**The Project Outline**

1. **The Problem Statement**
2. **Description of Data**
3. **Application Architecture and module division**
4. **Code Explanation**
5. **Data Ingestion**
6. **Data Validation**
7. **Data Preprocessing**
8. **Model Selection**
9. **Model Tuning**
10. **Prediction**
11. **Logging Framework**
12. **Deployment**

**The Problem Statement:**

1. Do a complete EDA in the python notebook file
2. Build a solution design architecture for end to end solution starting from data ingestion to deployment with a detail documentation.
3. Deploythe end to end automated solution to GCP.
4. Create a user interface for bulk testing uploaded through excel sheet and for a single record entry both.
5. Maintain log for each and every prediction request into any database
6. Create a low level documentation for end to end solution and deployment
7. Define a retraining approach in your documentation
8. Create an end to end video of the working of the project

**To build an application which predicts restaurant ratings based on given attributes.**

**Data Set Information:**

This dataset has been obtained by scraping the TA website for information about restaurants.

The data is a .csv file structured as follow:

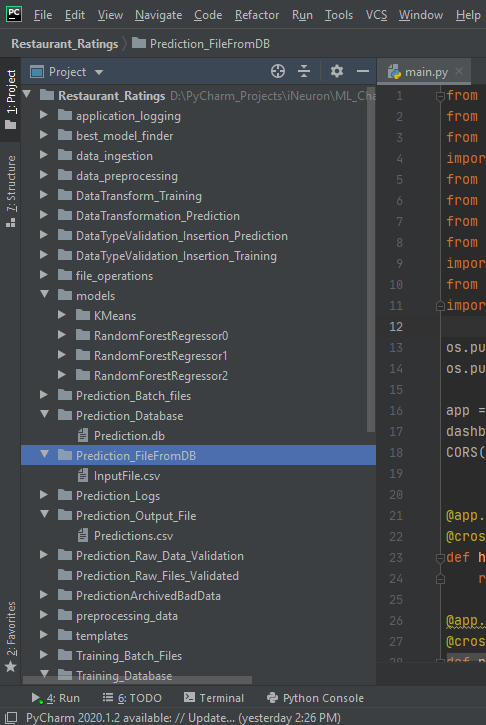
* + **Name: name if the restaurant**
  + **City**: city location of the restaurant
  + **Cuisine Style**: cuisine style(s) of the restaurant, in a Python list object (94 046 non-null)
  + **Ranking**: rank of the restaurant among the total number of restaurants in the city as a float object (115 645 non-null)
  + **Rating**: rate of the restaurant on a scale from 1 to 5, as a float object (115 658 non-null)(Target Column)
  + **Price Range**: price range of the restaurant among 3 categories , as a categorical type (77 555 non-null)
  + **Number of Reviews**: number of reviews that customers have let to the restaurant, as a float object (108 020 non-null)
  + **Reviews**: 2 reviews that are displayed on the restaurants scrolling page of the city, as a list of list object where the first list contains the 2 reviews, and the second le dates when these reviews were written (115 673 non-null)
  + **URL\_TA**: part of the URL of the detailed restaurant page that comes after 'www.tripadvisor.com' as a string object (124 995 non-null)

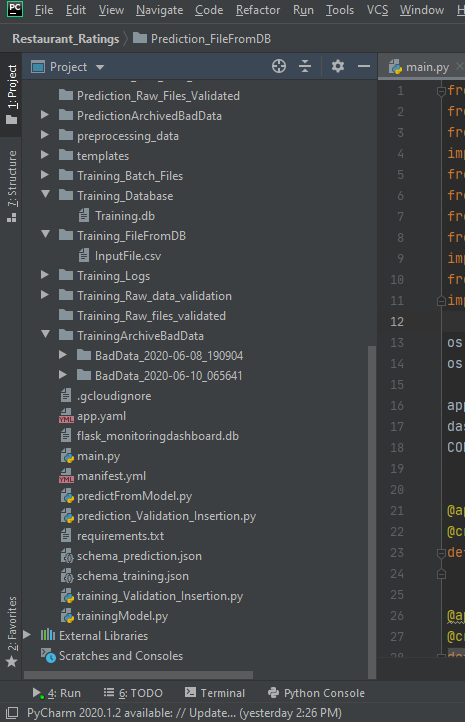
**ID\_TA**: identification of the restaurant in the TA database constructed a one letter and a number (124 995 non-null)

**Archiecture Application flow and module division**



**Project Folder Structure.**





**Training**

**Schema**

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.

* Schema for training in json format

{ "SampleFileName": "restaurants\_08012020\_120000.csv",  
 "LengthOfDateStampInFile": 8,  
 "LengthOfTimeStampInFile": 6,  
 "NumberofColumns" : 11,  
 "ColName": {  
 "City": "VARCHAR",  
 "Cuisine Style": "VARCHAR",  
 "Ranking": "FLOAT",  
 "Rating": "FLOAT",  
 "Price Range": "VARCHAR",  
 "Number of Reviews": "FLOAT"  
 }  
}

**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. If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder."
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, then the file is moved to "Bad\_Data\_Folder."
3. 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".
4. The datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad\_Data\_Folder".
5. 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".

**Code Snippet**:

with open(self.schema\_path, 'r') as f:  
 dic = json.load(f)  
 f.close()  
pattern = dic['SampleFileName']  
LengthOfDateStampInFile = dic['LengthOfDateStampInFile']  
LengthOfTimeStampInFile = dic['LengthOfTimeStampInFile']  
column\_names = dic['ColName']  
NumberofColumns = dic['NumberofColumns']  
  
file = open("Training\_Logs/valuesfromSchemaValidationLog.txt", 'a+')  
message ="LengthOfDateStampInFile:: %s" %LengthOfDateStampInFile + "\t" + "LengthOfTimeStampInFile:: %s" % LengthOfTimeStampInFile +"\t " + "NumberofColumns:: %s" % NumberofColumns + "\n"  
self.logger.log(file,message)  
  
file.close()

It will return the LengthOfDateStampInFile, LengthOfTimeStampInFile, column\_names, noofcolumns

# getting the regex defined to validate filename

regex = "['restaurants']+['\\_'']+[\d\_]+[\d]+\.csv"

# validating filename of prediction files

# delete the directories for good and bad data in case last run was unsuccessful and folders were not deleted.  
self.deleteExistingBadDataTrainingFolder()  
self.deleteExistingGoodDataTrainingFolder()  
  
onlyfiles = [f for f in listdir(self.Batch\_Directory)]  
try:  
 # create new directories  
 self.createDirectoryForGoodBadRawData()  
 f = open("Training\_Logs/nameValidationLog.txt", 'a+')  
 for filename in onlyfiles:  
 if (re.match(regex, filename)):  
 splitAtDot = re.split('.csv', filename)  
 splitAtDot = (re.split('\_', splitAtDot[0]))  
 if len(splitAtDot[1]) == LengthOfDateStampInFile:  
 if len(splitAtDot[2]) == LengthOfTimeStampInFile:  
 shutil.copy("Training\_Batch\_Files/" + filename, "Training\_Raw\_files\_validated/Good\_Raw")  
 self.logger.log(f,"Valid File name!! File moved to GoodRaw Folder :: %s" % filename)  
  
 else:  
 shutil.copy("Training\_Batch\_Files/" + filename, "Training\_Raw\_files\_validated/Bad\_Raw")  
 self.logger.log(f,"Invalid File Name!! File moved to Bad Raw Folder :: %s" % filename)  
 else:  
 shutil.copy("Training\_Batch\_Files/" + filename, "Training\_Raw\_files\_validated/Bad\_Raw")  
 self.logger.log(f,"Invalid File Name!! File moved to Bad Raw Folder :: %s" % filename)  
 else:  
 shutil.copy("Training\_Batch\_Files/" + filename, "Training\_Raw\_files\_validated/Bad\_Raw")  
 self.logger.log(f, "Invalid File Name!! File moved to Bad Raw Folder :: %s" % filename)  
  
 f.close()

# validating column length in the file

f = open("Training\_Logs/columnValidationLog.txt", 'a+')  
self.logger.log(f,"Column Length Validation Started!!")  
for file in listdir('Training\_Raw\_files\_validated/Good\_Raw/'):  
 csv = pd.read\_csv("Training\_Raw\_files\_validated/Good\_Raw/" + file)  
 if csv.shape[1] == NumberofColumns:  
 pass  
 else:  
 shutil.move("Training\_Raw\_files\_validated/Good\_Raw/" + file, "Training\_Raw\_files\_validated/Bad\_Raw")  
 self.logger.log(f, "Invalid Column Length for the file!! File moved to Bad Raw Folder :: %s" % file)

# validating if any column has all values missing

for file in listdir('Training\_Raw\_files\_validated/Good\_Raw/'):  
 csv = pd.read\_csv("Training\_Raw\_files\_validated/Good\_Raw/" + file)  
 count = 0  
 for columns in csv:  
 if (len(csv[columns]) - csv[columns].count()) == len(csv[columns]):  
 count+=1  
 shutil.move("Training\_Raw\_files\_validated/Good\_Raw/" + file,  
 "Training\_Raw\_files\_validated/Bad\_Raw")  
 self.logger.log(f,"Invalid Column Length for the file!! File moved to Bad Raw Folder :: %s" % file)  
 break  
 if count==0:  
 csv.rename(columns={"Unnamed: 0": "serial"}, inplace=True)  
 csv.to\_csv("Training\_Raw\_files\_validated/Good\_Raw/" + file, index=None, header=True)

# Starting Data Transforamtion!!

# replaceMissingWithNull

onlyfiles = [f for f in listdir(self.goodDataPath)]  
for file in onlyfiles:  
 csv = pd.read\_csv(self.goodDataPath+"/" + file, low\_memory=False)  
 csv.fillna("'NULL'", inplace=True)  
 csv.to\_csv(self.goodDataPath+ "/" + file, index=None, header=True)

# addQuotesToStringValuesInColumn

onlyfiles = [f for f in listdir(self.goodDataPath)]  
for file in onlyfiles:  
 data = pd.read\_csv(self.goodDataPath+"/" + file)  
 #dropping unnecessary column  
 data.drop(columns=['URL\_TA', "ID\_TA", "serial", "Reviews", "Name"], inplace=True)  
 columns = ["City", "Cuisine Style", "Price Range"]  
  
 for col in columns:  
 data[col] = data[col].apply(lambda x: '"' + str(x) + '"')  
  
 data.to\_csv(self.goodDataPath+ "/" + file, index=None, header=True)

**Data Insertion in Database**

1**) Database Creation and connection -** Create a database with the given name passed. If the database is already created, open the connection to the database.

conn = sqlite3.connect(self.path+DatabaseName+'.db')

2) **Table creation in the database -** Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" based on given column names and datatype in the schema file. If the table is already present, then the new table is not created and new files are inserted in the already present table as we want training to be done on new as well as old training files.

**Code Snippet**:

# create database with given name, if present open the connection! Create table with columns given in schema

conn = self.dataBaseConnection(DatabaseName)  
c=conn.cursor()  
c.execute("SELECT count(name) FROM sqlite\_master WHERE type = 'table'AND name = 'Good\_Raw\_Data'")  
if c.fetchone()[0] ==1:  
 conn.close()  
 file = open("Training\_Logs/DbTableCreateLog.txt", 'a+')  
 self.logger.log(file, "Tables created successfully!!")  
 file.close()  
  
 file = open("Training\_Logs/DataBaseConnectionLog.txt", 'a+')  
 self.logger.log(file, "Closed %s database successfully" % DatabaseName)  
 file.close()  
  
else:  
  
 for key in column\_names.keys():  
 type = column\_names[key]  
  
 #in try block we check if the table exists, if yes then add columns to the table  
 # else in catch block we will create the table  
 try:  
 conn.execute('ALTER TABLE Good\_Raw\_Data ADD COLUMN "{column\_name}" {dataType}'.format(column\_name=key,dataType=type))  
 except:  
 conn.execute('CREATE TABLE Good\_Raw\_Data ({column\_name} {dataType})'.format(column\_name=key, dataType=type))

conn.close()

**3) 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 and is moved to "Bad\_Data\_Folder".

# insert csv files in the table

for file in onlyfiles:  
 try:  
 with open(goodFilePath+'/'+file, "r") as f:  
 next(f)  
 reader = csv.reader(f, delimiter="\n")  
 for line in enumerate(reader):  
 for list\_ in (line[1]):  
 try:  
 conn.execute('INSERT INTO Good\_Raw\_Data values ({values})'.format(values=(list\_)))  
 self.logger.log(log\_file," %s: File loaded successfully!!" % file)  
 conn.commit()  
 except Exception as e:  
 raise e

**Model Training**

1. **Data Export from Db –**

The data in a stored database is exported as a CSV file to be used for model training.

# export data in table to csvfile

self.fileFromDb = 'Training\_FileFromDB/'  
self.fileName = 'InputFile.csv'  
log\_file = open("Training\_Logs/ExportToCsv.txt", 'a+')  
try:  
 conn = self.dataBaseConnection(Database)  
 sqlSelect = "SELECT \* FROM Good\_Raw\_Data"  
 cursor = conn.cursor()  
  
 cursor.execute(sqlSelect)  
  
 results = cursor.fetchall()  
 # Get the headers of the csv file  
 headers = [i[0] for i in cursor.description]  
  
 #Make the CSV ouput directory  
 if not os.path.isdir(self.fileFromDb):  
 os.makedirs(self.fileFromDb)  
  
 # Open CSV file for writing.  
 csvFile = csv.writer(open(self.fileFromDb + self.fileName, 'w', newline=''),delimiter=',', lineterminator='\r\n',quoting=csv.QUOTE\_ALL, escapechar='\\')  
  
 # Add the headers and data to the CSV file.  
 csvFile.writerow(headers)  
 csvFile.writerows(results)  
  
 self.logger.log(log\_file, "File exported successfully!!!")  
 log\_file.close()

**2) Data Preprocessing**

1. Rename Unstructure columns name to structure columns name

# rename columns  
self.data = self.data.rename(columns={'Cuisine Style': 'cuisines',  
 'Rating': 'rate',  
 'City': 'city',  
 'Number of Reviews': 'number\_of\_reviews',  
 "Price Range": "price\_range",  
 'Ranking': 'ranking'})

2. Clean text for objects data for (cuisines, price\_range, city) and replace empty string with NaN

# clean\_text for cuisines, price\_range, city and replace empty string with NaN  
str\_col = ['city', 'cuisines', 'price\_range']  
for col in str\_col:  
 self.data[col] = self.data[col].astype(str).str.replace('\[|\]|\'', '')  
 self.data[col] = self.data[col].str.replace('"', '')  
 self.data[col] = self.data[col].str.replace('NULL', '')  
 # to replace each empty string in a pandas DataFrame with NaN  
 self.data[col] = self.data[col].replace(r'^\s\*$', np.NaN, regex=True)

3. Droping the value which have NaN missing data in rate

# Droping the value which have NaN missing data in rate  
self.data.dropna(subset=['rate'], inplace = True)

4. Threshold for missing in rows which have more than 1 NaN value in rows

threshold = 1  
ls\_threshold = self.data.loc[  
 ((self.data.isnull().sum(axis=1)) <= threshold) & ((self.data.isnull().sum(axis=1)) != 0)]  
# Rows count for NaNs count more than threshold  
gr\_threshold = self.data.loc[(self.data.isnull().sum(axis=1)) > threshold]  
self.data.dropna(thresh=len(self.data.columns) - threshold, inplace=True)

5. Droping NaN missing data in cuisines if in further cases to deal with missing value for cuisines as in threshold all NaN value for cuisines has dropped.

# Droping the value which have NaN missing data in rate  
self.data.dropna(subset=['cuisines'], inplace = True)

6. lower text

# Data cleaning\_text for cuisines, and city to lowercase  
str\_col = ['city', 'cuisines']  
for col in str\_col:  
 # print(col)  
 self.data[col] = self.data[col].apply(lambda x : x.lower().strip())  
 self.data[col] = self.data[col].str.replace(' ', '')

7. Convert to interger (ranking, number\_of\_reviews) as it cant be in decimal form

# Convert to interger (ranking, number\_of\_reviews) as it cant be in decimal form  
float\_col = ['ranking', 'number\_of\_reviews']  
for col in float\_col:  
 self.data[col] = self.data[col].astype('Int64')

8. Dealing with price\_range renaming categorial and imputing missing values

# Convert categorise this 3 names as low, medium and high  
self.data = self.data.replace({'price\_range' : { '$$ - $$$' : 'low', '$' : 'medium', '$$$$' : 'high' }})  
  
# To deal missing value for price\_range will Assigning An Unique Category  
self.data['price\_range'].fillna('U', inplace=True)

9. Encode input variable

# Encode input variable  
for column in self.data.columns[~self.data.columns.isin(['ranking', 'rate', 'number\_of\_reviews'])]:  
 self.data[column] = self.data[column].factorize()[0]

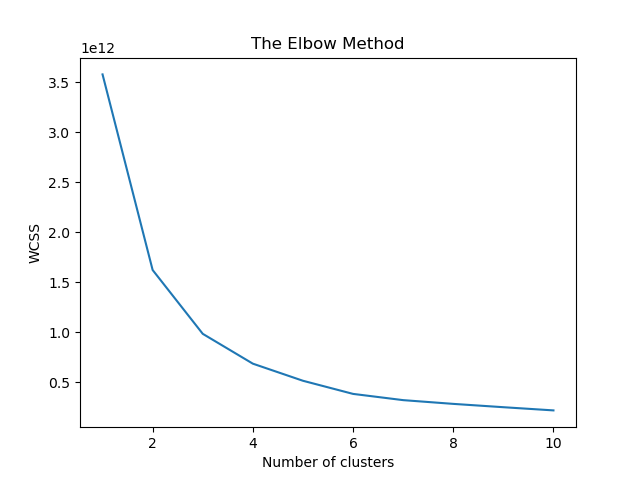
10. Null imputation

self.new\_data = self.data.fillna(self.data.median())

**3) Clustering –**

a) 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

wcss=[] # initializing an empty list  
try:  
 for i in range (1,11):  
 kmeans=KMeans(n\_clusters=i,init='k-means++',random\_state=42) # initializing the KMeans object  
 kmeans.fit(data) # fitting the data to the KMeans Algorithm  
 wcss.append(kmeans.inertia\_)  
 plt.plot(range(1,11),wcss) # creating the graph between WCSS and the number of clusters  
 plt.title('The Elbow Method')  
 plt.xlabel('Number of clusters')  
 plt.ylabel('WCSS')  
 #plt.show()  
 plt.savefig('preprocessing\_data/K-Means\_Elbow.PNG') # saving the elbow plot locally  
 # finding the value of the optimum cluster programmatically  
 self.kn = KneeLocator(range(1, 11), wcss, curve='convex', direction='decreasing')  
 self.logger\_object.log(self.file\_object, 'The optimum number of clusters is: '+str(self.kn.knee)+' . Exited the elbow\_plot method of the KMeansClustering class')  
 return self.kn.knee



Here I got the number of cluster 3 that can be seen through log files in ModelTrainingLog or in model folder i.e. number of model has been created.

b) To train data in different clusters. The Kmeans model is trained over preprocessed data and the model is saved for further use in prediction.

# Divide the data into clusters and save model

self.kmeans = KMeans(n\_clusters=number\_of\_clusters, init='k-means++', random\_state=42)  
#self.data = self.data[~self.data.isin([np.nan, np.inf, -np.inf]).any(1)]  
self.y\_kmeans=self.kmeans.fit\_predict(data) # divide data into clusters  
  
self.file\_op = file\_methods.File\_Operation(self.file\_object,self.logger\_object)  
self.save\_model = self.file\_op.save\_model(self.kmeans, 'KMeans') # saving the KMeans model to directory  
 # passing 'Model' as the functions need three parameters  
  
self.data['Cluster']=self.y\_kmeans # create a new column in dataset for storing the cluster information

**4) Model Selection –**

After clusters are created, we find the best model for each cluster. We are using Three algorithms, "Random forest Regressor", “Extra Tree Regressor” and “Xtreme boosting regressor”. For each cluster, the algorithms are passed with the best parameters derived from GridSearch. We calculate the Rsquared 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.

1. **Random forest Regressor**

# get\_best\_params\_for\_random\_forest

self.param\_grid = {"n\_estimators": [500,400],  
 "criterion": ['friedman\_mse'],  
 "min\_samples\_leaf": [.0001, 0.00001],  
 "random\_state": [329,369]}  
  
# Creating an object of the Grid Search class  
self.grid = GridSearchCV(estimator=self.clf, param\_grid=self.param\_grid, cv=5, verbose=3)  
# finding the best parameters  
self.grid.fit(train\_x, train\_y)  
# extracting the best parameters  
self.criterion = self.grid.best\_params\_['criterion']  
self.min\_samples\_leaf = self.grid.best\_params\_['min\_samples\_leaf']  
self.random\_state = self.grid.best\_params\_['random\_state']  
self.n\_estimators = self.grid.best\_params\_['n\_estimators']  
# creating a new model with the best parameters  
self.clf = RandomForestRegressor(n\_estimators=self.n\_estimators, criterion=self.criterion,  
 min\_samples\_leaf=self.min\_samples\_leaf, random\_state=self.random\_state)  
# training the mew model  
self.clf.fit(train\_x, train\_y)  
self.logger\_object.log(self.file\_object,  
 'Random Forest best params: ' + str(  
 self.grid.best\_params\_) + '. Exited the get\_best\_params\_for\_random\_forest method of the Model\_Finder class')  
return self.clf

1. **Extra Tree Regressor**

# get\_best\_params\_for\_extraTreeRegressor

# initializing with different combination of parameters  
self.param\_grid = {"n\_estimators": [500, 600],  
 "criterion": ['friedman\_mse'],  
 "min\_samples\_split": [7,6],  
 "min\_samples\_leaf": [4,5],  
 "max\_features": ['auto']}  
  
# Creating an object of the Grid Search class  
self.grid = GridSearchCV(estimator=self.etr, param\_grid=self.param\_grid, cv=5, verbose=3)  
# finding the best parameters  
self.grid.fit(train\_x, train\_y)  
# extracting the best parameters  
self.criterion = self.grid.best\_params\_['criterion']  
self.min\_samples\_split = self.grid.best\_params\_['min\_samples\_split']  
self.max\_features = self.grid.best\_params\_['max\_features']  
self.min\_samples\_leaf = self.grid.best\_params\_['min\_samples\_leaf']  
self.n\_estimators = self.grid.best\_params\_['n\_estimators']  
# creating a new model with the best parameters  
self.clf = ExtraTreesRegressor(n\_estimators=self.n\_estimators, criterion=self.criterion,  
 min\_samples\_split=self.min\_samples\_split,  
 min\_samples\_leaf= self.min\_samples\_leaf, max\_features=self.max\_features)  
  
# training the mew model  
self.etr.fit(train\_x, train\_y)  
self.logger\_object.log(self.file\_object,  
 'Extra Tree Regressor best params: '+str(self.grid.best\_params\_)+'. Exited the get\_best\_params\_for\_extraTreeRegressor method of the Model\_Finder class')  
  
return self.etr

1. **Xtreme boosting regressor**

self.param\_grid\_xgboost = {  
 "n\_estimators": [500,400],  
 'objective': ['reg:squarederror'],  
 'min\_child\_weight': [4,5],  
 "max\_depth": [5, 7],  
 'colsample\_bytree': [0.7, 1.0]  
  
}  
# Creating an object of the Grid Search class  
self.grid= GridSearchCV(XGBRegressor(),self.param\_grid\_xgboost, verbose=3,cv=5)  
# finding the best parameters  
self.grid.fit(train\_x, train\_y)  
  
# extracting the best parameters  
self.min\_child\_weight = self.grid.best\_params\_['min\_child\_weight']  
self.max\_depth = self.grid.best\_params\_['max\_depth']  
self.n\_estimators = self.grid.best\_params\_['n\_estimators']  
self.objective = self.grid.best\_params\_['objective']  
self.colsample\_bytree = self.grid.best\_params\_['colsample\_bytree']  
  
# creating a new model with the best parameters  
self.xgb = XGBRegressor(objective=self.objective, colsample\_bytree=self.colsample\_bytree,  
 min\_child\_weight=self.min\_child\_weight, max\_depth=self.max\_depth,  
 n\_estimators= self.n\_estimators, n\_jobs=-1)  
# training the mew model  
self.xgb.fit(train\_x, train\_y)  
self.logger\_object.log(self.file\_object,  
 'XGBoost best params: ' + str(  
 self.grid.best\_params\_) + '. Exited the get\_best\_params\_for\_xgboost method of the Model\_Finder class')  
return self.xgb

**RSquared**

In [statistics](https://en.wikipedia.org/wiki/Statistics), the **coefficient of determination**, denoted *R*2 or *r*2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

It is a [statistic](https://en.wikipedia.org/wiki/Statistic) used in the context of [statistical models](https://en.wikipedia.org/wiki/Statistical_model) whose main purpose is either the [prediction](https://en.wikipedia.org/wiki/Prediction#Statistics) of future outcomes or the testing of [hypotheses](https://en.wikipedia.org/wiki/Hypotheses), on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.[[1]](https://en.wikipedia.org/wiki/Coefficient_of_determination#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Coefficient_of_determination#cite_note-2)[[3]](https://en.wikipedia.org/wiki/Coefficient_of_determination#cite_note-3)

There are several definitions of *R*2 that are only sometimes equivalent. One class of such cases includes that of [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression) where *r*2 is used instead of *R*2. When an [intercept](https://en.wikipedia.org/wiki/Regression_intercept) is included, then *r*2 is simply the square of the sample [correlation coefficient](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient) (i.e., *r*) between the observed outcomes and the observed predictor values.[[4]](https://en.wikipedia.org/wiki/Coefficient_of_determination#cite_note-Devore-4) If additional [regressors](https://en.wikipedia.org/wiki/Regressor) are included, *R*2 is the square of the [coefficient of multiple correlation](https://en.wikipedia.org/wiki/Coefficient_of_multiple_correlation). In both such cases, the coefficient of determination normally ranges from 0 to 1.

self.xgboost= self.get\_best\_params\_for\_xgboost(train\_x,train\_y)  
self.prediction\_xgboost = self.xgboost.predict(test\_x) # Predictions using the XGBoost Model  
  
if len(test\_y.unique()) == 1: #if there is only one label in y, then roc\_auc\_score returns error. We will use accuracy in that case  
 self.xgboost\_score = accuracy\_score(test\_y, self.prediction\_xgboost)  
 self.logger\_object.log(self.file\_object, 'Accuracy for XGBoost:' + str(self.xgboost\_score)) # Log AUC  
else:  
 self.xgboost\_score = r2\_score(test\_y, self.prediction\_xgboost) # AUC for XGBoost  
 self.logger\_object.log(self.file\_object, 'AUC for XGBoost:' + str(self.xgboost\_score)) # Log AUC  
  
# create best model for extraTreeRegressor  
self.extraTreeRegressor=self.get\_best\_params\_for\_extraTreeRegressor(train\_x,train\_y)  
self.prediction\_extraTreeRegressor=self.extraTreeRegressor.predict(test\_x) # prediction using the extraTreeRegressor Algorithm  
  
if len(test\_y.unique()) == 1:#if there is only one label in y, then roc\_auc\_score returns error. We will use accuracy in that case  
 self.extraTreeRegressor\_score = accuracy\_score(test\_y,self.prediction\_extraTreeRegressor)  
 self.logger\_object.log(self.file\_object, 'Accuracy for ETR:' + str(self.extraTreeRegressor\_score))  
else:  
 self.extraTreeRegressor\_score = r2\_score(test\_y, (self.prediction\_extraTreeRegressor)) # AUC for extraTreeRegressor  
 self.logger\_object.log(self.file\_object, 'AUC for ETR:' + str(self.extraTreeRegressor\_score))  
  
# create best model for Random Forest  
self.random\_forest = self.get\_best\_params\_for\_random\_forest(train\_x, train\_y)  
self.prediction\_random\_forest = self.random\_forest.predict(test\_x) # prediction using the Random Forest Algorithm  
  
if len(test\_y.unique()) == 1: # if there is only one label in y, then roc\_auc\_score returns error. We will use accuracy in that case  
 self.random\_forest\_score = accuracy\_score(test\_y, self.prediction\_random\_forest)  
 self.logger\_object.log(self.file\_object, 'Accuracy for RF:' + str(self.random\_forest\_score))  
else:  
 self.random\_forest\_score = r2\_score(test\_y,  
 self.prediction\_random\_forest) # AUC for Random Forest  
 self.logger\_object.log(self.file\_object, 'AUC for RF:' + str(self.random\_forest\_score))  
  
# comparing the three models  
if(self.extraTreeRegressor\_score < self.xgboost\_score and self.random\_forest\_score < self.xgboost\_score):  
 return 'XGBoost',self.xgboost  
if(self.extraTreeRegressor\_score < self.random\_forest\_score and self.random\_forest\_score > self.xgboost\_score):  
 return 'RandomForestRegressor', self.random\_forest  
else:  
 return 'extraTreeRegressor\_score',self.extraTreeRegressor

**Prediction Data Description**

Client will send the data in multiple set of files in batches at a given location. Data will contain restaurants ratings in 10 columns.( restaurants\_08012020\_120000.csv)

Apart from prediction files, we also require a "schema" file from 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

* Schema for Prediction in json format

{ "SampleFileName": "restaurants\_08012020\_120000.csv",  
 "LengthOfDateStampInFile": 8,  
 "LengthOfTimeStampInFile": 6,  
 "NumberofColumns" : 10,  
 "ColName": {  
 "City": "VARCHAR",  
 "Cuisine Style": "VARCHAR",  
 "Ranking": "FLOAT",  
 "Price Range": "VARCHAR",  
 "Number of Reviews": "FLOAT"  
 }  
}

**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. If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder."
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, then the file is moved to "Bad\_Data\_Folder."
3. 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".
4. The datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad\_Data\_Folder".
5. 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".

**Code Snippet**:

with open(self.schema\_path, 'r') as f:  
 dic = json.load(f)  
 f.close()  
pattern = dic['SampleFileName']  
LengthOfDateStampInFile = dic['LengthOfDateStampInFile']  
LengthOfTimeStampInFile = dic['LengthOfTimeStampInFile']  
column\_names = dic['ColName']  
NumberofColumns = dic['NumberofColumns']  
  
file = open("Training\_Logs/valuesfromSchemaValidationLog.txt", 'a+')  
message ="LengthOfDateStampInFile:: %s" %LengthOfDateStampInFile + "\t" + "LengthOfTimeStampInFile:: %s" % LengthOfTimeStampInFile +"\t " + "NumberofColumns:: %s" % NumberofColumns + "\n"  
self.logger.log(file,message)  
  
file.close()

It will return the LengthOfDateStampInFile, LengthOfTimeStampInFile, column\_names, noofcolumns

# getting the regex defined to validate filename

regex = "['restaurants']+['\\_'']+[\d\_]+[\d]+\.csv"

# validating filename of prediction files

self.deleteExistingBadDataTrainingFolder()  
self.deleteExistingGoodDataTrainingFolder()  
self.createDirectoryForGoodBadRawData()  
onlyfiles = [f for f in listdir(self.Batch\_Directory)]  
try:  
 f = open("Prediction\_Logs/nameValidationLog.txt", 'a+')  
 for filename in onlyfiles:  
 if (re.match(regex, filename)):  
 splitAtDot = re.split('.csv', filename)  
 splitAtDot = (re.split('\_', splitAtDot[0]))  
 if len(splitAtDot[1]) == LengthOfDateStampInFile:  
 if len(splitAtDot[2]) == LengthOfTimeStampInFile:  
 shutil.copy(self.Batch\_Directory + '/' + filename,  
 "Prediction\_Raw\_Files\_Validated/Good\_Raw")  
 self.logger.log(f, "Valid File name!! File moved to GoodRaw Folder :: %s" % filename)  
  
 else:  
 shutil.copy(self.Batch\_Directory + '/' + filename, "Prediction\_Raw\_Files\_Validated/Bad\_Raw")  
 self.logger.log(f, "Invalid File Name!! File moved to Bad Raw Folder :: %s" % filename)  
 else:  
 shutil.copy(self.Batch\_Directory + '/' + filename, "Prediction\_Raw\_Files\_Validated/Bad\_Raw")  
 self.logger.log(f, "Invalid File Name!! File moved to Bad Raw Folder :: %s" % filename)  
 else:  
 shutil.copy(self.Batch\_Directory + '/' + filename, "Prediction\_Raw\_Files\_Validated/Bad\_Raw")  
 self.logger.log(f, "Invalid File Name!! File moved to Bad Raw Folder :: %s" % filename)  
  
 f.close()

# validating if any column has all values missing

f = open("Prediction\_Logs/missingValuesInColumn.txt", 'a+')  
self.logger.log(f, "Missing Values Validation Started!!")  
  
for file in listdir('Prediction\_Raw\_Files\_Validated/Good\_Raw/'):  
 csv = pd.read\_csv("Prediction\_Raw\_Files\_Validated/Good\_Raw/" + file, low\_memory=False)  
 count = 0  
 for columns in csv:  
 if (len(csv[columns]) - csv[columns].count()) == len(csv[columns]):  
 count+=1  
 shutil.move("Prediction\_Raw\_Files\_Validated/Good\_Raw/" + file,  
 "Prediction\_Raw\_Files\_Validated/Bad\_Raw")  
 self.logger.log(f,"Invalid Column Length for the file!! File moved to Bad Raw Folder :: %s" % file)  
 break  
 if count==0:  
 csv.rename(columns={"Unnamed: 0": "Index"}, inplace=True)  
 csv.to\_csv("Prediction\_Raw\_Files\_Validated/Good\_Raw/" + file, index=None, header=True)

csv.rename(columns={"Unnamed: 0": "serial"}, inplace=True)  
 csv.to\_csv("Training\_Raw\_files\_validated/Good\_Raw/" + file, index=None, header=True)

# Starting Data Transforamtion!!

# replaceMissingWithNull

onlyfiles = [f for f in listdir(self.goodDataPath)]  
for file in onlyfiles:  
 csv = pd.read\_csv(self.goodDataPath+"/" + file, low\_memory=False)  
 csv.fillna("'NULL'", inplace=True)  
 csv.to\_csv(self.goodDataPath+ "/" + file, index=None, header=True)

# addQuotesToStringValuesInColumn

onlyfiles = [f for f in listdir(self.goodDataPath)]  
for file in onlyfiles:  
 data = pd.read\_csv(self.goodDataPath+"/" + file)  
 #dropping unnecessary column  
 data.drop(columns=['URL\_TA', "ID\_TA", "serial", "Reviews", "Name"], inplace=True)  
 columns = ["City", "Cuisine Style", "Price Range"]  
  
 for col in columns:  
 data[col] = data[col].apply(lambda x: '"' + str(x) + '"')  
  
 data.to\_csv(self.goodDataPath+ "/" + file, index=None, header=True)

**Data Insertion in Database**

**1) Database Creation and connection -** Create a database with the given name passed. If the database is already created, open the connection to the database.

conn = sqlite3.connect(self.path+DatabaseName+'.db')

**2) Table creation in the database -** Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" based on given column names and datatype in the schema file. If the table is already present, then the new table is not created and new files are inserted in the already present table as we want training to be done on new as well as old training files.

**Code Snippet**:

# create database with given name, if present open the connection! Create table with columns given in schema

conn = self.dataBaseConnection(DatabaseName)  
conn.execute('DROP TABLE IF EXISTS Good\_Raw\_Data;')  
  
for key in column\_names.keys():  
 type = column\_names[key]  
  
 # we will remove the column of string datatype before loading as it is not needed for training  
 #in try block we check if the table exists, if yes then add columns to the table  
 # else in catch block we create the table  
 try:  
 #cur = cur.execute("SELECT name FROM {dbName} WHERE type='table' AND name='Good\_Raw\_Data'".format(dbName=DatabaseName))  
 conn.execute('ALTER TABLE Good\_Raw\_Data ADD COLUMN "{column\_name}" {dataType}'.format(column\_name=key,dataType=type))  
 except:  
 conn.execute('CREATE TABLE Good\_Raw\_Data ({column\_name} {dataType})'.format(column\_name=key, dataType=type))  
  
conn.close()

**3) 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 and is moved to "Bad\_Data\_Folder".

# insert csv files in the table

for file in onlyfiles:  
 try:  
 with open(goodFilePath+'/'+file, "r") as f:  
 next(f)  
 reader = csv.reader(f, delimiter="\n")  
 for line in enumerate(reader):  
 for list\_ in (line[1]):  
 try:  
 conn.execute('INSERT INTO Good\_Raw\_Data values ({values})'.format(values=(list\_)))  
 self.logger.log(log\_file," %s: File loaded successfully!!" % file)  
 conn.commit()  
 except Exception as e:  
 raise e

**Prediction**

**1) Data Export from Db -** The data in the stored database is exported as a CSV file to be used for prediction.

conn = self.dataBaseConnection(Database)  
sqlSelect = "SELECT \* FROM Good\_Raw\_Data"  
cursor = conn.cursor()  
  
cursor.execute(sqlSelect)  
  
results = cursor.fetchall()  
  
#Get the headers of the csv file  
headers = [i[0] for i in cursor.description]  
  
#Make the CSV ouput directory  
if not os.path.isdir(self.fileFromDb):  
 os.makedirs(self.fileFromDb)  
  
# Open CSV file for writing.  
csvFile = csv.writer(open(self.fileFromDb + self.fileName, 'w', newline=''),delimiter=',', lineterminator='\r\n',quoting=csv.QUOTE\_ALL, escapechar='\\')  
  
# Add the headers and data to the CSV file.  
csvFile.writerow(headers)  
csvFile.writerows(results)

**2) Data Preprocessing**

1. Rename Unstructure columns name to structure columns name

# rename columns  
self.data = self.data.rename(columns={'Cuisine Style': 'cuisines',  
 'Rating': 'rate',  
 'City': 'city',  
 'Number of Reviews': 'number\_of\_reviews',  
 "Price Range": "price\_range",  
 'Ranking': 'ranking'})

2. Clean text for objects data for (cuisines, price\_range, city) and replace empty string with NaN

# clean\_text for cuisines, price\_range, city and replace empty string with NaN  
str\_col = ['city', 'cuisines', 'price\_range']  
for col in str\_col:  
 self.data[col] = self.data[col].astype(str).str.replace('\[|\]|\'', '')  
 self.data[col] = self.data[col].str.replace('"', '')  
 self.data[col] = self.data[col].str.replace('NULL', '')  
 # to replace each empty string in a pandas DataFrame with NaN  
 self.data[col] = self.data[col].replace(r'^\s\*$', np.NaN, regex=True)

3. Droping NaN missing data in cuisines if in further cases to deal with missing value for cuisines as in threshold all NaN value for cuisines has dropped.

# Droping the value which have NaN missing data in rate  
self.data.dropna(subset=['cuisines'], inplace = True)

4. lower text

# Data cleaning\_text for cuisines, and city to lowercase  
str\_col = ['city', 'cuisines']  
for col in str\_col:  
 # print(col)  
 self.data[col] = self.data[col].apply(lambda x : x.lower().strip())  
 self.data[col] = self.data[col].str.replace(' ', '')

5. Convert to interger (ranking, number\_of\_reviews) as it cant be in decimal form

# Convert to interger (ranking, number\_of\_reviews) as it cant be in decimal form  
float\_col = ['ranking', 'number\_of\_reviews']  
for col in float\_col:  
 self.data[col] = self.data[col].astype('Int64')

6. Dealing with price\_range renaming categorial and imputing missing values

# Convert categorise this 3 names as low, medium and high  
self.data = self.data.replace({'price\_range' : { '$$ - $$$' : 'low', '$' : 'medium', '$$$$' : 'high' }})  
  
# To deal missing value for price\_range will Assigning An Unique Category  
self.data['price\_range'].fillna('U', inplace=True)

7. Encode input variable

# Encode input variable  
for column in self.data.columns[~self.data.columns.isin(['ranking', 'rate', 'number\_of\_reviews'])]:  
 self.data[column] = self.data[column].factorize()[0]

8. Null imputation

self.new\_data = self.data.fillna(self.data.median())

1. **Clustering –**

KMeans model created during training is loaded, and clusters for the preprocessed prediction data is predicted.

kmeans=file\_loader.load\_model('KMeans')  
clusters=kmeans.predict(data\_scaled)#drops the first column for cluster prediction  
data\_scaled['clusters']=clusters  
clusters=data\_scaled['clusters'].unique()

1. **Prediction –**

Based on the cluster number, the respective model is loaded and is used to predict the data for that cluster.

for i in clusters:  
 cluster\_data= data\_scaled[data\_scaled['clusters']==i]  
 cluster\_data = cluster\_data.drop(['clusters'],axis=1)  
 model\_name = file\_loader.find\_correct\_model\_file(i)  
 model = file\_loader.load\_model(model\_name)  
 for val in (model.predict(cluster\_data.values)):  
 if val >= 0 and val < 0.5:  
 result.append(0)  
 elif val >= 0.5 and val < 1:  
 result.append(0.5)  
 elif val >= 1 and val < 1.5:  
 result.append(1)  
 elif val >= 1.5 and val < 2:  
 result.append(1.5)  
 elif val >= 2 and val < 2.5:  
 result.append(2)  
 elif val >= 2.5 and val < 3:  
 result.append(2.5)  
 elif val >= 3 and val < 3.5:  
 result.append(3)  
 elif val >= 3.5 and val < 4:  
 result.append(3.5)  
 elif val >= 4 and val < 4.5:  
 result.append(4)  
 elif val >= 4.5 and val < 5:  
 result.append(4.5)  
 elif val >= 5:  
 result.append(5)

result = pandas.DataFrame(result,columns=['Predictions'])

**Once the prediction is made for all the clusters**, the predictions along with the original names before label encoder are saved in a CSV file at a given location and the location is returned to the client.

result.to\_csv("Prediction\_Output\_File/Predictions.csv", header=True, mode='a+') #appends result to prediction file

**Some Important files**

1. **requirements.txt** file consists of all the packages that you need to deploy the app in the cloud.

**Pip install -r requirements.txt**

1. **main.py** is the entry point of our application, where the flask server starts.

@app.route("/", methods=['GET'])  
@cross\_origin()  
def home():  
 return render\_template('index.html')  
  
@app.route("/predict", methods=['POST'])  
@cross\_origin()  
def predictRouteClient():  
 try:  
 # when we are sending the request in json used in postman  
 if request.json is not None:  
 path = request.json['filepath']  
 # Validation Part  
 pred\_val = Pred\_Validation(path) #object initialization  
 pred\_val.prediction\_validation() #calling the prediction\_validation function  
  
 # Training Part  
 pred = Prediction(path) #object initialization  
 # predicting for dataset present in database  
 path, json\_predictions = pred.predictionFromModel()  
  
 return Response("Prediction File created at !!!" + str(path) + 'and few of the predictions are ' + str(  
 json.loads(json\_predictions)))  
  
 # When we are sending the request in the form used in UserInterface (UI) part  
 elif request.form is not None:  
 path = request.form['filepath']  
 # Validation Part  
 pred\_val = Pred\_Validation(path) #object initialization  
 pred\_val.prediction\_validation() #calling the prediction\_validation function  
  
 # Training Part  
 pred = Prediction(path) #object initialization  
 # predicting for dataset present in database  
 path, json\_predictions = pred.predictionFromModel()  
 return Response("Prediction File created at !!!" + str(path) + 'and few of the predictions are ' +  
 str(json.loads(json\_predictions)))  
  
 except ValueError:  
 return Response("Error Occurred2! %s" %ValueError)  
 except KeyError:  
 return Response("Error Occurred3! %s" %KeyError)  
 except Exception as e:  
 return Response("Error Occurred4! %s" %e)  
  
  
  
@app.route("/train", methods=['POST'])  
@cross\_origin()  
def trainRouteClient():  
  
 try:  
 if request.json['folderPath'] is not None:  
 path = request.json['folderPath']  
 # Validation Part  
 train\_valObj = TrainValidation(path) #object initialization  
 train\_valObj.train\_validation()#calling the training\_validation function  
  
 # Training Part  
 trainModelObj = TrainModel() #object initialization  
 trainModelObj.trainingModel() #training the model for the files in the table  
  
 except ValueError:  
 return Response("Error Occurred! %s" % ValueError)  
 except KeyError:  
 return Response("Error Occurred! %s" % KeyError)  
 except Exception as e:  
 return Response("Error Occurred! %s" % e)  
 return Response("Training successfull!!")

1. This is the **predictionFromModel.py** file where the predictions take place based on the data we are giving input to the model.

class Prediction:  
  
 def \_\_init\_\_(self, path):  
 # path = r"Prediction\_Batch\_Files"  
 self.file\_object = open("Prediction\_Logs/Prediction\_Log.txt", 'a+')  
 self.log\_writer = logger.App\_Logger()  
 self.pred\_data\_val = Prediction\_Data\_validation(path)  
  
 def predictionFromModel(self):  
  
 try:  
 self.pred\_data\_val.deletePredictionFile() #deletes the existing prediction file from last run!  
 self.log\_writer.log(self.file\_object,'Start of Prediction')  
 data\_getter=data\_loader\_prediction.Data\_Getter\_Pred(self.file\_object,self.log\_writer)  
 data=data\_getter.get\_data()

1. **manifest.yml**:- This file contains the instance configuration, app name, and build pack language.

---  
applications:  
- name: restaurants  
 memory: 2GB  
 disk\_quota: 2GB  
 random-route: true  
 parameters:  
 memory: 2GB  
 buildpack: 'python\_buildpack'

1. **app.yaml** :- It contains the entry point of the app.

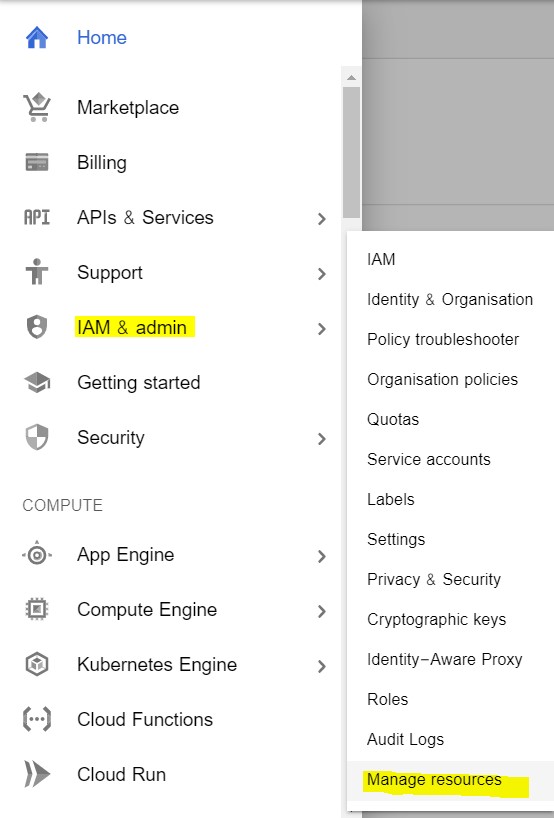
runtime: python37

**Deployment to G-cloud**

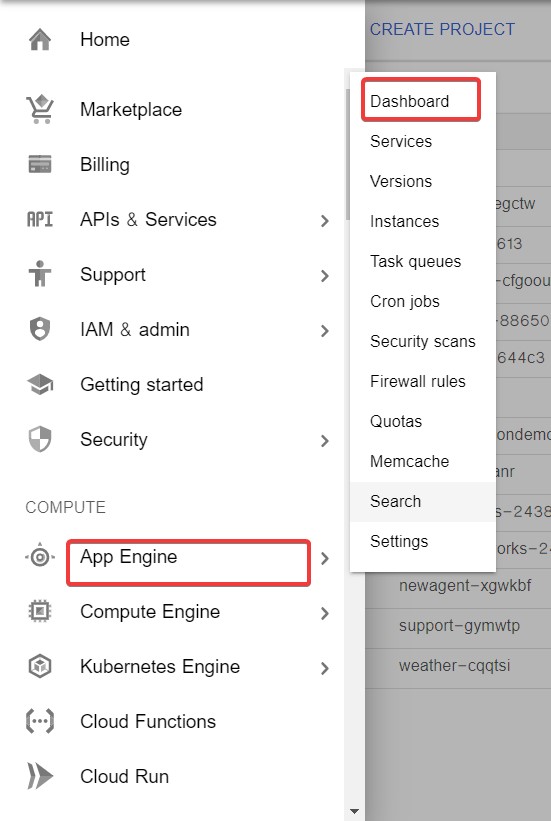
* Go to <https://cloud.google.com/>and create an account if already haven’t created one.

Then go to the console of your account.

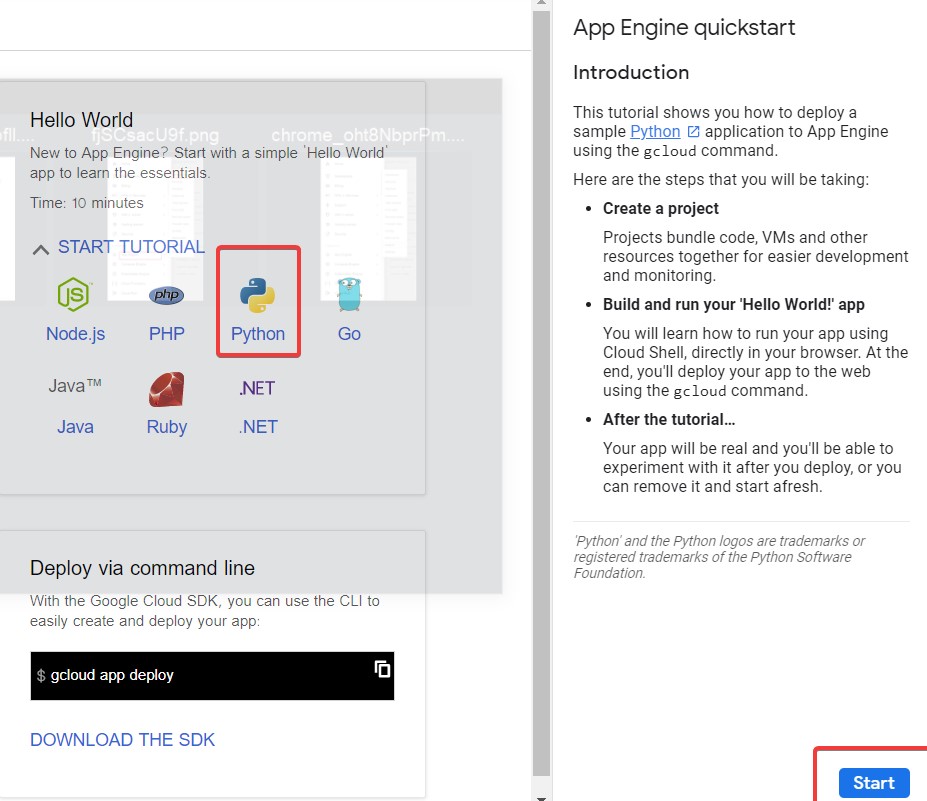
* Go to *IAM and admin*(highlighted) and click *manage resources*.



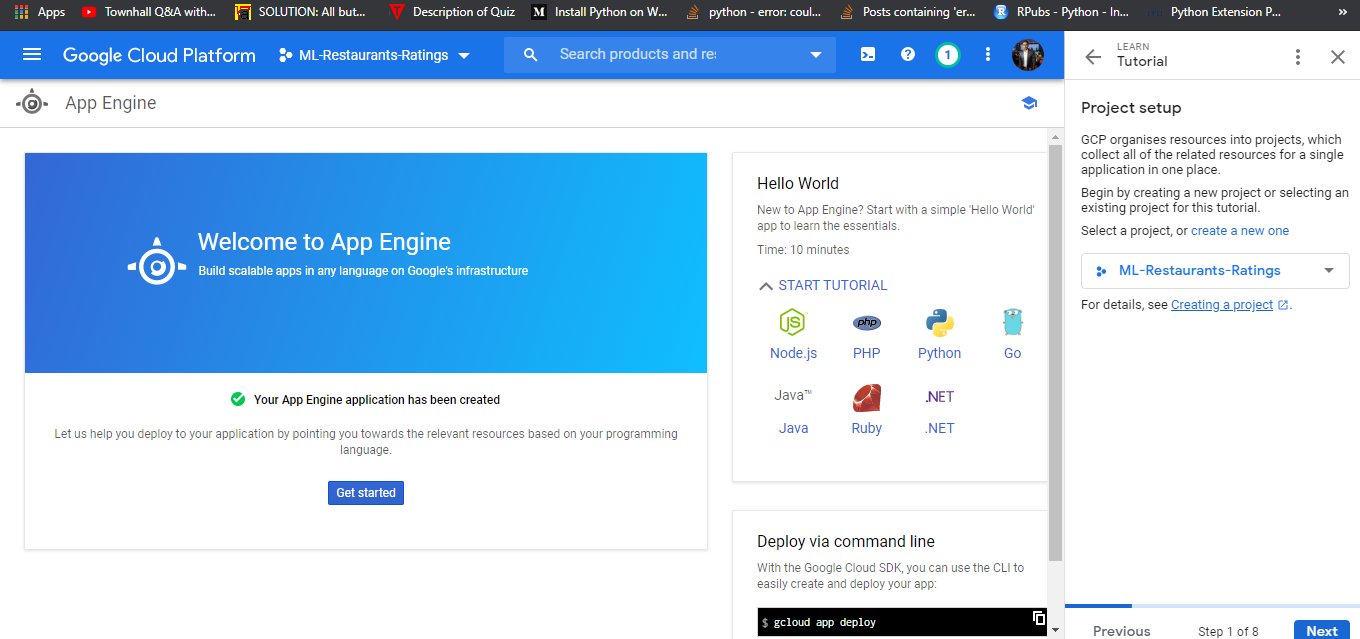
* Click *CREATE PROJECT* to create a new project for deployment.
* Once the project gets created, select *App Engine* and select *Dashboard.*



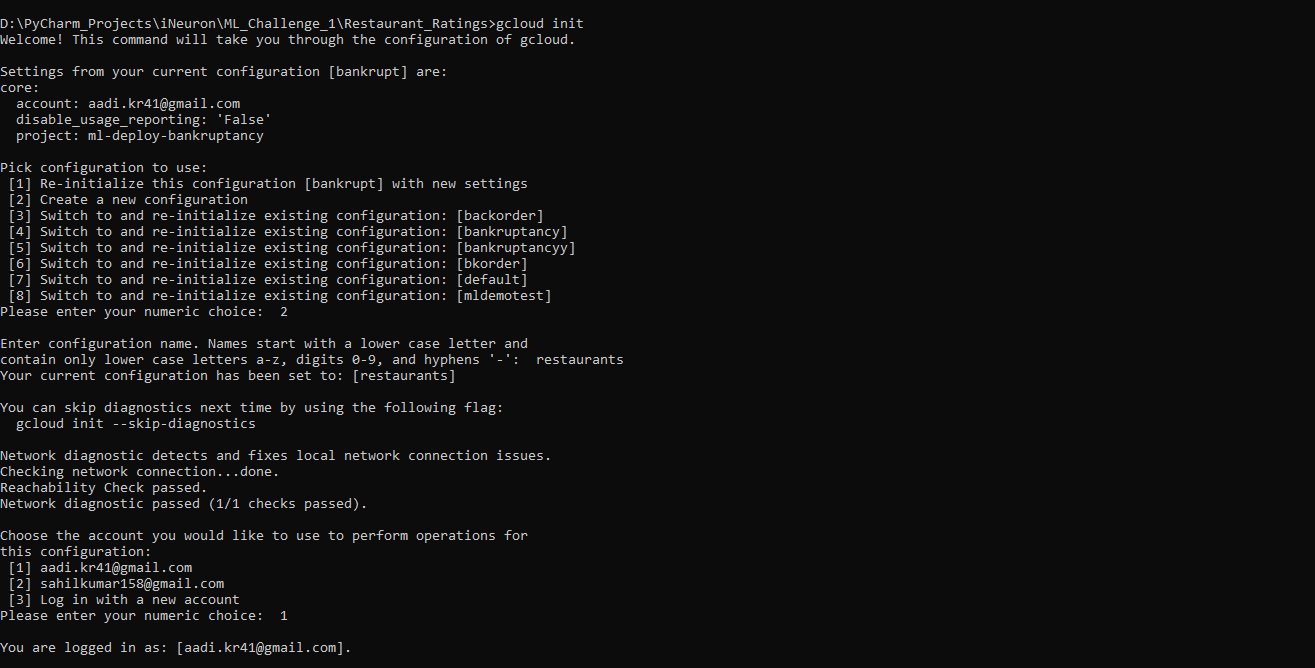
* Go to <https://dl.google.com/dl/cloudsdk/channels/rapid/GoogleCloudSDKInstaller.exe>to download the google cloud SDK to your machine.
* Click *Start Tutorial* on the screen and select Python app and click start.



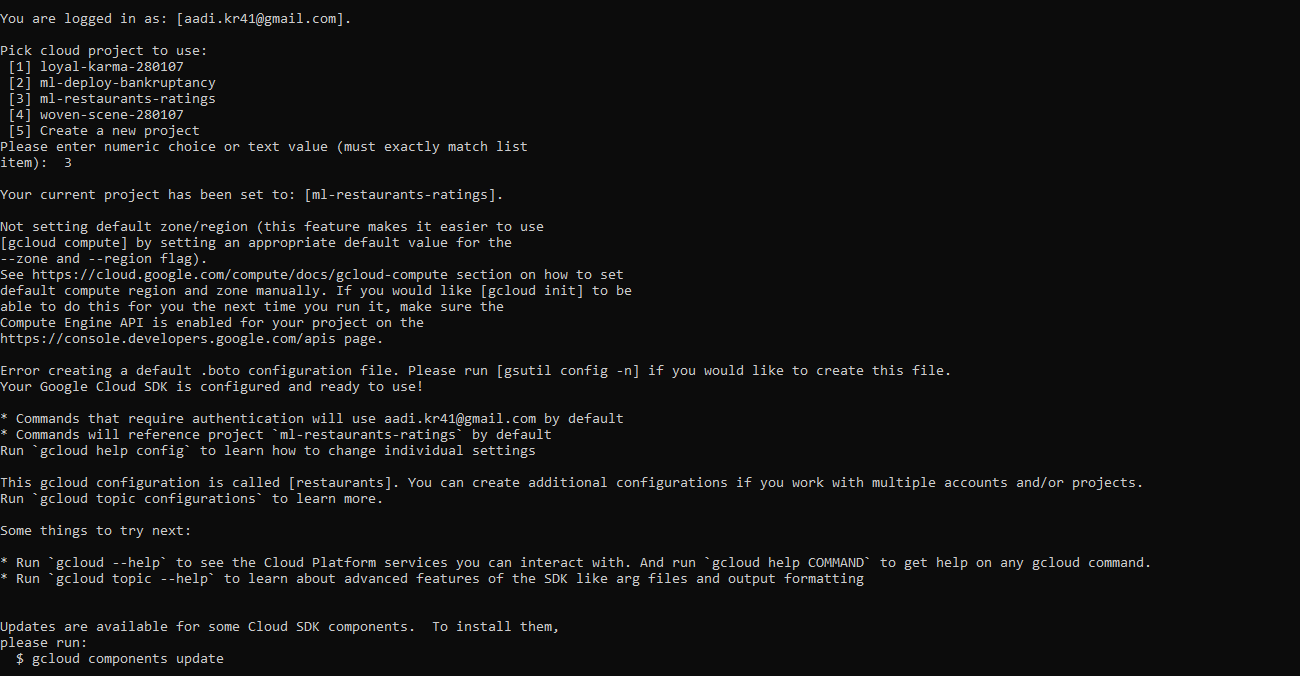
* Check whether the correct project name i.e ML-Restaurants-Ratings is displayed and then click next.



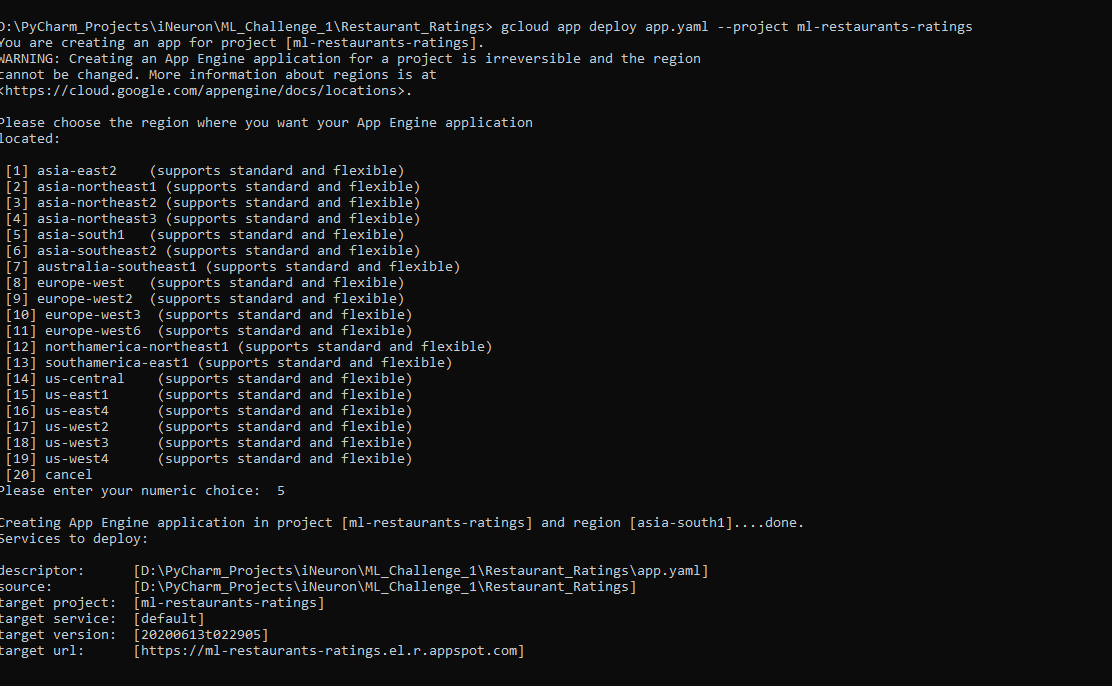
* Create a file ‘app.yaml’ and put ‘runtime: python37’ in that file.
* Create a ‘requirements.txt’ file by opening the command prompt/anaconda prompt, navigate to the project folder and enter the command ‘pip freeze > requirements.txt’. It is recommended to use separate environments for different projects.
* Your python application file should be called ‘main.py’. It is a GCP specific requirement.
* Open command prompt window, navigate to the project folder and enter the command *gcloud init* to initialise the gcloud context.
* It asks you to select from the list of available projects.

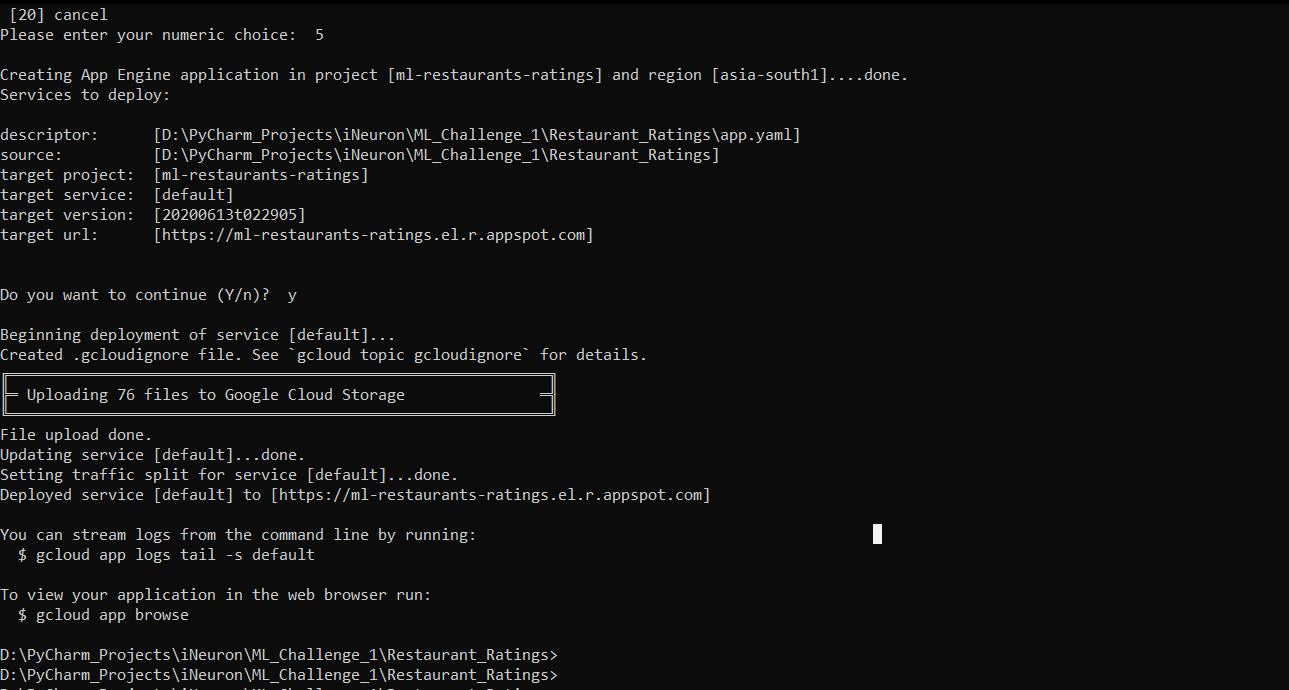


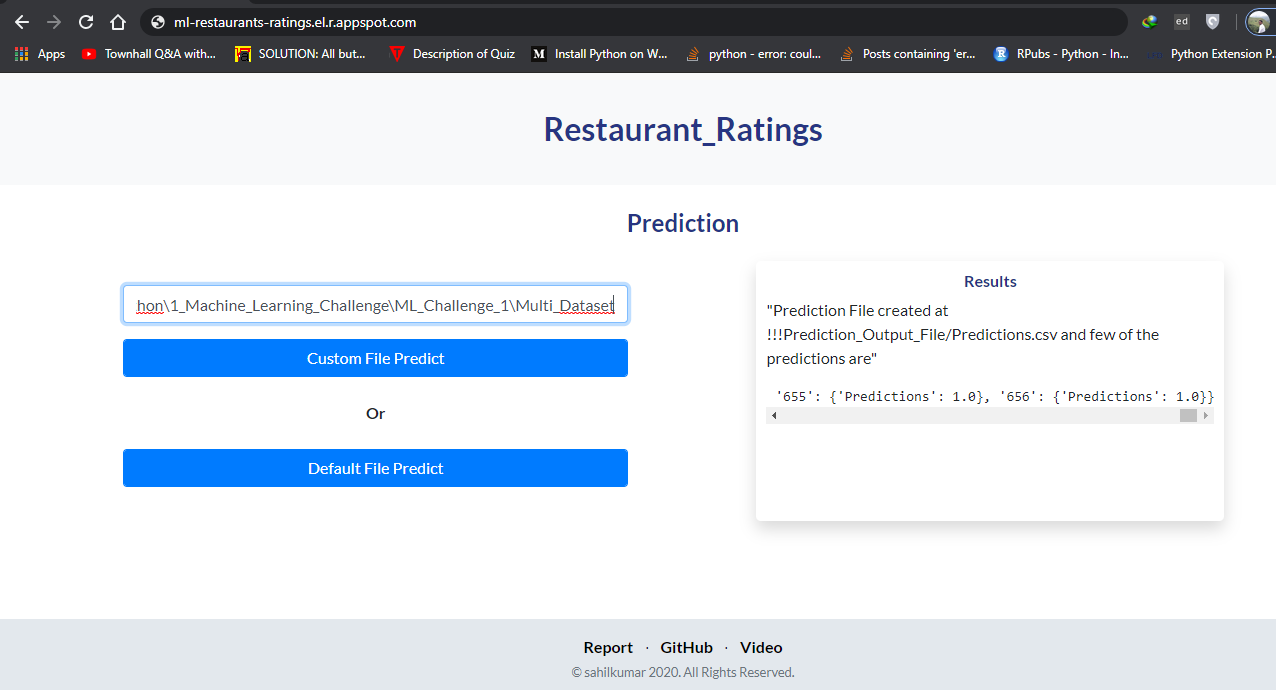
* I selected 2 option to create new configuration as restaurants.
* Select email id in which want to deploy the project.



* Once the project name is selected i.e ml-restaurants-ratings , enter the command gcloud app deploy app.yaml --project ml-restaurants-ratings.
* After executing the above command, GCP will ask you to enter the region for your application. Choose the appropriate one.



* I selected 5 option as asia-aouth1.
* GCP will ask for the services to be deployed. Enter ‘y’ to deploy the services.
* And then it will give you the link for your app, and the deployed app looks like:



* To save money, go to App Engine > settings and disable your app.

