A Semi-Supervised Active Learning Approach for Block-Status Classification.

*No Title-13 data is included in the presentation or the project.

10.6.24

xD
xd.gov

United States™ Census Bureau

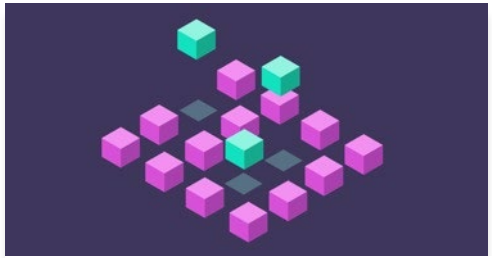
Atul Rawal, Ph.D.
James McCoy, Ph.D.
Elvis Martinez
Drew Duvall

Outline

- Goal
- Data
  - Data Engineering
  - Class Balancing
- Machine Learning
  - Supervised Learning
  - Semi-Supervised Learning
- Explainable AI
- Challenges
- Summary

xD Overview

xD is an emerging technologies group that's advancing the delivery of data-driven services through new and transformative technologies.

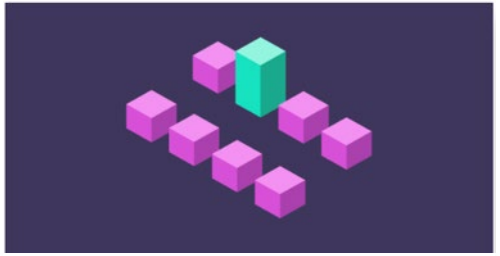


Responsible AI (RAI)

- AI/ML for labeling & classification of geographic data saving ~800,000 hours of manual labeling
- XAI & Causal learning for bias identification in geographic data
- Model Card Generator & AI Register
- Bias Toolkit

Privacy-Enhancing Technologies (PETs)

- UN Pilot for Secure Multi Party Computation
- Remote execution and Federated Learning
- Inter-Agency Multi Party Computation



Incubation & Transformation

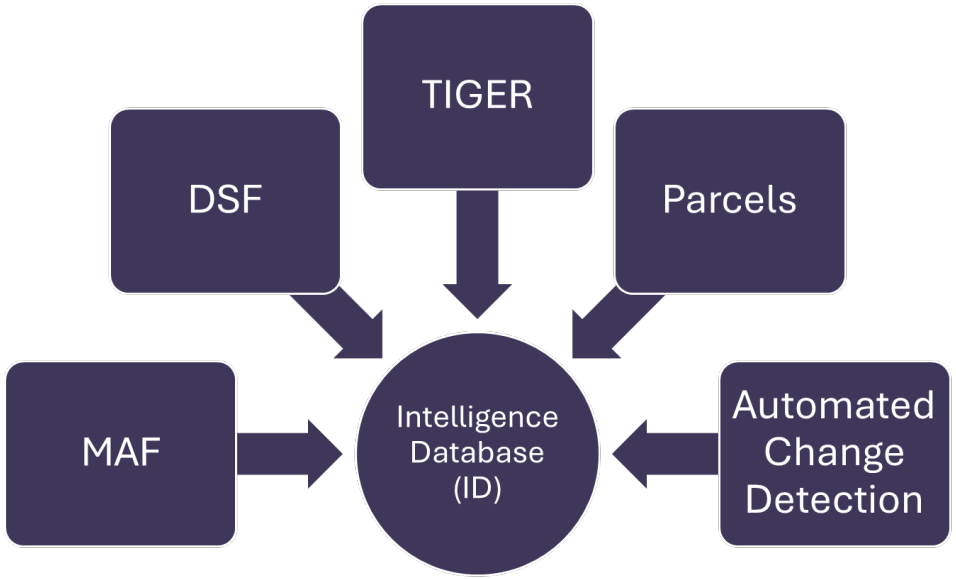
- Developer Experience at Census
- Bias in Infrastructure
- DAO for Equitable Government participation
- Privacy Preserving Record Linkage for Health

Goal

Develop a machine learning pipeline to improve both data labeling and classification of parcel data to enable new data-driven insight while reducing costs and effort for data assessment.

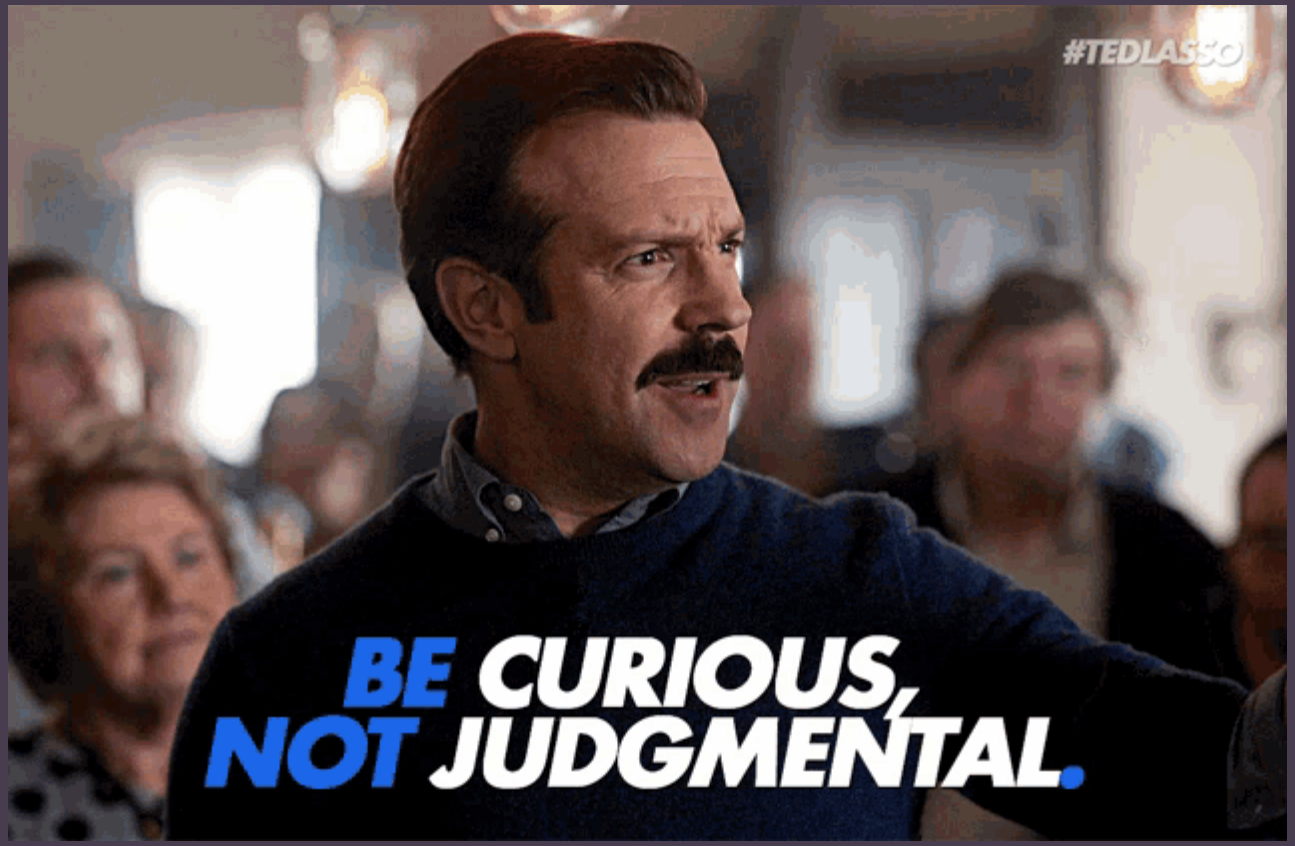
Data

- Census Bureau's Intelligence Database (ID)
  - ~40K labelled blocks
  - >8,000,000 unlabeled blocks.
  - Multi-class
    - Passive
    - Over-coverage
    - Under-coverage



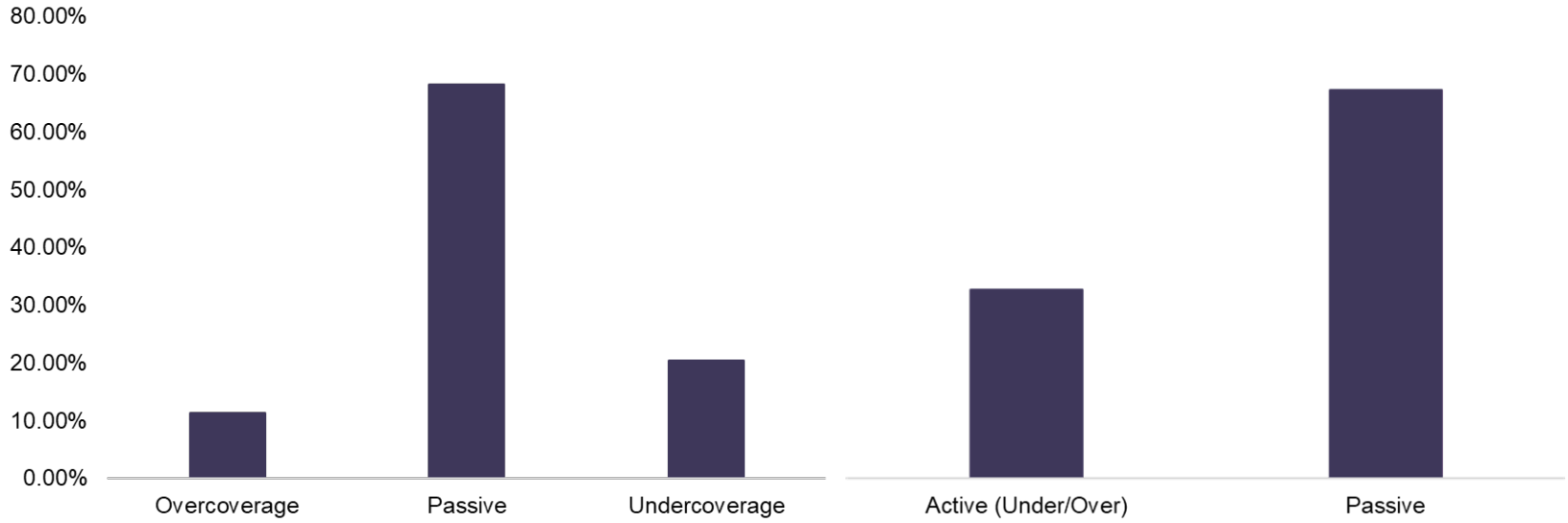
*No Title-13 data is included in the presentation or the project.

Data Curiosity



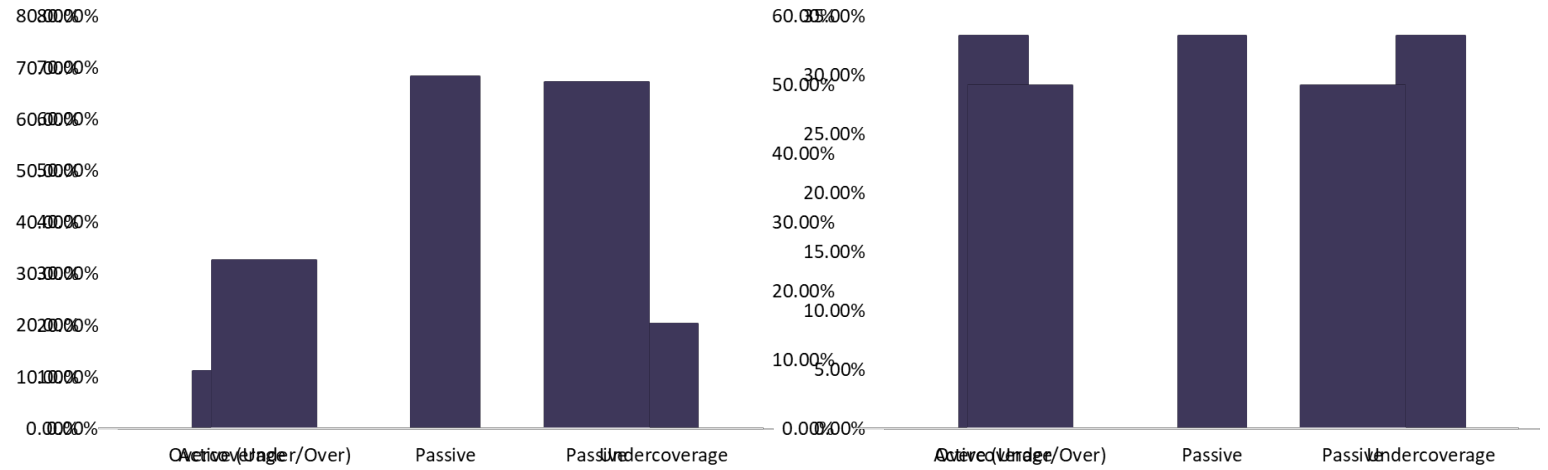
Data Engineering

The original dataset is heavily imbalanced towards the passive class.

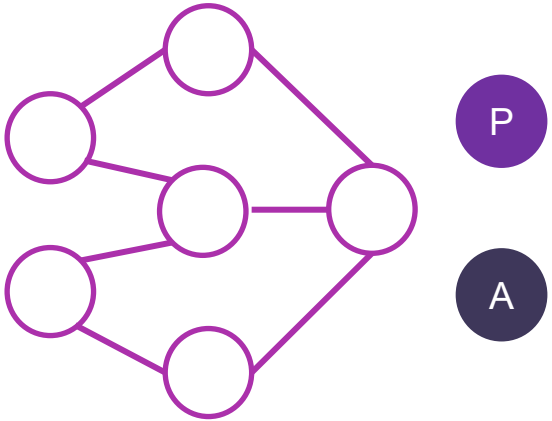


Class Balancing

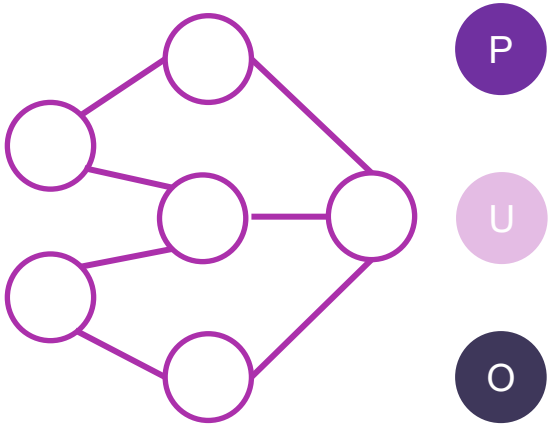
Using SMOTE and K-Means for data balancing



Binary Vs Multi-Class



Binary



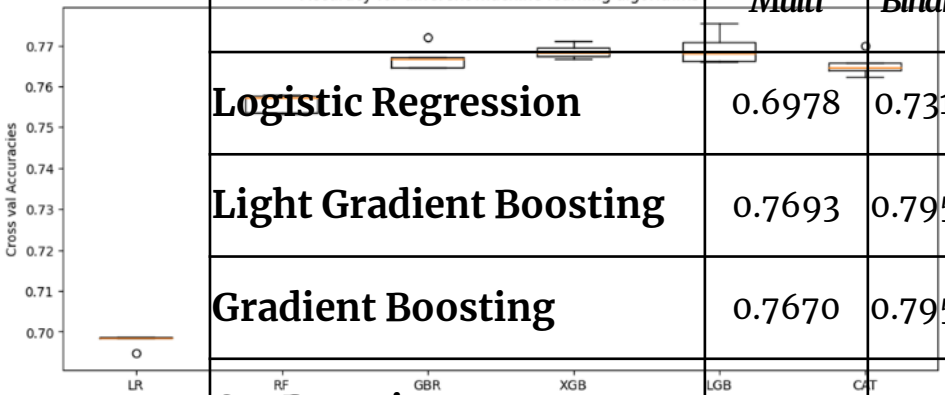
Multi Class

A = Active
P = Passive
O = Over-coverage
U = Under-coverage

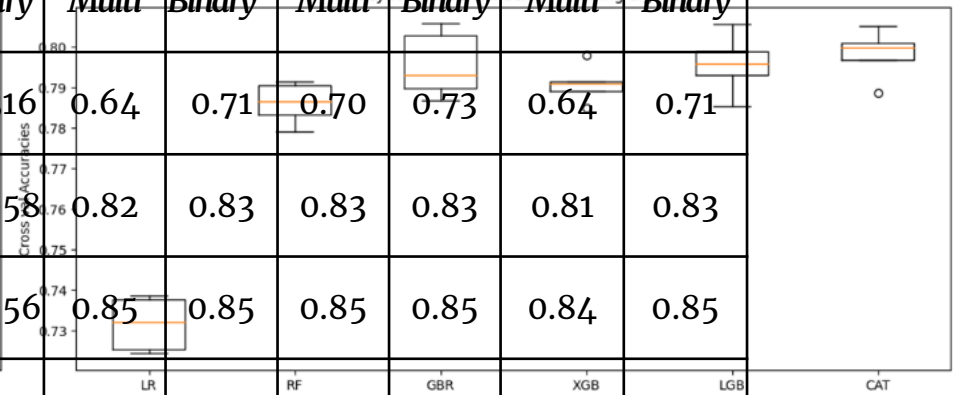
Let’s do Machine Learning

Supervised Learning - Unbalanced

Model	Accuracy		Precision		Recall		F1-Score	
	Multi	*Binary*	*Multi*	*Binary*	*Multi*	*Binary*	*Multi*	*Binary*
Logistic Regression	0.6978	0.7316	0.64	0.71	0.70	0.73	0.64	0.71
Light Gradient Boosting	0.7693	0.7958	0.82	0.83	0.83	0.83	0.81	0.83
Gradient Boosting	0.7670	0.7956	0.85	0.85	0.85	0.85	0.84	0.85
Cat Boosting	0.7653	0.7983	0.94	0.92	0.95	0.92	0.94	0.92
Extreme Gradient Boosting	0.7686	0.790	0.98	0.97	0.98	0.97	0.98	0.97
Random Forest	0.7559	0.7862	0.99	0.99	0.99	0.99	0.99	0.99

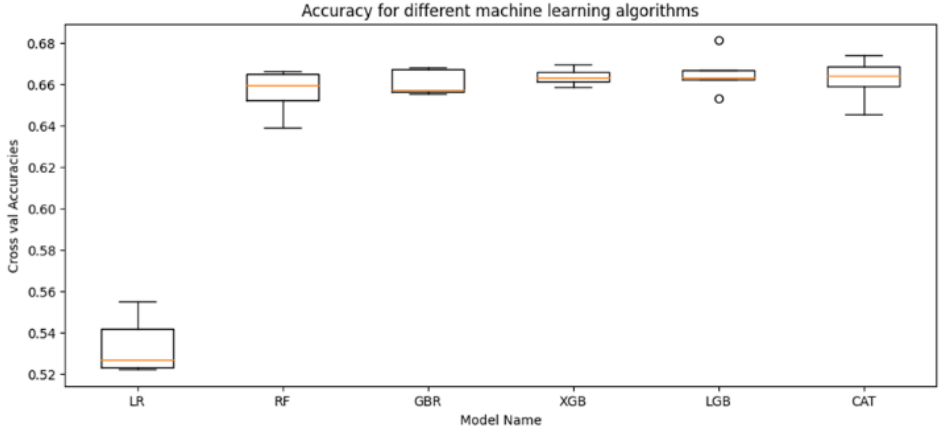


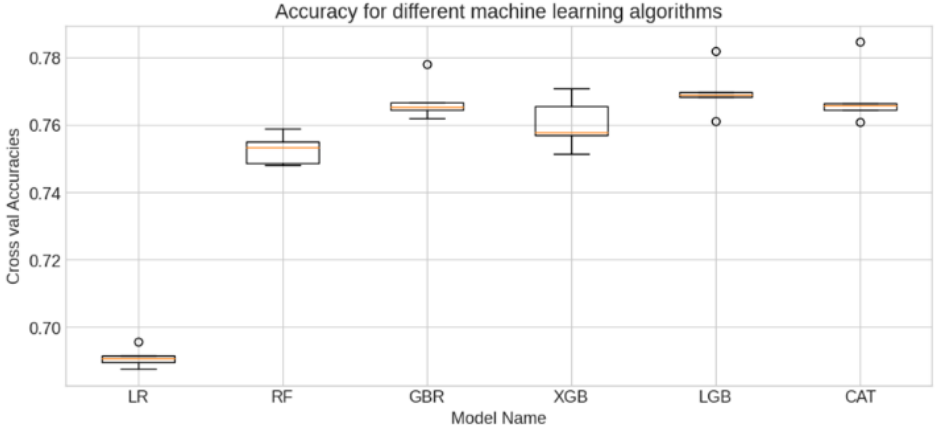
Multi-class



Binary

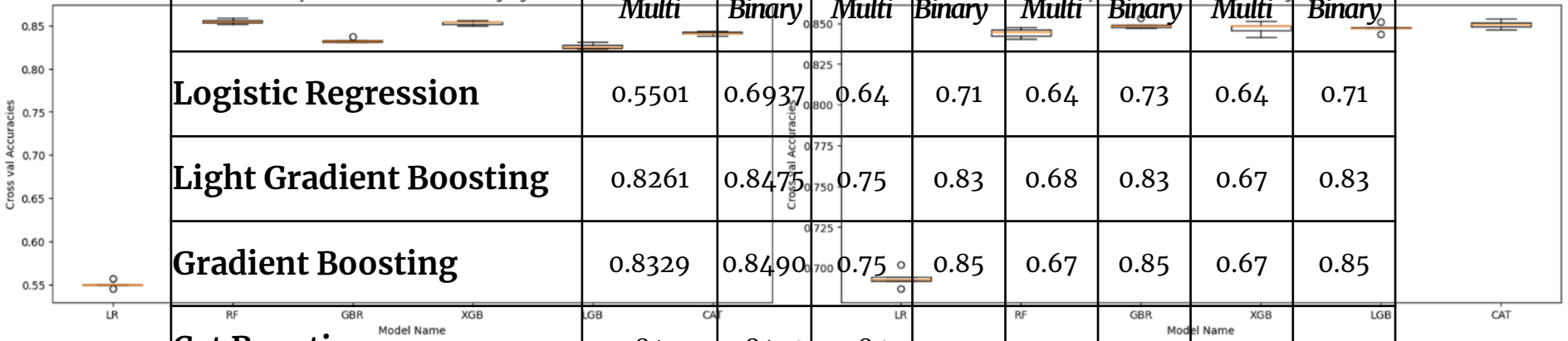
Supervised Learning - Undersampling





Model	Accuracy	Precision	Recall	F1-Score			
Cat Boosting	0.6622	0.7685	0.92	0.90	0.89	0.88	0.89	0.88
	Multi-class					Binary		
Extreme Gradient Boosting	0.6636	0.760	0.97	0.96	0.96	0.96	0.96	0.96
Random Forest	0.6564	0.7685	0.99	0.99	0.99	0.99	0.99	0.99

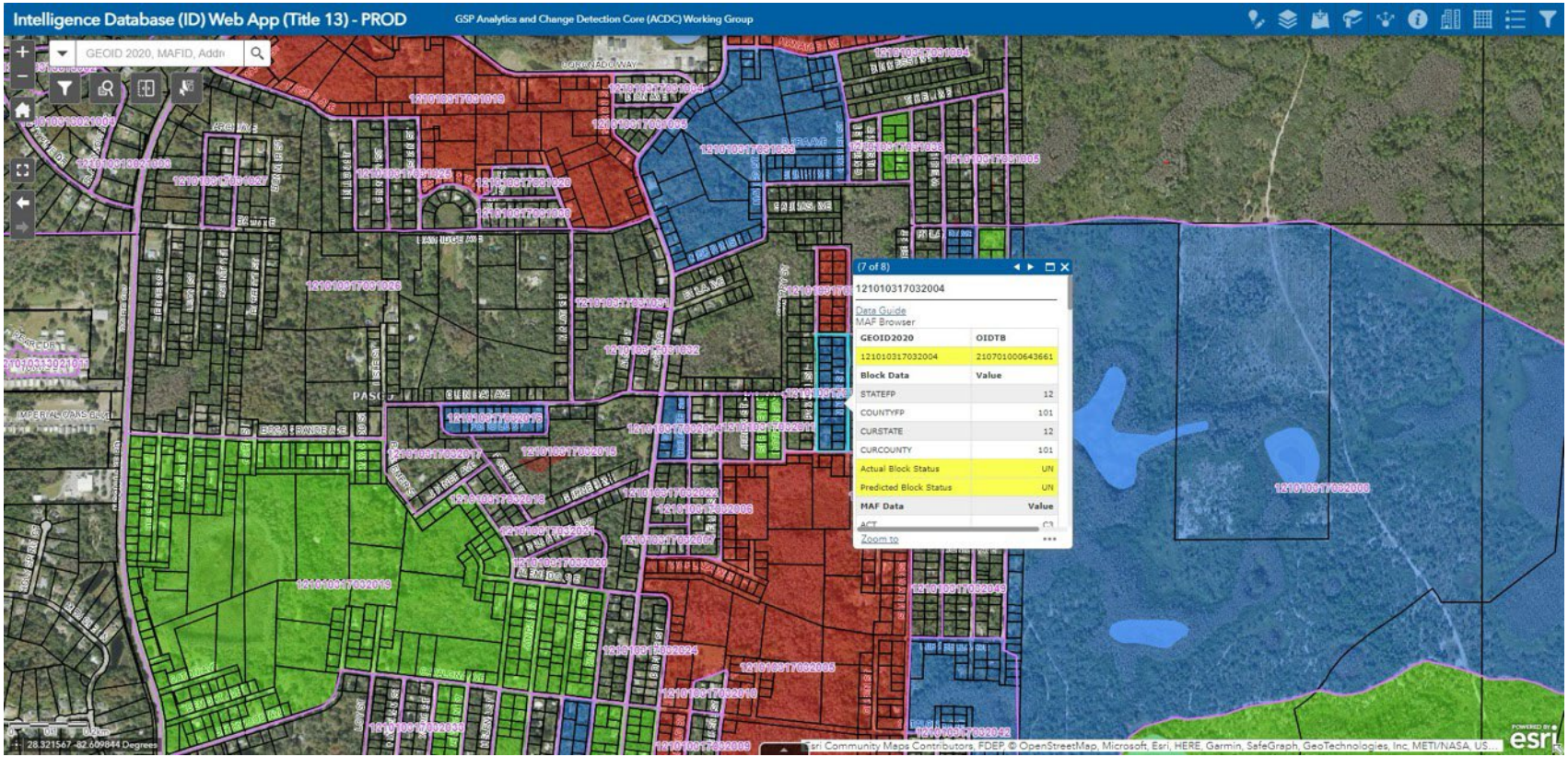
Supervised Learning - Oversampling



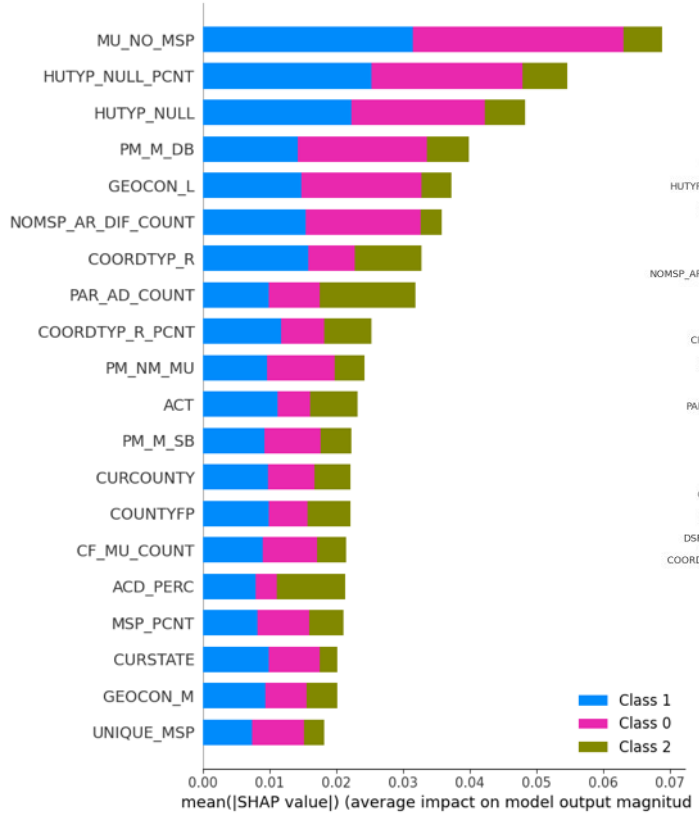
Model	Accuracy		Precision		Recall		F1-Score	
	Multi	*Binary*	*Multi*	*Binary*	*Multi*	*Binary*	*Multi*	*Binary*
Logistic Regression	0.5501	0.6937	0.64	0.71	0.64	0.73	0.64	0.71
Light Gradient Boosting	0.8261	0.8475	0.75	0.83	0.68	0.83	0.67	0.83
Gradient Boosting	0.8329	0.8490	0.75	0.85	0.67	0.85	0.67	0.85
Cat Boosting	0.8411	0.8493	0.85	0.92	0.79	0.92	0.79	0.92
Extreme Gradient Boosting	0.8534	0.8470	0.98	0.97	0.98	0.96	0.98	0.96
Random Forest	0.8547	0.8442	0.98	0.99	0.99	0.99	0.98	0.99

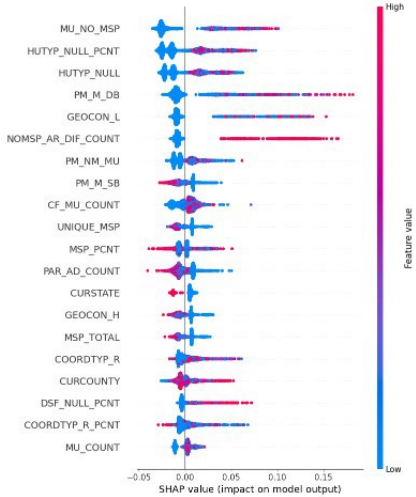
Intelligence Dashboard with SL

Blue – Undercoverage **Green – Passive** **Red – Overcoverage**

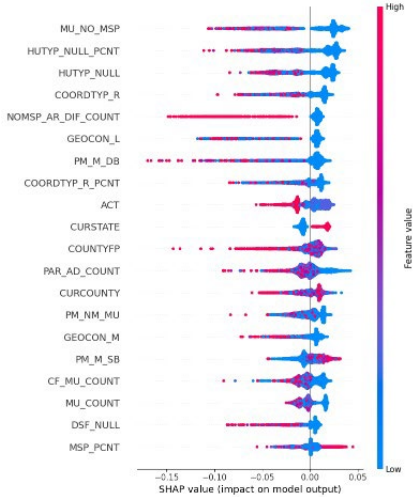


Explainable AI (XAI)

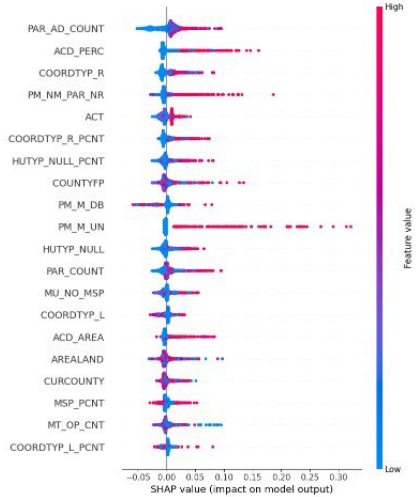




Over-coverage



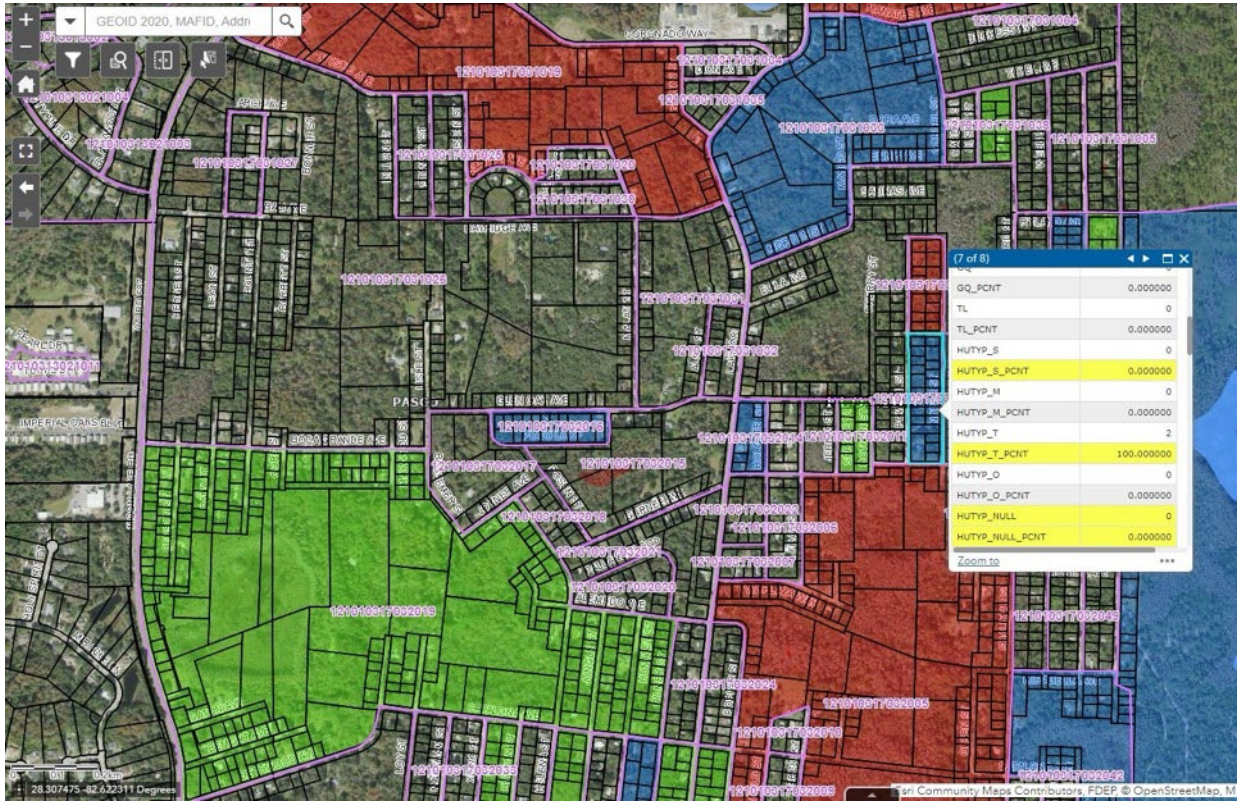
Passive



Under-coverage

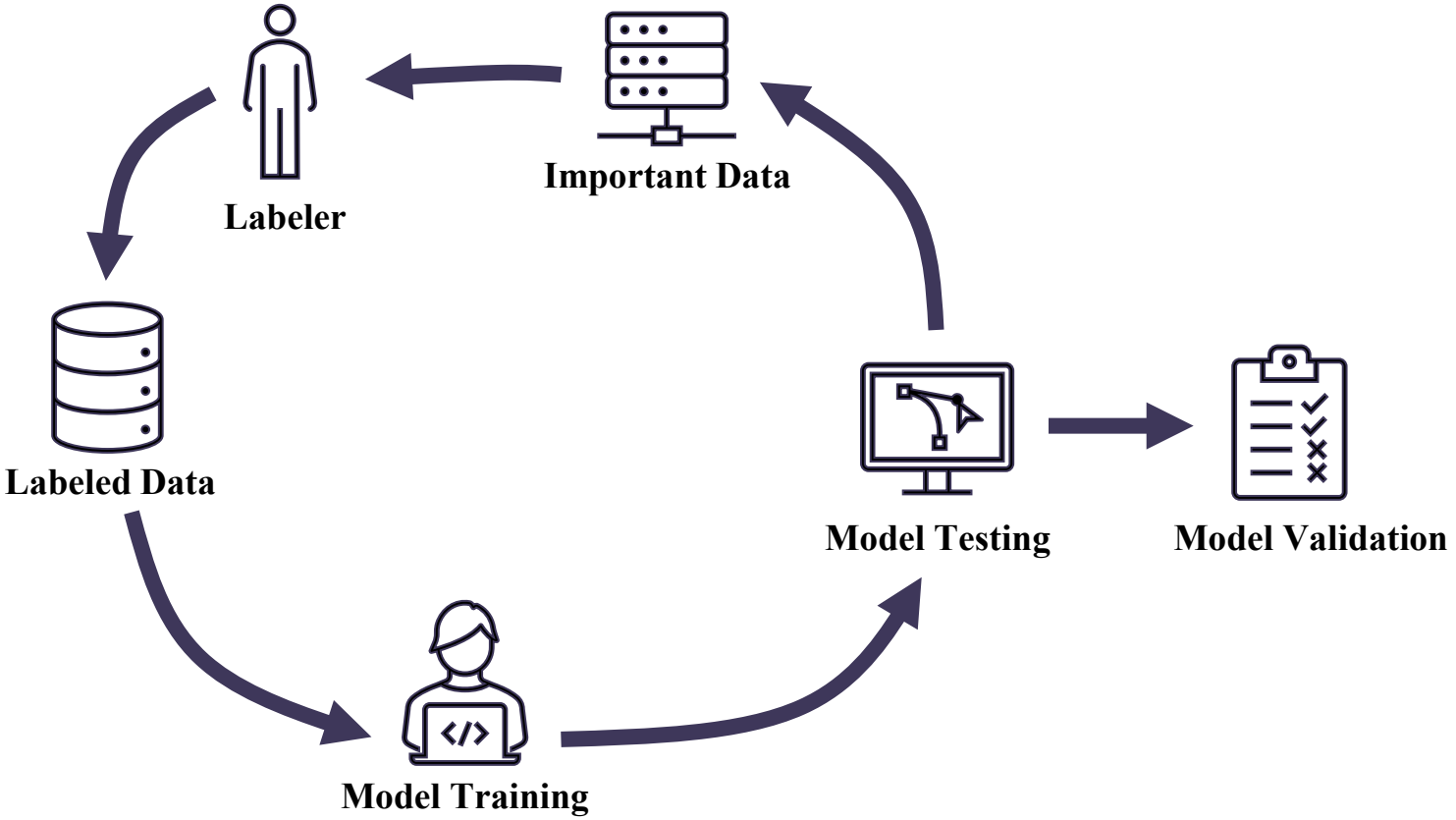
Intelligence Dashboard with XAI

Blue – Undercoverage **Green – Passive** **Red – Overcoverage**



Moving on to SSL.

Semi-Supervised Learning

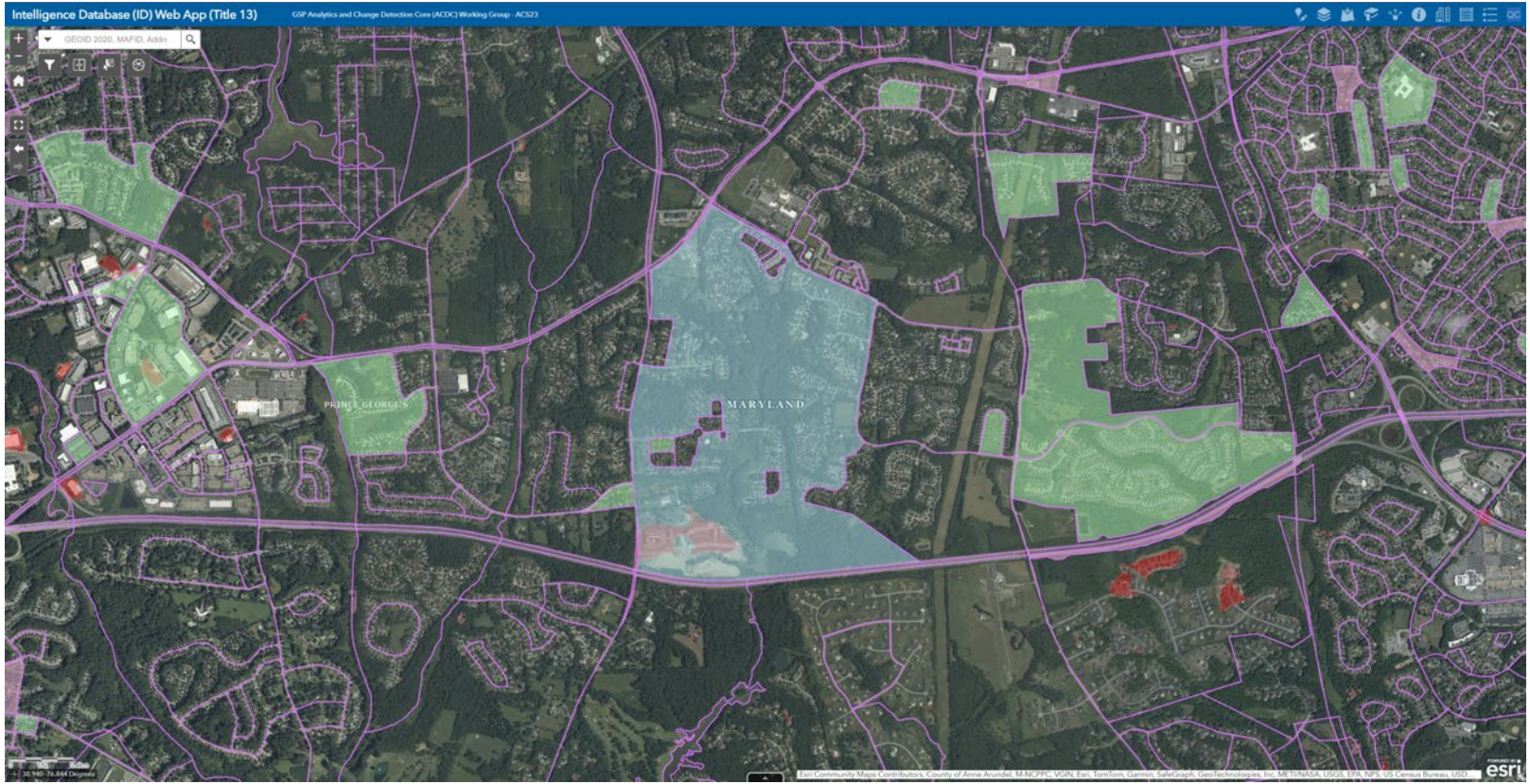


Semi-Supervised Learning

GEOID2020	STATEFP	COUNTYFP	CURSTATE	CURCOUNTY	LR	RF	GBR	XGB	LGB	CAT
121270909051056	12	127	12	127	0	0	0	0	0	0
210350101003007	21	35	21	35	1	1	1	1	1	1
211959316001030	21	195	21	195	1	1	0	1	0	1
121113816041026	12	111	12	111	1	1	1	1	1	1
120990003015011	12	99	12	99	1	1	0	1	1	1
120010009021002	12	1	12	1	1	1	1	1	1	1
50690009002007	5	69	5	69	1	1	1	1	1	1

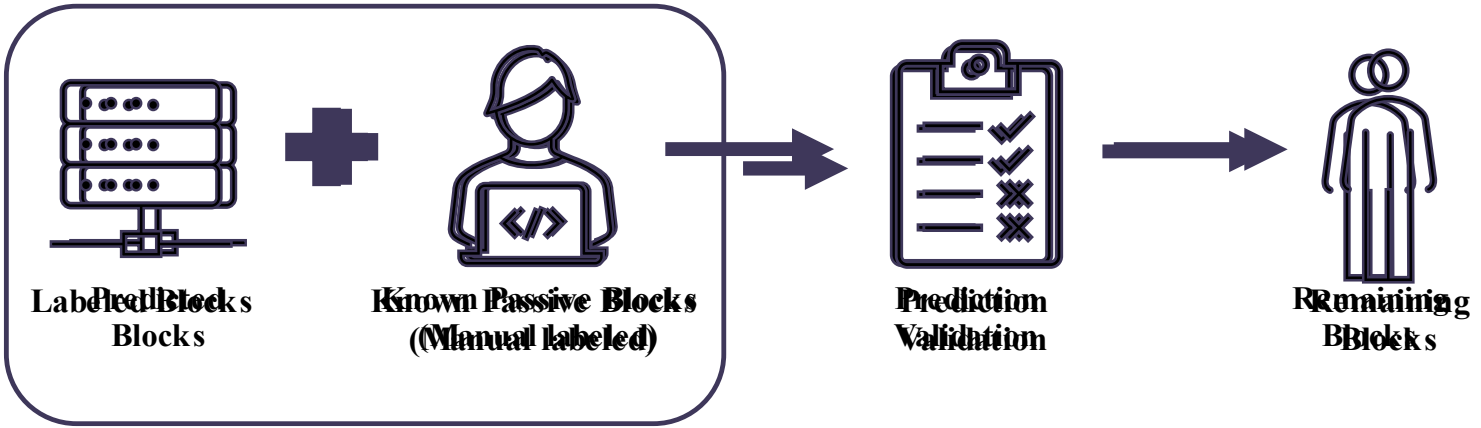
Intelligence Dashboard with SSL

Blue – Undercoverage **Green – Passive** **Red – Overcoverage**



Validation Decisions.

Validation with Passive Blocks



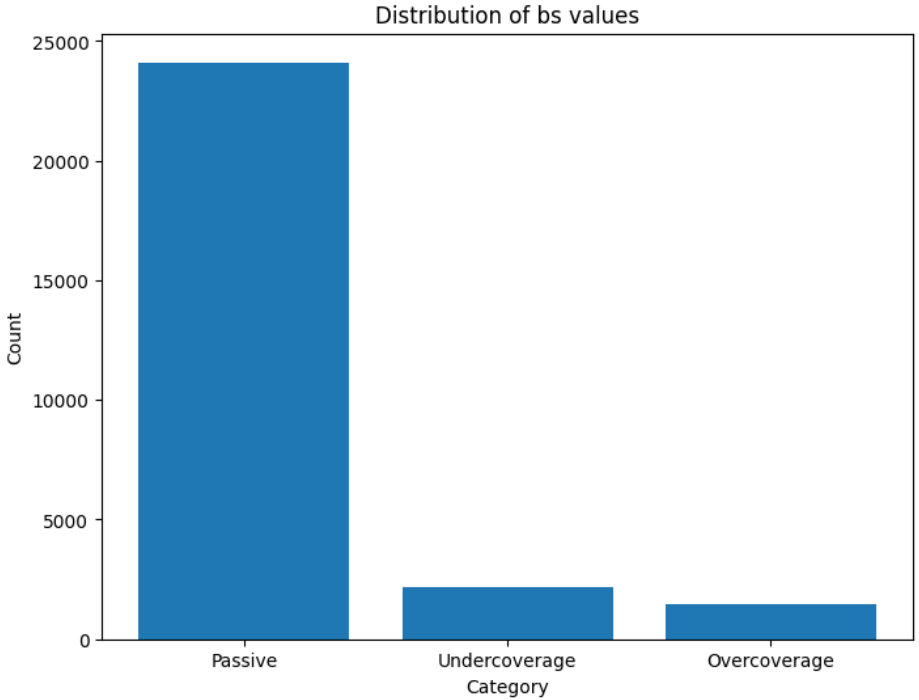
Validation with Passive Blocks

- 2,520,833 Blocks for SE.
  - Test set = 100K
- 4,255,172 Known Passive Blocks.
- Integrated test set:
  - 27710 known passive blocks.
  - 72290 unknown blocks.

```
There are matching GEOID2020 values in both datasets.
306128    1.208601e+14
90863     4.708907e+14
84811     3.703197e+14
157956    3.703501e+14
382669    1.206903e+14
             ...
43209     3.702100e+14
149477    1.205701e+14
4006      1.100101e+14
117247    5.069002e+13
145335    1.212709e+14
Name: GEOID2020, Length: 27710, dtype: float64
Number of matching GEOID2020 values: 27710
```

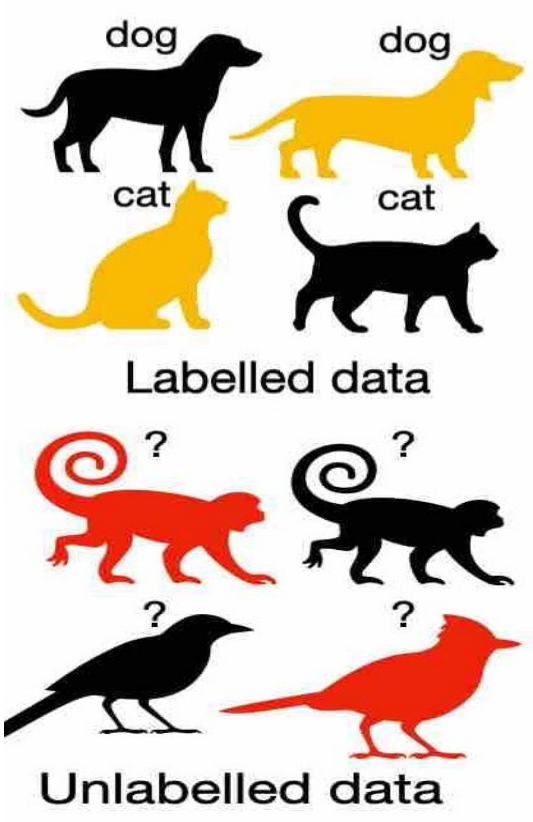
Validation with Passive Blocks

- Out of the 27710 known passive blocks
  - Passive = 24107 (87%)
  - Undercoverage = 2161 (7.80%)
  - Overcoverage = 1442 (5.20%)
- A solid validations set of 87% accuracy for the models.



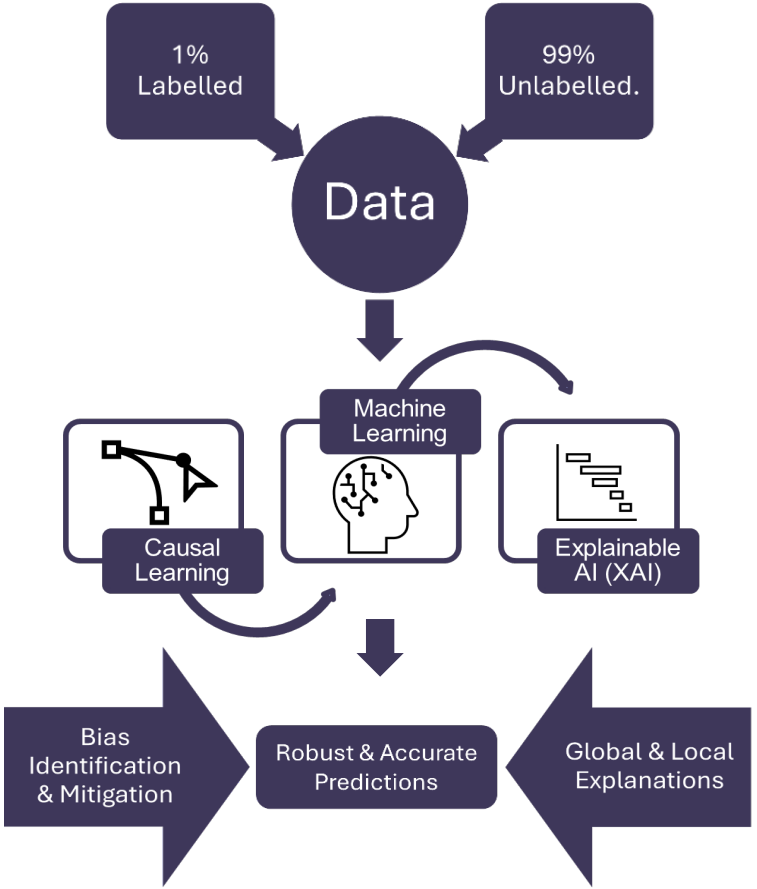
Challenges

- Administrative data is messy.
- <1% of the data is labeled.
- Data labelling takes time.
  - A lot of time!
- Validation for unlabeled data.



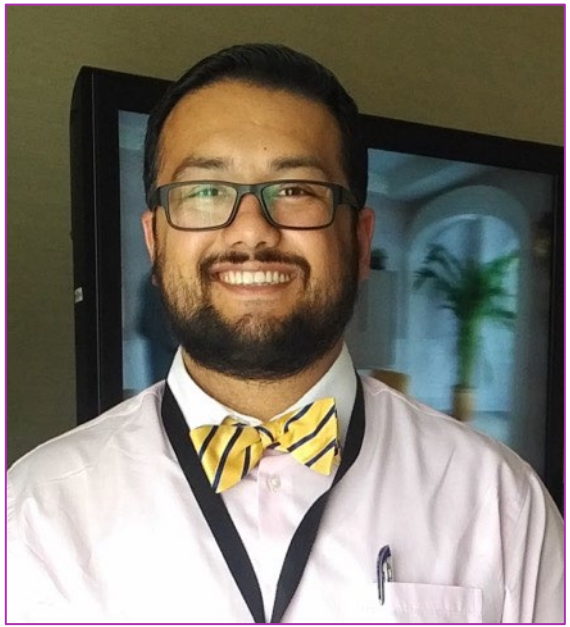
Summary

- Manual canvassing and labelling is cost and time ineffective.
- AI/ML can help alleviate some of the burden.
- Proof-of-concept for using AI/ML to core CB operational tasks.
    - Benefit of removal of human bias involved in interactive review.
- It showcases Census' expertise and evolution in applying AI/ML in geographic data.



—

Questions/Discussion



Atul Rawal, Ph.D.

Education

Albright College, B.Sc. in Theoretical Physics
Joint School of Nanoscience & Nanoengineering, Ph.D. in Nanoengineering
Towson University, Ph.D. in Computer Science (Fall 2024)
Howard University, Ph.D. in Electrical Engineering (On Leave)

Expertise

Artificial Intelligence/Machine Learning, Explainable AI (XAI), Causal learning, Computational Biology, Molecular Dynamics (MD), Quantum Mechanics (QM), Protein Engineering/Dynamics.

Hobbies

Soccer, Lacrosse, Motorcycles, Hiking/Chasing waterfalls with Red. Started a journey to visit all 63 National Parks by the time I turn 40.