Homework 3. Name: Joyce Xinyi Jiang

Part I

Description of Variables	# Variables Created
Age when Apply Contains age of the applicants when application was made.	1
Day of Week Target Encoding Contains target encoded day of week (DOW), where dow_risk demonstrates risk of fraud in the particular day (exp: Monday, Tuesday...)	1
Days Since Variable Amount of days since an application with the applicants occurred. **Velocity** Amount of records with the same attributes over the past [0,1,3,7,14,30] days.	161
Relative Velocity Demonstrates frequency where an attribute is appearing in the recent days [0,1] versus other days [3,7,14,30].	184
Count by entities Count the amount of unique values for one attribute in comparison to other attribute over the time period of [0,1,3,7,14,30,60] days.	3,542
Maximum Indicator The maximum count of the occurrences of each attribute within a certain period of time.	92
Age Indicator The age of the applicants when they applied, and it contains the min, max and mean age of the applicants by each attributes.	69
Total Variables Before removing duplicates. Including 2 orignial fields of 'Record' and 'fraud_label', however, it is not considered as a independent variable.	4,052

Homework 3. Name: Joyce Xinyi Jiang

Part II

- **Business problem:** Synthetic identity fraud refers to the type of fraud where criminal creates a fake identity using a mixture of real and fake information. For example, a real Social Security Number (SSN) and a fake name.

- **What events/things the algorithm will score for possible fraud:** Personal Identifiable Information (PII) of the applicants from documents or records. The algorithm will assign a numerical probability score for the risk of synthetic identity fraud, then flag potential fraud cases for further human investigation.

- **Likely data and fields:** There are several areas that the algorithm could consider when scoring the risk of synthetic identity fraud and flagging cases that requires more attention. Linkages can be built between fields to provide more identifiers and variables that may surface insights. Some examples of useful data are:
    - Personal Identifier Information: Such as Name, SSN, Address, Phone Number, etc
    - External Databases: Public records, credit reports, compliance databases are valid sources to cross check information
    - Days since last seen: Gaps between activity of the applicant
    - Velocity: Frequency of records received from the applicant over a period of time
    - Relative velocity: Ratio of short-term velocity to long-term velocity, it is also potentially comparable to trends observed in similar demographics
    - Combination of individual's identifier fields, such as name and address, to make the record more unique
    - Number of unique field or records for that group (exp: unique records of phone numbers, birthdate, etc)

- **What to look for:** We are looking for indicators of potential fraud in personal information that may appear to be irregular or inconsistent. We may also identify red flag based on trends or patterns based on individual or demographic. Some potential red flags for example:
    - A cluster of applicants with similar or identical personal information (address, SSN, DOB, phone numbers)
    - Inconsistencies or irregularities with similar personal information across multiple records
    - Similar/identical information that appears in high frequency
    - Inconsistent trends observed in entities' activity or last seen, such as spike, long pauses, etc
    - A phone number that is payphone or internet-based phone service
    - Irregular or inconsistent address or phone number across short time span
    - Mismatched DOB and SSN
    - Use of SSN/PII that belongs to a deceased person