iNeuron

Automated ML

Industrial Automation

Bharadwaja
Bhagyasree

Full Stack Data science Internship
ineuron

Overview

Objective

To create end to end automated machine learning solution where the user will only give dataset in recognizable formats and select the type of the problem, and the result will be the best performing hyper tuned machine learning model. The descriptive and graphical analysis of the data are also displayed.

- to use a given dataset for this project which is a Cassandra database.
- to use any cloud platform for this entire solution hosting like AWS, Azure or GCP.
- Logging is a must for every action performed by your code use the python logging library for this.

Benefits

- Automated machine learning (AutoML) helps to lower the barrier to entry for machine learning model building by streamlining the process thereby allowing non-technical users to harness the power of machine learning.
- Automated machine learning (AutoML) is the process of automating the time consuming, iterative tasks of machine learning. It allows data scientists and analysts to build machine learning models with efficiency while sustaining the model quality. The final goal of any AutoML solution is to finalize the best model based on some performance criteria.
- Traditional machine learning model development process is resource-intensive, requiring significant domain knowledge and time to produce and compare dozens of models. So we build an automated machine learning, where you will accelerate the time it takes to get production-ready ML models with great ease and efficiency.

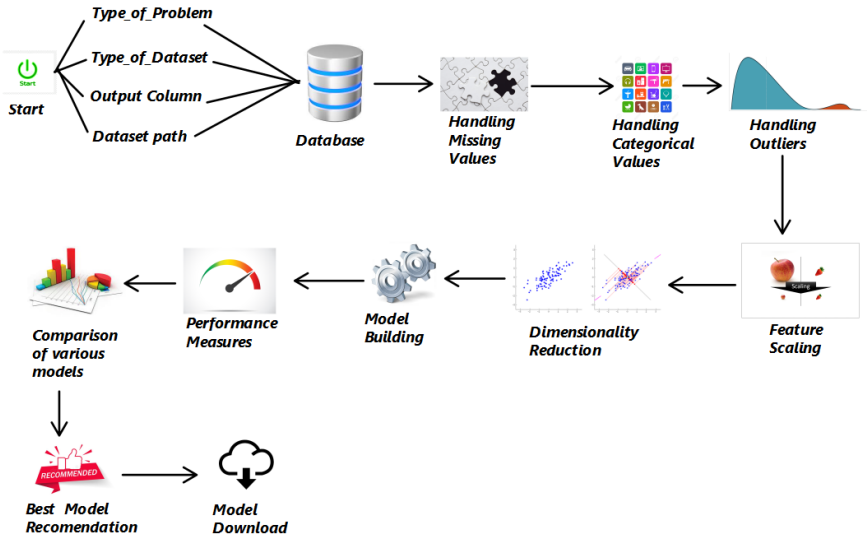
Tools

Figure: Tools used

Data Sharing Agreement

- The dataset should be in specified file formats , such as csv , xlsx , json , tsv , html , txt.
- Type of problem such as Regression , CLassification or Clustering should be specified.
- The output column for the dataset should be specified
- The dataset should be specified as url

Architecture



Data Insertion into Database

- Removing the special characters from the columns names using Regular Expressions ,because cassandra database won't allow special characters in column names.
- Custom table name is taken as type_of_problem + Number of Rows in the dataset.
- Checking the table name in the list of tables that exists in database.
- If table already exists then we will update the existing table.
- If table doesn't exists then we will create a new table with columns names as columns in dataset
- Then insert the data into Cassandra database

Export Data from Database

- While retrieving the data from the cassandra database , the retrieved data has columns not in the same order as inserted data.So we worked to get the retrieved data columns same as insterted data columns.This would help us in type casting data columns from cassandra data types to pandas data types.
- Later, type-casted the dataset columns and imported as pandas dataframe

Data Pre-processing

- Handled missing values or Nan values with Random sample imputation
- Removed those columns with object or int64 type categories greater then 20 unique categories
- For Regression and classification problems , Categorical features are handled with Target-Guided encoding.
- If classification output column is categorical,then handled it with unique values encoding.
- Lower boundary(LB) and Upper boundary(UB) are calculated with Inter Quartile Range(IQR) and those points outside these LB and UB are considered as outliers and removed.
  - $Q1 = 25thpercentile$, $Q3 = 75thpercentile$
  - $IQR = Q3 - Q1$
  - $LowerBoundary = Q1 - 1.5 * IQR$
  - $UpperBoundary = Q3 + 1.5 * IQR$
  - Data points $<$ Lower Boundary or Data points $>$ Upper Boundary are considered as outlier and are removed.

Data Transformation and Feature Extraction

- We applied Standardization or Z-score Normalization on the dataset to convert it with Mean = 0 and standard deviation = 1.This would help us while performing Feature Extraction.
- Principal Component Analysis is one of the way for Feature selection.Its working principle is based on Maximum variance. So we applied PCA for feature extraction.In order to find number of components , we used Knee method , (i,e Explained Variance Ratio and Number of components).

Model Training

- Various ML algorithms used for building different models
- Regression(Linear , Lasso , Ridge and ElasticNet) , Classification(Logistic , Random Forest , XGBoost) and Clustering(K-means , AffinityPropogation , Spectral Clustering , Agglomerative)
- Parameters grid is created for each model and given as input to GridSearchCV to get the hyper tuned parameters.
- Build different models using their respective hyper tuned parameters.

Model Training

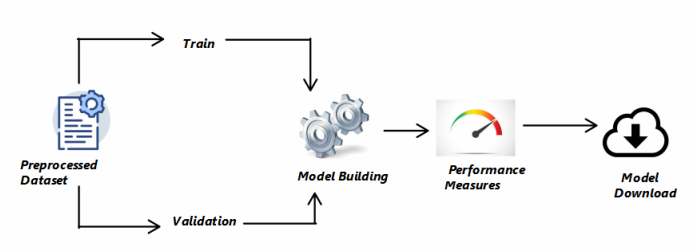


Figure: Model Training

Performance Measures and Model Recommendation

1. Regression : MSE , RMSE , MAE , R-squared ,Adjusted R-squared
2. Classification : Accuracy , Precision, Recall or Sensitivity , F1_Score, roc_auc_score.
3. Clustering : Silhouette_Score , calinski_harabasz_score , davies_bouldin_score
4. By comparing the performance measures of various models , we are recommending the best model.The performance measures are also displayed in the HTML page.

Test Cases

Test Case Description	Expected Result
1. Without file extension 2. Spelling mistake in file extension 3. File in other file format 4. File doesn't exist in specified path	Error while reading dataset. Either of the Reason Listed Below 1.Enter Dataset Path along with type of dataset extension 2.Check Spelling of type_of_dataset entered 3.Entered file may be in other File Format 4.Entered type_of_dataset File doesn't exists
Network issue or unable to connect to database	Error in connecting to cassandra Database. Following might be the reason 1.Failed to connect to database or Cluster 2.Check your Internet Connection or Network speed 3.Try to startover and resend your request
Regression 1. House Price Prediction Dataset 2. GPU performance Dataset	Performance measures such as MSE , RMSE , MAE , R-squared , Various models Download link , Highly recommended model
Classification 1. Titanic Dataset 2. Breast Cancer Dataset	Performance measures such as Accuracy , Precision, Recall or Sensitivity , F1_Score, roc_auc_score, Various models Download link , Highly recommended model
Clutering 1. CC General Dataset	Performance measures such as Silhouette_Score , calinski_harabasz_score , davies_bouldin_score , Various models Download link , Highly recommended model

Figure: Test Case

Test Cases

Verify whether the Application URL is accessible to the user	Application URL should be accessible to the user
Verify whether the Application loads completely for the user when the URL is accessed	The Application should load completely for the user when the URL is accessed
Verify whether user is able to successfully send request through HTML page	User should be able to successfully send request through HTML page
Verify whether user gets Get Best Model button to submit the inputs	User should get Get Best Model button to submit the inputs
Verify whether user is able to edit all input fields	User should be able to edit all input fields
Verify whether the recommended results are in accordance to best model	The recommended results should be in accordance to best model

Figure: Test Case

Q&A

- **Q.1 :** What is the source of data?
  - The data for training is provided in various file formats such as csv,xlsx,json,tsv,html and txt for building model.
- **Q.2 :** What was the type of data?
  - The data was the combination of numerical and Categorical values and it may also have missing values.
- **Q.3 :** How logs are managed ?
  - We are logging different log events as per the steps that we follow in validation and modeling like File validation log , Data Insertion ,Model Training log , prediction log etc.

Q&A

- **Q.4 :** 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