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

**Phase 1 - Requirement Document**



Table of Content

[Revision History 3](#_Toc63702628)

[Objective 4](#_Toc63702629)

[Scope 4](#_Toc63702630)

[Login 4](#_Toc63702631)

[User Actions 4](#_Toc63702632)

[Validations 4](#_Toc63702633)

[Navigation Panel 5](#_Toc63702634)

[Activities Timeline 6](#_Toc63702635)

[Data Ingestion Module 7](#_Toc63702636)

[Create Dataset Page 7](#_Toc63702637)

[Create Project Page 8](#_Toc63702638)

[All Projects Page 10](#_Toc63702639)

[Data clean-up and Visualization Module 11](#_Toc63702640)

[Data Detail Page 11](#_Toc63702641)

[Schema Mapping Page 12](#_Toc63702642)

[Data Exploration Page 14](#_Toc63702643)

[Data Clean-up Page 17](#_Toc63702644)

[Data Modelling 19](#_Toc63702645)

[Modelling Type 19](#_Toc63702646)

[Auto Modelling 20](#_Toc63702647)

[Manual Modelling 24](#_Toc63702648)

[Experiment List 27](#_Toc63702649)

[Formulas/Logic used 28](#_Toc63702650)

[1. Distribution Graph 28](#_Toc63702651)

[2. Datatype Determination 28](#_Toc63702652)

[3. Save/Save As functionality 28](#_Toc63702653)

## 

## 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 |

## 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

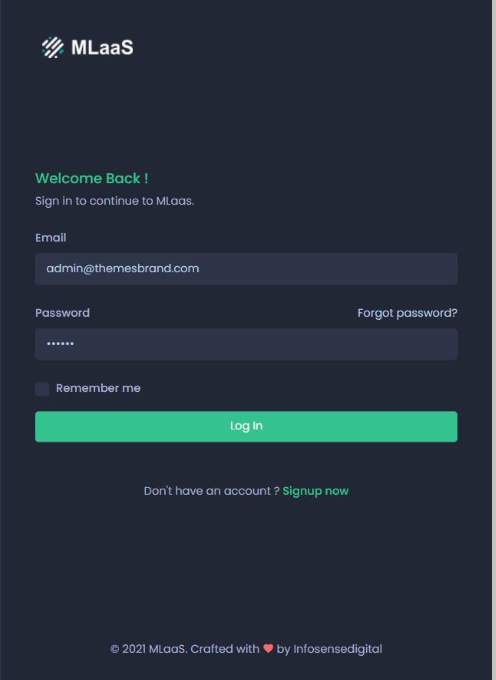


Figure 1: Login Page Mock-up

## 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

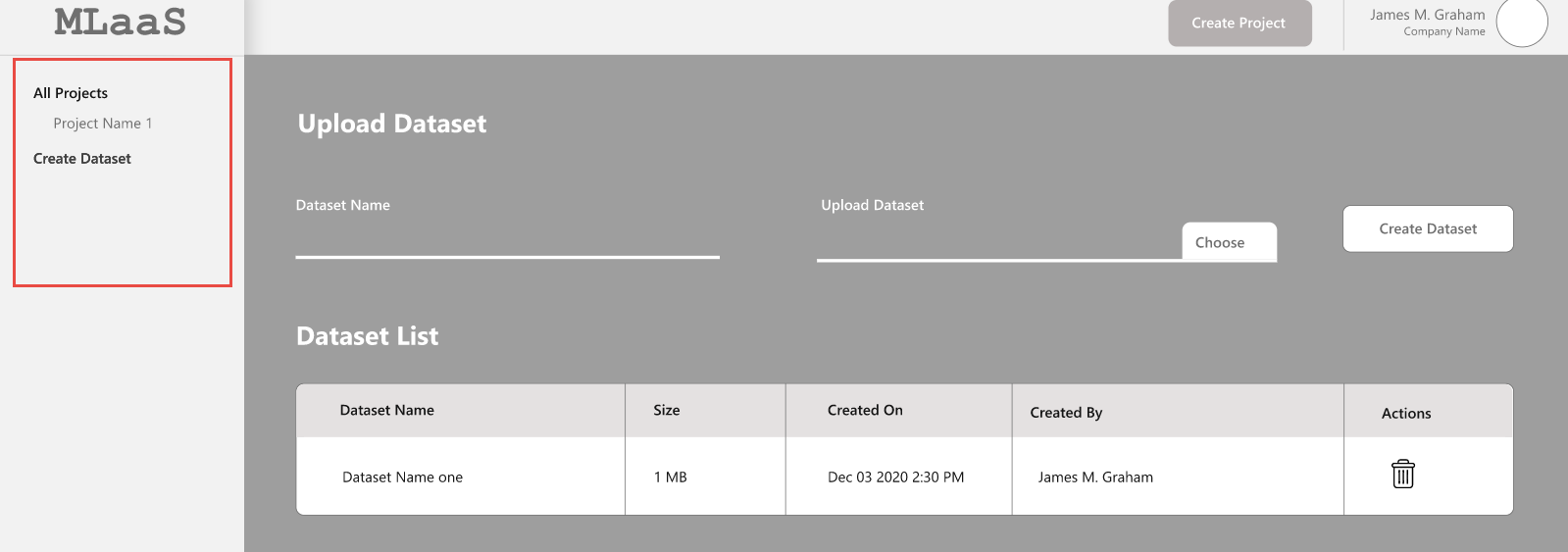


Figure 2: Navigation Panel Mock-up Screen

## 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” |

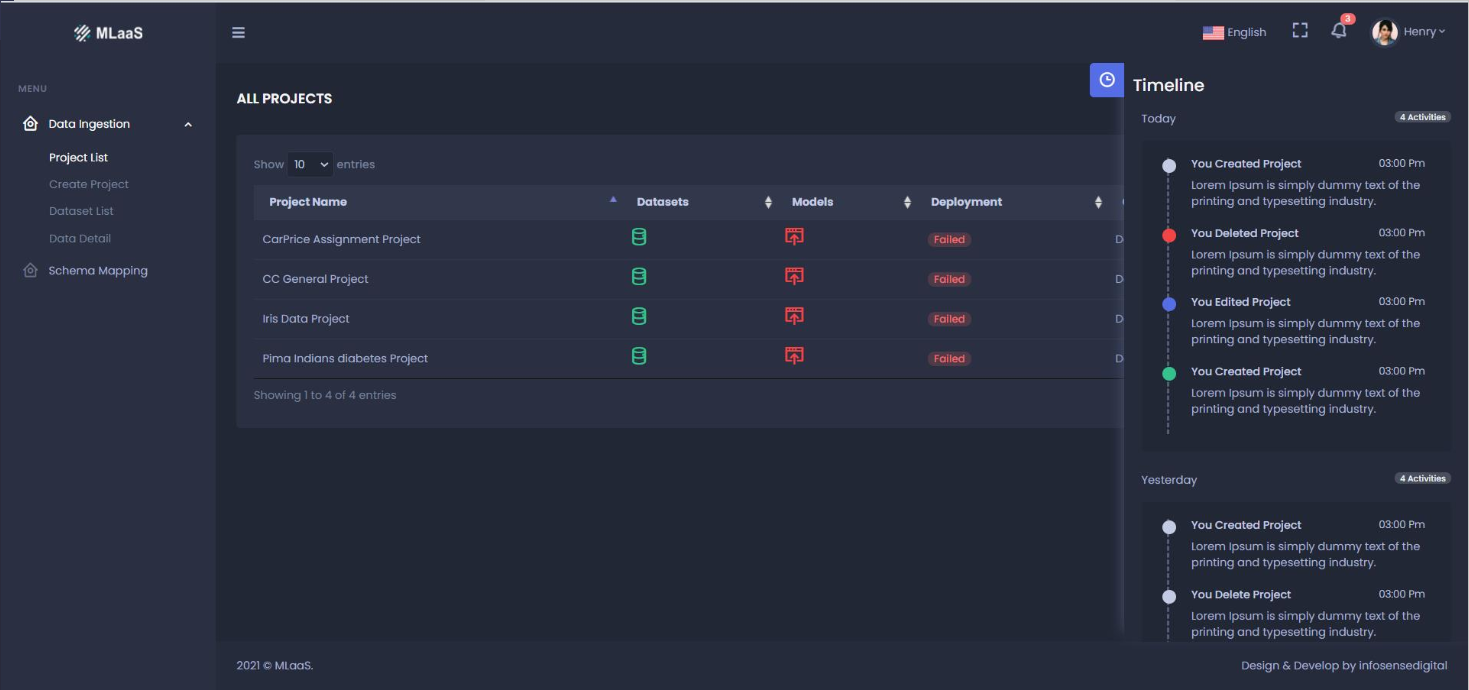


Figure 3: Activities Timeline Mock-up

## 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

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)

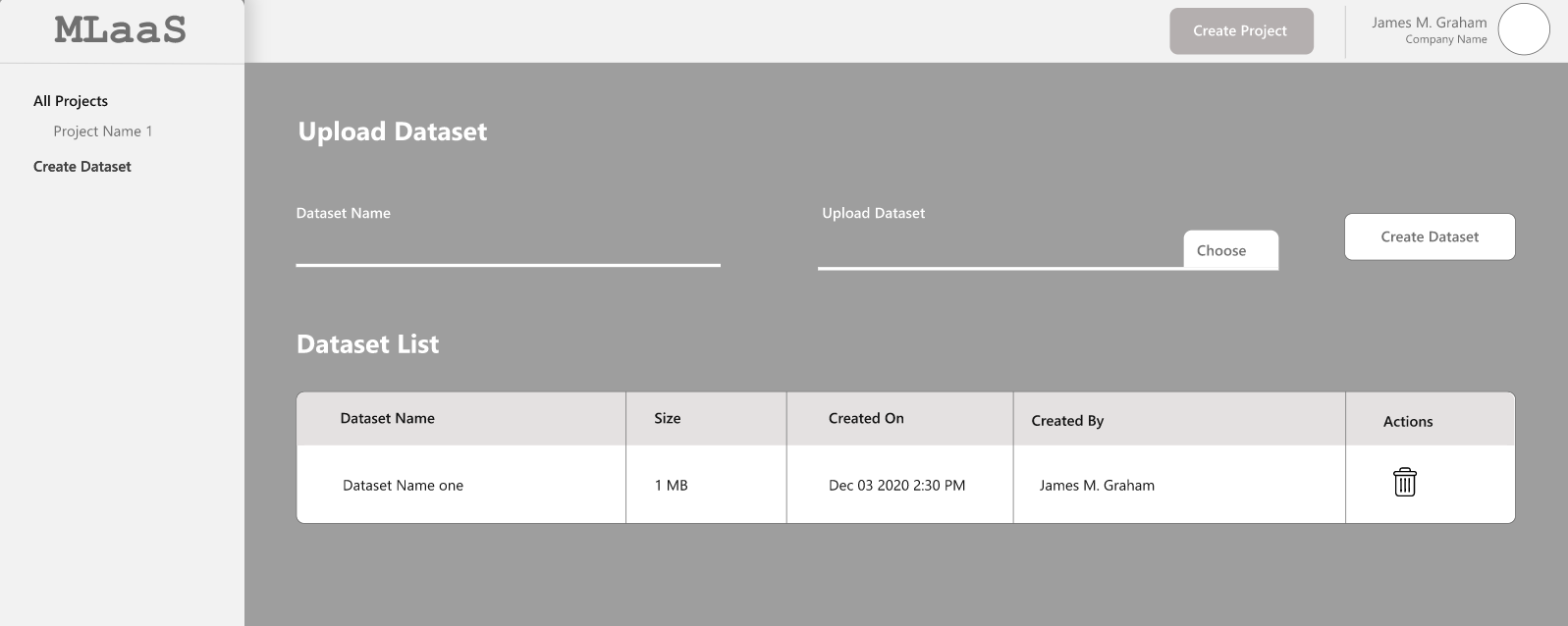


Figure 4: Create Dataset Mock-up Screen

### 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)

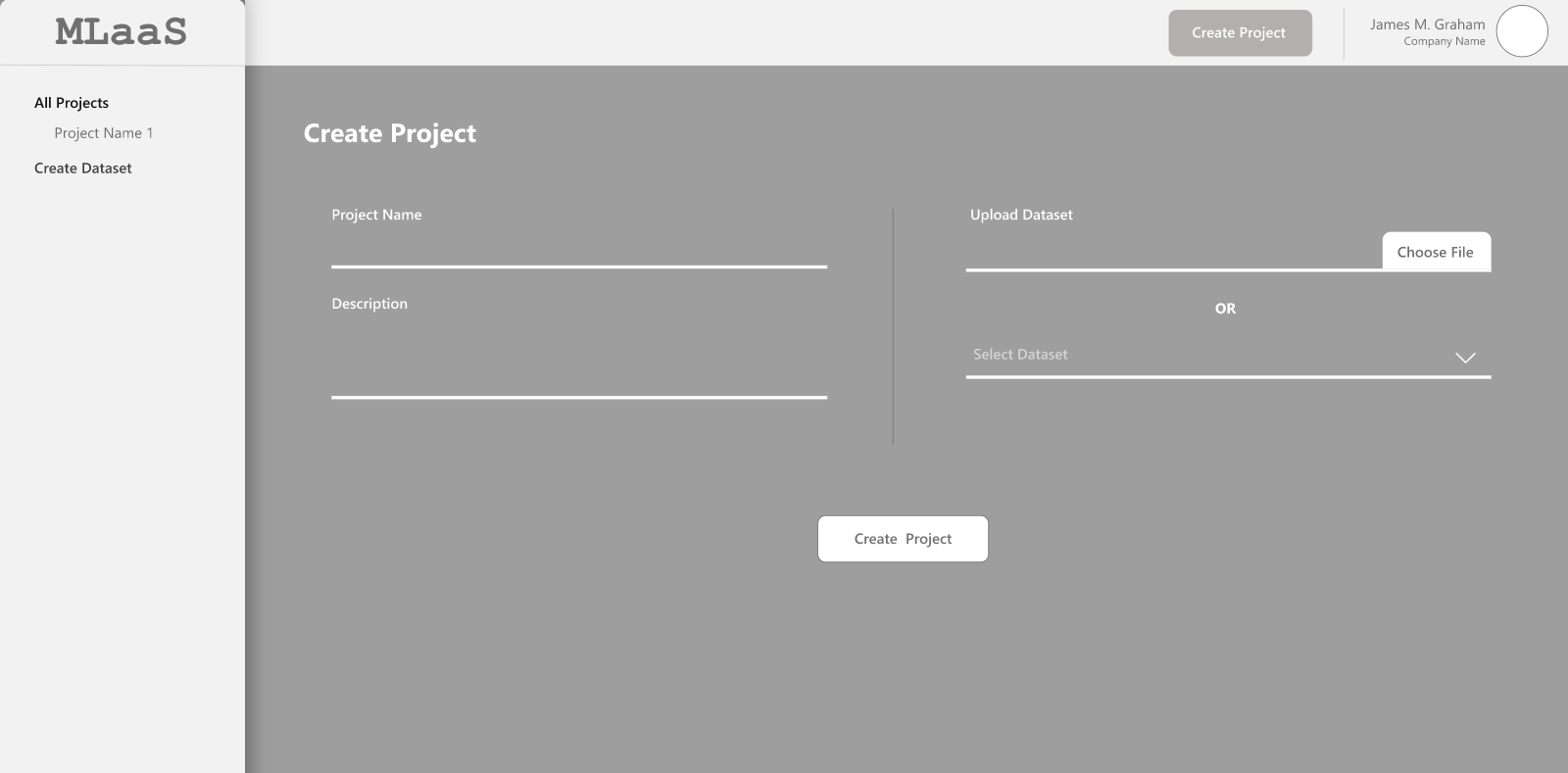


Figure 5: Create Project Mock-up Screen

### 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.
* **Deployment** – Shows thedeployment status of the project-
  + "Active" deployment
  + "Suspended" deployment
* **Created** – Project creation Date and Time
* **Actions** – Provide option to see Dataset details and delete the project

#### 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. 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

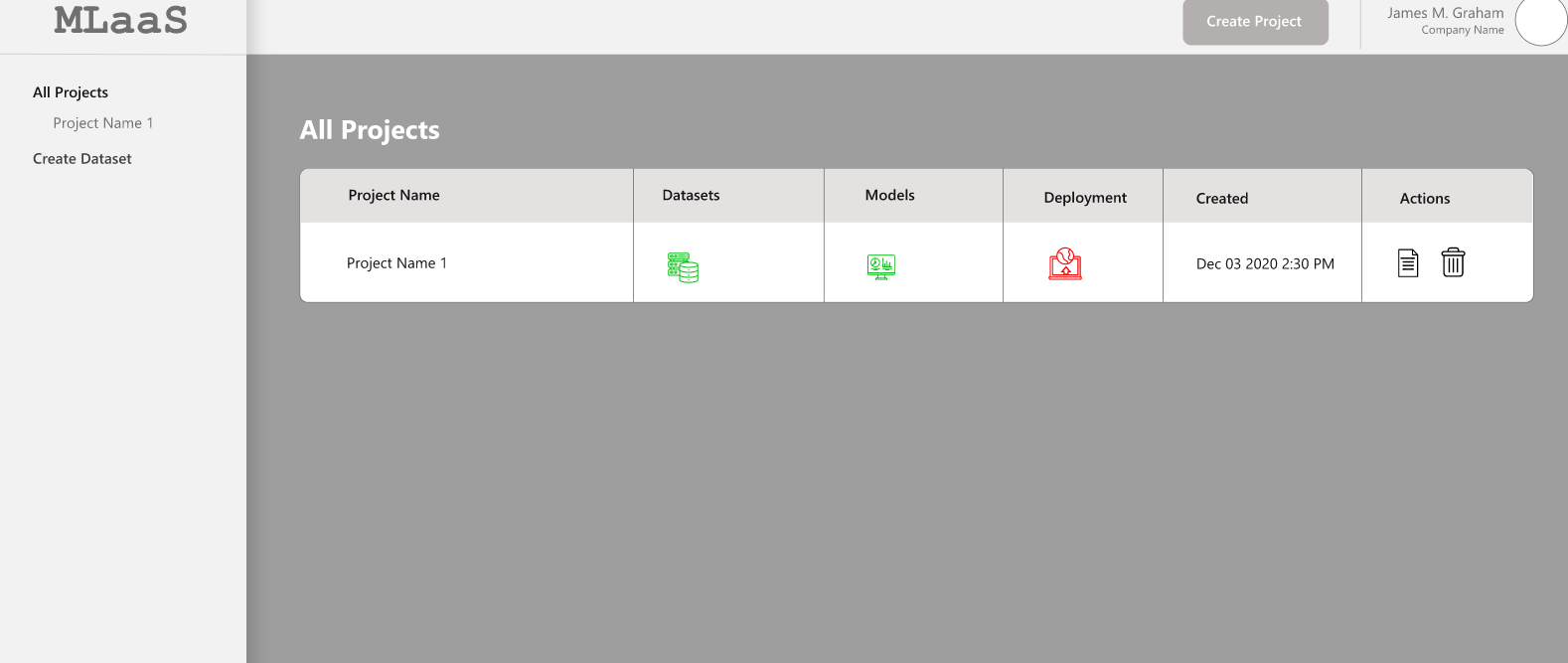


Figure 6: All Projects Mock-up Screen

## 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.

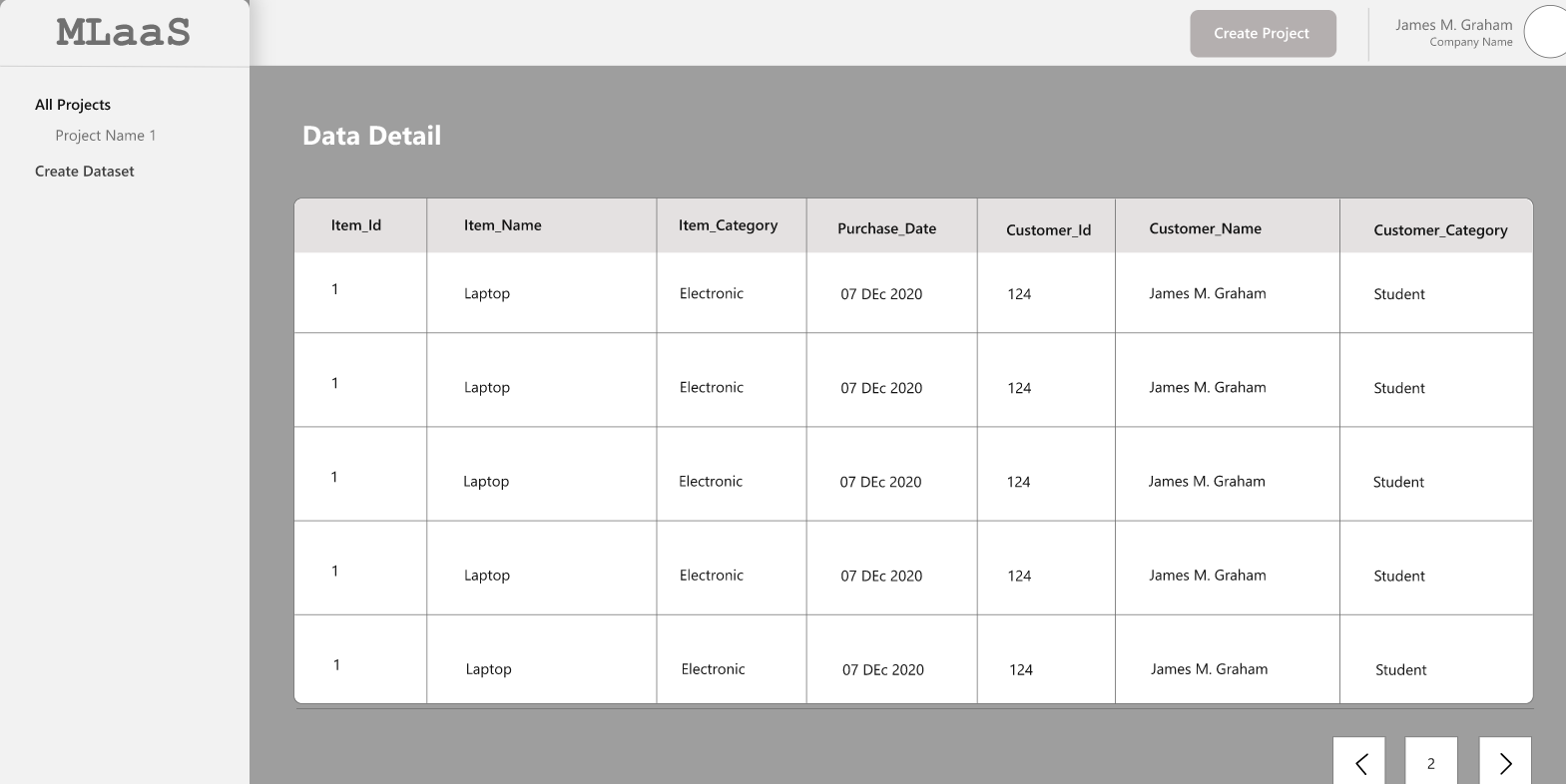


Figure 7: Data Detail Page Mock-up Screen

### 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

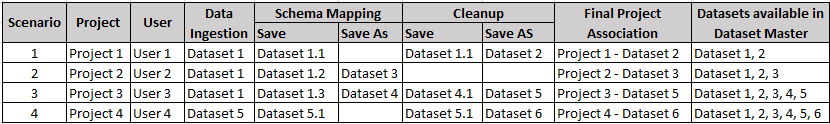
User can choose to update the name of the column from this page.

##### 4. Save/Save As the dataset

* **Save** - User can click Save to save the changes and proceed with Dataset clean-up
* **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) and clicked Save
  + Column name will get updated in table as a new dataset ID 1.1. 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) and clicked Save as
  + Dataset Id 2 will be created in the dataset master. This will have changed column name.
  + Dataset ID 2 will get associated with Project 1
  + Dataset Id 1 will also be shown in dataset master where column name is NOT changed



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

**Example**-

Scenario# 4 –

* At Data Ingestion Level, Dataset 5 is associated with Project 4
* In schema mapping page, user has applied some changes and clicked **Save**. At this point, Dataset 5.1 will get associated with Project 4
* In Data Clean-up page, user has applied some changes and clicked **Save**. At this point, dataset 5.1 will get updated with the new changes and this will be associated with Project 4
* User clicks **Save As**. Now, the dataset 6 will be created and get associated with Project. Dataset 6 will also be available in Dataset Master for selection by user.

#### 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
* On clicking Reset, application should initialize all user-input values
* It is NOT mandatory to select the Target column.

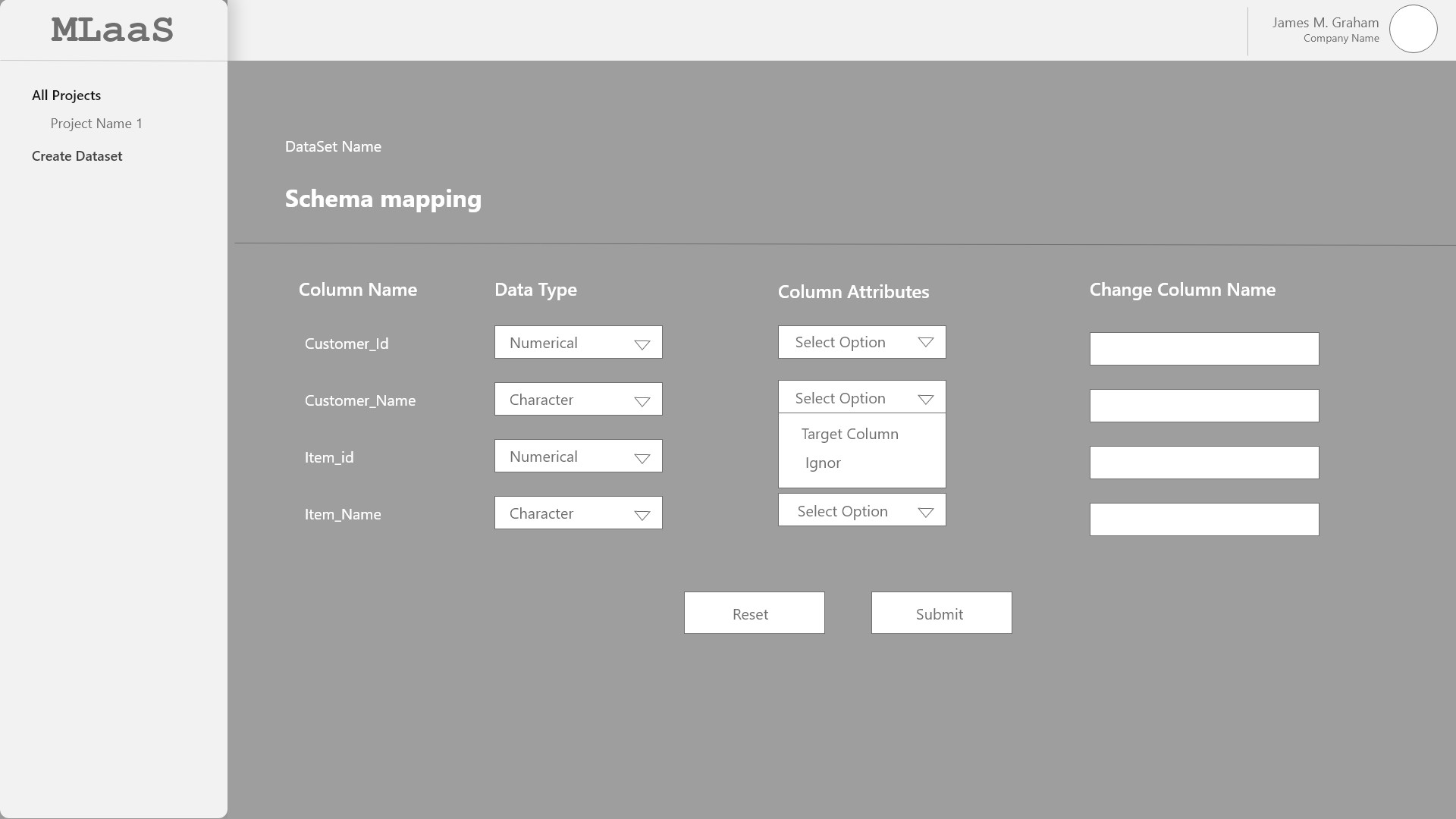


Figure 8: Schema Mapping Mock-up Screen

### 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

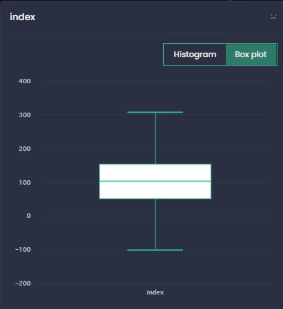


Figure 10: Box Plot Mock-up

* + Outliers plot
* For categorical datatype, Count plot is shown

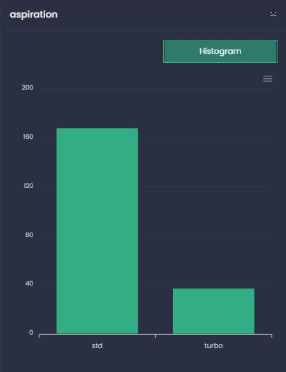


Figure 11: Count Plot Mock-up

##### 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

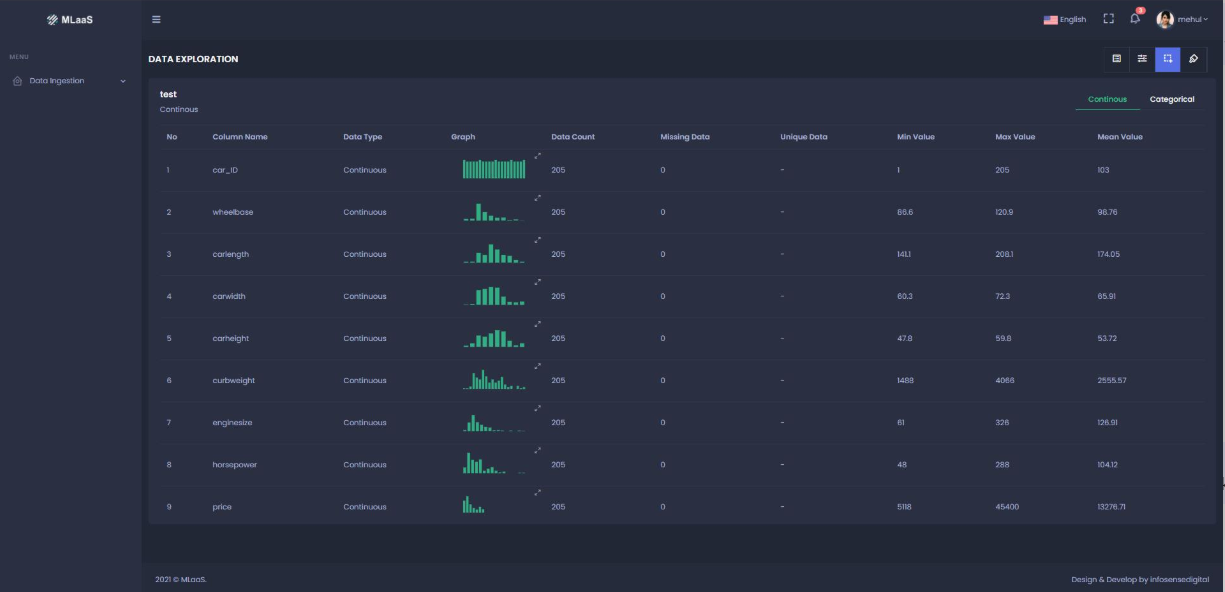


Figure 12: Data Exploration Mock-up

### Data Clean-up Page

#### User Actions

##### 1. Select column

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

##### 2. Missing values handling

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

* Replace Missing values with Median
* Replace missing values with Mean
* Drop the affected rows

##### 3. Rescaling

User can select one of the following methods for rescaling the data for selected columns-

* Min-Max
* Std. Scaler
* Robust
* Normalization

##### 4. Remove Noise

User will can choose to remove the noise form selected column by opting one of the following options-

* Regression
* Binning

##### 5. Remove duplicate data

User can select this checkbox to remove the duplicate data from the selected columns

##### 6. Select Encode type

User can select one of the following encoding types-

* One hot encoding
* Label encoding

##### 8. Replace outliers by

Users can select the column and choose to replace outliers by Median or Mean.

##### 9. [Save/Save As](#SaveFunctionality)

**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.

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

#### Validations

* Multiple columns can be selected to apply Data clean-up options
* Multiple options CANNOT be selected for Missing values handling, Rescaling, Remove Noise by, and Select Encode type
* Once the cleanup criteria are selected and user has clicked **Apply**, user should be able to reselect the columns and change the options
* The data cleanup options will get change based on the type of column selected. The set of cleanup activities will differ for Numerical and Categorical column type.

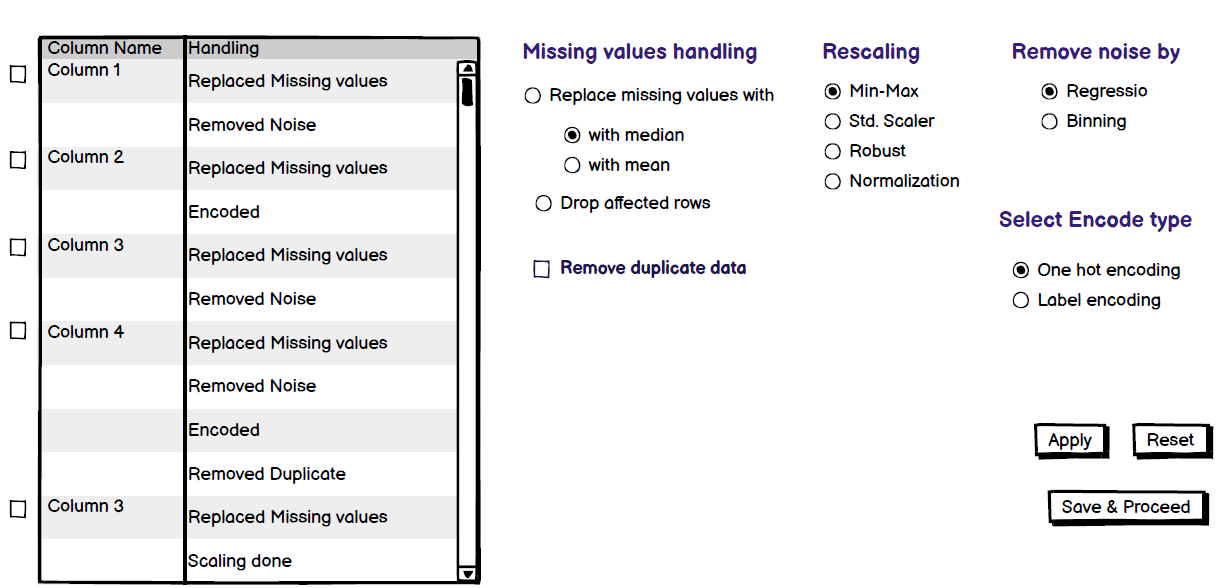


Figure 13: Data Clean-up Page Mock-up

## 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 Type

This page shows the name of Project, Dataset, and Target columns. User can select either Auto or Manual mode for the Modelling.

##### 2. 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.

#### Validations

* User can only select one of the modelling types (Auto or Manual) at a time

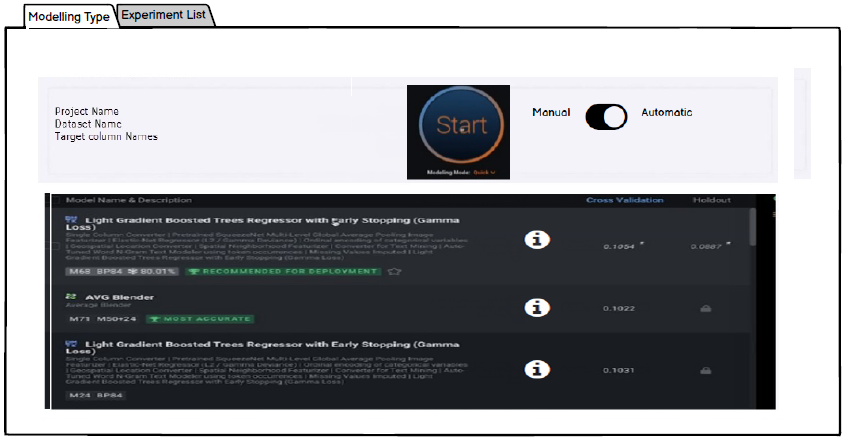


Figure 14: Modelling Type Mock-up

### Auto Modelling

When the user selects auto modelling, applications run the algorithm and shows following details-

* Model name and description
* Info icon – click this icon to see more details about the algorithm such as-
* [Learning curves](#LearningCurves)
* [Feature Importance](#FeatureImportance)
* [Performance Metrics](#PerformanceMetrics)
* [Model Summary](#ModelSummary)
* [Actual Vs Prediction](#ActualVsPrediction)
* 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

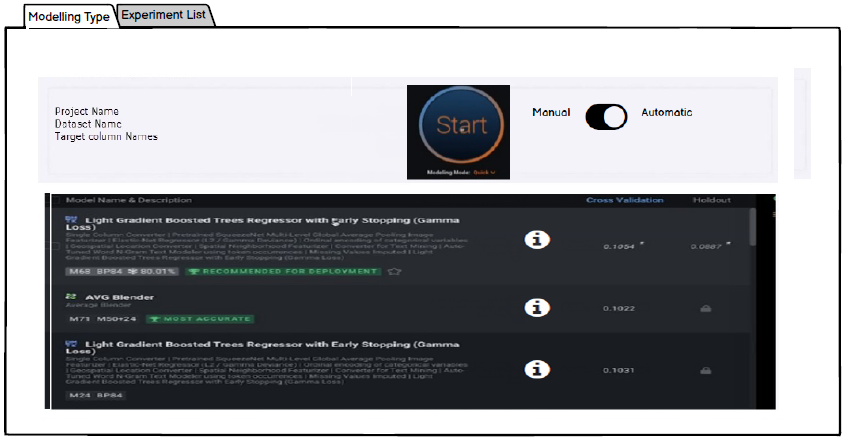


Figure 15: Auto Modelling Mock-up

#### 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**

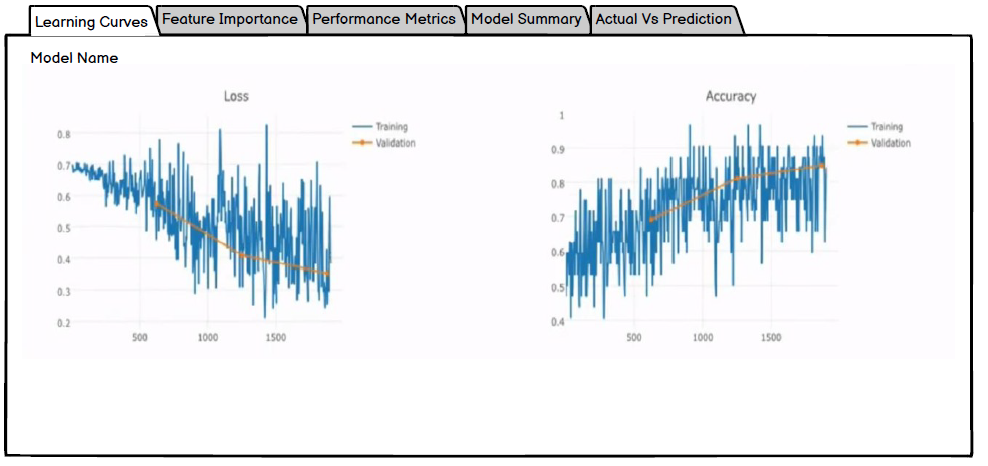


Figure 16: Learning Curve Mock-up

#### Feature importance

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

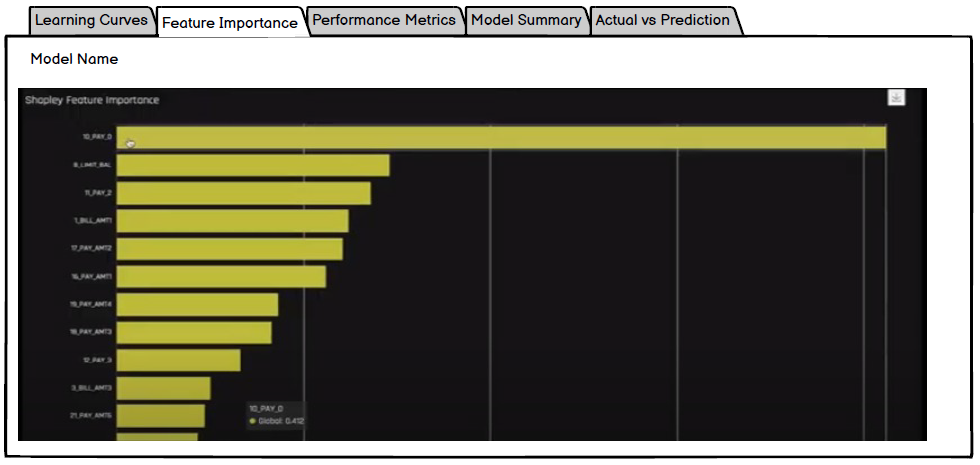


Figure 17: Feature Importance

#### Performance Metrics

This page will show following metrics for the model trained

* For regression model-
* R2-Score
* MSE
* RMSE
* MAE
* MAPE
* For Classification model-
* Accuracy
* Precision
* Recall
* Confusion Matrix
* AUC – ROC score

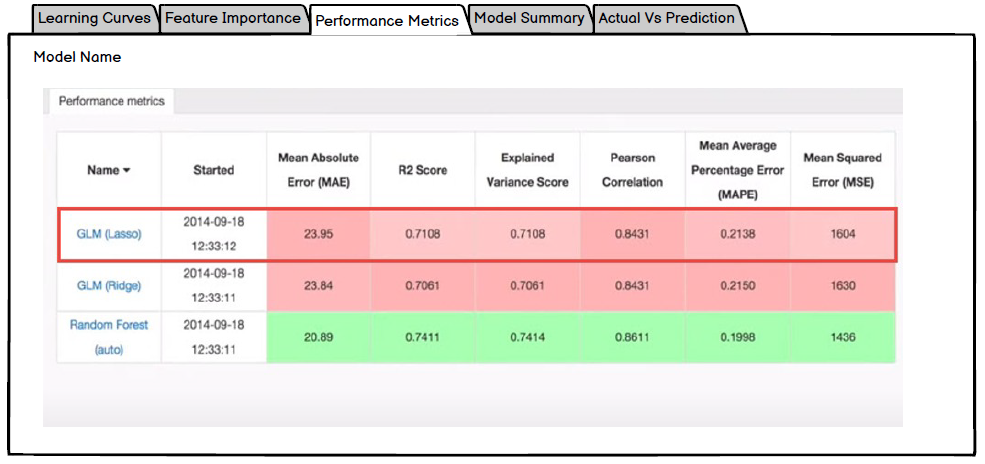


Figure 18: Performance Metrics Mock-up

#### Model summary

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

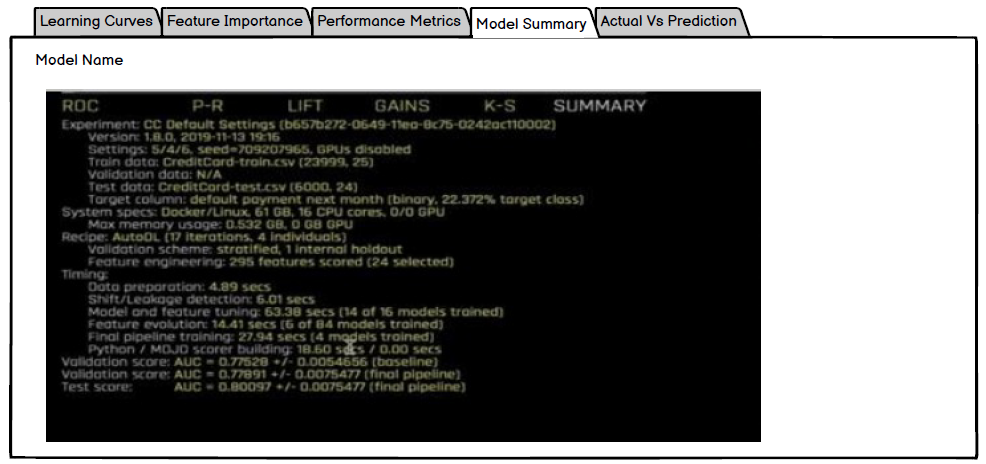


Figure 19: Model Summary Mock-up

#### 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

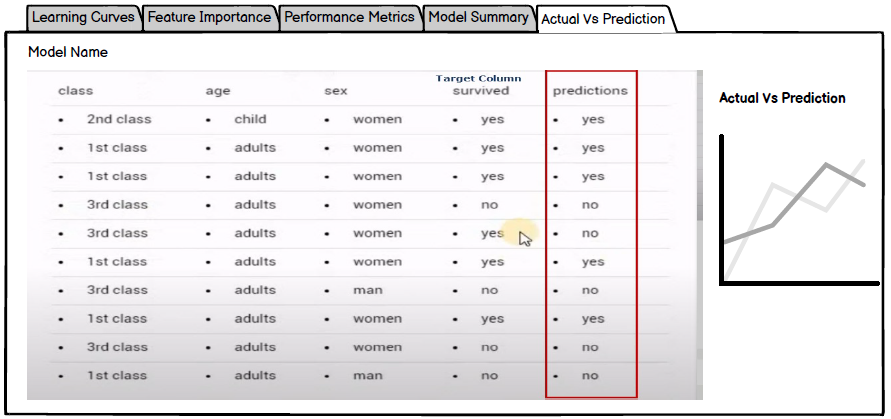


Figure 20: Actual Vs Prediction mock-up

### 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

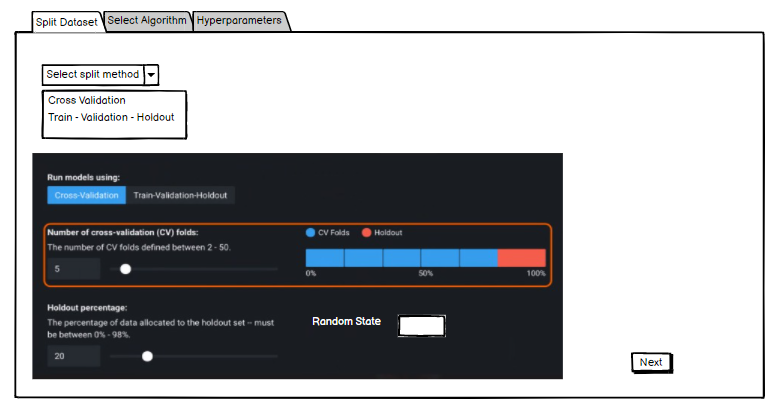


Figure 22: Split Dataset - Cross Validation Mock-up

**Train-Validation-Holdout**

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

* Split Ratio
* Random State

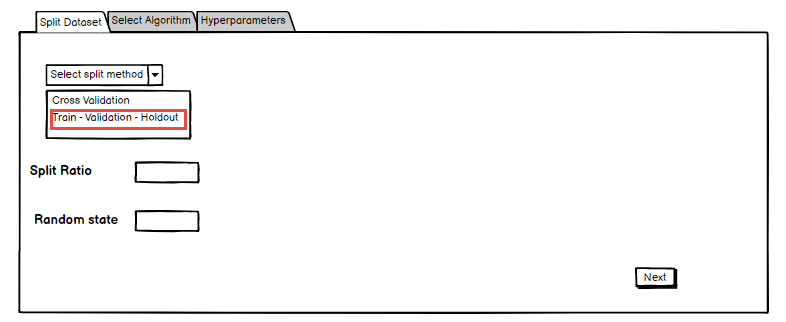


Figure 23: Split Dataset - Train Validation Holdout Mock-up

#### Select Algorithm

Select one of the following algorithms and click Next-

* Linear regression with SK Learn
* Linear Regression with Keras

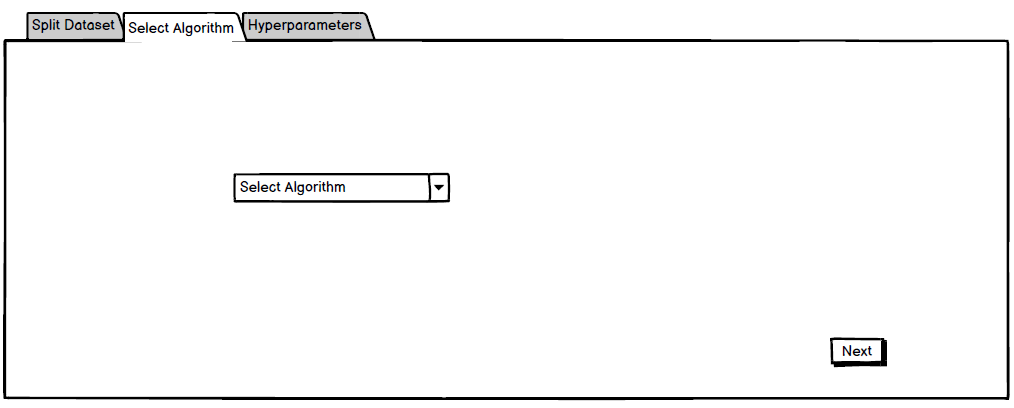


Figure 24: Select Algorithm Mock-up

#### 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

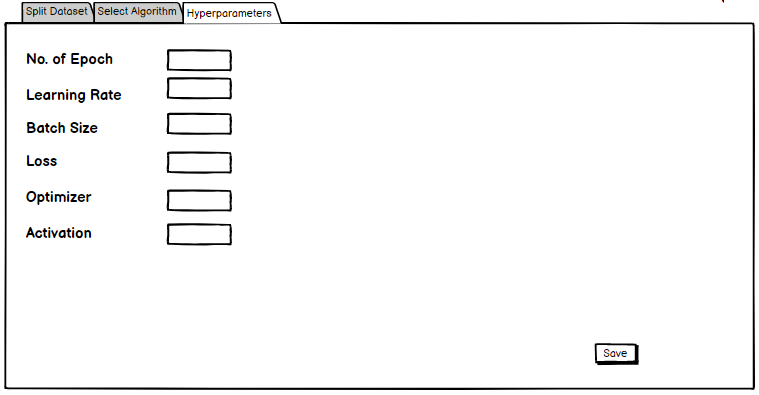


Figure 25: Hyperparameters Mock-up

After entering the parameters, user need to click Save. On clicking Save, user will be directed on modelling page where following details will be shown about the models-

* Model Name and Description
* Info icon – click this icon to see more details about the algorithm such as-
* [Learning curves](#LearningCurves)
* [Feature Importance](#FeatureImportance)
* [Performance Metrics](#PerformanceMetrics)
* [Model Summary](#ModelSummary)
* [Actual Vs Prediction](#ActualVsPrediction)
* Cross Validation Score
* Holdout Score

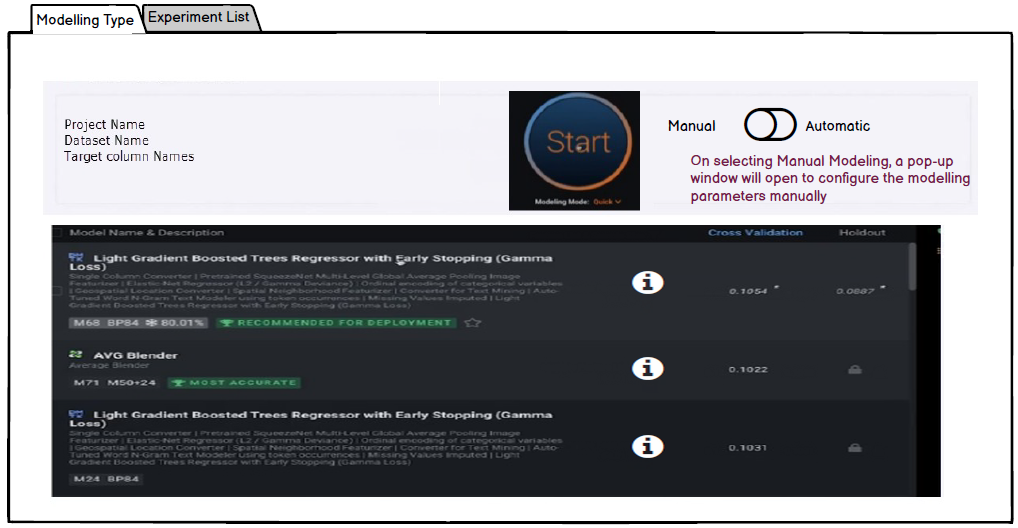


Figure 26: Manual Modelling - Model Details Mock-up

### Experiment List

This page shows the list of all the experiments done for the project. Following details are shown here-

* Experiment Name – Name of the experiment
* Model Name – Name of the model used in the experiment
* Dataset Name – Name of the dataset being used in the experiment
* Start Date and time – Date and time the experiment is started
* Accuracy – Accuracy of the experiment
* Type of experiment – Auto or Manual

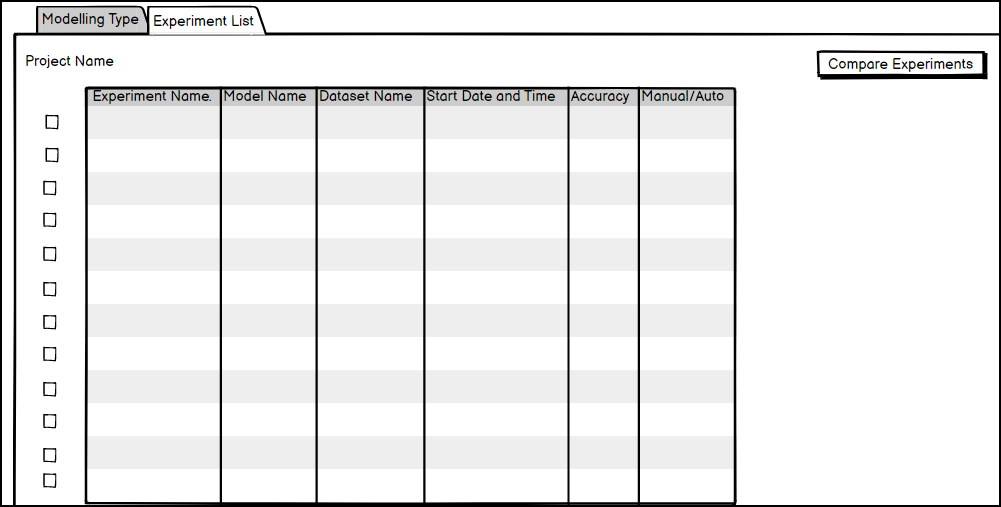


Figure 27: Experiment List Mock-up

#### 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

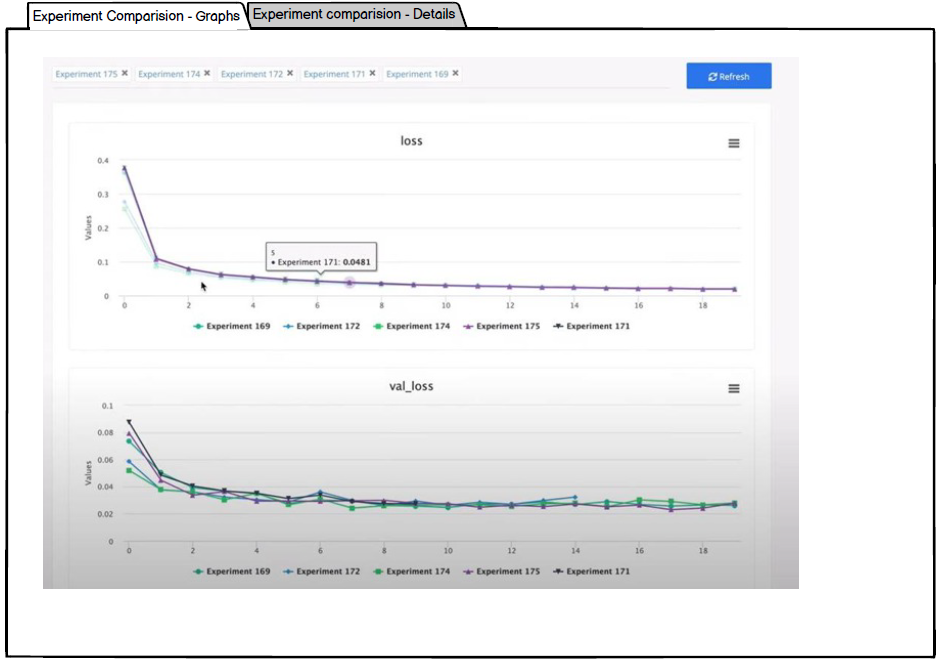


Figure 28: Experiment Comparison Graphs Mock-up

**Experiment Comparison Details-**

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

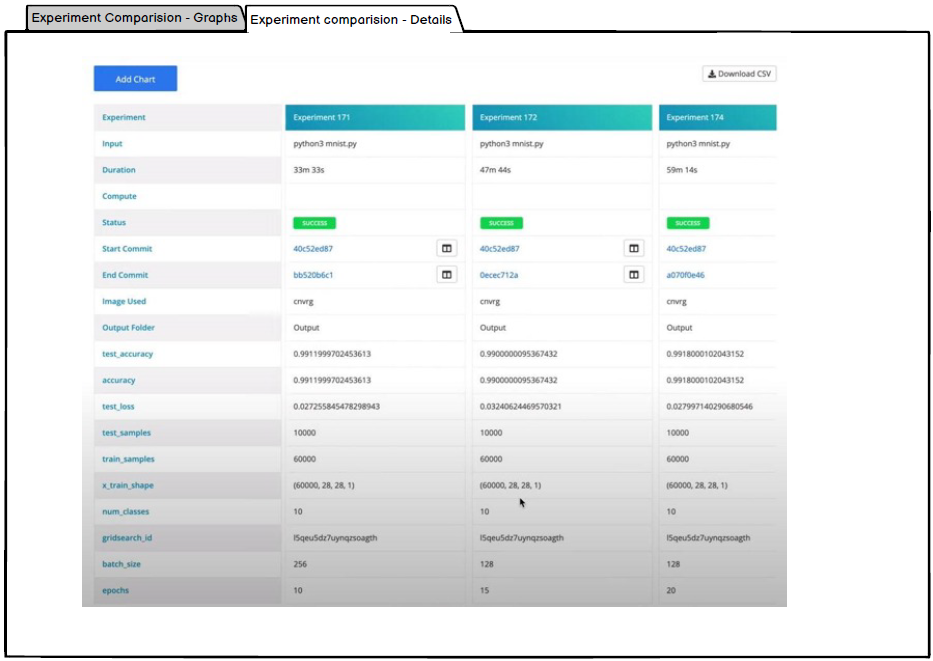
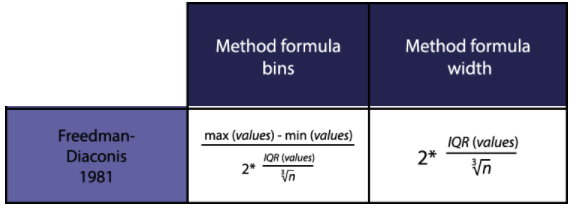


Figure 29: Experiment Comparison Details Mock-up

## 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. Save/Save As functionality

