are saved in csv format and shared.

Q & A:

Q1) What's the source of data?

The data for training is taken from Kaggle dataset.

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 4th for better Understanding

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been

shared with the client and we removed the bad data folder.

Q 5) How logs are managed?

   We are using different logs as per the steps that we follow in   validation and

modeling like File validation log , Data Insertion ,Model Training log , prediction log

etc.

Q 6) What techniques were you using for data pre-processing?

- ► Removing unwanted attributes

- ► Visualizing relation of independent variables with each other and output variables

- ► Checking and changing Distribution of continuous values

- ► Removing outliers

- ► Cleaning data and imputing if null values are present.

- ► Converting categorical data into numeric values.

- ► Scaling the data

Q 7) How training was done or what models were used?

▶ Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.

▶ As per cluster the training and validation data were divided.

▶ The scaling was performed over training and validation data

▶ Algorithms like Linear Regression, SVM, Random Forest and Gradient Boosting were used based on the recall final model was used for each cluster and we saved that model .

► Q 9) What are the different stages of deployment?

  ► When the model is ready we deploy it using Flask.

  ► Once it is done, we hosted the model in Amazon Web Services.