Final Project UW Boot Camp

Sonmi Kim

Role : Square

NOTE: Sample is often wrong because of initial testing was done

**STEP 1: Clean / Visualizing – Using Jupiter notebook**

Sample data segment:

Table

Description automatically generatedTable

Description automatically generated

Real data segment to run a model : input of vehicle ID expected from front end

NOTE: Vehicle id represents unique vehicle in the full table, dropped make model here

Table

Description automatically generated

Sample data description :

Table

Description automatically generated

Real Data Description:

Table

Description automatically generated

Sample Data correlation:

Table

Description automatically generated with medium confidence

Sample data correlation diagram:

Square

Description automatically generated with medium confidence

Real Data Correlation diagram:

Square

Description automatically generated with low confidence

Sample data graph

Sample data Graphs:Chart, bar chart, histogram

Description automatically generated

Real data graph

Chart, histogram

Description automatically generated

Sample data scattering:

Chart, scatter chart

Description automatically generated

Real data scattering:

Chart, scatter chart

Description automatically generated

**STEP 2 : Modeling – Using Colab**

**X-TRA – visual – Colab trying to visualize Real data**

Attempted to describe why the sample data is more effective to visualize than the real data, it is important to know that it is impossible to distinguish data from the diagrams that I attempted, although EDCF (Empirical Distribution Function ) seems a little better than others, I stopped attempting to see further. – This was part of the test BTW

A picture containing graphical user interface

Description automatically generated

Chart

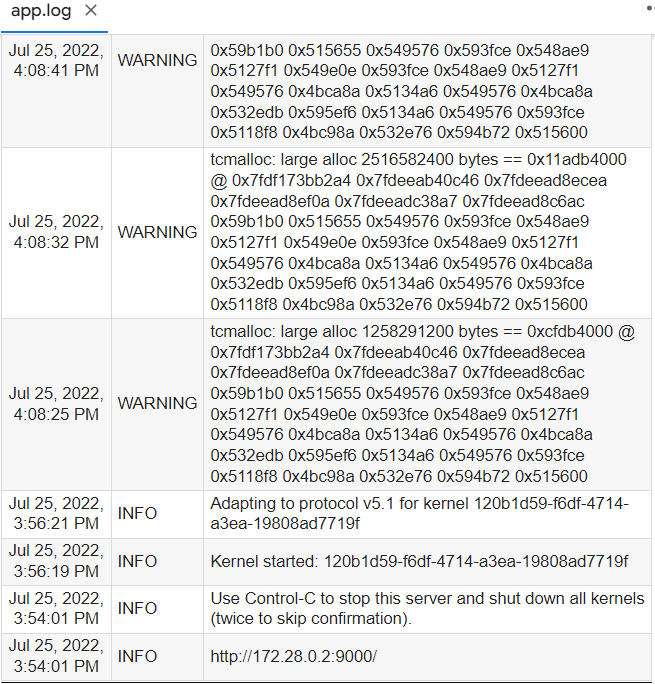
Description automatically generated

**MODELING :**

Logistic Regression Multiple Classes

* The variation of the credit score can be a lot. However, in the big picture, it can be categorized as a bunch of classes
* When Logistic Regression is used, it can predict like hand-written digits or sport, politics, stocks, etc.
* formular -> f(x:w) = 1/1+e**-w**t\*x
* Used sigmoid

I Can’t run the Model with full data – not capable of it. RAM crash



**First 1000:**

**RandomForestClassifier:**

Text

Description automatically generated

**KNN:**

A picture containing text

Description automatically generated

**First 5000:**

Text

Description automatically generated

Text

Description automatically generated

First 10,000:

Text

Description automatically generated

Text

Description automatically generated