**Machine Learning as a Platform (MLaaS)**

**Phase 1 - Requirement Document**



Table of Content

[Objective 6](#_Toc67070225)

[Scope 6](#_Toc67070226)

[Login 6](#_Toc67070227)

[User Actions 6](#_Toc67070228)

[Validations 6](#_Toc67070229)

[Navigation Panel 7](#_Toc67070230)

[Activities Timeline 8](#_Toc67070231)

[Validations 9](#_Toc67070232)

[Data Ingestion Module 9](#_Toc67070233)

[Create Dataset Page 9](#_Toc67070234)

[User Actions 9](#_Toc67070235)

[1. Upload Dataset 9](#_Toc67070236)

[2. Dataset Access 9](#_Toc67070237)

[3. Dataset List 9](#_Toc67070238)

[Validations 10](#_Toc67070239)

[Create Project Page 11](#_Toc67070240)

[User Actions 11](#_Toc67070241)

[1. Create Project 11](#_Toc67070242)

[Validations 11](#_Toc67070243)

[All Projects Page 12](#_Toc67070244)

[User Actions 12](#_Toc67070245)

[1. See Data Details 12](#_Toc67070246)

[2. Schema Mapping 12](#_Toc67070247)

[3. Delete Project 12](#_Toc67070248)

[Validations 13](#_Toc67070249)

[Data clean-up and Visualization Module 13](#_Toc67070250)

[Data Detail Page 13](#_Toc67070251)

[Schema Mapping Page 14](#_Toc67070252)

[User Actions 14](#_Toc67070253)

[1. Datatype 14](#_Toc67070254)

[2. Select column attribute 14](#_Toc67070255)

[3. Change column Name 14](#_Toc67070256)

[4. Save 14](#_Toc67070257)

[5. Reset 15](#_Toc67070258)

[Validations 15](#_Toc67070259)

[Data Exploration Page 15](#_Toc67070260)

[User Action 15](#_Toc67070261)

[1. Distribution graph 15](#_Toc67070262)

[2. Data Statistics 17](#_Toc67070263)

[Data Clean-up Page 18](#_Toc67070264)

[User Actions 18](#_Toc67070265)

[1. Select column 18](#_Toc67070266)

[2. Clean-up operations 18](#_Toc67070267)

[Missing values handling 18](#_Toc67070268)

[Noise Handling 22](#_Toc67070269)

[Transformation 23](#_Toc67070270)

[3. Scale and Split 29](#_Toc67070271)

[Scaling 29](#_Toc67070272)

[Split 30](#_Toc67070273)

[4. Modelling 30](#_Toc67070274)

[Validations 30](#_Toc67070275)

[Data Modelling 31](#_Toc67070276)

[Modelling Type 31](#_Toc67070277)

[User Action 31](#_Toc67070278)

[1. Select Modelling Mode 31](#_Toc67070279)

[2. Select modelling Type 31](#_Toc67070280)

[3. Start/Stop modelling 32](#_Toc67070281)

[4. Create Experiment 32](#_Toc67070282)

[Validations 32](#_Toc67070283)

[Auto Modelling 32](#_Toc67070284)

[Learning curves 33](#_Toc67070285)

[Feature importance 34](#_Toc67070286)

[Performance Metrics 34](#_Toc67070287)

[Model summary 35](#_Toc67070288)

[Actual Vs Prediction 36](#_Toc67070289)

[Compare Experiment 37](#_Toc67070290)

[Validations 38](#_Toc67070291)

[Manual Modelling 38](#_Toc67070292)

[Split Dataset 39](#_Toc67070293)

[Select Model 40](#_Toc67070294)

[Hyperparameters 41](#_Toc67070295)

[Validations 43](#_Toc67070296)

[Formulas/Logic used 43](#_Toc67070297)

[1. Distribution Graph 43](#_Toc67070298)

[2. Datatype Determination 43](#_Toc67070299)

[3. Upload Dataset 43](#_Toc67070300)

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Revised by** | **Date** | **Reason for Changes** | **Version** |
| Swati Srivastava | 12/11/2020 | First Draft | 1.0 |
| Swati Srivastava | 12/29/2020 | 1. Added requirements for **Schema Mapping** and **Data Exploration** page  2. Updated the validation for **Create Dataset** page | 2.0 |
| Swati Srivastava | 01/04/2021 | Added requirements for **Login** Page and **Activities Timeline** | 3.0 |
| Swati Srivastava | 01/19/2021 | 1. Added requirements for **Data Exploration**, **Data Visualization** and **Data clean-up** page  2. Updated validation for **Schema Mapping** page | 4.0 |
| Swati Srivastava | 01/27/2021 | Added Save/Save As functionality for **Schema Mapping** and **Data Clean-up Page** | 5.0 |
| Swati Srivastava | 02/08/2021 | 1. Added requirements for **Data Modelling** module  2. Removed **Data visualization** page details | 6.0 |
| Swati Srivastava | 02/17/2021 | 1. Removed Save/Save As functionality from **Schema Mapping** page  2. Updated the Data Clean-up requirements  3. Updated the screen design of Modelling page  4. Updated the screen design of All Projects page | 7.0 |
| Swati Srivastava | 03/10/2021 | 1. Added details about the clean-up operations | 8.0 |
| Swati Srivastava | 03/19/2021 | 1. Added details about the manual modelling and added validation for Compare Experiments | 9.0 |

## Objective

Objective of this POC is to develop a platform for Machine Learning as a Service. This platform will offer services for data clean-up, data visualization, and predictive analytics with machine learning algorithm.

## Scope

Scope of this POC is to develop an end-to-end pipeline to let users-

1. Upload the dataset in CSV format
2. Clean-up the data
3. Perform data exploration
4. Define data splitter parameter
5. Predict the data for target column

## Login

### User Actions

1. This page will let user login to the application by entering registered username and password

2. This page also provides feature to regenerate the password if the user forgets password.

### Validations

* Application should show error message when user enters invalid username or password

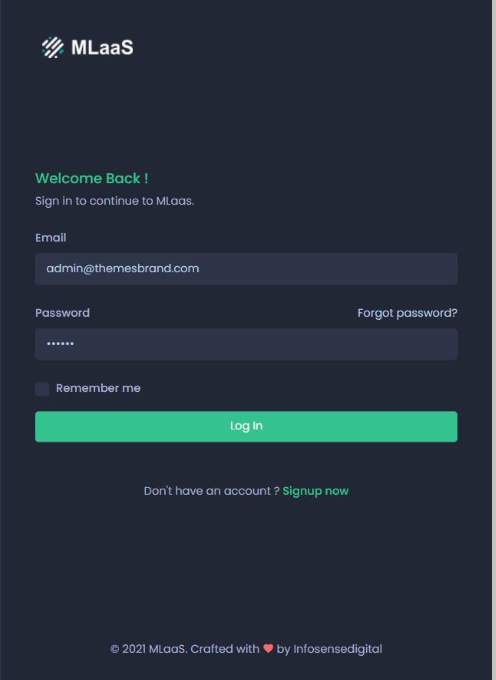


Figure 1: Login Page Mock-up

## Navigation Panel

* A navigation panel is provided at the left side of the application
* The panel will show the links of all the pages. User can navigate to the desired page by clicking the link
* The navigation panel will highlight the current page link
* The pages where user is not eligible to navigate at any given point of time will be shown as inactive link

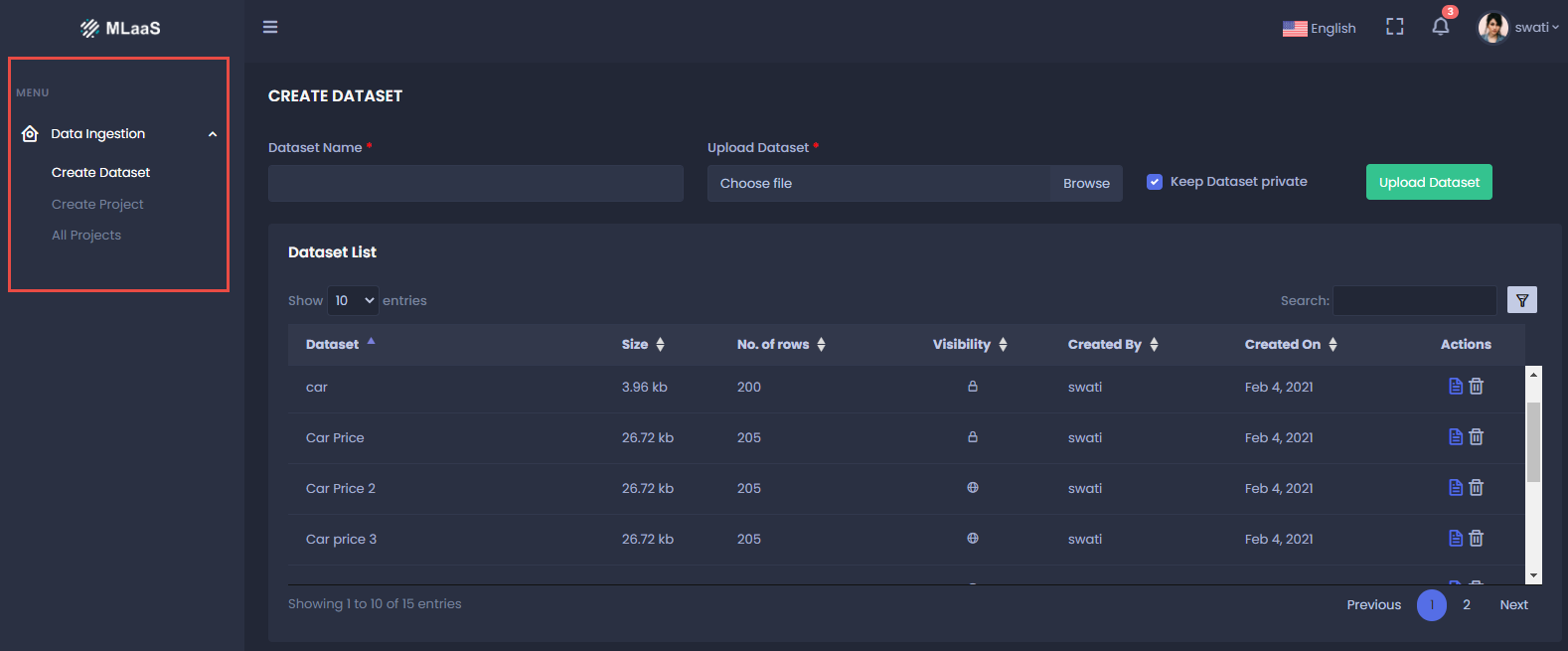


Figure 2: Navigation Panel Mock-up Screen

## Activities Timeline

Timeline panel of the application shows the history of recent activities performed by user.

Following are the activities and description to be shown for Data ingestion module-

|  |  |  |
| --- | --- | --- |
| **Sr. no** | **Activities** | **Description** |
| 1 | Created dataset | You have created dataset “ABC” |
| 2 | Deleted dataset | You have deleted dataset “ABC” |
| 3 | Created project | You have created project “ABC” |
| 4 | Deleted project | You have deleted project “ABC” |
| 5 | Column name updated | Column name “X” is updated to “Y” |
| 6 | Selected Target Column | You have selected columns X and Y as Target Columns |
| 7 | Ignored Column | You have ignored Columns “X, Y, and Z”. These columns will not be considered for the experiment |
| 8 | Created new dataset (Save As functionality) | You have saved new dataset “Updated Name” |

### Validations

* Activities timeline should appear on all the pages of application.

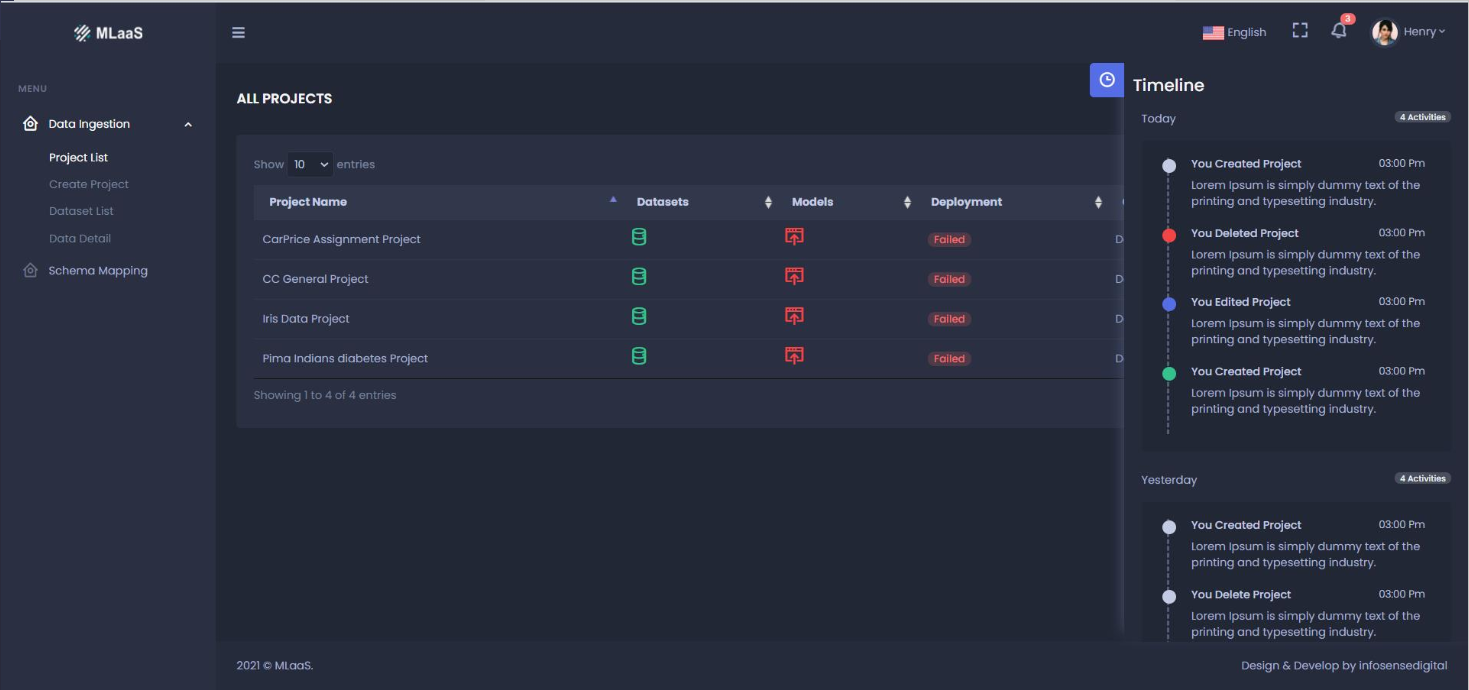


Figure 3: Activities Timeline Mock-up

## Data Ingestion Module

### Create Dataset Page

Create Dataset page will allow users to upload dataset in MLaaS page. These datasets can be used for multiple projects

#### User Actions

##### 1. [Upload Dataset](#UploadDataset)

To upload a dataset, user need to-

* Enter Dataset name
* Enter the description of Dataset
* Click **Choose file** to select the dataset from local system
* Click **Create Dataset** button to upload the dataset

##### 2. Dataset Access

User will get an option to either keep the dataset as **Public** or **Private**

##### 3. Dataset List

Once the dataset is uploaded, following details will be shown in Dataset List-

* Name of the dataset
* Description of the dataset
* No. of rows in the dataset
* Size of the dataset
* Date and time the dataset was uploaded on
* Name of the user who has uploaded the dataset
* User will get an option to remove the Dataset

#### Validations

* Datasets which are already being used by projects cannot be deleted.
* Application should show error message when user tries to delete a dataset being used by project
* User can only delete the dataset uploaded by him/her
* Confirmation window should appear when user deletes a dataset
* Users are only allowed to upload CSV dataset
* Application should not let user upload the CSV files with junk data.

Junk Data –

* Blank CSV is not allowed
* CSV with single column is not allowed
* System should not let user create datasets with duplicate name (Same as existing datasets)

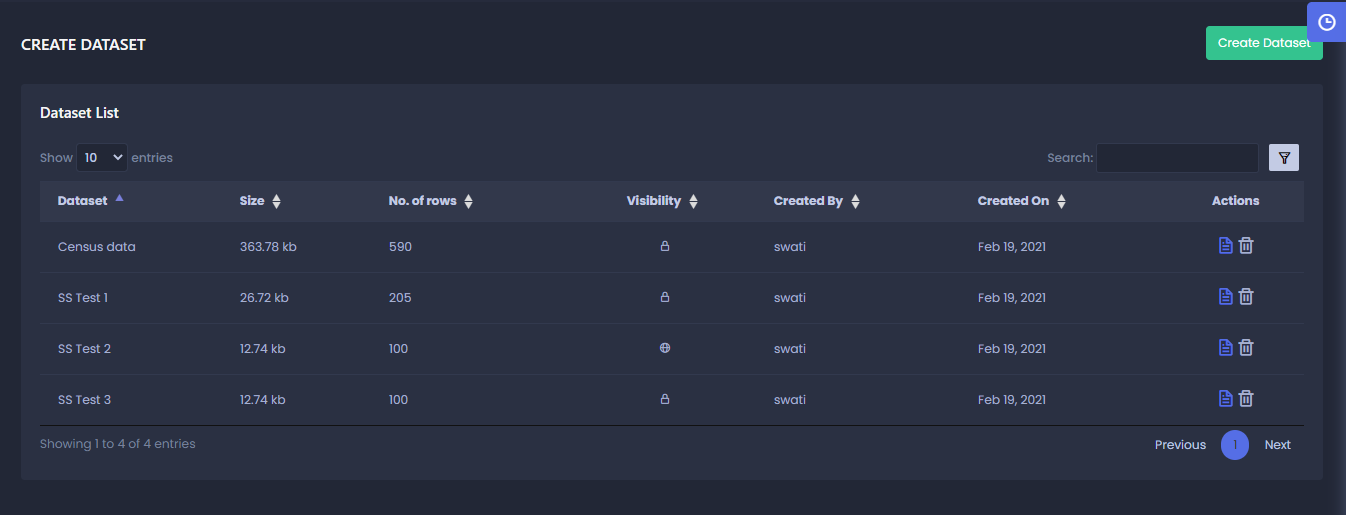


Figure 4: Create Dataset Mock-up Screen

### Create Project Page

This page let users create new projects and add dataset to the project.

#### User Actions

##### 1. Create Project

To create a project, user need to-

* Enter Project Name
* Provide a brief description of the project
* Add dataset to the project. A dataset can be added by-
  + Uploading a new dataset by clicking **Choose File,** defining the access- Private or Public**,** and provide the name of dataset
  + Selecting an existing dataset from the **Select Dataset** dropdown list
* Once the dataset is uploaded, click **Create Project**

#### Validations

* Select Dataset Dropdown list should display all the public dataset listed on the **Create Dataset** page and datasets private to the user
* On clicking **Create Project**, user should be directed to [**All Projects**](#AllProjectPage) page with the new project added to the list
* System should not let user create projects with duplicate name (Same as existing projects)

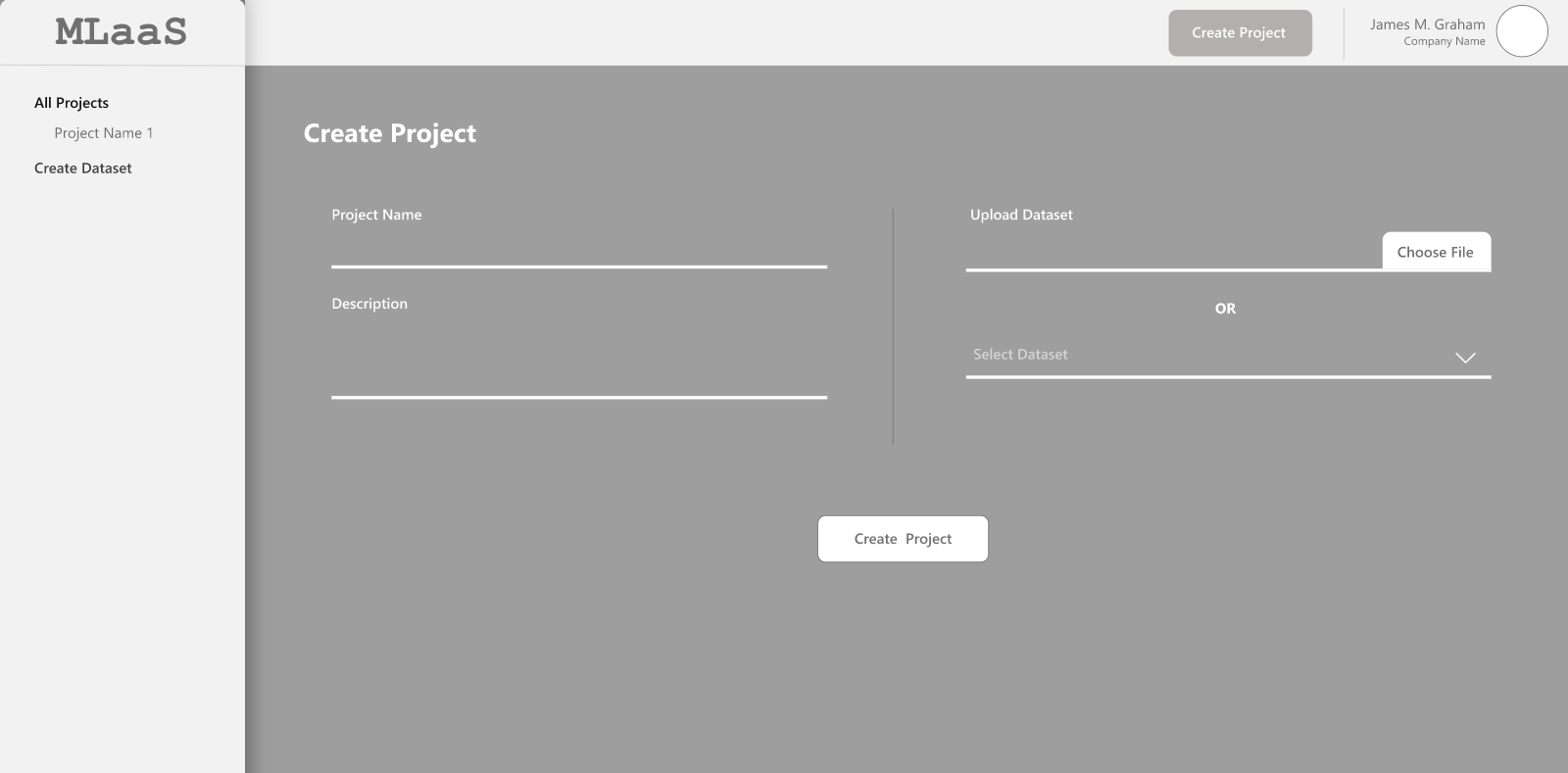


Figure 5: Create Project Mock-up Screen

### All Projects Page

This page shows all the existing projects in the system and its status. Following details are shown in this page-

* **Project Name** – Name of the Project
* **Dataset** - Shows the name of Dataset-
* **Model**- Shows the training status of the model
  + **Green** denotes a successfully trained model
  + **Yellow** denotes training in progress
  + **Red** stands for "Training Failed" status for a model.

On clicking the Model icon, user will be directed to Modelling page

* **Deployment** – Shows thedeployment status of the project-
  + "Active" deployment
  + "Suspended" deployment
* **Created** – Project creation Date and Time
* **Actions** – User can perform following actions by clicking respective icons-
  + Delete the project
  + Go to Schema Mapping page
  + Go to Data Detail page

#### User Actions

##### 1. See Data Details

User can see the details of the dataset uploaded for the project by clicking **See Detail** icon under Actions Tab.

Clicking this icon will direct the users to [**Data Detail**](#DataDetailPage) page.

##### 2. Schema Mapping

User can click the Schema Mapping icon to go to the Schema Mapping page of the dataset

##### 3. Delete Project

To delete a project, user need to click the **Delete** icon provided under the Actions tab

#### Validations

* Confirmation window should appear when user deletes a project
* User cannot delete a project if it’s deployment is going on. In such scenario, if user tries to delete a project, system should provide an error message.
* User cannot delete a project if the model training is in progress. System should display error message in such scenario
* When user deletes a project, its dataset will not get deleted

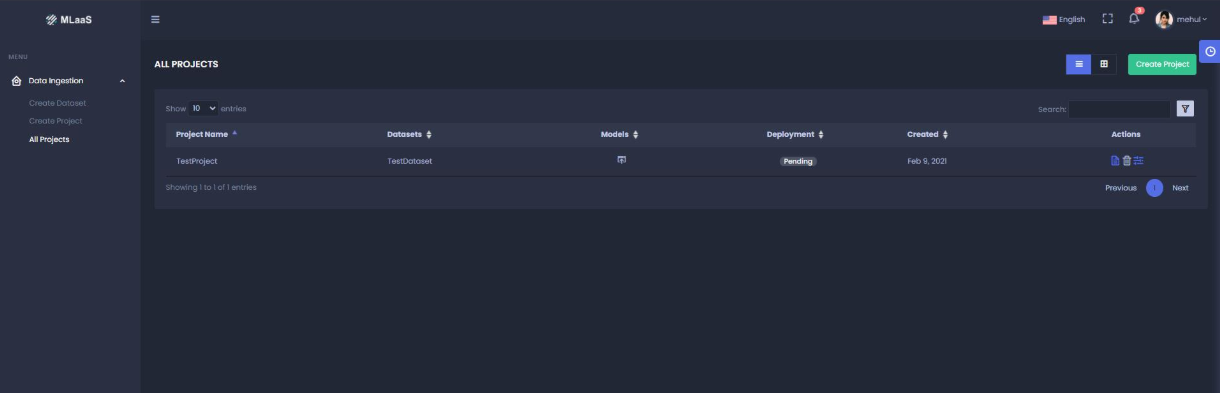


Figure 6: All Projects Mock-up Screen

## Data clean-up and Visualization Module

This Module is comprised of the processes of fixing/removing incorrect, corrupted, duplicate, or incomplete data within a dataset and provide the graphical representation of information.

### Data Detail Page

This page shows the raw data of the dataset uploaded for the project

User can navigate through the rows using page navigation and columns using scroll bar.

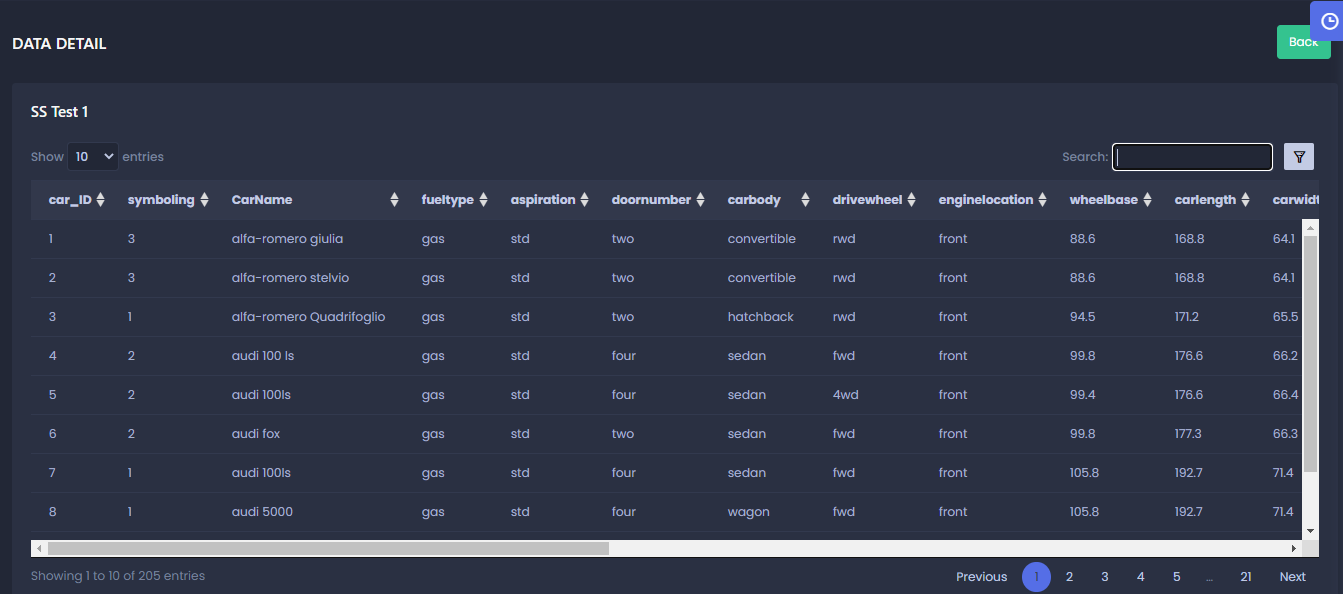


Figure 7: Data Detail Page Mock-up Screen

### Schema Mapping Page

#### User Actions

##### 1. [Datatype](#_2._Datatype_Determination)

Schema Mapping page shows the datatype of each column. Datatype can be-

* Categorical
* Categorical List
* Numerical
* Text
* Timestamp

##### 2. Select column attribute

User can select attributes for each column.

* **Target Column** -user can select a column as Target column on which the prediction is to be done
* **Ignore**- User can select Ignore for the columns to be not considered for the experiment

##### 3. Change column Name

User can choose to update the name of the column from this page.

##### [4. Save](#SaveInSchemaMapping)

* **Save** - User can click **Save** to save the changes and proceed with Dataset clean-up

##### 5. Reset

* On clicking Reset, application will initialize all user-input values

#### Validations

* Multiple columns can be selected as Target Column
* User cannot change the datatype
* Column Name and Changed Column Name cannot be same
* The target columns should get highlighted
* It is NOT mandatory to select the Target column.

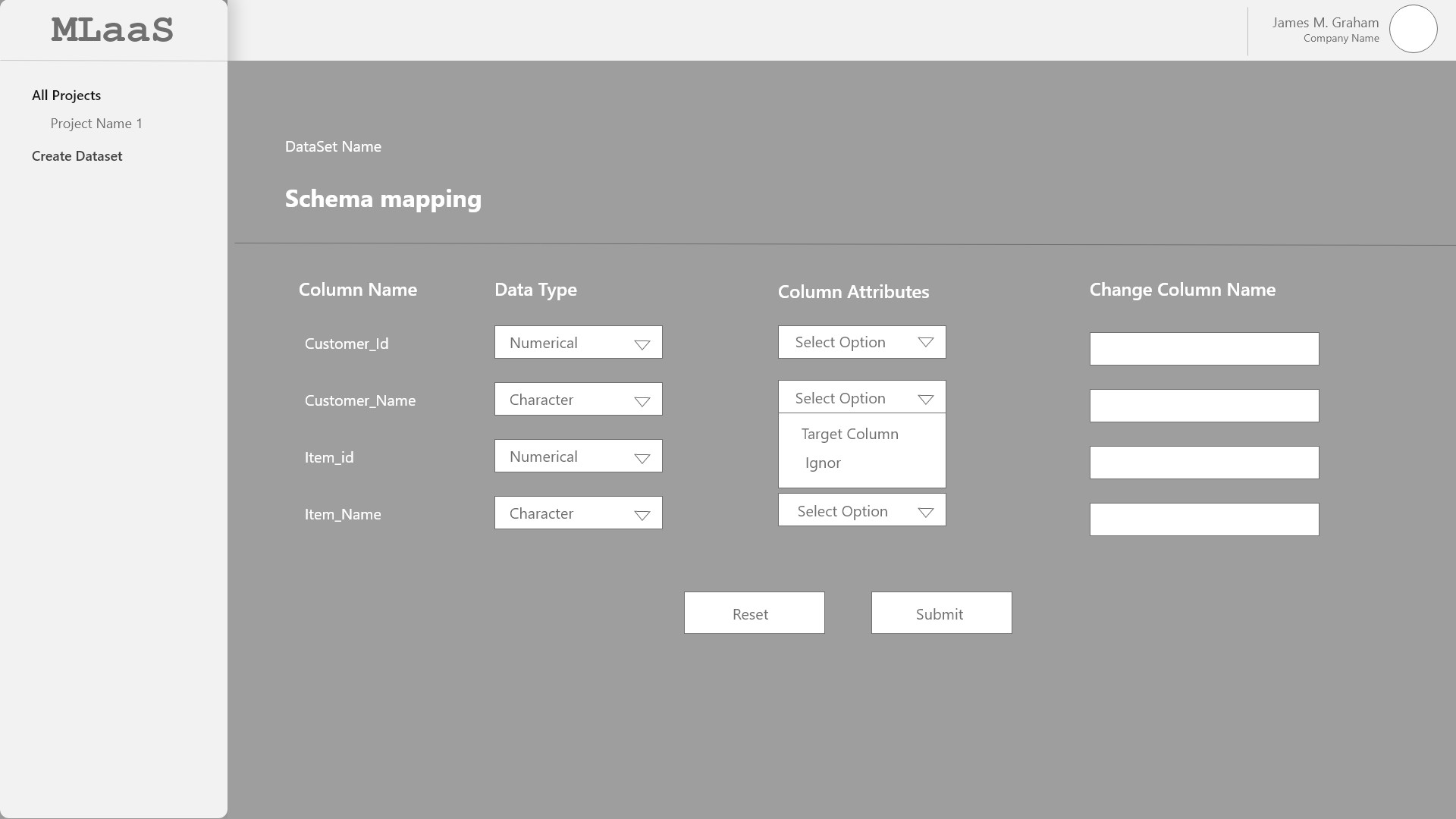


Figure 8: Schema Mapping Mock-up Screen

### Data Exploration Page

This page shows all the columns selected from the data Schema page and its statistics

#### User Action

##### 1. [Distribution graph](#_1._Distribution_Graph)

User can see the distribution graph for each of the column on this page. These distributions show the spread (dispersion, variability, scatter) of the data.

* For continuous datatype, following graphs are shown-
* Histogram



Figure 9: histogram Plot Mock-up

* Box plot

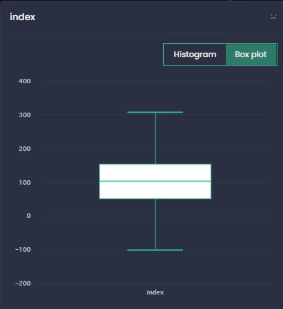


Figure 10: Box Plot Mock-up

Both Histogram and Box plot shows the outliers as well.

* For categorical datatype, Count plot is shown

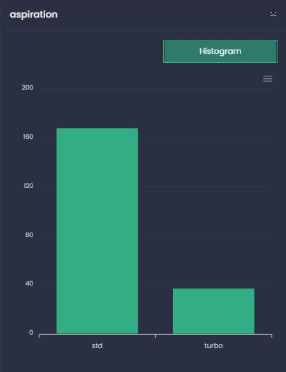


Figure 11: Count Plot Mock-up

##### 2. Data Statistics

User can see the data statistics based on the type of column-

For Continuous column-

* Data Type
* Data Count
* Missing data
* Mean value
* Min Value
* Max Value
* Std Deviation
* Unique Data

For Categorical column-

* Data Type
* Data count
* Missing Data
* Most Occurring
* Least occurring
* Unique data

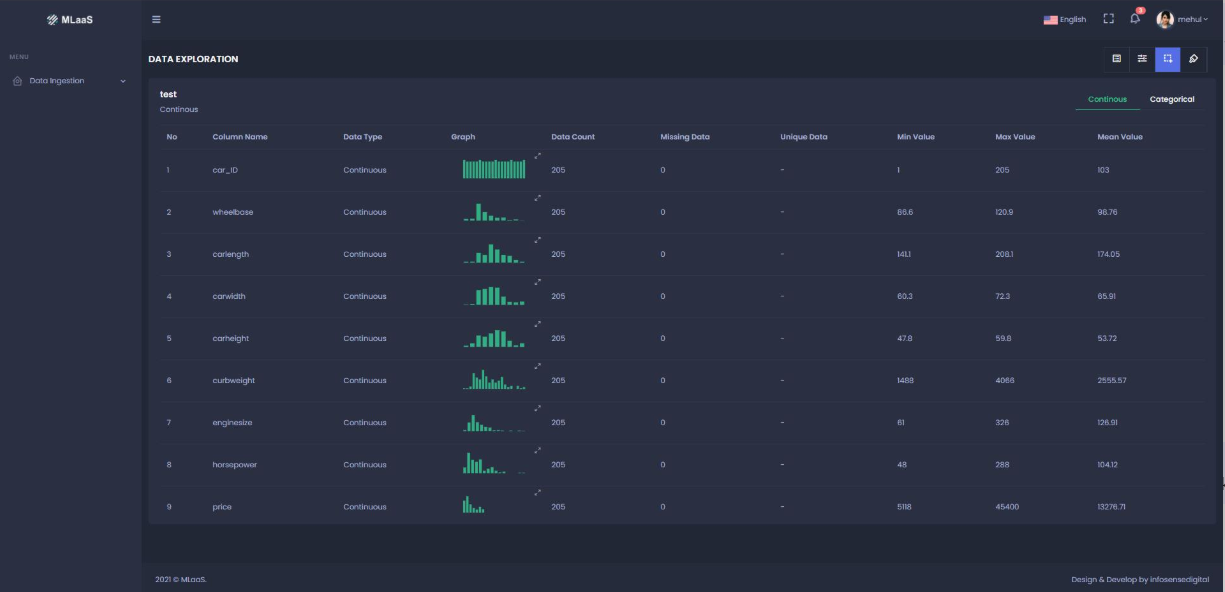


Figure 12: Data Exploration Mock-up

### Data Clean-up Page

#### User Actions

##### 1. Select column

User can select multiple columns of the dataset to perform the clean-up activities

##### 2. Clean-up operations

User will get several clean-up operations based on the selected columns. If multiple columns are selected by users, then common clean-up operations will appear for selection

###### Missing values handling

User can select the column and choose one of the following options to handle the missing values-

* **Discard missing values**
  + This option is available for all the columns
  + The entire rows having missing values will get deleted from the dataset

Ex. On selecting this option for Col 1, entire row of Col 1 having missing value will get deleted from the dataset

|  |  |  |
| --- | --- | --- |
| **Col 1** | **Col 2** | **Col 3** |
| 1 | A | 11 |
| 2 |  | 12 |
| 3 | C | 13 |
|  | D | 14 |
| 5 |  | 15 |
| 6 | F | 16 |
|  | G | 17 |
|  | H | 18 |
| 9 |  | 19 |
| 10 | J | 20 |

Note- If column 2 was selected, row 9 will be deleted as well

* **Mean imputation**
  + Applicable for Numeric columns only
  + The missing values will get replaced with the mean value of the selected column

Ex, in the column given below, Mean value is 3. So, the missing values will get replaced with 3

|  |  |
| --- | --- |
| **Col A** | **Output of Mean Imputation** |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
|  | 3 |
| 4 | 4 |
|  | 3 |
| 5 | 5 |

* **Median imputation**
  + Applicable for Numeric columns only
  + The missing values will get replaced with the Median value of the selected column

In the example below, Median value is 3, hence the missing values will get replaced with 3

|  |  |
| --- | --- |
| **Col A** | **Output of Median Imputation** |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
|  | 3 |
| 4 | 4 |
|  | 3 |
| 5 | 5 |

* **Mode Imputation**
  + Applicable for Numeric columns only
  + The missing values will get replaced with the Mode value of the selected column

In the example below, Mode of Col A is 4 as it is **most frequently appearing** value. Hence, the missing values will get replaced with 4

|  |  |
| --- | --- |
| **Col A** | **Output of Mode Imputation** |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
|  | 4 |
| 4 | 4 |
|  | 4 |
| 4 | 4 |
| 5 | 5 |
| 4 | 4 |

* **Arbitrary Value Imputation**
  + Applicable on both numeric and text columns
  + On selecting this option, user will need to enter a value. All the missing values of the selected column will get replaced with the value entered by user.

Ex, User has entered value 8, Hence, the missing values of Col A will get replaced with 8

|  |  |
| --- | --- |
| **Col A** | **Output of Arbitrary Imputation** |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
|  | 8 |
| 4 | 4 |
|  | 8 |
| 5 | 5 |

* **End or Tail Imputation**
  + Applicable for only Numeric columns
  + On selecting this option, the missing value of the selected column will get replaced with the value equivalent to 3 times the std. deviation of that column
* **Frequent Category Imputation** – This is applicable for Categorical columns only. On selecting this option, the missing values of the column will get replaced with most occurring categorical value

Ex, here A is the most frequent category, hence, the missing values will get replaced with A

|  |  |
| --- | --- |
| **Col A** | **Output for Frequent Category Imputation** |
| A | A |
| A | A |
| B | B |
| C | C |
|  | A |
| D | D |
| A | A |
|  | A |
| B | B |

* **Random Sample imputation**
  + Applicable to both Continuous and Categorical columns
  + On selecting this option, missing values of the selected column will get replaced with the random values of the same column.

Ex, Here, the missing value are replaced with random values from the column i.e., 1, 2, and 5

|  |  |
| --- | --- |
| **Col A** | **Output for Random Sample Imputation** |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
|  | 1 |
|  | 2 |
| 5 | 5 |
| 6 | 6 |
|  | 5 |
| 4 | 4 |

###### Noise Handling

* **Discard Noise**
  + On selecting this option ,the entire rows having missing values will get deleted from the dataset
* **Remove Noise**
  + On selecting this option, all the noise from the selected column will get replaced with Null values. Once the noise is removed, the column becomes eligible for Missing Value handling operation. User can then select one of the options to handle the missing values

|  |  |
| --- | --- |
| Col | Output for Remove Noise |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| / |  |
| / |  |
| 8 | 8 |
| 9 | 9 |
| 10 | 10 |

* **Replace noise with Mean**
  + On selecting this option, the noise from the selected column will get replaced with the mean value of the column
* **Replace noise with Median**
  + On selecting this option, the noise from the selected column will get replaced with the Median value of the column
* **Replace noise with Mode**
  + On selecting this option, the noise from the selected column will get replaced with the Mode value of the column
* **Replace noise with End or Tail value**
  + On selecting this option, the Noise of the selected column will get replaced with the value equivalent to 3 times the std. deviation of that column
* **Replace noise with Arbitrary value**
  + On selecting this option, user will need to enter a value. All the noise of the selected column will get replaced with the value entered by user.
* **Replace noise with Arbitrary value**
  + On selecting this option, noise values of the selected column will get replaced with the random values of the same column.

###### Transformation

**Math Operation**

This option is applicable only for Numeric columns

* **Add to column** – On selecting this option, user will need to enter a value. The entered value will get added to each value of the selected column

Ex. Value entered by user – 3

|  |  |
| --- | --- |
| Col A | Output of Add to Column |
| 2 | 5 |
| 3 | 6 |
| 11 | 14 |
| 8 | 11 |
| 1 | 4 |

* **Subtract from column** – On selecting this option, user will need to enter a value. The entered value will get subtracted from each value of the selected column

Ex. Value entered by user – 2

|  |  |
| --- | --- |
| Col A | Output of Subtract from Column |
| 5 | 3 |
| 14 | 12 |
| 11 | 9 |
| 8 | 6 |
| 10 | 8 |

* **Divide from column** – On selecting this option, user will need to enter a value. The entered value will get divided from each value of the selected columns

Ex. Value entered by user – 2

|  |  |
| --- | --- |
| Col A | Output of Divide from Column |
| 5 | 2.5 |
| 14 | 7 |
| 11 | 5.5 |
| 8 | 4 |
| 10 | 5 |

* **Multiply into column** – On selecting this option, user will need to enter a value. The entered value will get multiplied with each value of the selected column

Ex. Value entered by user – 2

|  |  |
| --- | --- |
| Col A | Output of Multiply to Column |
| 5 | 10 |
| 14 | 28 |
| 11 | 22 |
| 8 | 16 |
| 10 | 20 |

**Encoding**

Encoding is applicable for String columns. Encoding allows the representation of categorical data to machine-readable form. It is **mandatory** to select one of the following encoding methods for all the String columns

* **Label encoding** – On selecting Label Encoding the Categories are assigned with labels (numeric form)

Ex,

In the example below, we have three categories – France, Germany, and Spain

On selecting, Label Encoding, a new column will get created replacing each category with unique number such as

France – 1

Germany – 2

Spain - 3

|  |  |
| --- | --- |
| **Col** | **Col1** |
| France | 1 |
| France | 1 |
| Germany | 2 |
| Spain | 3 |
| Germany | 2 |
| France | 1 |
| Spain | 3 |
| Spain | 3 |
| Germany | 2 |

* **One Hot encoding** – On selecting One-hot encoding, the selected column gets split into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value.

If there are a lot of categories, then the least occurring categories will be combined to form a new category and will be shown in a single column.

Ex- In our example, we’ll get three new columns, one for each country — France, Germany, and Spain.

For rows which have the first column value as France, the ‘France’ column will have a ‘1’ and the other two columns will have ‘0’s. Similarly, for rows which have the first column value as Germany, the ‘Germany’ column will have a ‘1’ and the other two columns will have ‘0’s.

|  |  |  |  |
| --- | --- | --- | --- |
| **Col** | **France** | **Germany** | **Spain** |
| France | 1 | 0 | 0 |
| France | 1 | 0 | 0 |
| Germany | 0 | 1 | 0 |
| Spain | 0 | 0 | 1 |
| Germany | 0 | 1 | 0 |
| France | 1 | 0 | 0 |
| Spain | 0 | 0 | 1 |
| Spain | 0 | 0 | 1 |
| Germany | 0 | 1 | 0 |

**Outlier Handling**

Users can select columns and select one of the following options to handle outliers

* **Delete above –** to remove the outlier above a certain value, user can enter the value and all the entries higher than the entered value will get deleted from the selected column

Ex, User has entered value 300 for Delete Above option. So all the value of col a higher than 300 will get deleted

|  |  |
| --- | --- |
| Col A | Output of Delete Above |
| 12 | 12 |
| 44 | 44 |
| 32 | 32 |
| 386 |  |
| 302 |  |
| 45 | 45 |

* **Delete below–** to remove the outlier below a certain value, user can enter the value and all the entries lesser than the entered value will get deleted from the selected column

Ex, User has entered value 45 for Delete Above option. So all the value of col A lesser than 45will get deleted

|  |  |
| --- | --- |
| Col A | Output of Delete Above |
| 12 |  |
| 44 |  |
| 51 | 51 |
| 56 | 56 |
| 45 | 45 |
| 91 | 91 |

* **Replace Outliers –** Outliers can be replaced by one of the following options-
  + Replace Outliers with Mean using Extreme Value Analysis Detection
  + Remove outliers with Z-score Detection Method
  + Replace Outliers with Median using Extreme Value Analysis Detection
  + Replace Outliers with Median using Z-score Detection

**Note**-

**Extreme Value analysis** - In Extreme value analysis we declare all those values which exceeds a certain threshold as an outlier.

In our case the threshold is 1.5 \* IQR. Where IQR stands for the Interquartile Range.

Each value that is above 1.5\* IQR or below -1.5\*IQR will be considered as an outlier and will be eligible for the outlier handling operations.

**Z Score Detection -**

Z-Score = (value - mean)/std, where std stands for standard deviation.

In the Z- Score detection method, a value will be considered as an outlier if it's Z-Score value is greater than 3 or lesser than -3.

For all the values with >3 or <-3, Z-Score will be eligible for the outlier handling operations.

* **Apply Log Transformations**

**Reset** - On clicking Reset, application should initialize all user-input values.

**Save-**

On clicking **Save**, the selected clean-up criteria should get applied for the selected column. All the options must be visible under the **Handling** section for respective columns.

**Save As-**

On clicking **Save As,** a new dataset (with the changes applied) will get saved. This dataset will be available for selection in the Create Dataset page

* **Public/Private** – After clicking **Save As**, user can choose to keep the dataset public or Private.
* **Dataset Name and Description** – While saving the changes as a new dataset, it is mandatory to provide the name and description of the dataset.

**Example -**

**Save** –

* User has used dataset ID 1 for Project 1
* Applied some transformations (changed the column name in Schema Mapping) and clicked Save
* Column name will get updated in table as a new dataset ID 2. This dataset will get associated with Project 1

**Save As-**

* User has used dataset ID 1 for project 1
* Applied some transformation (changed the column name in Schema Mapping) and clicked Save
* Column name will get updated in table as a new dataset ID 2. This dataset will get associated with Project
* User has applied some cleanup activities in the Data Clean-up page and clicked Save. The changes will get updated in the Dataset ID 2 and this dataset will be associated with Project
* User has applied some cleanup activities in the Data Clean-up page and clicked Save As. The changes will get updated in the new Dataset ID 3
* Dataset ID 3 will get associated with Project 1 and will be available in dataset master for selection

**Note** – Whenever user clicks **Save** or **Save As** the project and dataset relationship gets re-associated

##### 3. Scale and Split

Once the clean-up operation are applied, user can scale and split the dataset in this section.

###### Scaling

Selected scaling method applies on the entire dataset

* **Min-Max Scaling** – On selecting this option, the data is scaled to a fixed range - 0 to 1. Where, Max Value = 1 & Min Value = 0
* **Std. Scaler -**On selecting this technique, the values are centered around the mean with a unit standard deviation. This means it scale the values to have a mean of zero and a standard deviation of one.
  + Standardization is calculated by subtracting the mean value and dividing by the standard deviation.
  + value = (value – mean) / std dev
* **Robust** – Standardizing input variables in the presence of outliers is to ignore the outliers from the calculation of the mean and standard deviation, then use the calculated values to scale the variable. This is called robust standardization or robust data scaling.
  + This can be achieved by calculating the median (50th percentile) and the 25th and 75th percentiles. The values of each variable then have their median subtracted and are divided by the interqurtile range (IQR) which is the difference between the 75th and 25th percentiles.
  + value = (value – median) / (p75 – p25)
  + The resulting variable has a zero mean and median and a standard deviation of 1, although not skewed by outliers and the outliers are still present with the same relative relationships to other values.

###### Split

##### 4. Modelling

After performing the clean-up activities, user can click **Modelling** to go to the modelling page.

#### Validations

* Multiple columns can be selected to apply Data clean-up options
* On selecting multiple columns, only common clean-up operations (which are applicable to all the selected columns) will be shown on screen
* Once the cleanup criteria are selected and user has clicked **Save**, user should be able to reselect the columns and change the options
* On clicking the operations, it will show up in the handling section
* On clicking save, the selected operations will get applied on the columns, and will no longer show in handling section.
* On clicking save as, the selected operations will get applied on the columns, and will no longer show in handling section, and a new dataset will be saved. Here, user will have to provide the new name and description to the dataset
* Reset button will be enabled prior to save/save as. On clicking Reset, the selection operations will get removed from the handling section.
* Activities timeline will show an activity when the cleanup operation is started and another activity will be shown when the clean-up activity is completed
* Modelling button will remain disabled when the cleanup operations are being performed in the backend
* When the cleanup operations get applied, application will show a flash message stating “Dataset saved successfully”
* While the data cleanup operation is getting performed, the data clean up screen will remain locked

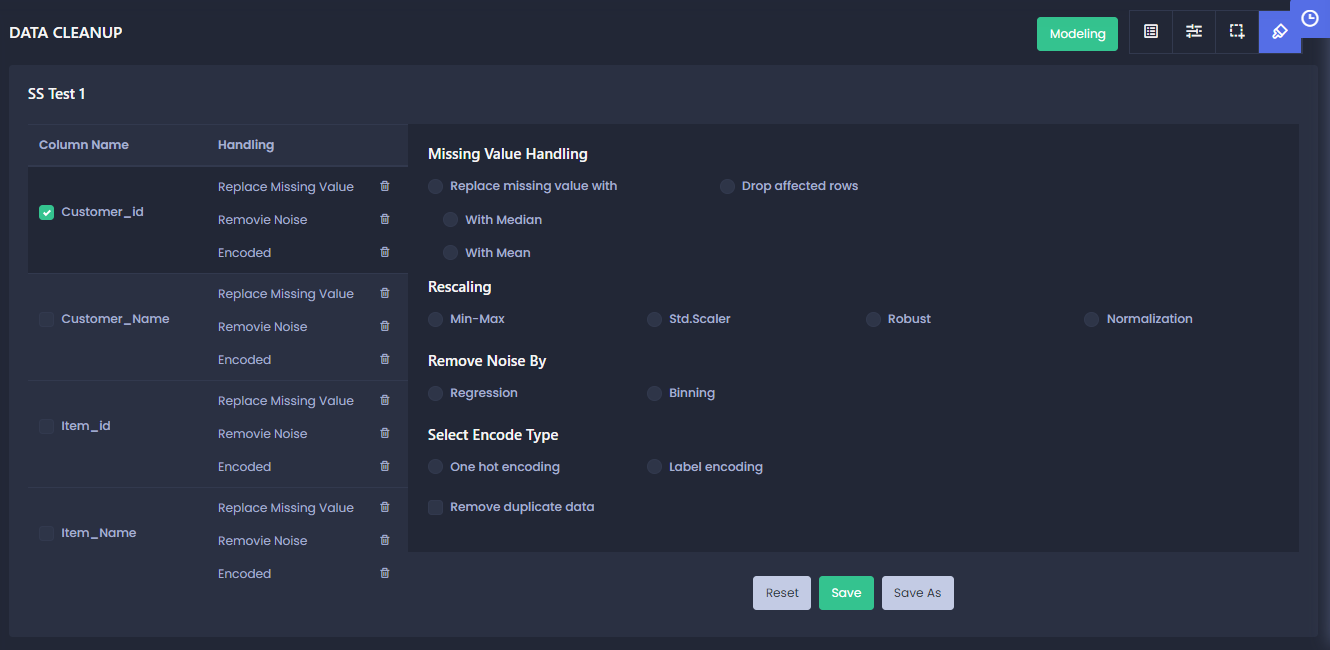


Figure 13: Data Clean-up Page Mock-up

## Data Modelling

### Modelling Type

This page shows the Project Name, Dataset Name, Target columns and let user select the type of modelling: Auto or Manual

#### User Action

##### 1. Select Modelling Mode

User can select either Auto or Manual mode for the Modelling.

##### 2. Select modelling Type

User can select **Regression** or **Classification** as Modelling type.

Regression is recommended if the target column is of Continuous type and Classification is for Categorical target columns

Application will identify the type of modelling based on the target column and the modelling type will be selected accordingly. User will have an option to change the modelling type, if required.

##### 3. Start/Stop modelling

Once the user selects the modelling type, he can click this button to start the modelling.

Once the modelling is started, this button will show the option to Stop the modelling.

##### 4. Create Experiment

On clicking Start button, user will need to create experiment by providing Name and Description.

#### Validations

* User can only select one of the modelling types (Auto or Manual) at a time
* When the model training gets completed, application should show a message stating “Modelling of project ABC has been completed successfully”
* Experiment name should be unique

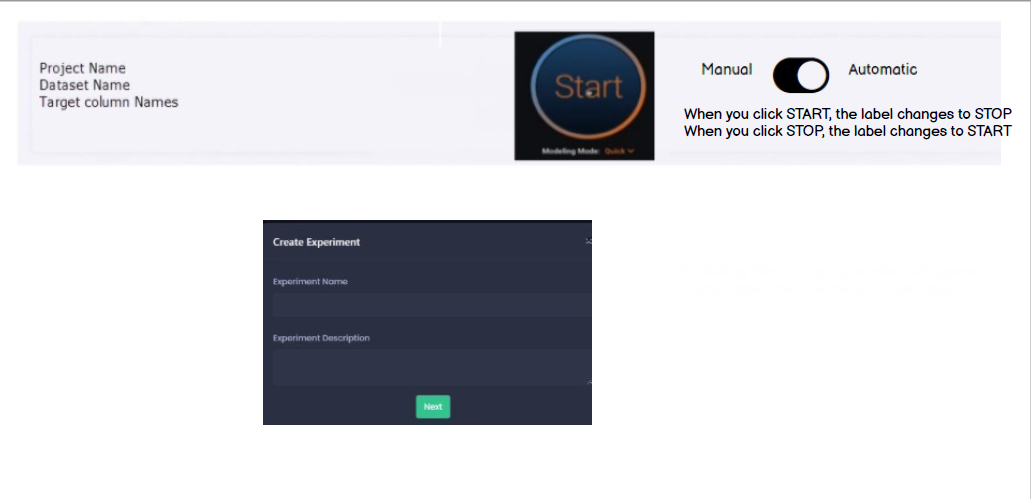


Figure 14: Modelling Type Mock-up

### Auto Modelling

When the user selects auto modelling, application run the algorithm and shows the list of experiment. Following details are shown-

* Experiment Name – On clicking the experiment name, Experiment Details window will open showing following details-
  + [Learning curves](#LearningCurves)
  + [Feature Importance](#FeatureImportance)
  + [Performance Metrics](#PerformanceMetrics)
  + [Model Summary](#ModelSummary)
  + [Actual Vs Prediction](#ActualVsPrediction)
* Model Name
* Dataset Name
* Modelling Type – Auto or Manual
* Status – Running, Completed, Aborted
* Cross Validation Score – This is the combined score of K fold (K fold will be by default 5 for auto modelling)
* Holdout Score – This will be by default set as 80:20 (80 is for training and 20 is for testing) for auto modelling
* Start Date and time

**Note** – User can choose to see only running or All the experiments by setting up the filer

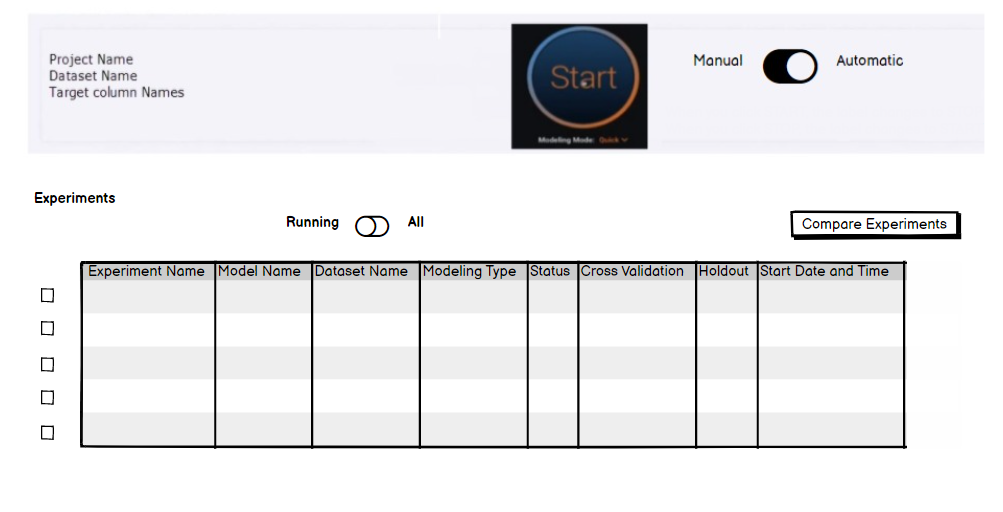


Figure 15: Auto Modelling Mock-up

#### Learning curves

This page shows the graphical representation of model performance (Accuracy and LOSS)

* Accuracy
  + No. Of Epoch Vs Accuracy – if the model is developed in **Keras**
  + Size of dataset Vs Accuracy – if the model is developed in **SK learn**
* LOSS
  + No. of Epoch Vs LOSS – if the model is developed in **Keras**
  + Size of dataset Vs LOSS – if the model is developed in **SK learn**

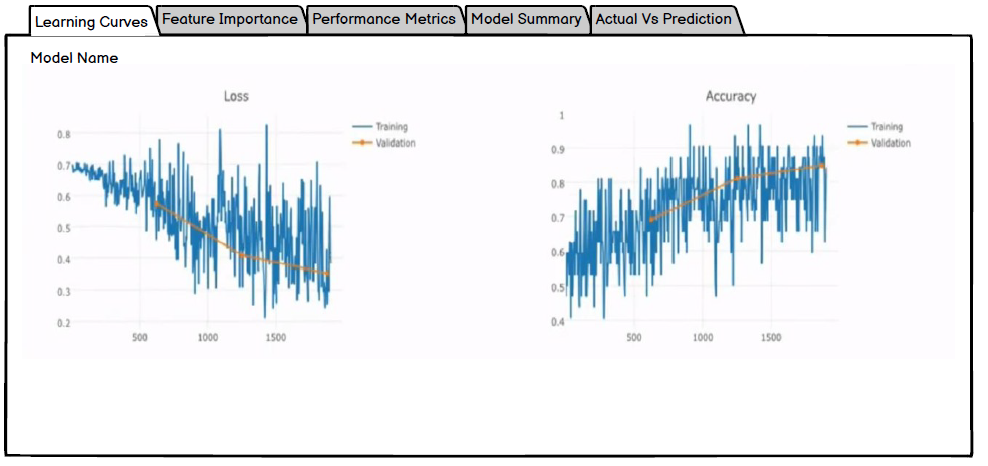


Figure 16: Learning Curve Mock-up

**Note** - For now, the learning curve is shown only for those experiments which are in completed status

#### Feature importance

This page will show the list of features used in modelling in order of their importance.

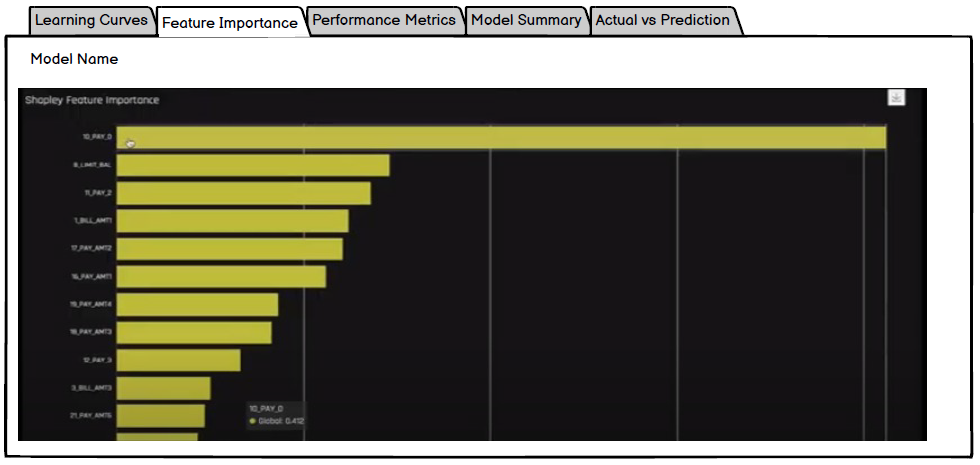


Figure 17: Feature Importance

#### Performance Metrics

This page will show following metrics for the model trained

* For regression model-
  + R2-Score – if more than 0.70 – Average performance.

More than 0.90 - best performance this goes more than 0.90

* + MSE – value nearest to zero denotes best performance (Actual - Prediction), calculated on each column
  + RMSE
  + MAE - nearest to zero denotes good performance
  + MAPE- nearest to zero denotes good performance
* For Classification model-
  + Accuracy – shown in percentage
  + Precision – Nearest to 1 value denotes good performance
  + Recall – Nearest to 1 value denotes good perfomance
  + Confusion Matrix – Statistics about the prediction
  + AUC – ROC score

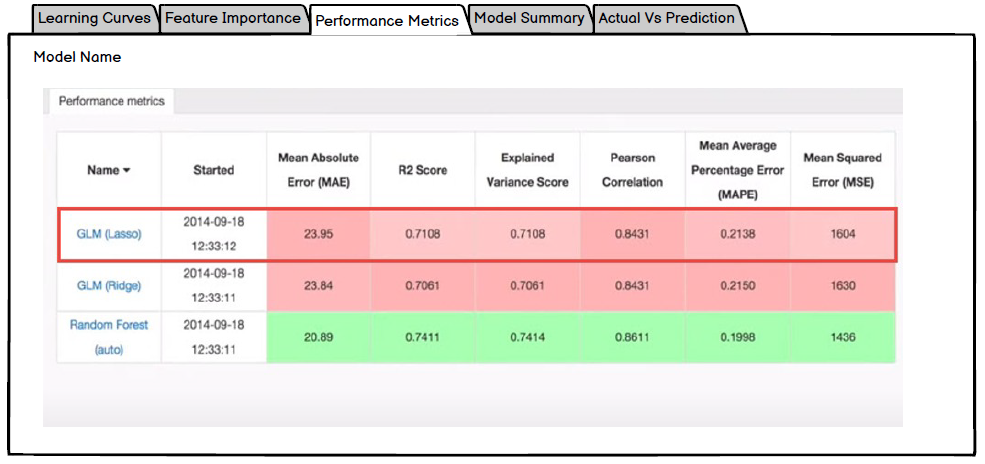


Figure 18: Performance Metrics Mock-up

#### Model summary

This page shows the detail information of model such as, Model name, description, size, parameters used, etc

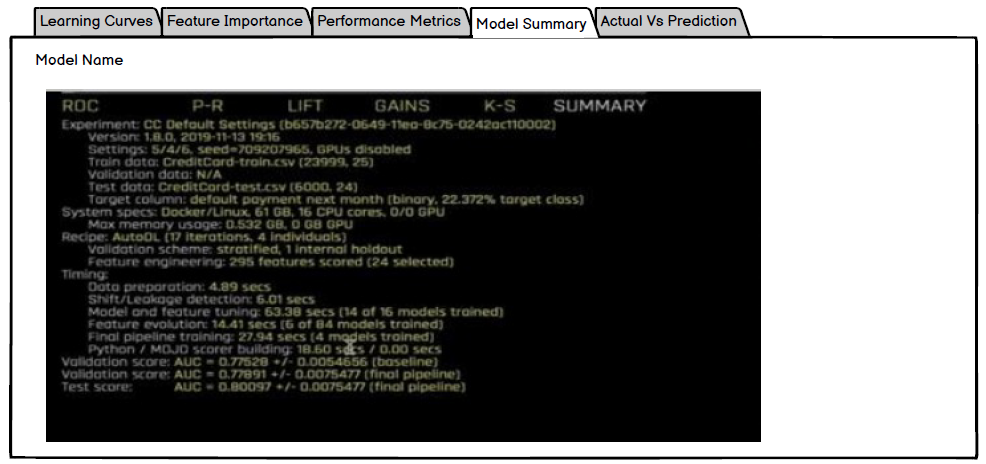


Figure 19: Model Summary Mock-up

#### Actual Vs Prediction

This page will show the comparison of actual and predicted result of the target column. Also, the input features based on which the predicted value is calculated is shown here.

A comparison graph (Line chart) will also be shown to compare the Actual Vs Predicted value of Target Column

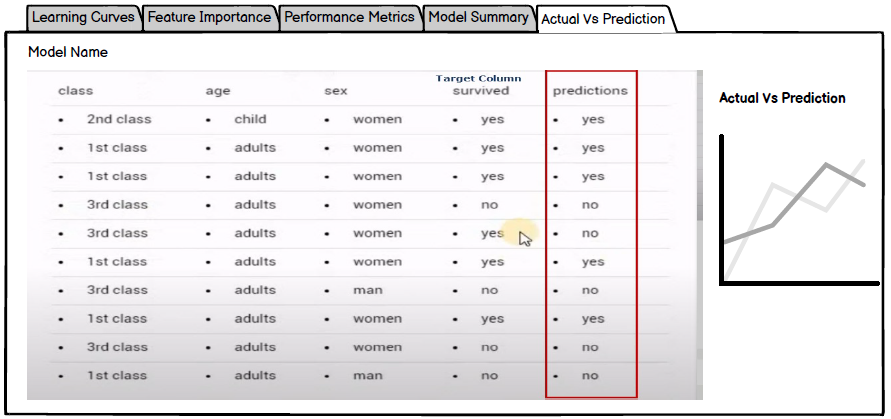


Figure 20: Actual Vs Prediction mock-up

### Compare Experiment

To compare multiple experiments, select the checkboxes and click **Compare Experiments**

**Experiment Comparison Graphs-**

This page will show the comparison Graphs for selected experiments-

* Accuracy Vs No. Of Epoch
* Loss Vs No, of Epoch

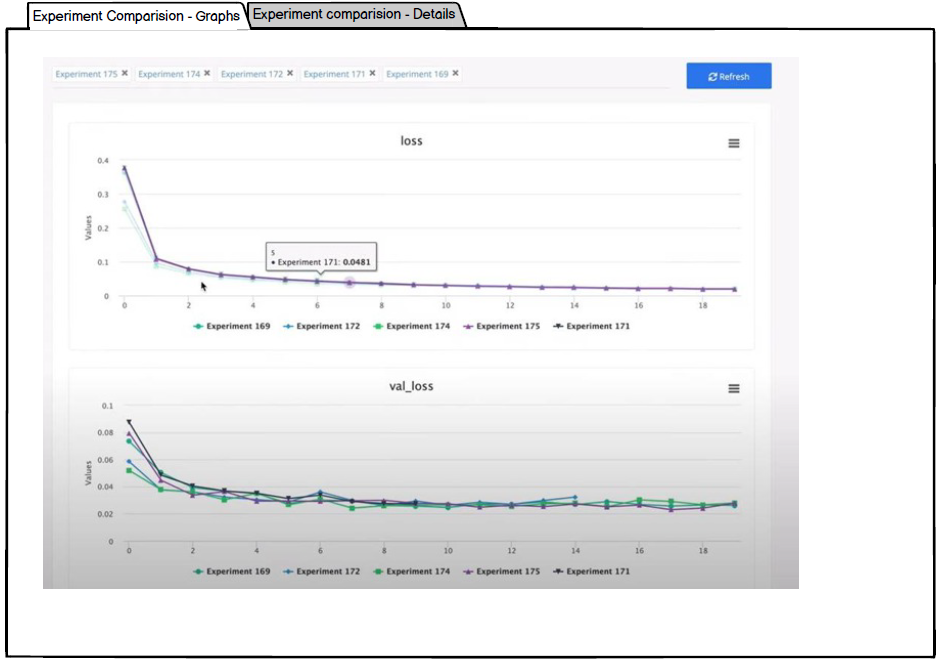


Figure 28: Experiment Comparison Graphs Mock-up

**Experiment Comparison Details-**

This page will show the comparison of Performance metrics and Accuracy score of the selected experiments

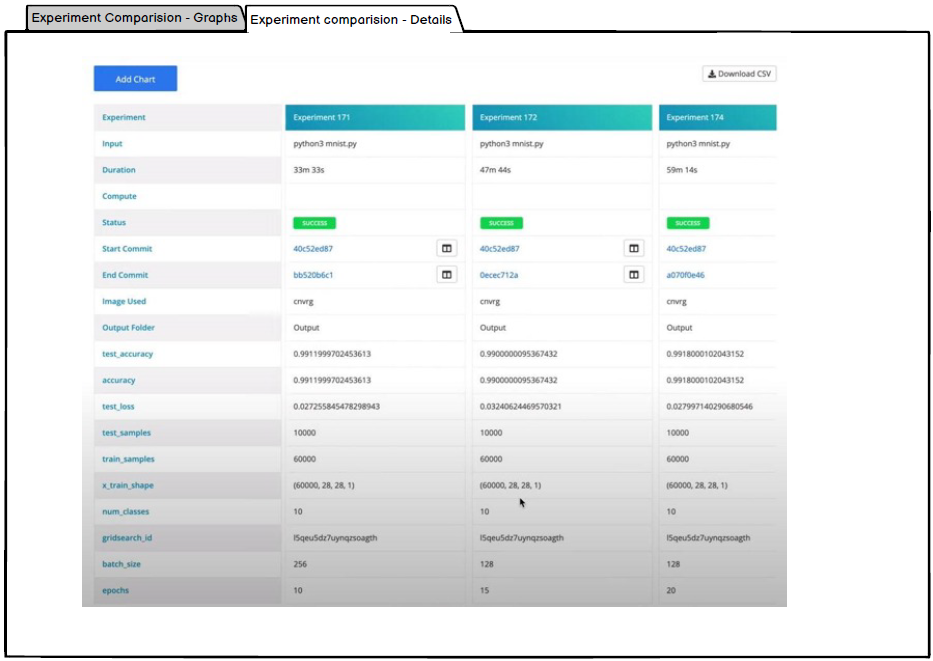


Figure 29: Experiment Comparison Details Mock-up

#### Validations

* Only those experiments which are in Completed status can be compared
* Only those experiments can be compared which have same target columns. When a user selects one experiment, all the other experiments (in completed status) with same target column will get enabled for selection.

### Manual Modelling

To opt for modelling, user need to select the **Manual** option on Modelling Type page and click Start.

On selecting Manual Modelling, a parameter window will appear to configure the modelling parameters manually

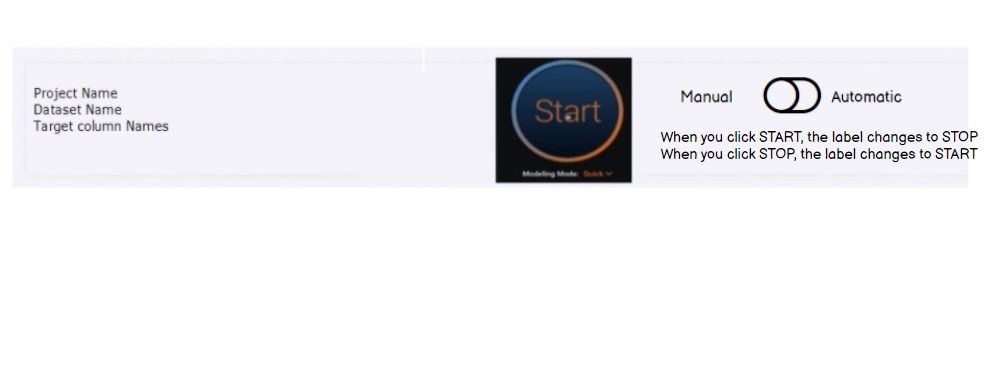


Figure 21: Manual Modelling Mock-up

#### Split Dataset

To configure the modelling manually, user need to first select one of the below Split method-

* Cross Validation
* Train-Validation-Holdout

**Cross Validation**

For cross validation split method, user need to provide following parameter-

* No. of cross-validation (CV) folds
* Holdout percentage
* Random state

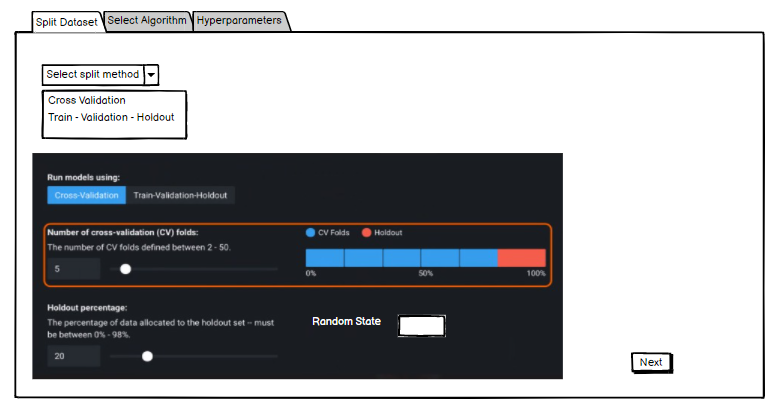


Figure 22: Split Dataset - Cross Validation Mock-up

**Train-Validation-Holdout**

For Train-Validation-Holdout method, user need to select

* Split Ratio
* Random State – Mandatory field

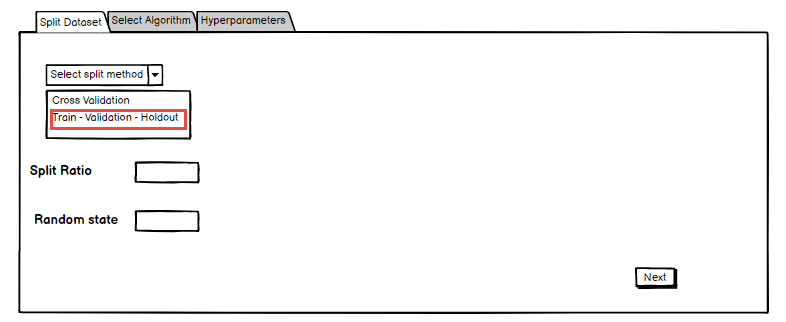


Figure 23: Split Dataset - Train Validation Holdout Mock-up

#### Select Model

Select one of the following Models and click Next-

* Linear regression with SK Learn
* Linear Regression with Keras

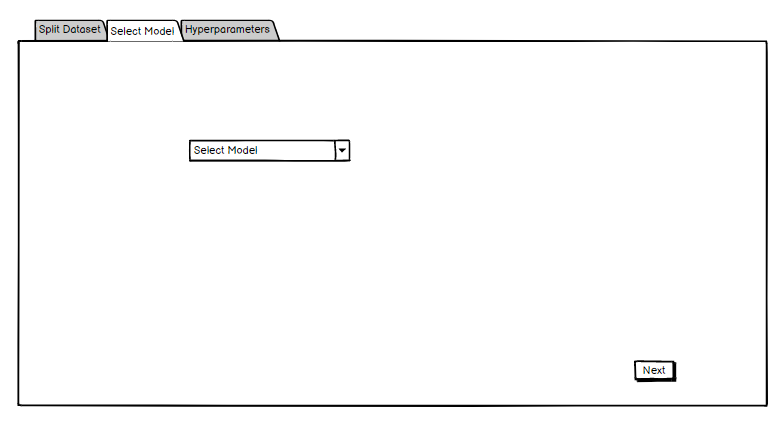


Figure 24: Select Model Mock-up

#### Hyperparameters

If the user has selected, **Linear Regression with SK Learn method, then no hyperparameters are required**. In case of Linear regression with Keras, user need to provide following hyperparameters-

* Learning Rate
* No. of Epoch
* Batch Size
* Loss
* Optimizer
* Activation

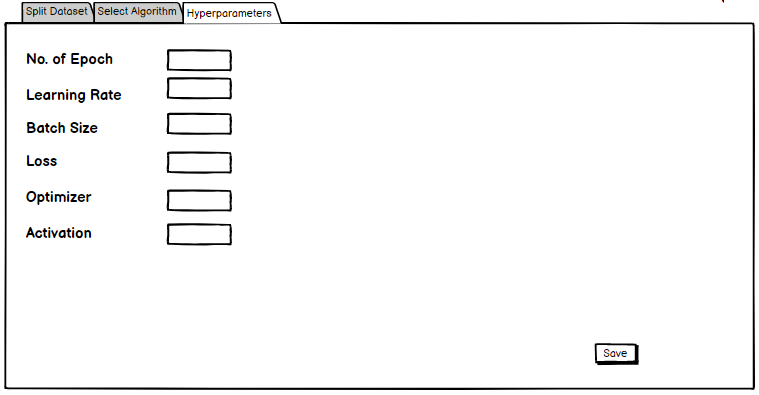


Figure 25: Hyperparameters Mock-up

After entering the parameters, user need to click Save. On clicking Save, user will be directed on modelling page. Here, on clicking Start button, to start Modelling, user will be required to Create experiment by providing name and description.

Following details of thew experiments will be shown-

* Experiment Name – On clicking the experiment name, Experiment Details window will open showing following details-
  + [Learning curves](#LearningCurves)
  + [Feature Importance](#FeatureImportance)
  + [Performance Metrics](#PerformanceMetrics)
  + [Model Summary](#ModelSummary)
  + [Actual Vs Prediction](#ActualVsPrediction)
* Model Name
* Dataset Name
* Modelling Type – Auto or Manual
* Status – Running, Completed, Aborted
* Cross Validation Score
* Holdout Score

**Note** – User can choose to see only Running or All the experiments by setting up the filer

User can select the experiments in Completed Status and click [Compare](#CompareExperiment) to compare the experiments.

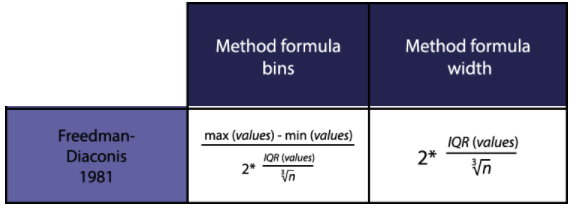
#### Validations

* All the hyperparameters are mandatory to be entered by user

## Formulas/Logic used

### 1. Distribution Graph

To plot the graph, bins are decided on based on **Freedman – Diaconis formula**

* Histogram graphs will be plotted with the bins in the range of 2-20
* Count graph will be plotted for all the unique values of the column

### 2. Datatype Determination

Application is reading the CSV file (dataset) using pandas library and based on that datatype will be defined-

|  |  |
| --- | --- |
| **Datatype based on pandas library** | **Datatype to be assigned on Schema Mapping** |
| float64, float32, int32, int64 | Numerical |
| object (String) | Text |
| datetime64[ns] | Timestamp |
| string of comma separated values | Categorical list |
| (unique column values/total rows) < 20% | Categorical |

### 3. Upload Dataset

When user uploads a dataset in the system, it’s duplicate will be created in backend. This duplicate dataset will be used for further operations in Schema Mapping.

Ex-

* User has uploaded Dataset 1 with column A, B, and C
* Dataset 2, which is duplicate of dataset 1, will be created in backend
* User selects Dataset 1 from frontend for Project 1
* From Backend, Dataset 2 (duplicate of dataset 1) will get associated with Project 1
* In data cleanup page, user has cleaned up dataset and clicked **Save**, Dataset 2 will get updated and get associated with Project 1
* In data cleanup page, user has cleaned up data and clicked **Save As**, Dataset 3 will be created with new changes and will be available in Dataset Master to be used for other projects