



A QUICK USER GUIDE FOR SEDGE

September 2020

Solution EDGE - Quick User Guide

SEDGE is a cloud based predictive analytics platform for building models. Users can load structured data, which may comprise numerical, categorical, Date, Boolean and text data. The process for analysis is done in 8 stages-

Stage 1 – Uploading & Previewing data

Stage 2 – Profiling and Data protecting

Stage 3 – Generating Statistics, feature creation









Stage 4 – Data cleaning & Transformation

Stage 5 – Data Visualization

Stage 6 – Data Balancing and Model creation

Stage 7 – Model Saving and Deploying

Stage 8 – Monitoring Model Performance

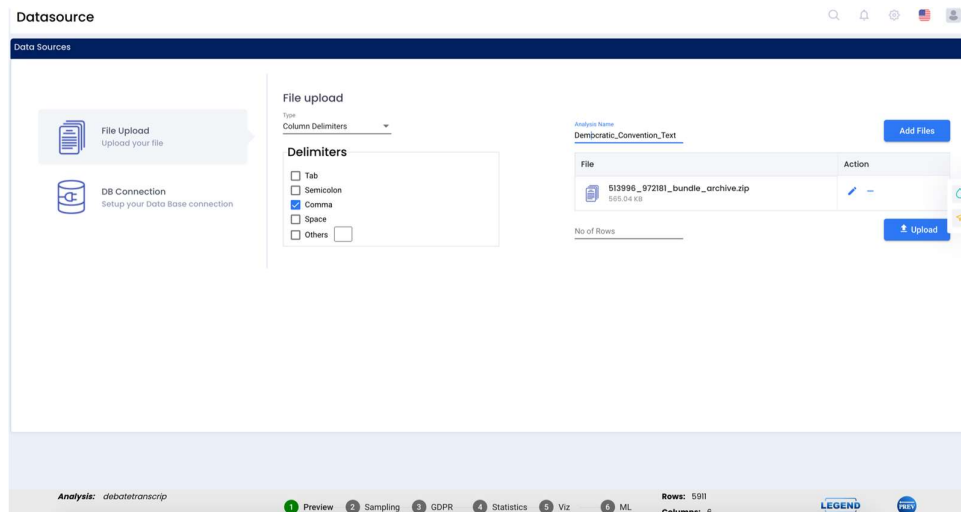
 <p>Connecting to different database, CSV files or data in cloud</p>	 <p>Data cleaning, feature creation</p>	 <p>Data transformation</p>	 <p>Feature Engineering (Features influencing target variable)</p>
 <p>Data Visualization</p>	 <p>Build models and evaluate models</p>	 <p>Deploy Models in cloud platform (AWS, Azure and Google cloud)</p>	 <p>Predict using model on streaming or batch data</p>

Stage 1 – Uploading & Previewing data

One of the first stages is to upload the data into SEDGE platform. There are different ways by which the data can be uploaded into SEDGE. The two main methods by which the data can be uploaded into SEDGE are –

1. Loading CSV / TSV or other delimited or other format files

One of the most common ways to load a file is using Comma Separated Value (CSV) format to upload. There is no limitation with respect to the size of the CSV file that can be uploaded in SEDGE. The limitation of file will be governed only by how much memory the server has. The maximum size that SEDGE has been tested is up to 200 million records by 20 columns. Since many a client's have limitation with respect to the bandwidth, we have taken care of the fact and SEDGE can upload data which is compressed and this helps to speed the data transfer across, from client side to SEDGE server, and users can just zip the CSV file and upload it into the server.



The other formats that are supported are Tab Separated (TSV), Semi colon, Space, JSON, Parquet, Avro, S3, HDFS files. These varied file formats give the flexibility to users to upload multiple file formats.

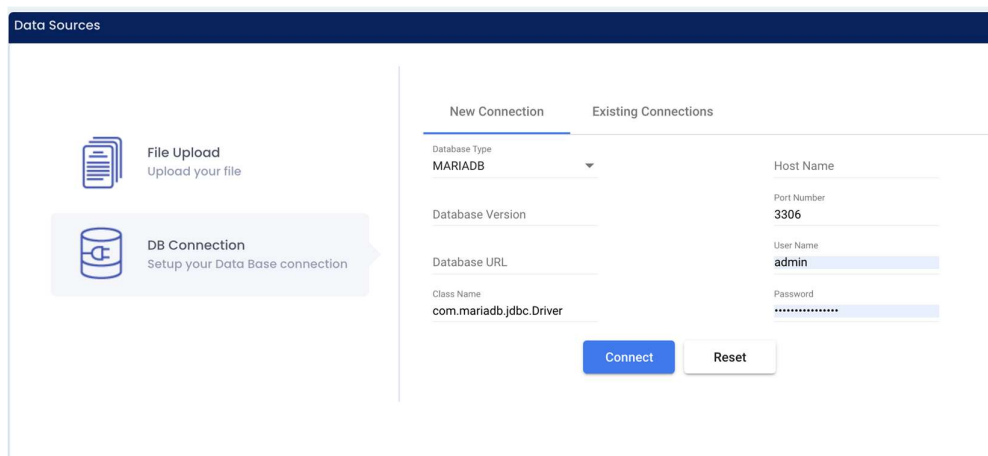
Loading of multiple files (same format)

In addition to the above, users can also load multiple CSV files and can join them similar to Join query used in SQL.

2. Uploading the data from a database

Table with database can also be loaded into SEDGE, to handle the same the user needs to select the type of Database (MYSQL, MariaDB, Oracle, MS SQL, Postgres and SQL servers).

User can select the database and enter the details such as IP address, DB instance name, password and log into the database, and select the table which has the data. In addition, the SQL query builder can be used to connect multiple tables and create a normalized data, which can also be loaded into SEDGE.



The screenshot shows the 'Data Sources' section of the SEDGE interface. On the left, there are two options: 'File Upload' (Upload your file) and 'DB Connection' (Setup your Data Base connection). The 'DB Connection' option is selected. On the right, there are two tabs: 'New Connection' and 'Existing Connections'. The 'New Connection' tab is active, showing a form to configure a new database connection. The form includes fields for 'Database Type' (set to MARIADB), 'Database Version', 'Database URL', 'Class Name' (set to com.mariadb.jdbc.Driver), 'Host Name', 'Port Number' (set to 3306), 'User Name' (set to admin), and 'Password' (masked with asterisks). There are 'Connect' and 'Reset' buttons at the bottom of the form.

3. Previewing the loaded data-

Once the data is loaded into SEDGE, the preview screen displays the columns which has been loaded and the datatype of each columns. The contents of each column is automatically identified to display the datatype such as categorical, numerical, Boolean, date and Text based data types. The preview displays top 100 data rows in order for users to understand the structure of the loaded data.

Data preview

Analysis: mentalhealth

Preview Sampling GDPR Statistics Viz ML Rows: 496 Columns: 65

Stage 2 – Profiling and Data protecting

The design of SEDGE is kept simple and self-intuitive, which ensures that the users with minimum training can move across various screens. By clicking on the next button, the user moves to the profiling screen. The purpose of the screen profiling serves 3 main purpose

EDGE Profile your dataset

Home

Data Explorer

Analysis

Statistics

Visual Analytics

Predictive Analytics

Target

Select the target to predict

Severity[4]

Sampling

Do you wish to do sampling on your data?

YES NO

Type Random

Size % of a variable

Percentage 1 to 99

Variable Select

Apply

Data Protection

Do you wish to protect personalized data?

YES NO

Machine Learning

Would you like EDGE to run automatic prediction?

YES NO

Data Overview

Dataset Info

Number of variables 38

Number of observations 35130

Total Missing (%) 9.14%

Numeric 14

Categorical 35

Date 4

Text (Unique) 1

Warnings

Features Target

Column	Type	Unique	Missing	Missing %	Mean	Median	SD
ID	N	35130	0	0%			
Source	C	3	0	0%			
TMC	N	17	10316	29.37%	208.273	201	21.743
Severity	C	4	0	0%			
Start_Time	D	34317	0	0%			
End_Time	D	34275	0	0%			
Start_Lat	N	32921	0	0%	36.592	36.04	4.895
Start_Lng	N	32398	0	0%	-95.821	-91.057	17.399
End_Lat	N	9537	24814	70.63%	37.601	37.816	4.837
End_Lng	N	9814	24814	70.63%	-100.355	-96.918	18.526
Distance(mi)	N	2445	0	0%	0.289	0	1.54
Description	T	29655	0	0%			
Number	N	5218	22774	64.83%	6132.04	2805	13223.941
Street	C	13414	0	0%			
Side	C	2	0	0%			
City	C	4028	0	0%			
County	C	913	0	0%			
State	C	49	0	0%			
Zipcode	C	15918	10	0.03%			

Analysis: USAccidents.June

Target: Severity[4]

Preview Sampling GDPR Statistics Viz ML Rows: 34864 Columns: 54

1. Selecting the Target variable

The screen profiling as the name suggests profiles the data which has been uploaded and displays the high-level data overview, basic statistical information for categorical, numerical data types. The purpose is to understand the data types and the basic statistical information. The user selects the target variable for which the predictive analytics has to be demonstrated. On selection a column chart is created with the count of the classes within the target variable. It is mandatory to select the target variable as this will allow the user to proceed to the next stage.

2. Data Sampling

Sampling is another optional function which the users can utilize here, as this helps to sample out the data from the population set. There are various mechanisms available for users to sample the data set out. We have created various different means by which the data sampling can be performed, and the users can use these different sampling techniques to sample out the dataset.

Why do we need to sample the dataset? Sampling is an optional feature, and users can ignore sampling, however sampling helps to reduce the large dataset to a smaller dataset which helps in the performance, especially with data transformation, data imputation, model building etc. SEDGE has a pipeline feature built into it, which ensures that whatever set of data imputation, transformation, data cleaning is done with the sample data set, the same pipeline functionality will be done with the dataset which is subsequently passed in from production to SEDGE for predictive analytics.

3. Data Protection

General Data Protection Regulation (GDPR) rules in Europe are in force and data protection plays an important role in ensuring that data related to individual privacy is protected by the use of pseudonymization. SEDGE has a built in feature to identify columns which need data protection. Apart from the above, the user also has the option to select the columns which need to be

pseudonymized, which helps to protect the sensitive fields from data processors.

The screenshot displays the EDGE application interface. On the left is a sidebar with navigation icons for Home, Data Explorer, Analysis, Statistics, Visual Analytics, and Predictive Analytics. The main content area is titled 'General data protection regulation' and contains a 'Questionnaire for General Data Protection Regulation'. The questionnaire has several sections with radio button options for 'yes' or 'no':

- 1. The Personal data being uploaded does it have any special category fields-
 - a. Personal data revealing racial or ethnic origin.
 - b. Political opinions.
 - c. Religious or philosophical beliefs.
 - d. Trade union membership.
 - e. Genetic data and biometric data processed for the purpose of uniquely identifying a natural person.
 - f. Data concerning health.
 - g. Data concerning a natural person's sex life or sexual orientation.
- 2. For Personal data, I have explicit consent from data subjects whose personal data is being processed?
- 3. If the data contains Personal data, will it be pseudonymized during processing?
- 4. What is the source of data?
- 5. For Personal data under which category mentioned below are you allowed to process the Personal Data
 - a. expressly authorised by Union or Member State law to which the controller is subject,
 - b. including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and
 - c. to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller,
 - e. the data subject has given his or her explicit consent.
- 6. Who will be accessing this data?
- 7. How long will this data be stored? (Set to 3 Months)
- 8. What is the purpose of processing personal data?

Below the questionnaire, there are buttons for 'Auto ML' and 'Custom ML'. At the bottom, a status bar shows 'Analysis: USAccidentsJune', 'Target: Severity[4]', and navigation buttons for Preview, Sampling, GDPR, Statistics, Viz, ML, and Legend. On the right side of the interface, there is a table with columns 'Personal Data FI...', 'Values', and 'Detail'. The table lists various data fields such as Side, City, County, State, Zipcode, Country, Timezone, Airport_Code, Weather_Time, Temperature, Wind_Chill, Humidity, Pressure, Visibility, Wind_Direction, Wind_Speed, Precipitation, Weather_Condition, Amenity, Bump, Crossing, Give_Way, and Junction, each with corresponding values and a detail icon.

4. Auto ML

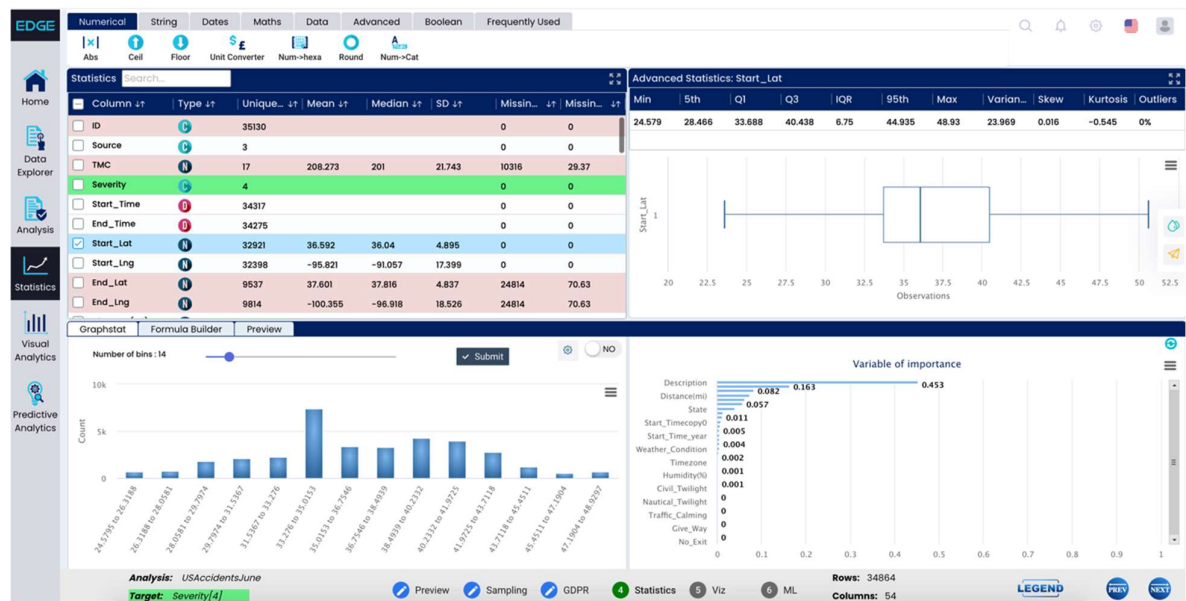
The system has the feature of automatically cleaning the data, performing data transformation and running the models. The Auto ML feature helps users who want to quickly create the model from the sampled data. This is an optional feature and can be skipped by the users, for cases where they want to perform, data transformation, imputation and data cleaning.

Stage 3 – Generating Statistics, feature creation

This section has multiple functionalities, such as

1. Generating detailed statistics

Navigation from profile page to statistics can be done by clicking on the Next button which takes the users to the statistics page. Here the user can view the descriptive statistics of each data type, such as mean, median, missing row count / percentage, distinct counts, minimum, maximum, 1st quartile, 3rd quartile, inner fence, outer fence, kurtosis, skewness and outlier data. These data's, help the user to get a detailed understanding of each column.



2. Identifying variables of importance

The variables which are important and influence the target variables are sorted based on the importance value. This gives a good understanding of which variables are important and influence the outcome of predictive analytics. There are a lot of imputation and data transformations that can be performed in this section. There are close to 60+ built in functions which can be used for transformation, data cleaning, imputation, changing data type, string, numerical and date operations.

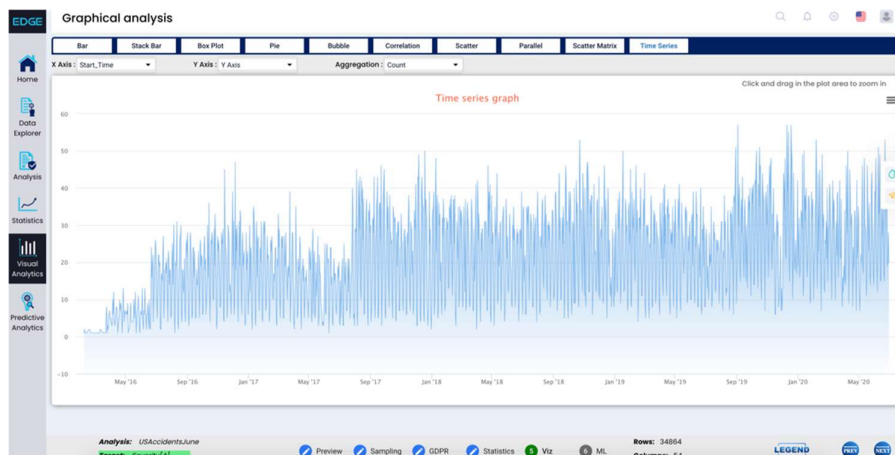
Stage 4 – Data cleaning & Transformation

3. Data Cleaning and Data Transformation

One of the key functionality of data models is the ability to clean the data, so that the existing features can be strengthened by removing data and ignoring columns which may not be of value. SEDGE has the ability to perform data transformation by using the formula builder, which is a powerful feature which uses the built-in functions to transform data. Examples of some of the functions are string manipulation, numerical functions, date / month / year and time stripping, use of case statements, if-else conditions. The power of the programming is in the hands of the users. With little or no python programming experience, the power of programming, using the built-in functions have been empowered to users.

Stage 5 – Data Visualization

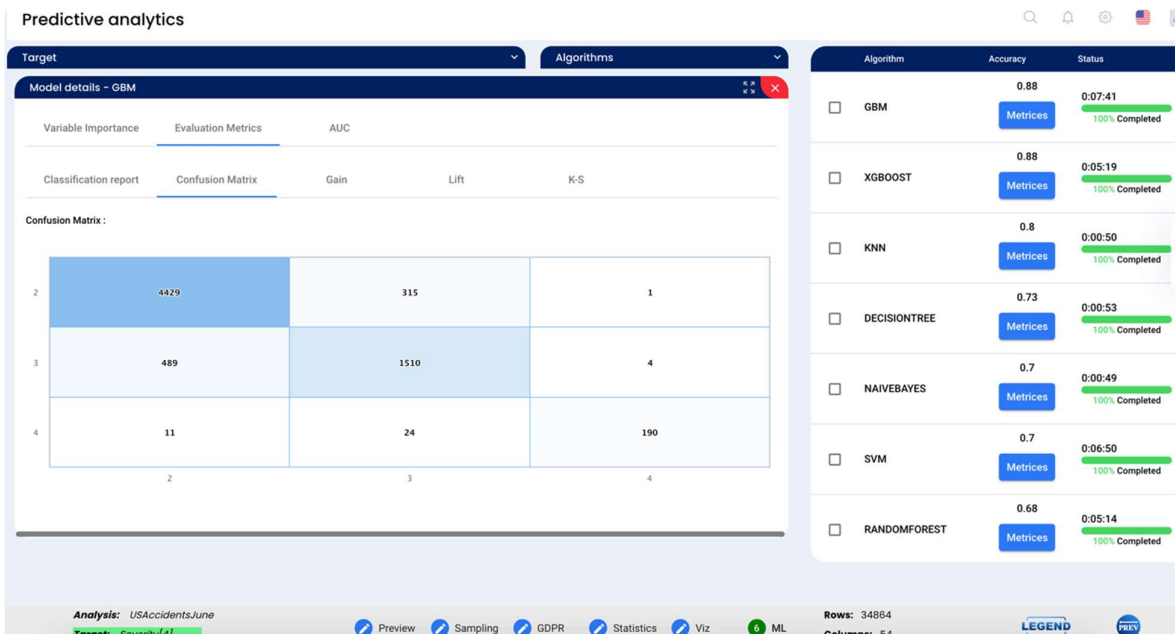
A quick way to visualize the data is to create charts of different fields, such as for categorical data visualization through column, stack and pie chart. For Numerical fields a good way to visualize the data is to view it through boxplot, scatter, bubble chart and correlation matrix. Date fields can be viewed through Time series plots. These are some of the charts, there are many more visualization charts that users can utilize to explore the data.



Stage 6 – Data Balancing and Model creation

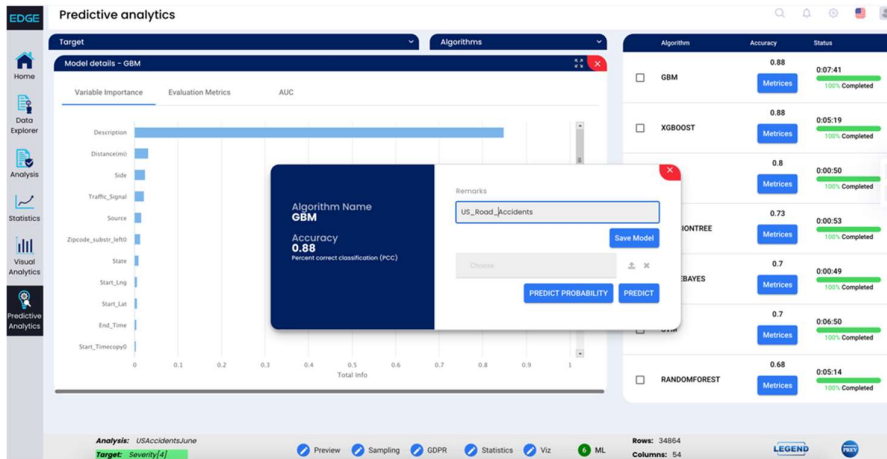
After performing data transformation, data cleaning, and exploring the data through data visualization, the next step is to create the model. Before running the model, the users should check if the classes within the target variable needs to be balanced. If the data needs to be balanced users can balance them by up sampling or downsampling as the case may be.

The users can select the list of models which are provided and click on start Learning. This will trigger the learning of the target variable using the models provided. Once the models have been learnt the accuracy of each model will be displayed. Clicking on the metrics, will display the confusion matrix, Area under the curve (AUC) and other charts such as gain, lift and other charts.



Stage 7 – Model saving and deploying

Once the model with high accuracy and maximum AUC is selected, the user can save the model and give the name to the saved model. The model can be deployed by clicking in the deploy button.



Stage 8 – Monitoring the model performance

Once the model is deployed, the model needs to be monitored to measure the performance, as to whether the data drift would affect the accuracy of the model.