Homework 3. Name: Joshua Chang

Part 1:

Description of variables	# Variables Created
Original fields from the dataset excluding ‘record’ and ‘fraud label’	8
Date of week target encoded (average fraud percentage of that day)	1
New entities combining/concatenating different original fields	9
Days since Variables: # days since an application with that entity has been seen	161
Velocity: # records with the same entity over the last 0, 1, 3, 7, 14, 30 days	184
Frequency Variables: This set of variables include the number of applications submitted with this various combinations between the fields and the combinations of fields over the past 0, 1, 3, 7, 14, 30, 60 days	3542
Risk Variable: The likelihood of fraud for any day of the week	1
Maximum Indicator Variables: This set of variables outputs the maximum number of times an attribute shows up over the past 0, 1, 3, 7, 14, 30 days	92
Age Indicator Variables: This set of variables include maximum, mean, and minimum age when application was submitted	69

Part 2:

- **Business problem**: Synthetic identity fraud. When someone combines real and fake personally identifiable information (PII) to create an identity and commit fraud.

- **What events/things the algorithm will score for possible fraud**: We will score applications for a product or service. Assign a numerical probability that the application is synthetic identity fraud.

- **Likely data and fields**: The algorithm will look at and assign a probability to each application. Full name, date of birth, SSN, full address, home phone number, and other identifying information. It would be best to have many examples of known synthetic identity fraud so we could build supervised models. Other auxiliary data we could use include other identity data about people, for example:
    - Known fraudulent or risky locations, people, emails (e.g. known compromised emails)
    - Source data (e.g. IP address or Device ID in which the application was sent from)
    - Phone book data (name, address, phone number)
    - Credit bureau header data (lists of all adults with their name, date of birth, SSN, address, phone numbers)

- **What to look for**: Indications of a made-up, or synthetic, identity. Do the identity components of each application seem to go together? Have we seen any of these identity components in suspicious activities before? Have we seen a lot of applications coming out of the same SSN or address? Have we seen a lot of suspicious applications coming from the IP address or Device ID in which the application was submitted? Has the address or home phone number in which the application was submitted been known to be fraudulent before? Has there been many applications submitted from the same address (IP or physical) in a short period of time? After verifying other suspicious patterns such as relatively high frequency of applications, what is the most common day of the week in which these applications are being submitted?