Describe what we did

FEATURES

Model training UI

Model training takes input sample size

Pre processing, cleaning and formatting

Model Predict UI

Maps Data point visuals with clustering

Data analysis

EDA charts

Histograms

Heatmaps Correlation matrix

ALGORITHIMS

Clustering (Unsupervised)

-We implemented auto clustering when the sample size is more than 5000.

-We implemented Fast marker clusters function on the map

Decision Trees (Supervised)

-We implemented decision trees with decision tree classifier function

Linear Regression (Unsupervised)??

Naïve Bayes (Supervied)??

SVM?

Random forest?

PROCEDURES

CREATING THE MODEL

* We take a random sample
* We preprocess and clean up and format
* We split train and test based on input
* We pick a model
* We fit the model
* We predict the test group to determine the accuracy

PREDICTING

* We input parameters to predict
* We use input parameters and predict using the model
* The prediction is shown

DATA ANALYSIS

* While training, the application also analyses the random number samples
* Histogram, correlation, and EDA is shown

INTERACTIVE DATA VISUALIZATION USING GEOMAPPING

* The data points are plotted in the map using their geo coordinates
* Details of the points are shown in a popup
* Automatic Clustering is implemented using fastmarkercluster function from folium to optimize speed of rendering the map.
* Automatic clustering is set for sample size of greater than 5,000 points

LOG OF UPDATES:

March 27

1. Created folder Heirarchy
2. Created dummy code and pushed

April 1

1. Created Geomaps with datapoints

April 2, 2020

1. Created Pyqt5 application and Linked Geomap with datapoints
2. Used Qwebview, Folium by Leaflet for the Geo maps
3. Added buttons to GUI and better looking Geo map markers (circle markers)
4. Created EDA and Graphs

April 3,2020

1. Preprocessing, can handle nan values in the map for displaying marker descriptions
2. Cleaned up and segmented code into different scripts for easier management
3. Edited and modified paths to work on different computers. Added installs for unrecognized packages
4. Faster load times for the map. Paths and preprocessing of data created and modified
5. Added tabs and EDA statistics script and connected images to the GUI
6. Automatic clustering of maps for training samples of morethan 5k
7. Created summary of statistics and attached to gui with buttons to load
8. Fixed gui colors

April 4

1. Updated GUI Layout and style
2. Group proposal created
3. EDA graphs modified

April 5

1. Edited Groupbox 1 with input boxes for prediction and training

April 6

1. connected training and prediction into script classes
2. model still to be determined.
3. removed button veiw stat summary
4. made traning button take random sample of given number
5. automatic updates of historam screen, stat table and geo maps
6. screen updates happen after model building
7. building new feature soon.
8. Fixed Fast cluster markers through java script callback
9. colored individual points yellow with radius 100
10. edited histogram size and color
11. geo map visible to the right groupbox
12. rounded off summary to 4decimals
13. Decision tree model created.
14. cleaned up data for processing.
15. filled in missing nans with means of column
16. coneccted inputs from screen.
17. train input range from 10-90
18. sample size range from 100-200,000
19. accuracy of model in lcd box connected

April 7

1. Correlation Heatmap created and added to the tabs

April 8

1. Updated color maps
2. fixed gui layout
3. Color coded severity in the map
4. Fixed text pop up in maps (severity, address, description)
5. created a status bar with Qlcd number
6. color coded status while before and after model training

April 9

1. Created content of predict functions
2. changed the shape to be able to use predict function
3. Connected predict class to button and severity output
4. reset predict severity to 0 if train model is pressed
5. Created a message box to output when there is no model
6. or when there are missing values when predict button is clicked
7. moved test codes into a folder

April 10

1. Created Random Forest
2. Created ordinal logistic regression
3. Connected random forest model

April 12

1. Updated accuracy calculation for rf model and Logistic regression

April 14

1. Presentation PDF created
2. Created more training algorithm functions
3. Connected them to the application
4. Logistic Regression, KNN, SVM, and Naive Bayes

April 15

1. Added histogram functions and saved images
2. connected to application tabs
3. update after training connecteed

April 19

1. changed status bar
2. synced geomap with bar
3. using predict algorithm
4. Updated Knn default to 1
5. Creted popup for knn input values
6. PCA added for knn for increased accuracy

April 20

1. created status bar buttons to visualize models

April 21, 2020

1. Geo Map load times not in sync with status bar. Created different prediction models. SVM with poly kernel, Linear regression, Polyfit function, Logistic regression, and Curve fit with sinusoidal function. The data set looks sinusoidal. Best model was SVM with poly kernel and Curvefit. Curve fit was better in prediction but looking for a better model.
2. created a fit curve function to estimate map render times
3. now have a more accurate status bar update
4. the function chosen is sigmoid since the data points look it
5. the prediction can be displayed by clicking the geomap button

April 22, 2020

1. Decision tree colors converted to dark mode previously by using plot\_tree color was good but look was bad. Decided to switch back to graphviz. Better look but color is bad. Inverted colors for dark mode and set black background into grey color to match application. Tree node colors have to be changed by setting inverted colors and inverting it back to retain node colors
2. Scroll bar tab with concatenated histogram PNG all combined in a single tab
3. The histogram PNGS were previously all separated in different tabs
4. Decision tree Visualization dark mode updated beeter look
5. Create scatter plots for the features
6. Connected to the tabs
7. Automatic gradient scatterplot for sample size greater than 1000. Used sns.jointplot
8. Manually created dark mode colors

April 23