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

High Level Design (HLD)
Automated ML

Document Version Control

Date Issued	Version	Description	Author
15-Aug-2021	1	Initial HLD - 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

1 Introduction 3

1.1 What is High-Level design document? 3

1.2 Scope 3

1.3 Definitions 3

1.4 Overview 4

2 General Description 4

2.1 Product Perspective 4

2.2 Problem Statement 5

2.3 Proposed Solution 5

2.4 General Constraints 5

2.5 Assumptions 5

2.6 Tools used 6

2.7 Special Design aspects 6

2.8 Further Improvements 7

2.9 Data Requirements 7

3 Design Details 7

3.1 Process Flow 7

3.2 Training , Evaluation and Model Download 8

3.3 Event log 8

3.4 User Interface 8

3.5 Reports 8

3.6 Error Handling 8

3.7 Performance 9

3.8 Reusability 9

3.9 Application compatibility 9

3.10 Resource utilization 9

3.11 Security 9

3.12 Maintainability 9

3.13 Portability 9

4 KPIs (Key Performance Indicators) 10

5 Conclusion 10

List of Figures

Fig. 1 Tools used 6

Fig. 2 Processflow 7

Fig. 3 Training ,Evaluation and Download 8

1. Introduction

1.1 What is High-Level design document?

The purpose of this High Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.3 Definitions

Term	Description
AutoML	Automated Machine Learning
Database	A database is an organized collection of data
IDE	Integrated Developement Environment
AWS	Amazon Web Services
CQL	Cassandra Query Language
PCA	Principal Component Analysis

1.4 Overview

The HLD will :

- present all of the design aspects and define them in detail
- describe the user interface being implemented
- describe the hardware and software interfaces
- describe the performance requirements
- include design features and the architecture of the project
- list and describe the non-functional attributes like:
  - security
  - reliability
  - maintainability
  - portability
  - reusability
  - application compatibility
  - resource utilization
  - serviceability

2. General Description

2.1 Product Perspective

The Automated machine learning (AutoML) will be comprised of several different components.The language implemented will be dictated by it's purpose. The administrative and user interfaces will be using HTML and CSS to display the pages, and CQL to retrieve, insert, delete, and update the database. Python will be used to submit CQL commands for the automated part of the project such as updating and retrieving the user dataset.Model deployed using AWS , this setup will allow for multiple users to interact with the program at the same time.

2.2 Problem Statement

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.

2.3 Proposed Solution

The solution proposed here is AutoML , can be implemented to perform above mentioned use cases.In first use case to insert or retrieve data from cassandra database datastax astra is used.In second use case AWS is used to deploy the model.Explored various Machine Learning algorithms in

- Regression(Linear,Lasso,Ridge,ElasticNet)
- Classification(Logistic,Random Forest,XGBoost)
- Clustering(K-means,AffinityPropogation,Spectral,Agglomerative)

to build models and logging is performed.

2.4 General Constraints

The AutoML must be user friendly and as automated as possible. Users should not be required to do anything besides the initial requirements, and users should not be required to know any of the workings.To get best ML model, the user should select appropriate requirements, give a dataset link and click on Get best Model. The statistics , performance measures and model download link will be displayed in the results section.

2.5 Assumptions

This project is based on the idea of a Automating ML, and the goal of any AutoML solution is to finalize the best model based on some performance criteria. In doing so, many documents are created, and it is assumed that design flaws will be found early on. It is also assumed that all aspects of this project have the ability to work together in the way the designer is expecting. Another assumption is that the current intended documentation will suffice to make this project count towards the Software Engineering Subtract. There is also an assumption that none of the work or hardware will be stolen or sabotaged. The final assumption is that a demonstration and presentation will be possible at the end of the Final submission.

2.6 Tools used

1. PyCharm is used as IDE
2. For visualization of plots , Plotly is used
3. AWS is used for deployment of the model
4. Cassandra DB is used to retrieve , insert , delete and update the database
5. Front end Development is done using HTML/CSS
6. Python Flask is used for back-end development
7. GitHub is used as version control system

Fig. 1. Tools used

2.7 Special Design aspects

One special design aspect is the understanding that without significant modification, this system will only work when the int or float columns have no special characters and which leads them to object type column. This means model will give results , but it is not known how to handle each object column , remove special characters and to convert them to int or float type respectively.

2.8 Further Improvements

AutoML can be added with more use cases like anomaly detection and Time series problems. AutoML can be added with recent research trends in Machine Learning algorithms.

2.9 Data Requirements

- AutoML model accepts datasets in various file formats such as csv,xlsx,json,tsv,html and txt for building model.
- For Regression or Classification problems the output column name should also be give as input and for clustering either None or empty field can also be accepted.

3. Design Details

3.1 Process Flow

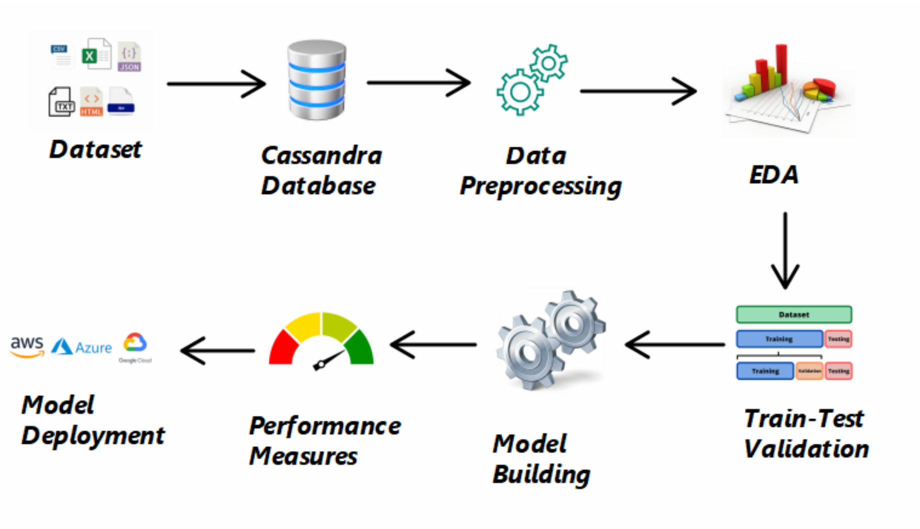

Fig. 2. Processflow

3.2 Training , Evaluation and Model Download

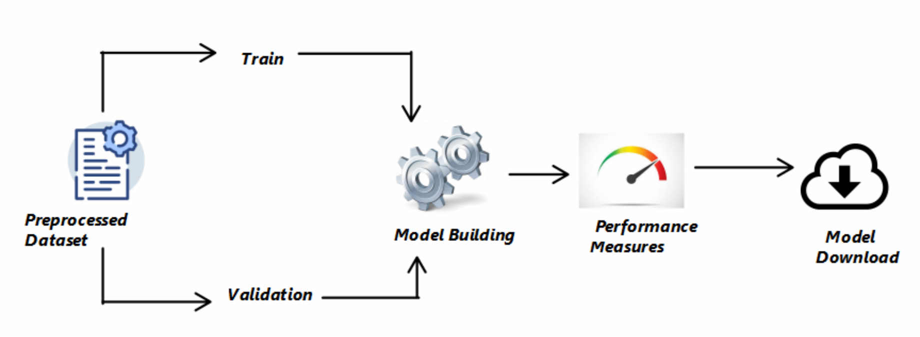

Fig. 3. Training ,Evaluation and Download

3.3 Event log

Event logs record events taking place in the execution of a system in order to provide an audit trail that can be used to understand the activity of the system and to diagnose problems. They are essential to understand the activities of complex systems, particularly in the case of applications with little user interaction (such as server applications).So we logged every event starting from Request till that request ends.

3.4 User Interface

The user interface is a very simple plain layout . It will display information very clearly for the user and will primarily output information to the user through HTML pages. The user need to give type of dataset , type of file , output column and dataset file path .

3.5 Reports

Performance Measures of various algorithms , recommendations for the best model and various models download link will be provided in results section.

3.6 Error Handling

Exception handling is the process of responding to the occurrence of exceptions during the execution of a program.In general, an exception breaks the normal flow of execution and executes a pre-registered exception handler.So along with built in exceptions, we also raised custom exceptions in order to handle the errors and makes the system more robust.

3.7 Performance

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. Hence the recommended model should be as accurate as possible. So we tried with various validation techniques, hyper tuning on various ML algorithms and compared the model's with their performance indices. Thus we tried to optimize the entire process to fetch us the best model.

3.8 Reusability

The code written should have the ability to be reused with no problems. Should time allow, and detailed instructions are written on how to create this project, everything will be completely reusable to anyone.

3.9 Application compatibility

The different processes for this project will be using Python as an interface between them. Each process will have its own task to perform, and it is the job of the Python code to ensure proper transfer of information.

3.10 Resource utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

3.11 Security

Because security is not the prime focus of this project, only the minimal aspects of security will be implemented. A host name and IP address will be required to log into an administrative log file. For now, all data will be sent in plain text.

3.12 Maintainability

Very little maintenance should be required and probability of performing a successful repair action within a given time.

3.13 Portability

Code and program portability should be possible between kernel-recompiled Linux distributions. For everything to work properly, all processes should be compiled from server.

4. KPIs (Key Performance Indicators)

- Key indicators displaying statistics , recommendations and summary
- Time and workload reduction by automating the time consuming, iterative tasks of machine learning.
- Event logs for every event along with host name and IP address

5. Conclusion

The designed AutoML will deliver the best model as we used various validation , hyper tuning and preprocessing techniques and there is no loss of data as data is already inserted into database.