Data Quality Monitoring for the 2022 SCF

Identifying Falsified Data & Introduction of NORC ProofPoint

04.12.23

Jimmy Herdegen, Research Associate I
Kate Bachtell, Senior Research Director II
Jason Keller, Senior Data Scientist II
Cathy Haggerty, Vice President
Lisa Blumerman, Senior Vice President

NORC at the University of Chicago

Agenda

01 Introduction: Challenges & Goal

02 Detection Methods at the Beginning of 2019 Cycle

03 Falsification Problems & Responses During 2019 Cycle

04 Solution for the 2022 Cycle: ProofPoint & Its Results

05 Looking Ahead

Introduction

Problem Statement

Problems Addressed

- There is no one way falsified data can happen, which poses challenges on how to detect invalid cases
- Creating a framework to filter out falsified data and enhance data quality is critical for large scale surveys like the SCF

Goal of Presentation

- By learning lessons from previous cycles, and creating and implementing a set of metrics to filter cases through, we can be smarter and faster in detecting falsification

Field interviewer falsification comes in many forms

Survey Format

- Interviewers with invalid data often use filter branches in surveys to intentionally shorten an interview (Walzenbach, 2021)
- Fewer missing answers (SCF 2019)

Survey Content

- Certain behavioral & attitudinal variables could predict falsified data (Menold, Kemper, 2014)

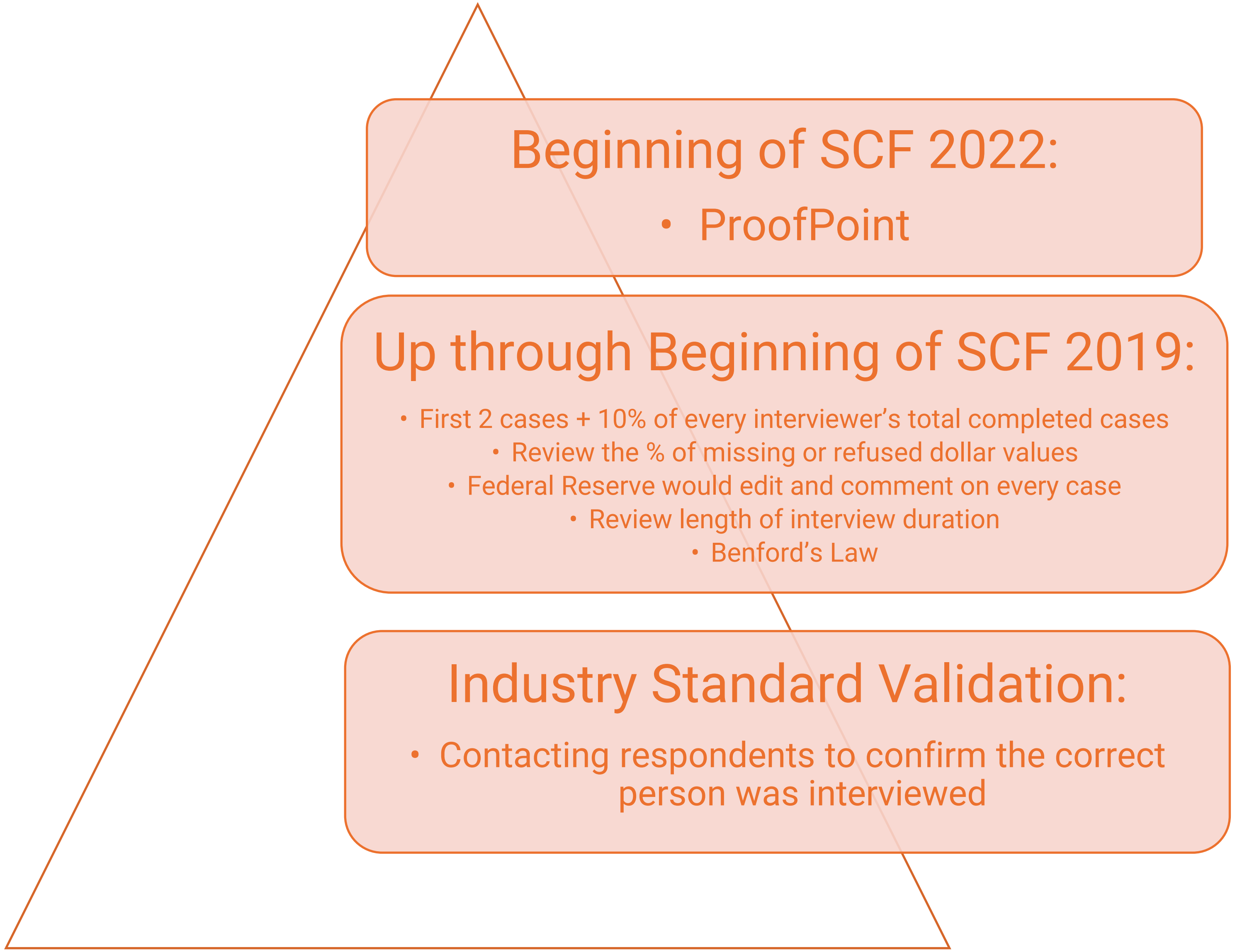
Curb-Stoning

- Fabricating the entire interview at the time of the interview (Thissen, Myers, 2016)
- Omitting interviews by reporting them as refusals or unlocatable, when little effort was done to contact the respondent (Thissen, Myers, 2016)

Detection Methods at the Beginning of the 2019 Cycle

NORC at the University of Chicago

Multiple layers to validation procedure



Beginning of SCF 2022:

- ProofPoint

Up through Beginning of SCF 2019:

- First 2 cases + 10% of every interviewer's total completed cases
- Review the % of missing or refused dollar values
- Federal Reserve would edit and comment on every case
- Review length of interview duration
- Benford's Law

Industry Standard Validation:

- Contacting respondents to confirm the correct person was interviewed

Falsification Problems & Responses During the 2019 Cycle

NORC at the University of Chicago

We discovered bad actors later in the field period

- Additional measures were implemented during the midpoint of data collection during the 2019 cycle:
  - Paradata Analysis:
    - Missing phone numbers via CAPI and NORCSuite
    - Looking at the record of calls (ROCs)
  - Reviewing GPS coordinates from 2 different apps associated with interviewer activity
  - Electronic signatures for payment receipts
  - CAPI data quality analysis:
    - Quex timings
    - Missing dollar values
    - Data quality flags
    - Feedback from the Federal Reserve

Characteristics of Falsified Cases During the 2019 Cycle

Shorter

- Without genuine human interaction between the field interviewer and respondent, survey times will tend to be shorter

Less Descriptive

- Invalid case comments are shorter and not as descriptive

Lack Contact Info

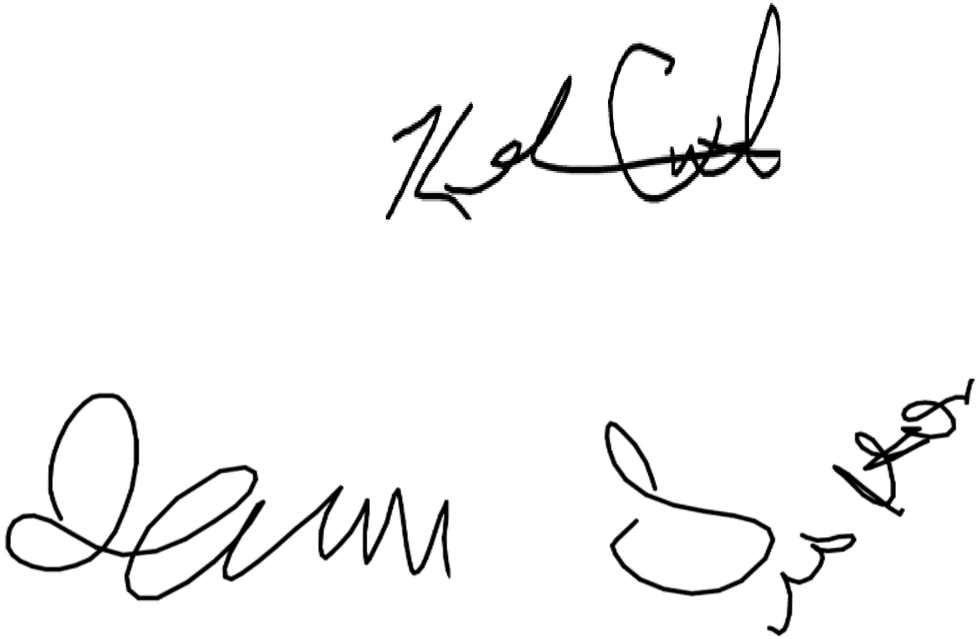
- Interviews that were falsified were more likely not to include contact info for respondent

Fewer Missing Answers

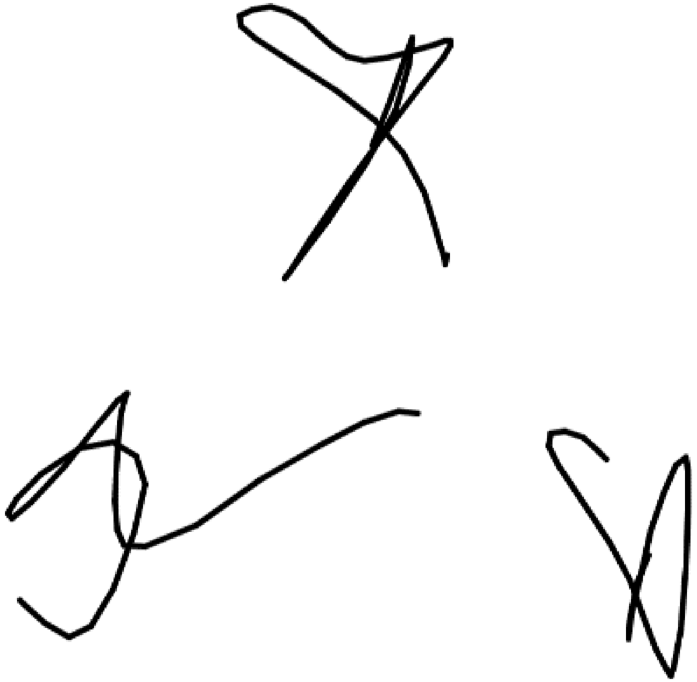
- Field Interviewers would tend to provide fewer missing responses compared to valid data

Examples of Respondent Signatures from Valid & Invalid Cases

Example of Signatures from Valid Cases



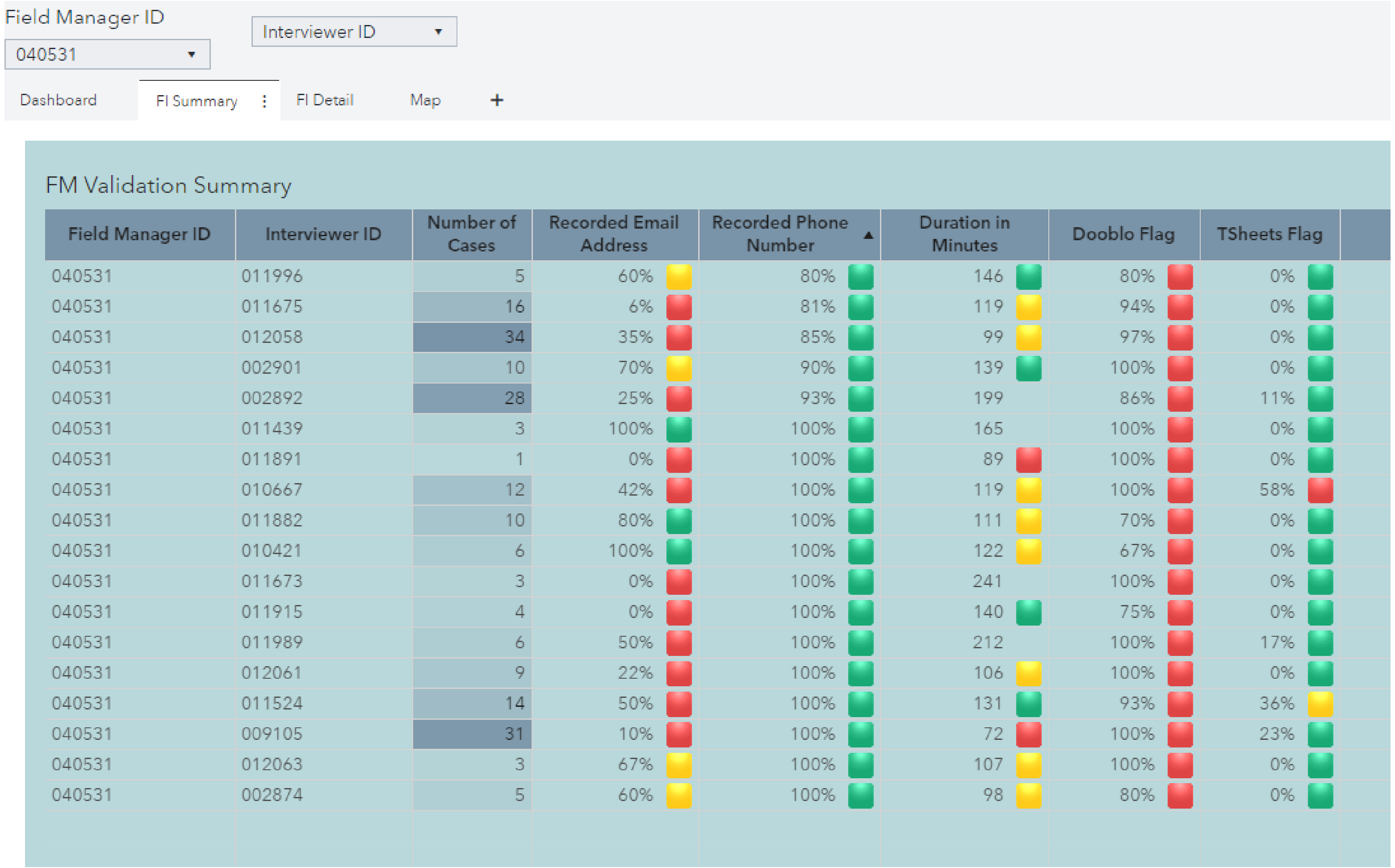
Example of Signatures from Invalid Cases



Solution for SCF 2022: ProofPoint

NORC at the University of Chicago

Interviewer Summary View



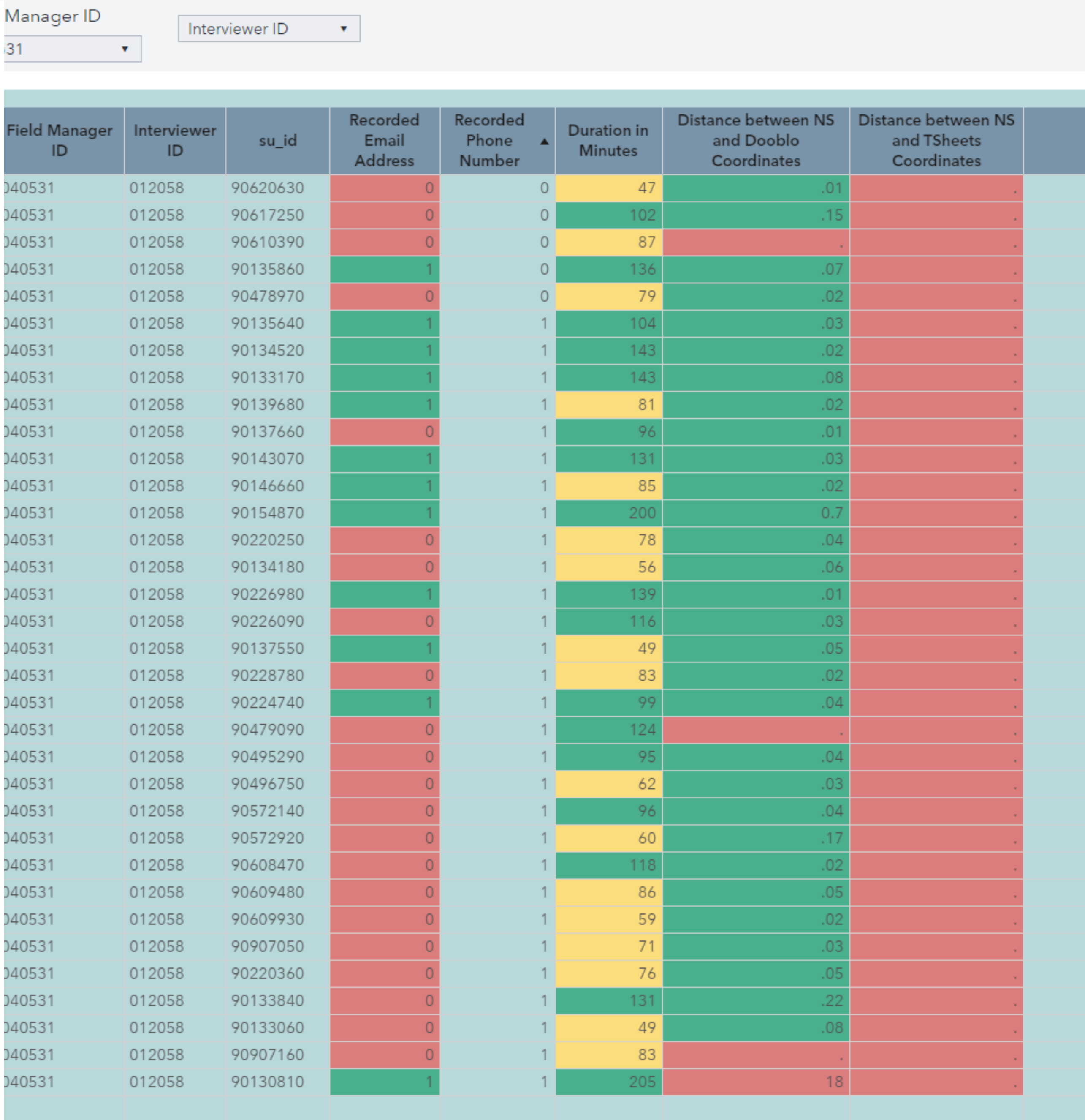
FM Validation Summary

Field Manager ID	Interviewer ID	Number of Cases	Recorded Email Address	Recorded Phone Number ▲	Duration in Minutes	Dooblo Flag	TSheets Flag
040531	011996	5	60% 🟨	80% 🟩	146 🟩	80% 🟥	0% 🟩
040531	011675	16	6% 🟥	81% 🟩	119 🟨	94% 🟥	0% 🟩
040531	012058	34	35% 🟥	85% 🟩	99 🟨	97% 🟥	0% 🟩
040531	002901	10	70% 🟨	90% 🟩	139 🟩	100% 🟥	0% 🟩
040531	002892	28	25% 🟥	93% 🟩	199	86% 🟥	11% 🟩
040531	011439	3	100% 🟩	100% 🟩	165	100% 🟥	0% 🟩
040531	011891	1	0% 🟥	100% 🟩	89 🟥	100% 🟥	0% 🟩
040531	010667	12	42% 🟥	100% 🟩	119 🟨	100% 🟥	58% 🟥
040531	011882	10	80% 🟩	100% 🟩	111 🟨	70% 🟥	0% 🟩
040531	010421	6	100% 🟩	100% 🟩	122 🟨	67% 🟥	0% 🟩
040531	011673	3	0% 🟥	100% 🟩	241	100% 🟥	0% 🟩
040531	011915	4	0% 🟥	100% 🟩	140 🟩	75% 🟥	0% 🟩
040531	011989	6	50% 🟥	100% 🟩	212	100% 🟥	17% 🟩
040531	012061	9	22% 🟥	100% 🟩	106 🟨	100% 🟥	0% 🟩
040531	011524	14	50% 🟥	100% 🟩	131 🟩	93% 🟥	36% 🟨
040531	009105	31	10% 🟥	100% 🟩	72 🟥	100% 🟥	23% 🟩
040531	012063	3	67% 🟨	100% 🟩	107 🟨	100% 🟥	0% 🟩
040531	002874	5	60% 🟨	100% 🟩	98 🟨	80% 🟥	0% 🟩

- Summary of each interviewer’s quality metrics
- Percentage of critical contact info:
  - Email address
  - Phone number
- Percentage of validation passed:
  - Email
  - Phone
  - Mail
- Average interview duration time
- Average falsification score
]

Interviewer Detail

- Displays case-level metrics for every interviewer
- Indicates if email address or phone number were captured
- Interview duration time
- Distance between interviewer and respondent’s home
- How a case has been validated
- Case’s falsification score



Manager ID: 31 ▾

Interviewer ID ▾

Field Manager ID	Interviewer ID	su_id	Recorded Email Address	Recorded Phone Number ▲	Duration in Minutes	Distance between NS and Dooblo Coordinates	Distance between NS and TSheets Coordinates
040531	012058	90620630	0	0	47	.01	.
040531	012058	90617250	0	0	102	.15	.
040531	012058	90610390	0	0	87	.	.
040531	012058	90135860	1	0	136	.07	.
040531	012058	90478970	0	0	79	.02	.
040531	012058	90135640	1	1	104	.03	.
040531	012058	90134520	1	1	143	.02	.
040531	012058	90133170	1	1	143	.08	.
040531	012058	90139680	1	1	81	.02	.
040531	012058	90137660	0	1	96	.01	.
040531	012058	90143070	1	1	131	.03	.
040531	012058	90146660	1	1	85	.02	.
040531	012058	90154870	1	1	200	0.7	.
040531	012058	90220250	0	1	78	.04	.
040531	012058	90134180	0	1	56	.06	.
040531	012058	90226980	1	1	139	.01	.
040531	012058	90226090	0	1	116	.03	.
040531	012058	90137550	1	1	49	.05	.
040531	012058	90228780	0	1	83	.02	.
040531	012058	90224740	1	1	99	.04	.
040531	012058	90479090	0	1	124	.	.
040531	012058	90495290	0	1	95	.04	.
040531	012058	90496750	0	1	62	.03	.
040531	012058	90572140	0	1	96	.04	.
040531	012058	90572920	0	1	60	.17	.
040531	012058	90608470	0	1	118	.02	.
040531	012058	90609480	0	1	86	.05	.
040531	012058	90609930	0	1	59	.02	.
040531	012058	90907050	0	1	71	.03	.
040531	012058	90220360	0	1	76	.05	.
040531	012058	90133840	0	1	131	.22	.
040531	012058	90133060	0	1	49	.08	.
040531	012058	90907160	0	1	83	.	.
040531	012058	90130810	1	1	205	18	.

Falsification Score Calculation

True Negatives

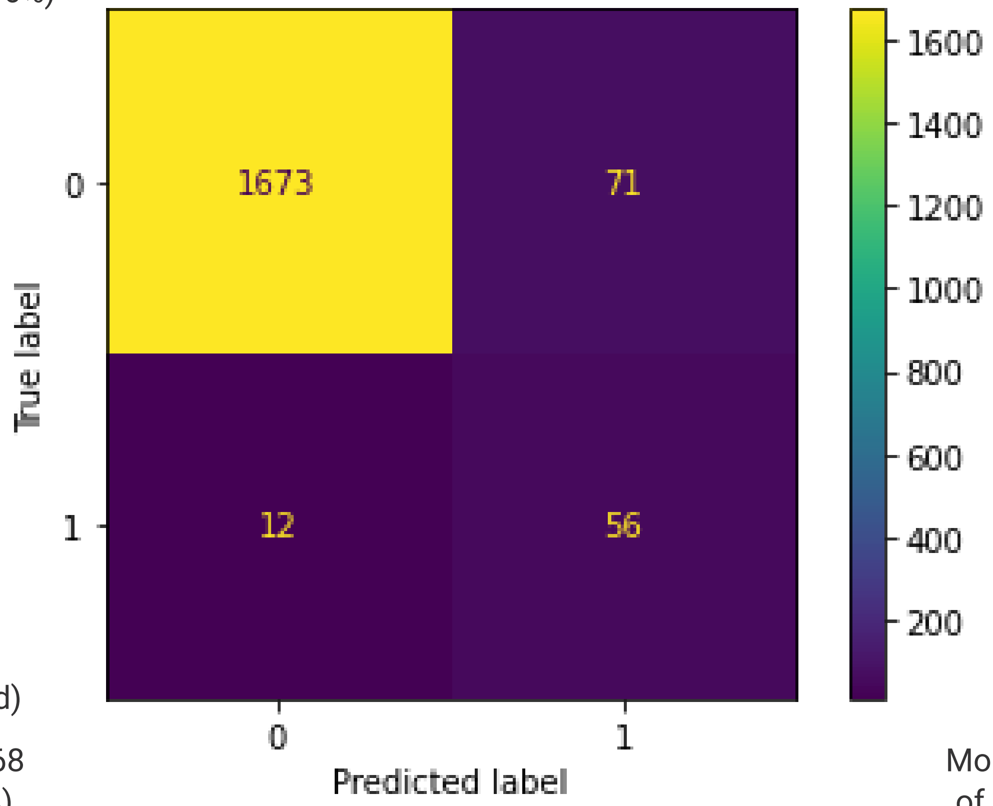
(Not falsified and not flagged)

Model correctly identifies 1,673 out of 1,744 valid interviews (96%)

False positives

(Not falsified but flagged)

Model incorrectly identifies 71 out of 1,744 valid interviews as falsified (4%)



False Negatives

(Falsified and not flagged)

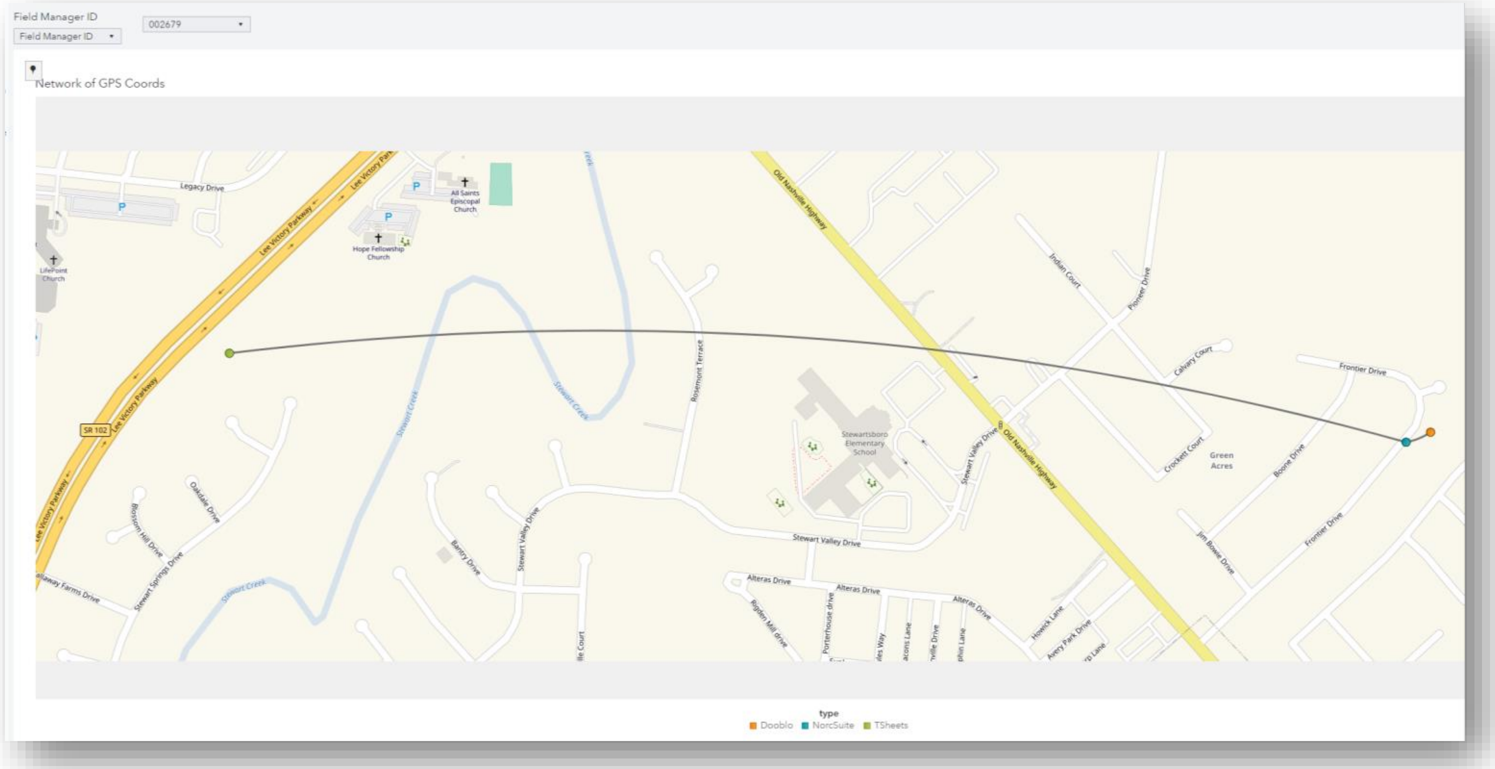
Model misses 12 out of 68 falsified interviews (18%)

True Positives

(Falsified and flagged)

Model correctly identifies 56 out of 68 falsified interviews (82%)

Proofpoint has been effective in isolating potential falsifiers, though improvements will strengthen it



Field Manager ID: 002679

Field Manager ID

Network of GPS Coords

type: ■ Dooblo ■ NorcSuite ■ TSheets

- Falsifiers were identified sooner compared to 2019
  - Falsification scores were able to identify some falsified cases, though they were unable to capture others
- GPS locating helpful for reviewing in-person cases
- ProofPoint's data complements phone and mail validation efforts
- Data helped enlighten the average for key metrics
  - Email and phone number recorded
  - Survey duration time
  - Phone validation rate

Looking Ahead

NORC at the University of Chicago

How can we take full advantage of Proofpoint for the future?

Examine falsification score

Retrain model

Better Proofpoint integration

Thank you.

Jimmy Herdegen
Research Associate I
Herdegen-Jimmy@norc.org

Kate Bachtell
Senior Research Director II
Bachtell-Kate@norc.org

Jason Keller
Senior Data Scientist II
Keller-Jason@norc.org

Cathy Haggerty
Vice President
Haggerty-Cathy@norc.org

Lisa Blumerman
Senior Vice President
Blumerman-Lisa@norc.org

Research You Can Trust™

NORC at the University of Chicago