EasyChair Preprint

№ 9111

Web Based Portal for Complete Data Engineering

Prashant Dwivedi, Shreyas Jadhav, Shruti Jadhav and Vidyadhari Singh

October 24, 2022

Web Based Portal for Complete Data Engineering

Prashant Dwivedi
Department of Computer Engineering, Thakur College of Engineering &Technology, Mumbai, India
prashantdwivedi194@gmail.com

Shreyas Jadhav
Department of Computer Engineering, Thakur College of Engineering &Technology, Mumbai, India
jshreyas12@gmail.com

Shruti Jadhav
Department of Computer Engineering, Thakur College of Engineering &Technology, Mumbai, India
shrutijadhav2845@gmail.com

Dr. Vidyadhari Singh
Department of Computer Engineering, Thakur College of Engineering &Technology, Mumbai, India
vidyadhari.rks@thakureducation.org

Abstract—***Machine Learning has become the need of today's world. Any intelligent system is incomplete without Machine Learning. People use ML models in various applications. Every ML-based project has to go through some of the steps during the development process. This process consists of certain common steps which every ML project needs to follow. The process of building ML models is tedious and time consuming. There are some common problems which people face while developing ML-based projects. That includes lack of resources and infrastructure, lack of uniformity in codebase and writing repetitive code for every project. These problems are very serious as building ML models needs a lot of effort and time. This paper aims to describe some of the problems that people face during the development of ML-based projects. It also provides a solution to reduce many problems and save efforts in the development process. The system proposed in this paper can save a lot of time and effort in building ML models. It will help to build projects faster and increase efficiency.***

Keywords—***Machine Learning, ML pipeline, ML models, Web based Portal.***

I. Introduction

Currently, to make any project involving machine learning or data science[4] one thing that comes at the forefront of the project is the data preprocessing[5] part. Without the proper data preprocessing, no project can be made that can give the desired results. Whenever, a database is picked up to work on, in its raw form it has many flaws, irregularities and redundant data as well. All these irregularities need to be overcome and the data must be made feasible to serve machine learning models or to make any future assumptions from this data. This process of removing the irregularities is nothing but the preprocessing.

Presently, the preprocessing part is carried out by using various coding techniques and making use of various data science libraries. Examples of such libraries are Pandas, Numpy, Matplotlib, etc. Below image shows the typical coding snap of any data preprocessing part.

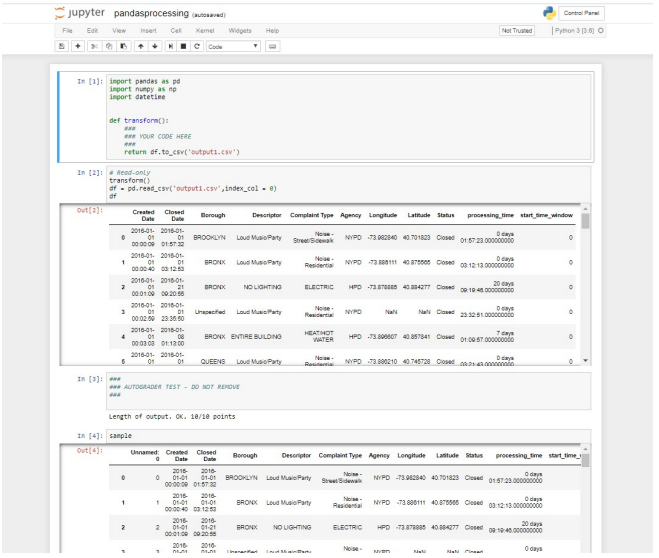

Figure 1: Typical Preprocessing Process

Any data cleaning or data preprocessing part involves few standard steps:

a. Exploratory Data analysis
b. Data Preprocessing
c. Feature Engineering

Presently, these steps are implemented by coding using the libraries as discussed earlier.

However, since these steps of data preprocessing are something very standard and need to be done every time you start with a fresh dataset. When one has to do it every time with every new project, then it feels the requirement of some easier process to do so. At the same time, for the people who do not want to code, or don't know how to code, for them this becomes a big hurdle to get the data done to start any fresh project.

Here, this portal is proposed, which can conduct all these steps of data preprocessing without using a single line code. A user of the portal has to just upload the data set and get his job done by just a few clicks.

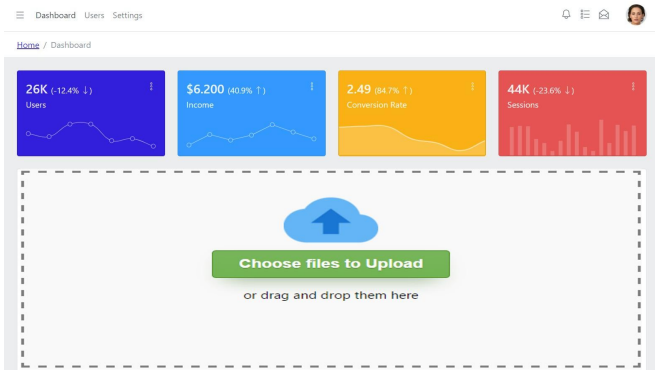

Figure 2: Upload the Database

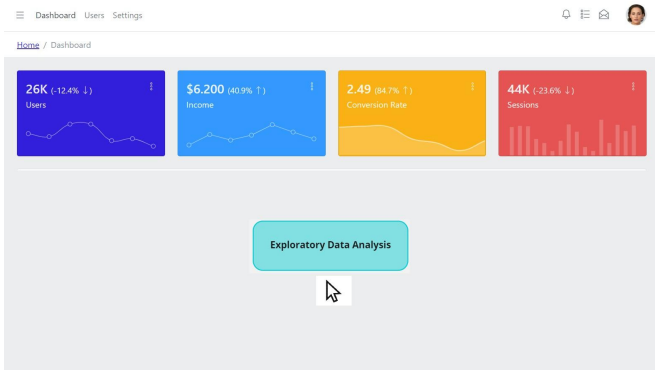

Figure 3: Select the Functionality

Such a portal will prove to be handy for the frequent coders who don't want to code every single time the same steps and also it's going to save a lot of time in the development of bigger projects.

II. Literature Survey

Machine learning is one of the popular domains on which people write articles, technical papers and books. In this literature people write about various architectures, different techniques to implement models, various machine learning algorithms and problems that people face during its implementation.

There are certain challenges that people face during practical implementation of machine learning models. In the research paper "Challenges In The Deployment And Operation Of Machine Learning In Practice" by Bair, Jöhren and Seebacher[1] , many such problems are discussed in detail. This paper contains a complete analysis of various challenges that ML practitioners face. That includes pre deployment issues like encryption of data, data format,etc and deployment issues like scalability and resource availability issues to train on big data. It also talks about some non technical challenges such as interpretability and transparency of complex machine learning models. It also suggests some solutions to solve some of the problems that are described in this paper.According to this paper, there are many problems which are unknown to ml theretitians. In order to get complete data they conducted interviews and surveys of industry experts who are working in this field for many years. It is focused on studying the cause of these problems in order to solve them.

According to Ashmore [2], for developing any ML-based project there are four stages:

- **Data management**, means collecting and transforming data into correct format
- **Model learning**, means learning features used in prediction
- **Model verification**, means testing accuracy of results from model
- **Model deployment**, means using a trained model in a production environment.[6]

There are many small steps under each of four stages. In the paper "Challenges in Deploying Machine Learning: a Survey of Case Studies" by Andrei Paleyes, Raoul-Gabriel Urma and Neil D. Lawrence [3], detailed description about various problems in each of these steps is mentioned. It also describes main causes for these problems.

There are more such papers and articles describing problems related to ml pipeline, workflow and deployment of models. It shows that developing and managing ML models is a difficult task and a lot of resources are used to do this task. There is a need for some generalized pipeline to reduce some of the problems and fasten up the process of developing and deploying ML models.

III. Problem Description

Every ML project goes through some common steps in the development phase. Some of the common steps are:

1. Gathering Data
2. Generating Hypothesis
3. Exploratory Data Analysis (EDA)
4. Data Preprocessing
5. Feature Engineering
6. Model Selection
7. Model Training and Evaluation
8. Hyperparameter Tuning
9. Model Deployment

Traditional way of writing code for these steps is very tedious. There is no uniformity in the code. People write repetitive code for various steps in the ML pipeline in every project. It is time consuming and requires a lot of resources to build. There is a need to automate some of the steps of this pipeline. This can save a lot of time and resources. People can build their project in less time and effort. It can boost efficiency and improve the quality of the project.

3

IV. Proposed System

The proposed system for the problems and hurdles discussed so far in the field of data engineering, would be this "Web Based Portal For Complete Data Engineering". The overview of features of this system can be seen pictorially as;

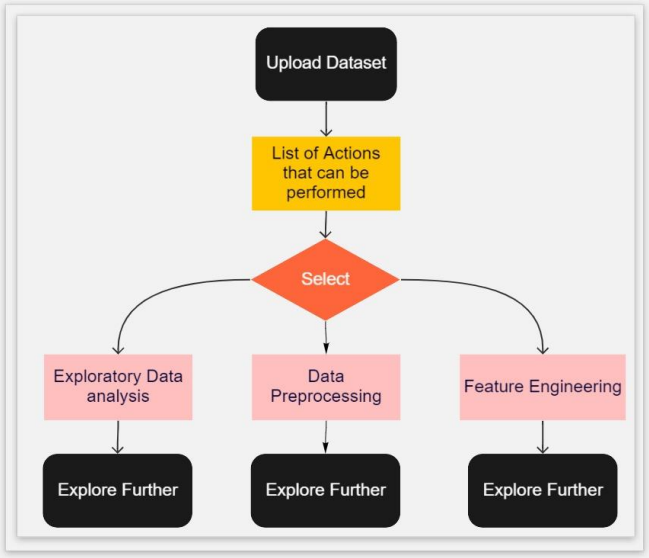

Figure 4: Features Overview

The detailed features of this system are as follows:

1. Users must login to the account.
2. Users should be able to load a dataset from source(File, Database, Cloud Storage).
3. Users should be able to choose a component from the below 3 categories.
    a. Exploratory Data analysis
    b. Data Preprocessing
    c. Feature Engineering
4. User should select any options like If user select Exploratory Data Analysis, the user will get
    a. All column details like missing values in specific columns, mode, median, mean, memory size, etc.
    b. Correlations between dependent and independent features.
    c. Missing values Plotting as well as % of missing values in each column.
    d. Sample of First rows and last rows
5. In EDA we have a lot of Graphs (count plot, scatter plot, box plot, etc) and we have to provide the ability to plot graphs based on the dataset. Eg: If a user wants to perform some specific task like outliers detection then here user will get an option to plot a scatter plot or box plot.
6. If the user select Data Preprocessing, the user will get 4 options:
    a. Data Cleaning
    b. Data Integration
    c. Data Reduction
    d. Data Transformation

    Example: If user select data cleaning
    - User will get a percentage (%) of missing values in each column
    - If in our columns we have 95% of missing values user should drop this column
    - If in our columns we have 40% of missing values and our columns are categorical columns, users can apply mode here and so on.
7. If the user select Feature engineering, the user will get 4 options
    a. Handling Imbalanced data
    b. Handling categorical data
    c. Based on categorical data, users should do the operations like One-Hot Encoding, Label encoding, Target encoding, etc.
    d. Users should also add a manual process to handling categorical data.

V. Conclusion

Machine Learning is being used in almost all the projects. There are some problems that people face while building, deploying and maintaining ML models.These problems are very common and serious. The portal described in this paper can help to reduce some of the problems that people face.The portal will be helpful to perform some important steps of the ML pipeline such as Exploratory Data Analysis, Data Preprocessing, Feature Engineering, etc with less effort and in less time. It will improve the speed of development and will reduce repetitive work.

VI. References

[1] Lucas Baier, Fabian Jöhren and Stefan Seebacher. "Challenges In The Deployment And Operation Of Machine Learning In Practice", 2019.

[2] Rob Ashmore, Radu Calinescu, and Colin Paterson. "Assuring the machine learning lifecycle: Desiderata, methods, and challenges", 2019.

[3] Andrei Paleyes, Raoul-Gabriel Urma and Neil D. Lawrence. "Challenges in Deploying Machine Learning: a Survey of Case Studies", 2020.

[4] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall. “Software Engineering for Machine Learning:
A Case Study” - 2019
[5] Cheng Fan, Meiling Chen, Xinghua Wang, Bufu Huang, Jiayuan Wang. “Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data” 2021
[6] Andrei Paleyes, Raoul-Gabriel Urma, Neil D. Lawrence. “Challenges in Deploying Machine Learning: a Survey of Case Studies” 2019