<**TEMPLATE NOTE –** **Anything between “< >” should be removed or changed**.>

**Data Audit Report**

Prepared by

<The Analytics Team>

In support of

<Predictive model for the variable annuity marketing campaign>

Requested by

< Requestor>

March 28, 2018

**Introduction**

<**TEMPLATE NOTE –** **The first section through the timeline should be changed to reflect the project at hand**.>

The analytical team has been asked by <Requestor> to build a predictive model to identify current bank customers who would be likely to respond to a marketing campaign for a variable annuity product. The deliverable will be a scored marketing file.

<**TEMPLATE NOTE –** **It is not always the case the sample qualifications will be so thoroughly spelled out.**>

The target sample qualifications provided by <Requestor> are as follows:

1. <Insurance account opened within the last 2 years;
2. FICO score >= 700;
3. Bank customer for at least 1 year;
4. At least 25 years old;
5. At least 2 active credit tradelines with oldest opened in 2010 or earlier.>

Non-target customers will have qualifications 2 – 5.

The timeline for this project is as follows:

|  |  |
| --- | --- |
| **Milestone** | **Timeline** |
| Data Audit/Aggregation | Week 1 |
| Data Cleansing and Preparation | Weeks 2 - 3 |
| Modeling Construction | Week 4 |
| Scoring of Marketing File | Week 5 |
| Marketing Campaign Commencement | Week 6 |

It is essential that the analytical team has a full understanding of the quality and quantity of data provided to it in support of the analytical request.

Hence, the purpose of this data audit is to ensure that:

* all data received by the analytical team for the project are consistent with the team’s understanding of the requested analytical deliverable;
* that the team is reading and interpreting these data correctly;
* that the team has received all data intended to be supplied;
* that the data are functionally usable for modeling purposes.

The data audit is broken into four main sections:

1. Dataset Summary – A list and description of all datasets received.
2. Dataset Detail – For each dataset, tables showing all data variables received. It is important that this section be reviewed to ensure that the analytical team has all the data sent, the data are being read correctly and the data have reasonable values.
3. Modeling Sample – Based on the requestor’s sample requirements, a determination is necessary as to whether adequate sample is available to support modeling.
4. Questions – Specific questions that the analytical team needs answered to ensure that the team fully understands the data and that the data can support the requested analytical deliverable.

**Dataset Summary**

The analytical team has received <n> data files from <data provider> as listed in Table 1.

Table 1. Data Files Received

|  |  |  |
| --- | --- | --- |
| Filename | File Type | File Contents |
| <file name> | <csv, Excel, SAS, DBF, etc.> | <brief description of contents> |
|  |  |  |
|  |  |  |
|  |  |  |

**Dataset Detail**

Each dataset contains the data fields as shown in the following tables.

**<TEMPLATE NOTE – not all datasets will have all datatypes. Repeat for each dataset. File contents refer to the raw files (before any merging is done). >**

Dataset #1: <file name>

File Contents: <brief description of contents>

**<TEMPLATE NOTE – add any notes about the dataset, if necessary. >**

Table 2. Numeric Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **Count** | **Missing** | **Min** | **Mean** | **Median** | **Max** |
| DDABal | 26316 | 0 | -774.83 | 2,660.49 | 887.01 | 278,093.83 |
|  |  |  |  |  |  |  |

Table 3. Categorical Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Frequency** | | |
| **Field Name** | **Missing** | **Value** | **Count** | **Percent** |
| Area Classification | 0 | U | 12,681 | 39.3 |
|  |  | S | 11,506 | 35.7 |
|  |  | R | 8,077 | 25.0 |

Table 4. Character Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Length of Character Field** | | | | | |
| **Field Name** | **Count** | **Missing** | **Min** | **Mean** | **Median** | **Max** |
| fst\_nm | 32264 | 0 | 2 | 6 | 6 | 11 |
| lst\_nm | 32264 | 0 | 2 | 6 | 6 | 13 |

Table 5. Date Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | **Min Freq** | | **Max Freq** | |
| **Field Name** | **Count** | **Missing** | **Oldest** | **Most Recent** | **Year** | **Count** | **Year** | **Count** |
| BIRTH\_DT | 25962 | 6302 | 3-May-18 | 31-Oct-14 | 2014 | 1 | 1975 | 968 |
| ORIG\_DT | 32264 | 0 | 3-Apr-65 | 30-Dec-14 | 1965 | 1 | 2014 | 5,790 |

**Modeling Sample**

|  |  |
| --- | --- |
| **Segment** | **Count** |
| Customers with insurance product |  |
| Opened account within last 2 years |  |
| etc. |  |
|  |  |
| **Available target sample** |  |
|  |  |
| Customers without insurance product |  |
| FICO score >= 700 |  |
| etc. |  |
|  |  |
| **Available non-target sample** |  |
|  |  |
| **Total sample** |  |

**<TEMPLATE NOTE – is the modeling sample of adequate size? If not, what recommendations can be made for increasing the size?>**

**Questions**

**<TEMPLATE NOTE – these are questions from YOU to whomever provided you with the data.>**

1. Does the above information appear to be correct? Specifically:
   * Does the analytical team have all the data that was meant to be sent?
   * Is the team interpreting the data correctly?
   * Do the data appear to have reasonable values?
2. Here is a list of the data integrity issues the analytical team uncovered:
3. The following are specific questions the analytical team has about the data: