Project Title	Neuro Data Engineering
Technologies	Flask/ Django
Domain	Utility
Project Dificulties level	Advance

Problem Statement :

Create a web app application to perform Data cleaning, Feature engineering, and EDA. The web application should allow the user to perform various data transformation operations on the dataset with help of prebuild component. Users must be able to drag and drop the existing component at UI to perform any operation.

Example: Consider a component that can perform standard scaling on the dataset. Standard Scaling component required dataset as input and it will produce an output after applying standard scaling on a dataset called as Scaled Dataset. Refer to figure 1.

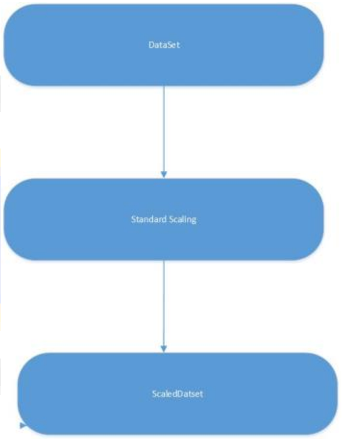

Figure 1

Requirement:

1. User must login to the account.

2. User should be able to load a dataset from source(File, Database, Cloud Storage).

3. User 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

    1. All column details like missing values in specific columns, mode, median, mean, memory size, etc.

    2. Correlations between dependent and independent features

    3. Missing values Plotting as well as % of missing values in each column

    4. 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

1. User will get a percentage (%) of missing values in each column

2. If in our columns we have 95% of missing values user should drop this column

3. If in our columns we have 40% of missing values and our columns are categorical columns user 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

   Example: If the user selects Imbalanced data

   1. User will get options to plot a graph with percentage of imbalanced data and based on graph user should be able to choose any methods like Undersampling, Oversampling, Resampling, etc.

   2. User should select an option to do under-sampling, Oversampling, Resampling

   If the user selects handling categorical data,

   a. user should add 50+ types of categorical data like state columns, Educational level (B.tech, M.tech, Bsc, Msc, etc.), Sex, Age, etc

   b. User should add 50+ common types of categorical data

   c. Based on categorical data user should do the operations like One-Hot Encoding, Label encoding, Target encoding, etc.

   d. User should also add a manual process to handling categorical data.

Note: Design other data preprocessing features like StandardScaling, MinMaxScaling, Dataset Splitter, String operation, Adding a new feature in the dataset based on the existing feature.

8. Every project has to be isolated from another project. Project checkpoint has to be saved so that user can continue their project later as well.

We have listed a detailed component that can be designed to build a tool to design various components:

1. Data flow task
1.1 Common Tasks:

1. Aggregate component: The Aggregate transformation is used to perform aggregate operations/functions on groups in a dataset. The aggregate functions available are- Count, Count Distinct, Sum, Average, Minimum, and Maximum. The Aggregate transformation has one input and one or more outputs.

2. Balanced Data Distributor: It takes a single input and distributes the incoming rows to one or more outputs uniformly via multithreading.

3. Conditional Split: The Conditional Split transformation can route data rows to different outputs depending on the content of the data. The implementation of the Conditional Split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified output. This transformation also provides a default output, so that if a row matches no expression it is directed to the default output.

4. Data Conversion: The Data Conversion transformation converts the data in an input column to a different data type and then copies it to a new output column.

5. Data Streaming Destination:

6. Derived Column: It is used to add a new column to the data pipeline.

7. HDFS File Destination: The HDFS File Destination component enables writing data to a HDFS file.

8. HDFS FIle Source: The HDFS File Source component enables to read data from a HDFS file.

9. Lookup component: The Lookup transformation performs lookups by joining data in input columns with columns in a reference dataset. You use the lookup to access additional information in a related table that is based on values in common columns.

10. Merge component: The Merge transformation combines two sorted datasets into a single dataset. The rows from each dataset are inserted into the output

based on values in their key columns.

11. Merge Join component: The Merge Join Transformation is used to perform Joins such as Inner Join, Left Outer Join, Full Outer Join, and Right Outer Join.

12. Multicast component: The Multicast transformation distributes its input to one or more outputs. This transformation is similar to the Conditional Split transformation. Both transformations direct an input to multiple outputs. The difference between the two is that the Multicast transformation directs every row to every output, and the Conditional Split directs a row to a single output.

13. Row Count component: Returns number of row in dataset
14. Script Component: Allow to write custom python script.
15. Sort component: The Sort transformation sorts input data in ascending or descending order and copies the sorted data to the transformation output. You can apply multiple sorts to an input;
16. Union All component: Union All Transformation is used to combine data from multiple sources (excel files, flat files, databases etc.). Or multiple SQL tables and produce one output to store in the destination table or file. Union All Transformation does not follow any particular order while merging the data and storing in the destination table or file.
17. Duplicate Resolver component: Remove duplicate row (You can specify column to identify row as duplicate).

1.2. Other transformations

1. Audit: The Audit transformation enables the data flow in a package to include data about the environment in which the package runs. For example, the name of the package, computer, and operator can be added to the data flow.

2. CDC Splitter (Change data Capture): The CDC splitter splits a single flow of change rows from a CDC source data flow into different data flows for Insert, Update and Delete operations.

3. Character Map: Character Map transformation applies string functions, such as a conversion from lowercase to uppercase for character data. This transformation works only on columns which have a string data type. As per the transformation output, it creates a new column or changes the converted data into the existing column.

4. Copy Column: The Copy Column transformation creates new columns by copying input columns and adding the new columns to the transformation output.

5. DQS Cleansing: The DQS Cleansing transformation uses Data Quality Services (DQS) to correct data from a connected data source, by applying approved rules that were created for the connected data source or a similar data source.
Export Column

6. Import Column: The Import Column transformation reads data from files and adds the data to columns in a data flow. For example, a data flow that loads data into a table that stores product information can include the Import Column transformation to import customer reviews of each product from files and add the reviews to the data flow.

7. Percentage Sampling: The Percentage Sampling transformation creates a sample data set by selecting a percentage of the transformation input rows. The sample data set is a random selection of rows from the transformation input, to make the resultant sample representative of the input.

8. Pivot: The Pivot transformation makes a normalized data set into a less normalized but more compact version by pivoting the input data on a column value.

9. Row sampling: The Row Sampling transformation is used to obtain a randomly selected subset of an input dataset.

10. Unpivot: The Unpivot transformation makes an unnormalized dataset into a more normalized version by expanding values from multiple columns in a single record into multiple records with the same values in a single column.

1.3. Other Sources

Excel Source
Flat File Source
Raw File Source
XML Source
MongoDB
Cassandra DB
SQLite
MYSQL
S3 bucket
Blob Storage
Oracle Database
CSV File
TSV File
Google Cloud Storage

1.4. Other destinations

Excel Source
Flat File Source
Raw File Source
XML Source
MongoDB
Cassandra DB
SQLite

MYSQL
S3 bucket
Blob Storage
Oracle Database
CSV File
TSV File
Google Cloud Storage

1.5: Other

1. Bulk Insert Task: Inserting massive dataset into database,
2. File System Task: Creating, Updating, Removing, Copying
3. Custom Script Task: Allow to write customer task
4. Send mail task:
5. Web service task: Allow to call web Service
6. for each Loop task: Allow to write loop statement

1.6 Descriptive Statistics:

Sr	Function	Description
1	count()	Number of non-null observations
2	sum()	Sum of values
3	mean()	Mean of Values
4	median()	Median of Values
5	mode()	Mode of values
6	std()	Standard Deviation of the Values
7	min()	Minimum Value
8	max()	Maximum Value
9	abs()	Absolute Value
10	prod()	Product of Values
11	cumsum()	Cumulative Sum
12	cumprod()	Cumulative Product

The describe() function computes a summary of statistics about the DataFrame columns.

1.7 Function Application:

- Table wise Function Application: pipe()
- Row or Column Wise Function Application: apply()
- Element wise Function Application: applymap()

1.8 Reindexing: Reindexing changes the row labels and column labels of a DataFrame.

1.9 Indexing & Description:

1 .loc()
Label based
2 .iloc()
Integer based
3 .ix()
Both Label and Integer based

2 Statistical Functions:

1. Percent_change:

Series, DatFrames and Panel, all have the function pct_change(). This function compares every element with its prior element and computes the change percentage.

Covariance, Correlation and Data Ranking

2. Window Functions
3. Rolling
4. Expanding

Technology:

- **Django**
- **NoSQL DB**
- **UI (HTML CSS, or Angular or React)**

Project Evaluation metrics :

Code:

- You are supposed to write a code in a modular fashion
- Safe: It can be used without causing harm.
- Testable: It can be tested at the code level.
- Maintainable: It can be maintained, even as your codebase grows.
- Portable: It works the same in every environment(operating system)
- You have to maintain your code on Github.
- You have to keep your GitHub repo public so that anyone can check your code.
- Proper readme file you have to maintain for any project development.
- You should include basic workflow and execution of the entire project in the readme file on GitHub
- Follow the coding standards: https://www.python.org/dev/peps/pep-0008/

Database:

- You are supposed to use a given dataset for this project which is a Cassandra database.
- https://astra.dev/ineuron

Cloud:

- you can use any cloud platform for this entire solution hosting like AWS, Azure or GCP

API Details or user interface:

- you have to expose your complete solution as an API or try to create a user interface for your model testing. Anything will be fine for us.

Logging:

- Logging is a must for every action performed by your code use the python logging library for this.

Ops Pipeline:

- If possible, you can try to use AI ops pipelining for project delivery Ex. DVC, Mlflow , segmaker , Azure machine learning studio, Jenkins, Circle CI, Azure DevOps , Tfx, Travis CI

Deployment:

- You can host your model in the cloud platform, edge devices, or maybe local, but with a proper justification of your system design.

Solutions Design:

- you have to submit complete solution design strategies in HLD and LLD document

System Architecture:

- You have to submit a system architecture design in your wireframe document and architecture document.

Latency for model response:

- you have to measure the response time of your model for a particular input of a dataset.

Optimization of solutions:

- Try to optimize your solution on code level, architecture level and mention all of these things in your final submission.
- Mention your test cases for your project.

Submission requirements:

High-level Document:

You have to create a high-level document design for your project. You can reference the HLD form below the link.

Sample link:

HLD Document Link

Low-level document:

You have to create a Low-level document design for your project; you can refer to the LLD from the below link.

Sample link

LLD Document Link

Architecture: You have to create an Architecture document design for your project; you can refer to the Architecture from the below link.

Sample link

Architecture sample link

Wireframe: You have to create a Wireframe document design for your project; refer to the Wireframe from the below link.

Demo link

Wireframe Document Link

Project code:

You have to submit your code Github repo in your dashboard when the final submission of your project .

Demo link

Project code sample link :

Detail project report:

You have to create a detailed project report and submit that document as per the given sample.

Demo link

DPR sample link

Project demo video:

You have to record a project demo video for at least 5 Minutes and submit that link as per the given demo.

Demo link

Project sample link :

The project LinkedIn a post:

You have to post your project detail on LinkedIn and submit that post link in your dashboard in your respective field.

Demo link

Linkedin post sample link :