Using Machine Learning for Quality Assessments of Call Center Interactions – A Case Study

Elizabeth Nichols[1], Monica Puerto[2], Brian Sadacca[2], Shaun Genter[1], Kevin Zajac[1]

U.S. Census Bureau[1] and Accenture Federal Services[2]

2023 FCSM Research and Policy Conference

October 26, 2023

United States® Census Bureau

Case study: 2020 Census Bureau call center

- Customer Service Representatives (CSRs) took calls
- Utilized an agent desktop
  - Record calls
  - Scripted text
- Follow read-as-worded rule
  - Answers to frequently asked questions (FAQs)
  - Questions in the census questionnaire

United States® Census Bureau

Quality control current procedure and challenges

- CSRs
  - Quality monitors listened to CSR audio recordings to ensure the read-as-worded rule is followed.
    - Labor intensive/Costly
    - Ad hoc
- Scripts
  - Pretesting and expert review
  - Behavior coding is used to evaluate questions and text from recorded calls
    - Labor intensive
    - Post-production and not real time

United States® Census Bureau

Today

- Share our research investigating whether Machine Learning (ML)/Natural Language Processing (NLP) could improve

- the quality control for read-as-worded rule
  - making it more systematic
- the scripts
  - today's focus is on the FAQ scripts

United States® Census Bureau

Data

- 2020 Census call center audio files (English language only)
  - Collected between March and October 2020
  - Approximately 5 million recorded calls
    - 65K transcripts of General Assistance Calls + Technical calls
      - This is what we used in the analysis.
      - Wav2vec transcription model
- 7,000+ CSRs across 11 call centers (nationwide)
  - ~21% bilingual (mostly English and Spanish)
  - We knew the calls each CSR took – dataset was labeled with CSR ID
- Scripts
  - 300+ FAQ answer scripts
  - We did not know the FAQ used for each call – dataset was not labeled with FAQ read

United States® Census Bureau

ML method

- String search called Fuzzy Match
  - Partial ratio methodology
  - Took 2.8 seconds / .26 Std. for each transcript
- Compare the CSR's transcribed first 300 words against the 300+ FAQ scripts
- A score is given between 0 and 100 for each FAQ
- Higher scores indicate the script matches the text read aloud more closely than lower scores

United States® Census Bureau

Fuzzy match score on CSR's transcribed words

How do I answer the Hispanic origin question?	99
what is Hispanic origin?	94
how do I answer the relationship question?	73
Report a vacant home nonresidential address or…	40

CSR A

" hello you have rich the twenty twenty census question or assistance lane my name is [redacted] do i have your permission to continue recording this call for quality assurance purposes thank you how may i help you okay i'll be happy to assist you with this yes okay so you want to know how to answer the **hispanic question** correct okay i'll be happy to assist do with that or just give me a few seconds please you're welcome okay your response to this question should be based on **how you identify each person can decide how to answer would you like more information about the response categories** okay so that would be a **hispanic latino or spanish orgia includes all individuals who identify with one or more nationalities or ethnic groups originating in mexico portico cuba central and south american and other spanish cultures examples of these groups** "

United States® Census Bureau

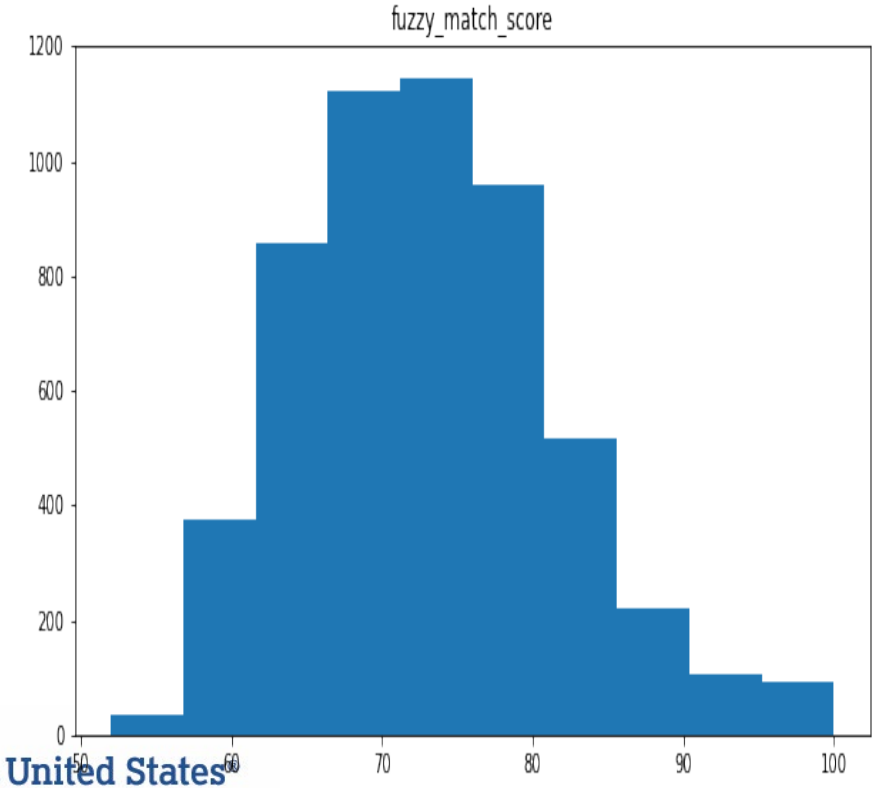
Using Fuzzy Match scores for Quality Control

- Idea: Look for below average scores
- Evaluating CSRs
  - Take average Fuzzy match score
    - By CSR (across a particular question or FAQ)
  - CSRs that have lower Fuzzy match scores might not read as worded as well as other CSRs
- Evaluating scripts
  - Take average Fuzzy match score
    - By Script (for a particular question or FAQ but across CSRs)
  - Scripts that have lower Fuzzy match scores might be harder to read as worded and therefore need modifying

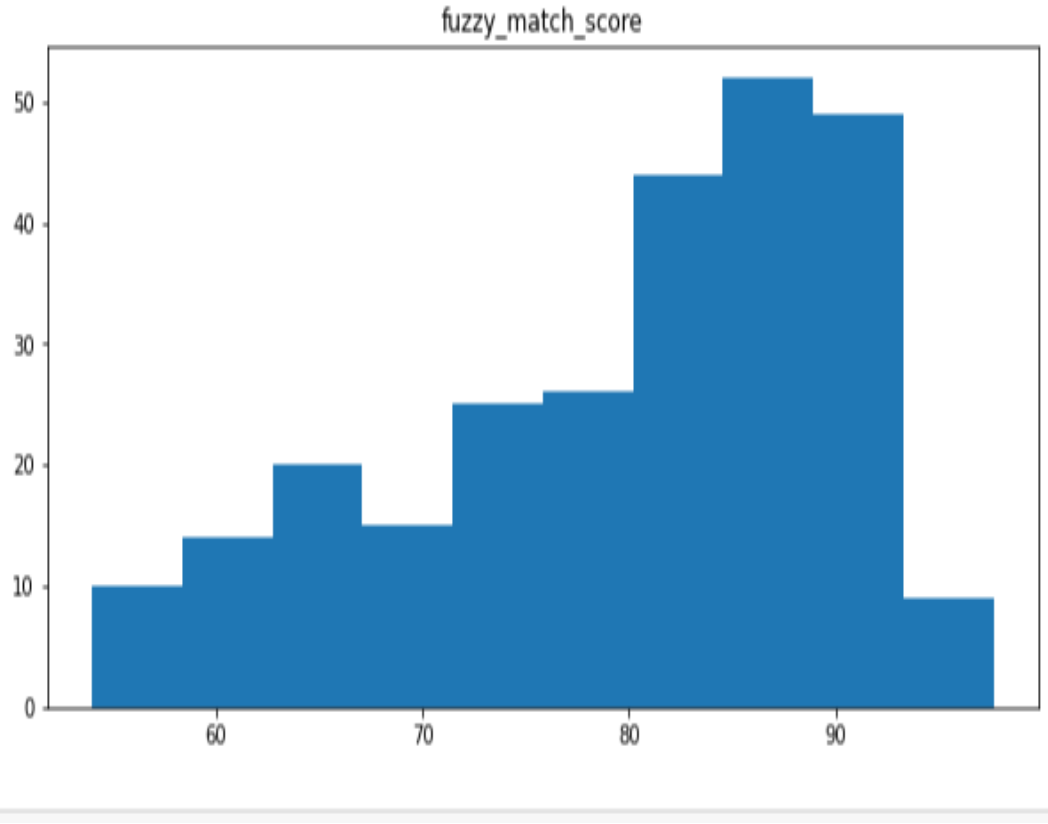
United States® Census Bureau

Average Fuzzy match score by CSR (left) by FAQ (right) for 65K transcripts

Avg Fuzzy score of CSR: 73 , STD of 8

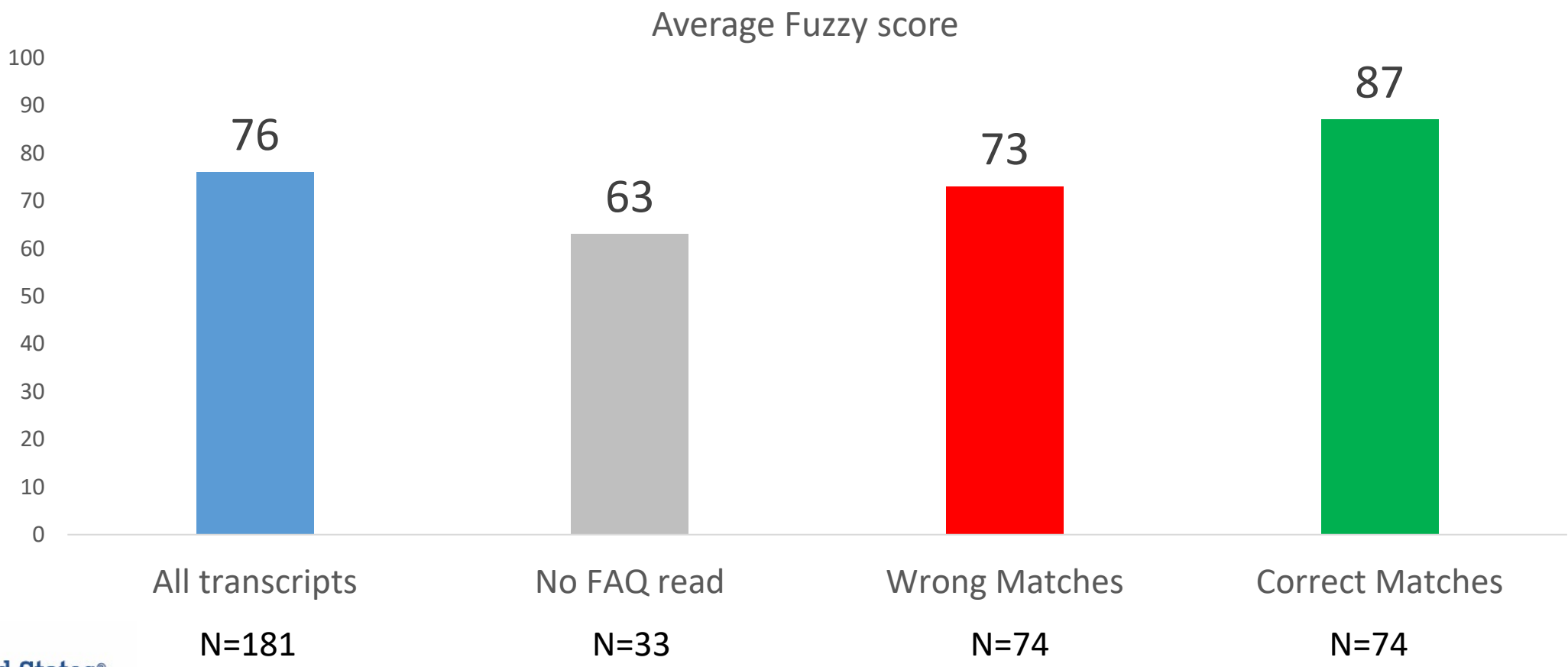


Avg Fuzzy score of FAQ : 79 , STD of 10



United States Census Bureau

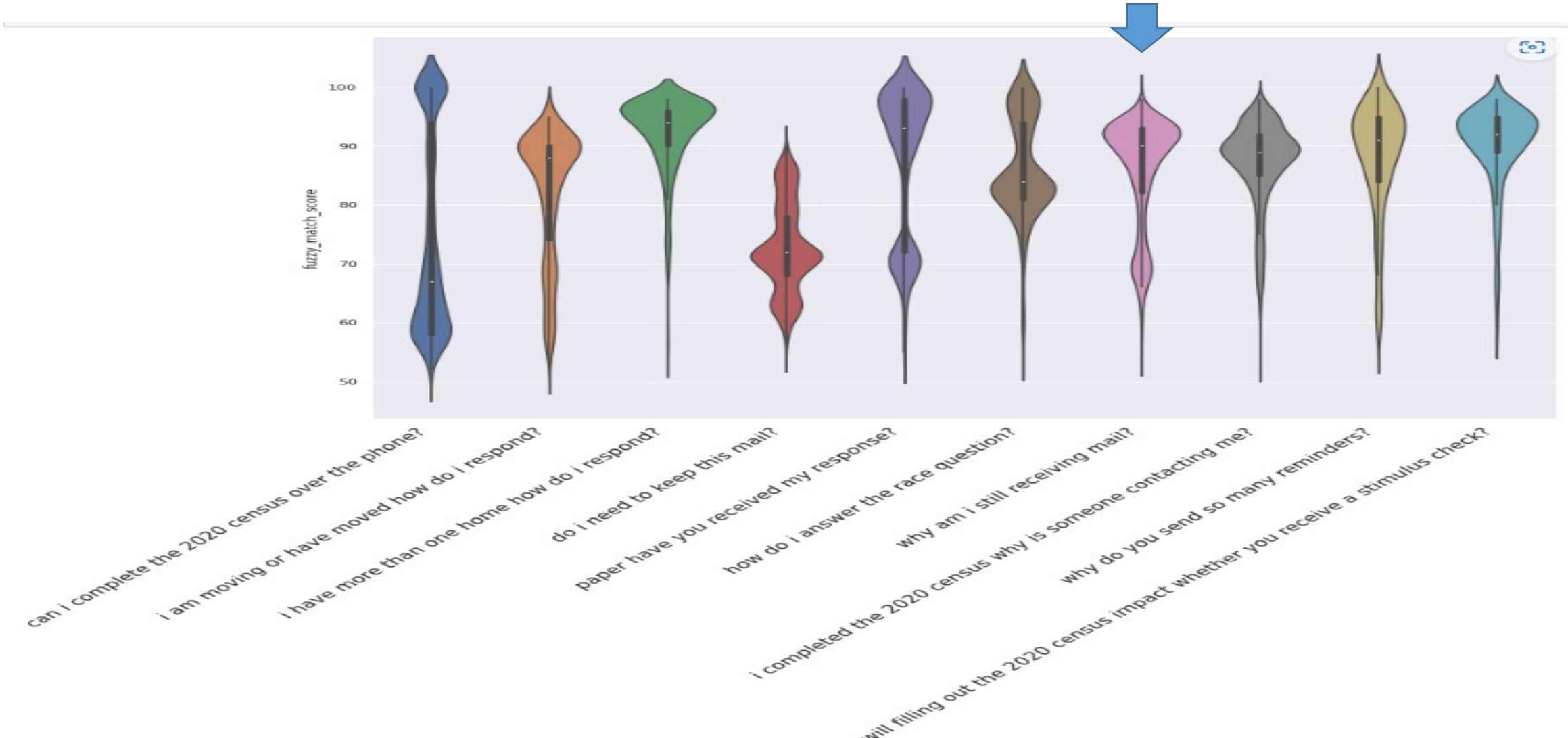
Fuzzy match check for FAQs



United States® Census Bureau

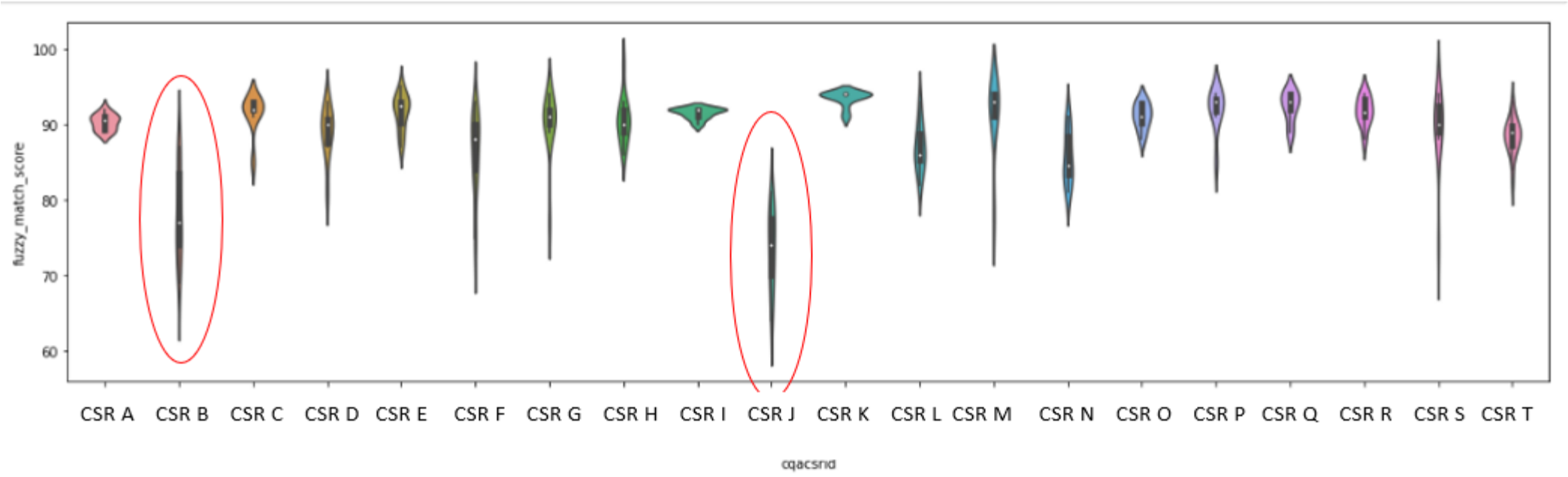
Fuzzy match score by top FAQs read aloud

We can see variation by FAQ which ones are more consistent versus have more range in fuzzy scores



Fuzzy match scores by CSR

Why am I still receiving mail?



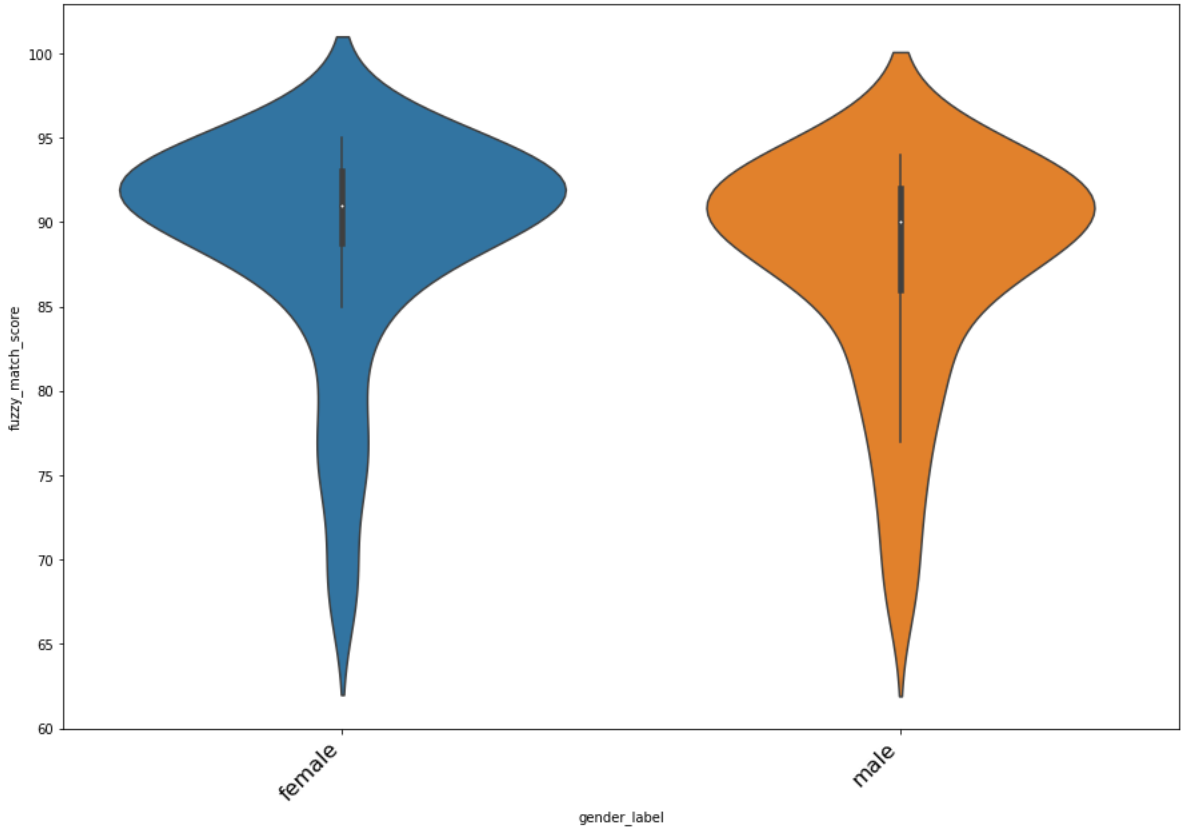
This FAQ had a very high overall Fuzzy match score. All CSRs presented had a minimum of nine calls where they read that FAQ (according to the Fuzzy Match link) with a maximum of 18 calls.

Is there a bias using ML for Quality Control?

United States® Census Bureau

Fuzzy match scores by sex

Why am I still receiving mail



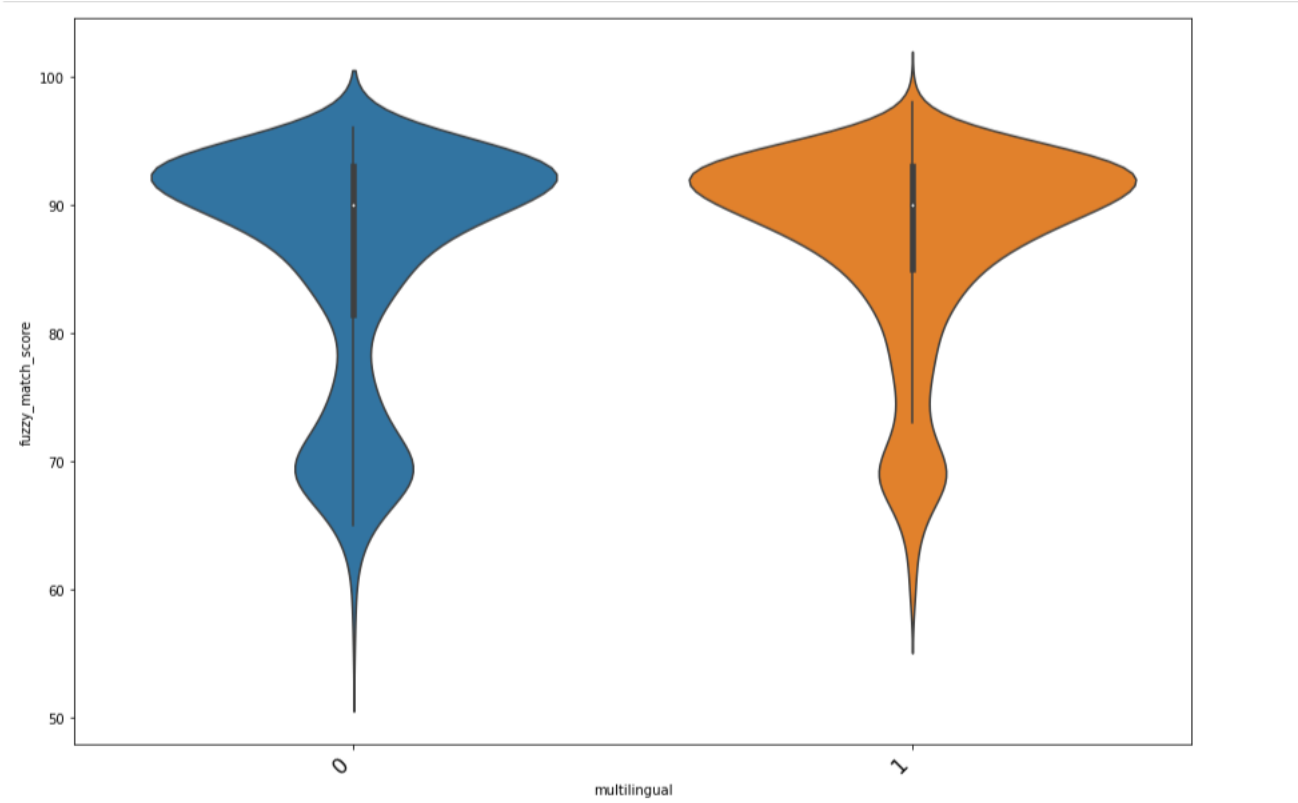
N=97, split fairly equally

United States® Census Bureau

Fuzzy match scores by CSRs languages spoken

(0=English only speaker; 1=bilingual)

Why am I still receiving mail



N=97, split fairly equally

United States® Census Bureau

Take aways

- Overall results
  - Fuzzy match method was standardized/not ad hoc
  - Efficient/possible cost savings
- CSR Quality Control
  - Positive results
  - We had a CSR code assigned to each call
  - Identified some CSRs with lower scores to investigate further
    - Ideally, we would listen to the calls for those 2 CSRs to determine if in fact they were not reading as worded.
  - Did not detect any obvious bias for the two criteria we used – but consider looking at different characteristics
    - The transcript model probably matters here. We used wav2vec
- Script Quality Control
  - The violin plots were not as clear cut for FAQ
    - We did investigate some questions in the enumeration
  - Perhaps the FAQ scripts in English were okay based on all the testing
  - Better if assigned the FAQ for each call
- Going forward
  - Try this method in a test
  - Develop a dashboard to compare by FAQ and by CSR and run the report daily or weekly

United States® Census Bureau

Thank you

Elizabeth Nichols

elizabeth.may.nichols@census.gov

United States® Census Bureau

Different Fuzzy Matching Methodology

Partial Ratio does best for Fuzzy Matching
(100 of the 200 labeled transcripts)

Fuzzy Methodology	Avg Seconds per Transcript & STDEV	Avg Fuzzy Score & STDEV	Accuracy	Accuracy > 70 fuzzy score
Ratio	.09 seconds / .00	46 / 5	5%	0%
Partial Ratio	2.8 seconds / .26	73 / 12.7	30%	65%
Token Sort Ratio	.12 seconds / .01 seconds	52 / 3.4	3%	0%
Token Set Ratio	.09 seconds / .00	86 / .004	24%	24%
Weighted Ratio	5.4 seconds / .61	86 / .72	1%	1%