Technology	:	Machine Learning
Domain	:	Industrial Automation
Project	:	Automated ML

Low Level Design (LLD)
Automated ML

Document Version Control

Date Issued	Version	Description	Author
15-Aug-2021	1	Initial LLD - V1.0	Bharadwaja,Bhagyasree

Abstract

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.

Table of Contents

List of Figures

1. Introduction

1.1 What is Low-Level design document?

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual program code. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

1.2 Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work

2. Architecture

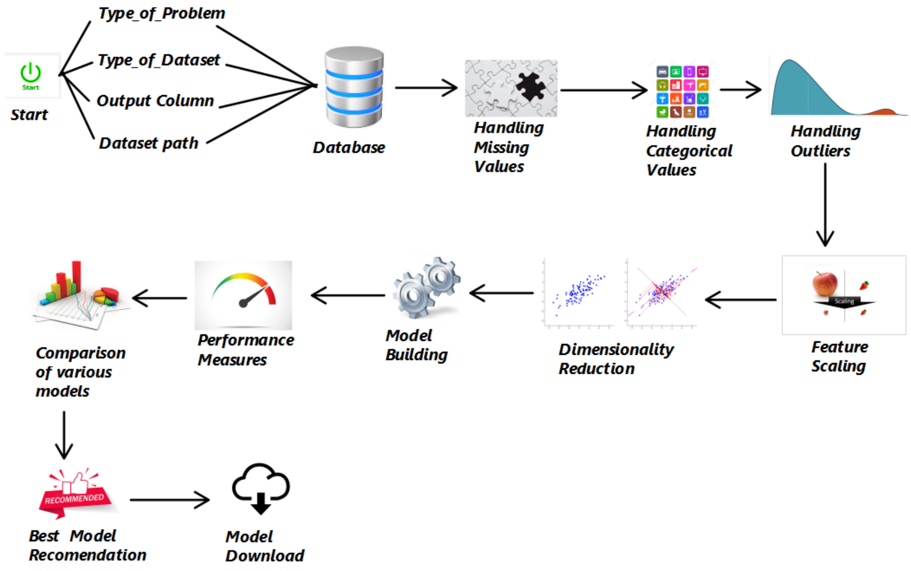


Fig. 1. Architecture

3. Architecture Description

3.1 Data Description

AutoML model performs various ML tasks such as Regression , Classification and clustering.So the dataset should according to the problem statement that we are performing.For supervised Learning problems(i,e Regression and classification) , the output column name should be specified before sending the request.But for clustering problems , we can either send Output column name as None or can be left empty.

3.2 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 coulmns names as columns in dataset
- Then insert the data into Cassandra database

3.3 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

3.4 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.

3.5 Data Transformation or Feature Scaling

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.

3.6 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).

3.7 Model Building

- 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.

3.8 Performance Measures or Model Evaluation

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

3.9 Model Recommendation

By comparing the performance measures of various models , we are recommending the best model.The performance measures are also displayed in the HTML page.

3.10 Data from user

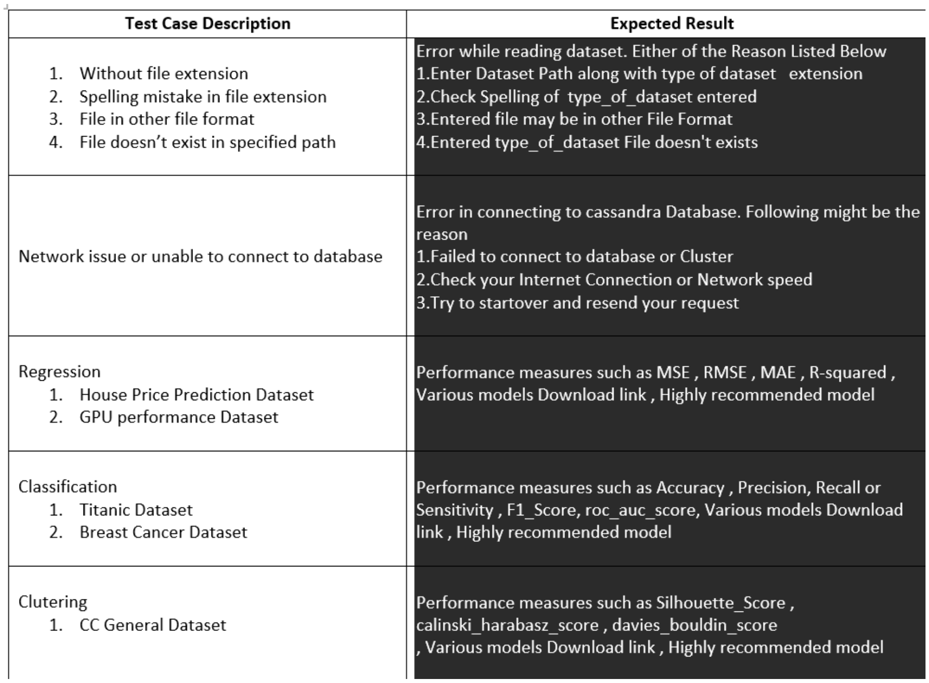
The user should give the following details as input in HTML page

1. Type_of_Dataset
2. Type_of_Problem
3. Output column name
4. Dataset path

3.11 Deployment

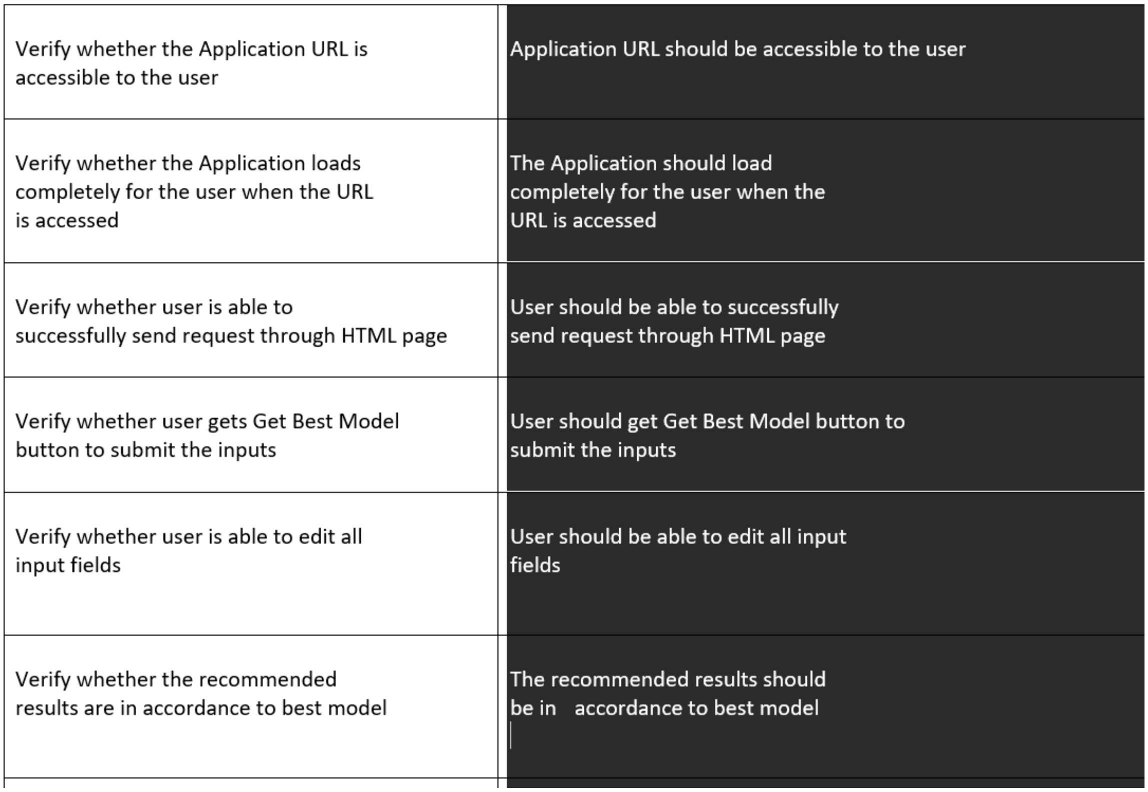
We will be Deploying the workflow in Amazon Web Services(AWS)

4. Unit 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

Fig. 2. Test case 1



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

Fig. 3. Test case 2