**California Community College Technology Center**

**Spam Filter Project: Data Management Strategy**

****

August 21, 2017

1. Objective

Review data requirements for Spam filter project and agree on overall data management requirements including security/privacy requirements for PII data.

As discussed in the kick-off meeting, Infiniti/SeyVu will have a two-stage approach.

* 1. Stage 1 (Non-PII Data Analysis)

Require only non-PII data. Goal is for Infiniti/SeyVu team is to familiarize ourselves with the data and do exploratory data analysis.

**Next steps:**

1. Review and approve what constitutes PII? [**Owner:** Jeff/Patty]. Our initial stab at the PII list is provided below (Section 1.4.1; Jeff’s suggestions after security review annotated in table)
2. Extract non-PII data from 1 college (that had high incidence of fraud) for the last 12-24 months and provide to machine learning team [**Owner**: CCCTC]

**Timeline:**Can we shoot for **Aug 30rd, 2017**? (1 week from meeting)

* 1. Stage 2 (Hashed PII Data Inclusion)

Post kick-off meeting, machine learning team held an internal review and have come up with **a proposal that drastically reduces the need for access to raw PII data**.

Anonymize/hash raw PII data prior to hand-off to Infiniti/SeyVu. From a machine learning perspective, that is a reasonable place to start and in many instances, this could be sufficient to get a high-quality model.

~~In order to accomplish this, Infiniti/SeyVu will need access to a very small subset of PII data. We will develop the scripts necessary to do the hashing. CCCTC can run these scripts on extracted data and provide the complete anonymized/hashed dataset.~~ CCCTC has routine that can enable HMACs on PII data.

**Required decisions:**CCCTC team reviews proposal and provide feedback to proceed.

**Next steps:**

* ~~Step 1: Provide access to very small subset of PII data (500 fraud data-points and 500 non-fraud). From same data-set as non-PII data provided in stage 1 [~~**~~Owner:~~** ~~CCCTC].~~ Not required.
* ~~Step 2: Develop hashing scripts~~~~[~~**~~Owner~~**~~:~~~~Infiniti/SeyVu].~~ Not required.
* Step 3: Extract data, run hashing scripts and provide output to ML team [**Owner**: CCCTC]
	1. Stage 3 (Contingent Stage)

If results using above step is not satisfactory, Infiniti/SeyVu might need access to PII data to better understand fraud trends.

We believe this might not be necessary and should first wait for initial results (~6 weeks)

* 1. Specifics

* + 1. What data constitutes PII?

Below is Infiniti/SeyVu teams’ first draft of what we think constitutes PII data. (From [CCCApply Data Dictionary](https://cccnext.jira.com/wiki/download/attachments/67043586/CCCApplyDataDictionary-Pilot.V2017.1.pdf?version=2&modificationDate=1489601535029&cacheVersion=1&api=v2))

After review, the following do not need to be counted as PII: Date of Birth, Military Discharge, IP address.

|  |  |  |
| --- | --- | --- |
| **Section** | **Field** | **Comments** |
| Account Data | Legal Name: First |  |
| Account Data | Legal Name: Middle |  |
| Account Data | Legal Name: Last |  |
| Account Data | Legal Name: Suffix |  |
| Account Data | Other/Previous First Name |  |
| Account Data | Other/Previous Name: Middle |  |
| Account Data | Other/Previous Name: Last |  |
| Account Data | Preferred First Name |  |
| Account Data | Preferred Middle Name |  |
| Account Data | Preferred Last Name |  |
| Account Data | Social Security Number |  |
| Account Data | Social Security Number/Taxpayer Identification Number |  |
| ~~Account Data~~ | ~~Date of Birth~~ | ~~Can we convert this to age?~~  |
| Account Data | Mailing Address: Street |  |
| Account Data | Permanent Address: Street1 and Street2 |  |
| Account Data | Main Phone Number |  |
| Account Data | Second Phone Number |  |
| Account Data | Email Address |  |
| Personal Information | Parent/Guardian Name: First |  |
| Personal Information | Parent/Guardian Name: Last |  |
| Citizenship/Military | Alien Registration Number |  |
| ~~Citizenship/Military~~ | ~~Military Discharge Date~~ | ~~Is this considered PII?~~ |
| ~~Submission~~ | ~~IP address~~ | ~~Is this considered PII? There is valuable information for fraud detection here.~~ |

* + 1. Security Design

As shown in the tables in the next section data will be stored on S3. Data will be stored in encrypted format.

We will use Python and related packages for much of our work. Some of the statistical analysis may rely on R. During early development, we expect to store intermediate files such as logs and pickle files to optimize the development process. (The table captures these aspects when considering the visibility of data)

* + 1. Data Life Cycle: Training Phase

Outlined below is the data life-cycle during the training phase. The focus of this table is on the development phase of the project. Once deployed, there is no direct human involvement in the process.

Only during initial model development

Stages exist only during initial model development

Data Extraction

CCCTC

DB

PII Data

Hashing

Exploratory Analysis

Data Integrity Analysis

Data Preparation

Learning

Phase 1

Learning Phase 2

Model

(S3)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ownership** | CCCTC | CCCTC ~~(Scripts developed by Infiniti/SeyVu)~~ | Infiniti/SeyVu | Infiniti/SeyVu | Infiniti/SeyVu | Infiniti/SeyVu | Infiniti/SeyVu | Infiniti/SeyVu |
| **Data Visibility** | Only CCCTC | Only CCCTC | All Involved | All Involved | All Involved | All Involved | All Involved | None |
| **Hashing****(For PII data)** | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Data Storage** | S3<Infiniti to fill> | S3<Infiniti to fill> | S3<Infiniti on AWS security> | S3<Infiniti on AWS security> | S3 (early dev);In-memory for rest of dev & production<Infiniti on AWS security> | S3 (early dev)<Infiniti on AWS security> | S3 (early dev)In-memory for rest of dev & production<Infiniti on AWS security> | Model in S3.Source data destroyed |

**Raw PII (un-hashed) data only visible to CCCTC**

* + 1. Data Life Cycle: Prediction Phase

Outlined below is the data life-cycle during the prediction phase. The focus of this document is on the development phase of the project. Once deployed, there is no direct human involvement in the process (except for the suspend/final decision process owned by CCCTC)

Only on applications identified as fraud

CCCApply

Application

Data Extraction

PII Data Hashing

API Request

API Response

Suspend Process

CCCApply

Decision

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ownership** | CCCTC | CCCTC (Scripts developed by Infiniti/SeyVu) | CCCTC | Infiniti/SeyVu | CCCTC | CCCTC |
| **Data Visibility** | Only CCCTC | Only CCCTC | All Involved\*(early dev/deploy) | All Involved\*(early dev/deploy) | Only CCCTC | Only CCCTC |
| **Hashing (For PII data)** | No | No | Yes | Yes | Yes\* | N/A |
| **Data Storage** | S3 <Infiniti to fill> | S3 <Infiniti to fill> | Early Dev/Deploy: S3Post-deploy: No Storage  | Early Dev/Deploy: S3Post-deploy: No Storage | S3<Infiniti on AWS security> | N/A |

**Raw PII (un-hashed) data only visible to CCCTC**