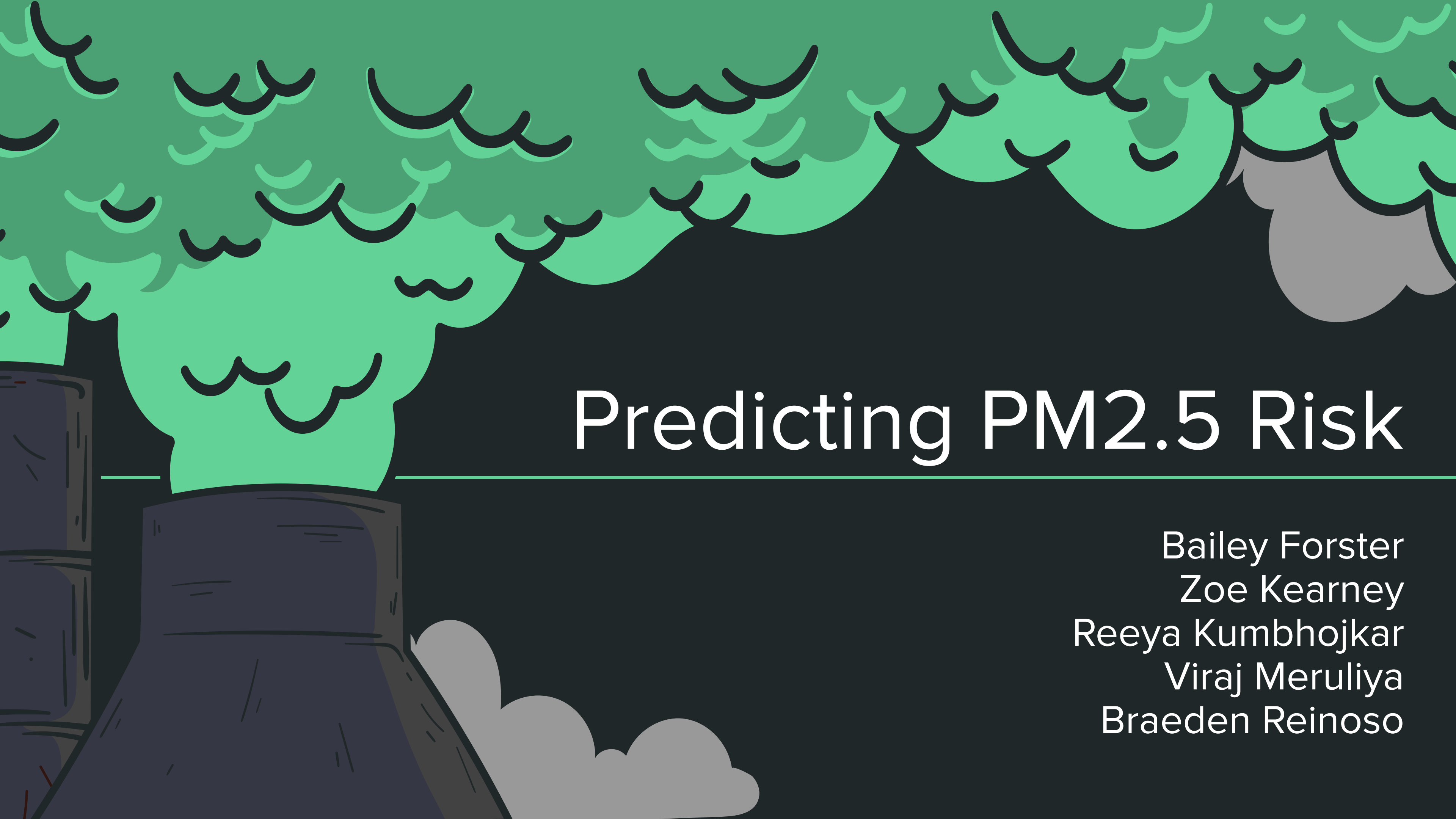
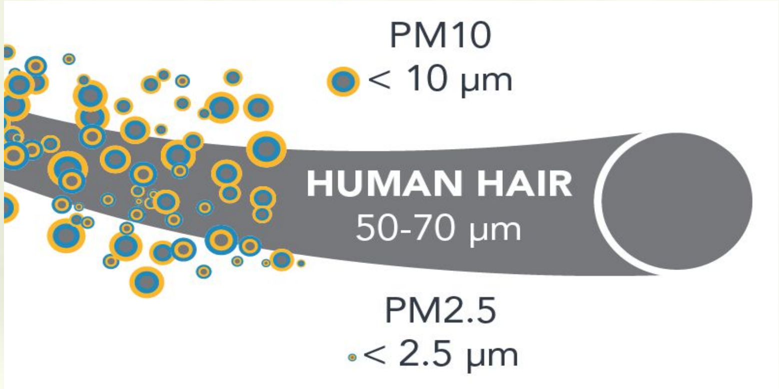


Predicting PM2.5 Risk

Bailey Forster
Zoe Kearney
Reeya Kumbhojkar
Viraj Meruliya
Braeden Reinoso

PM2.5: Overview

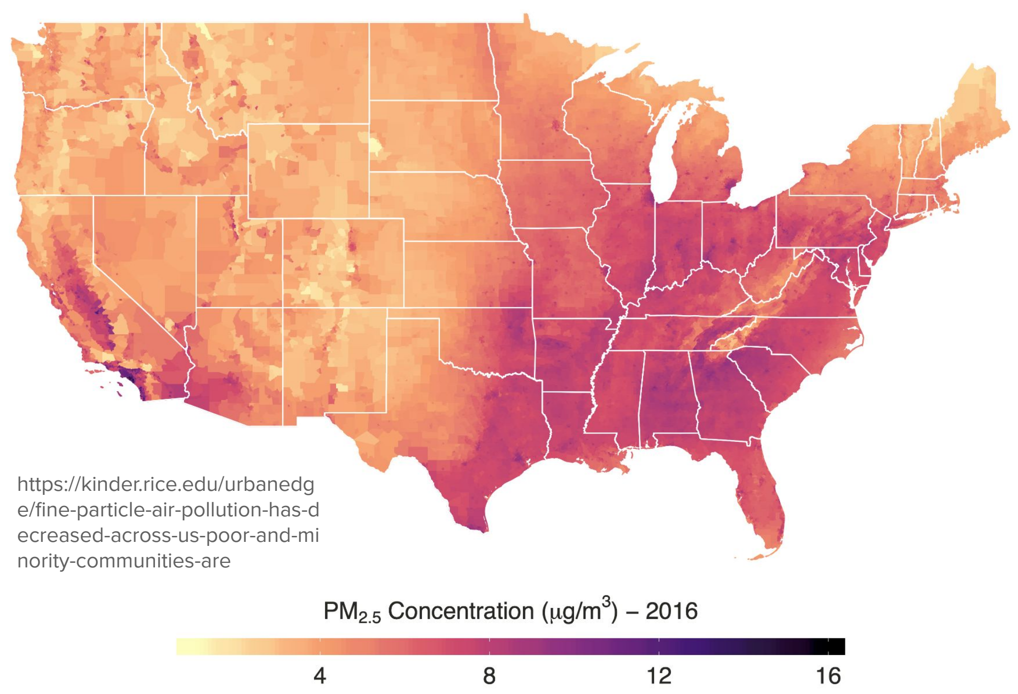
- PM2.5 = inhalable particulate matter in the air
- Risks: cancer, heart attacks, respiratory diseases, low visibility
- Main causes: construction, factories, power plants, cars, natural factors
- WHO Standard: concentration < 5 µg/m^3
- **EPA Standard: concentration < 9 µg/m^3**



https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health

PM2.5: Distribution of Risk

- Problem: PM2.5 risk is distributed highly unequally
- Previous research:
  - People of color at higher risk
  - Urbanization increases risk
  - Focus: large geographic areas (cities, counties, states)



https://kinder.rice.edu/urbanedge/fine-particle-air-pollution-has-decreased-across-us-poor-and-minority-communities-are

Our Project: Goals and Results

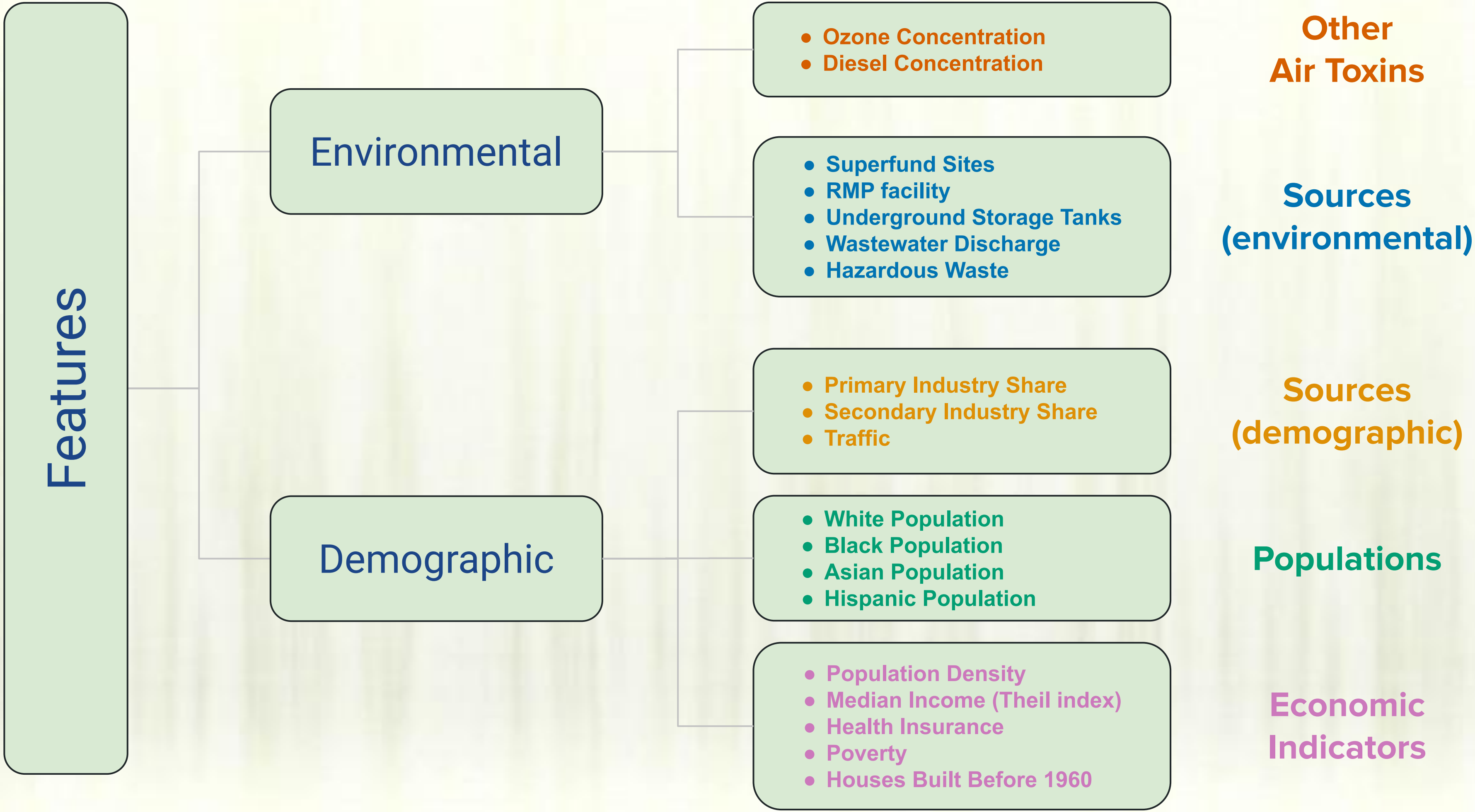
- **Our Goal: Predict high-risk for urban areas based on demographic and environmental data, at the highly local (census tract) scale**
- Motivations:
  - Compare sources of PM2.5 risk to make informed policy decisions
  - Understand which populations are at increased risk, and from which PM2.5 sources
  - Identify key risk predictors at highly local scale
- Results:
  - Model predicts high-risk areas with **93% accuracy**
  - Identified **clear patterns of risk** among demographic groups and man-made sources

Data: Collection, Cleaning, and Analysis

Data Collection and Cleaning

- Collected data at census tract level
  - Environmental data: U.S. Environmental Protection Agency (EPA)
  - Demographic data: U.S. Census Bureau

- Some hurdles along the way:
  - Lack of granularity in key variables
  - Missing data in rural and non-continental areas
  - Tract boundaries - all data must be post-2020



Features

Environmental

- Ozone Concentration
- Diesel Concentration

Other Air Toxins

- Superfund Sites
- RMP facility
- Underground Storage Tanks
- Wastewater Discharge
- Hazardous Waste

Sources (environmental)

Demographic

- Primary Industry Share
- Secondary Industry Share
- Traffic

Sources (demographic)

- White Population
- Black Population
- Asian Population
- Hispanic Population

Populations

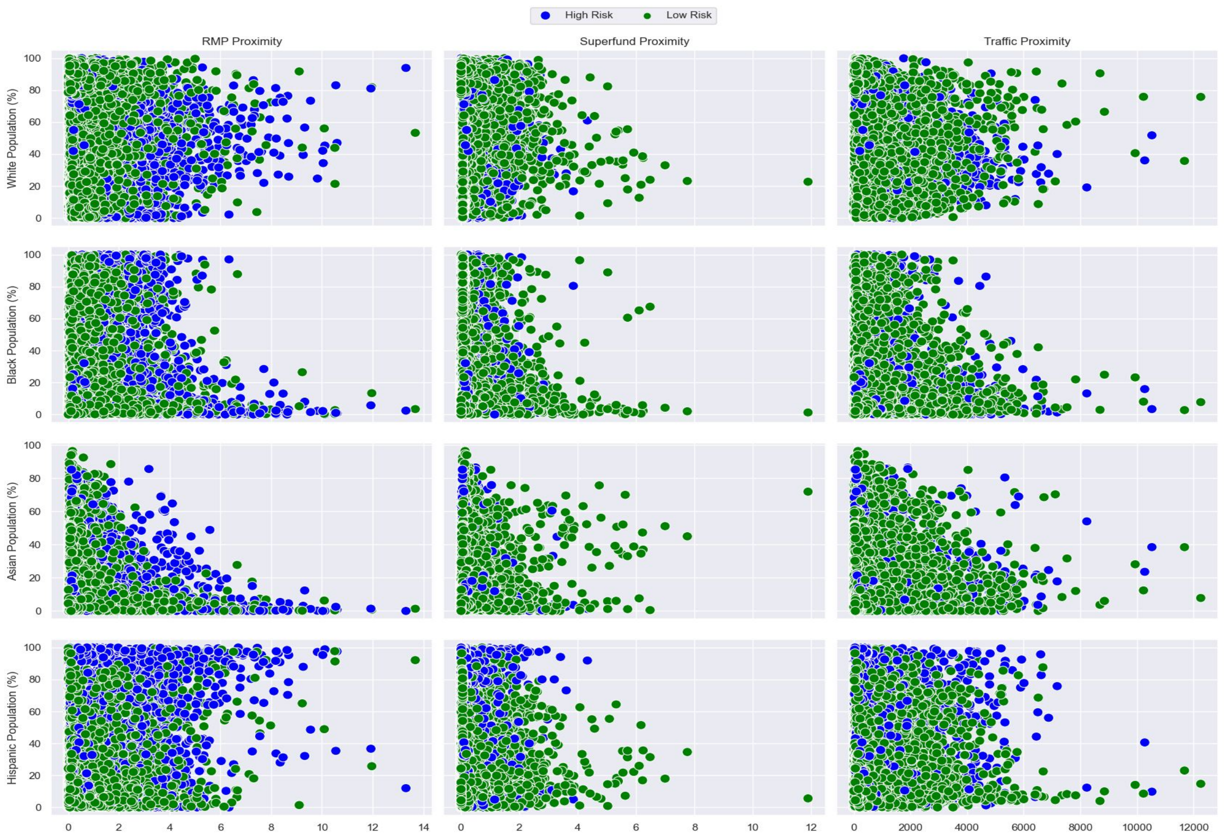
- Population Density
- Median Income (Theil index)
- Health Insurance
- Poverty
- Houses Built Before 1960

Economic Indicators

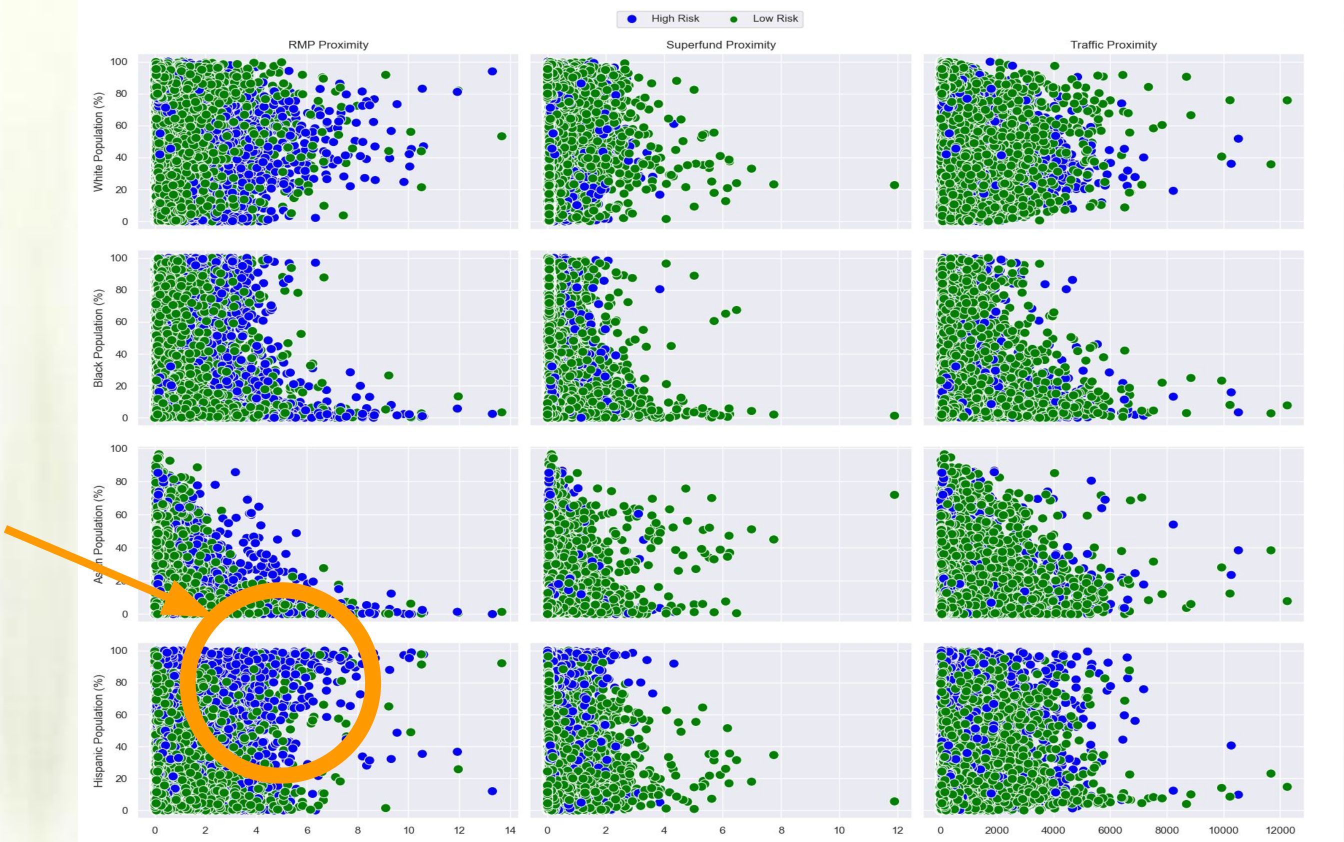
Feature Comparison: Environmental vs Demographic

- High or low risk (EPA standard: PM2.5 < 9 µg/m^3)

- Imbalanced data: 34% high-risk / 66% low-risk



Feature Comparison: Environmental vs Demographic

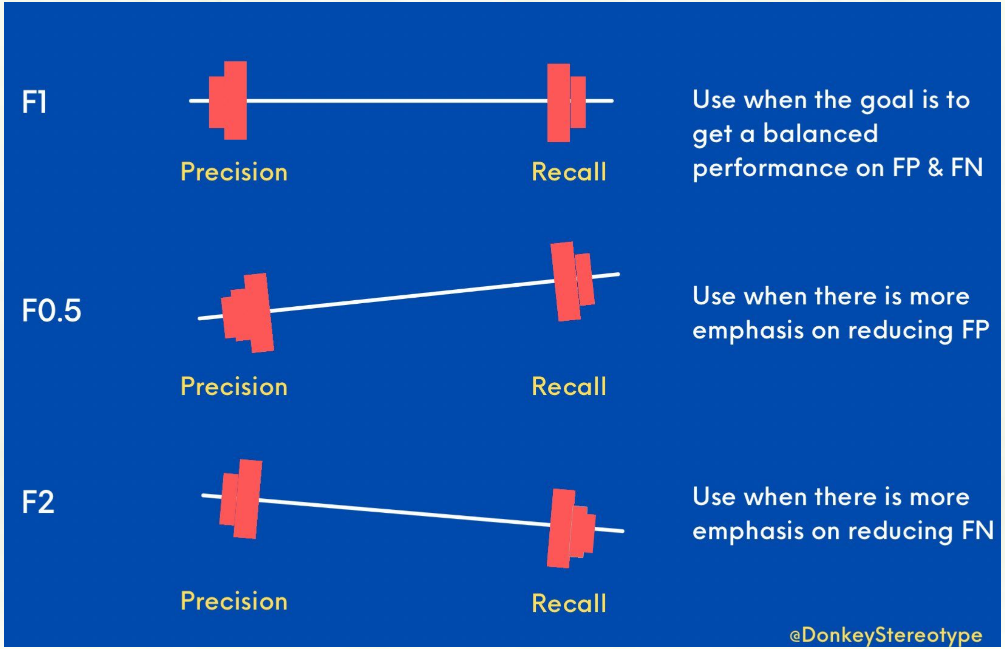


Hispanic Population at more risk at high RMP proximity (PM2.5 source)

Modeling: Approach and Comparison

Modeling: Metrics and Baseline

- Trade-off: recall (reduce false negatives) vs precision (reduce false positives)
- **Prioritize correctly identifying high-risk areas**
- **Baseline model:** predict all tracts as high risk
  - Perfect recall (100%) - No false negatives!
  - Poor accuracy and precision (both 34%)
- **Evaluation metric:** f2 score



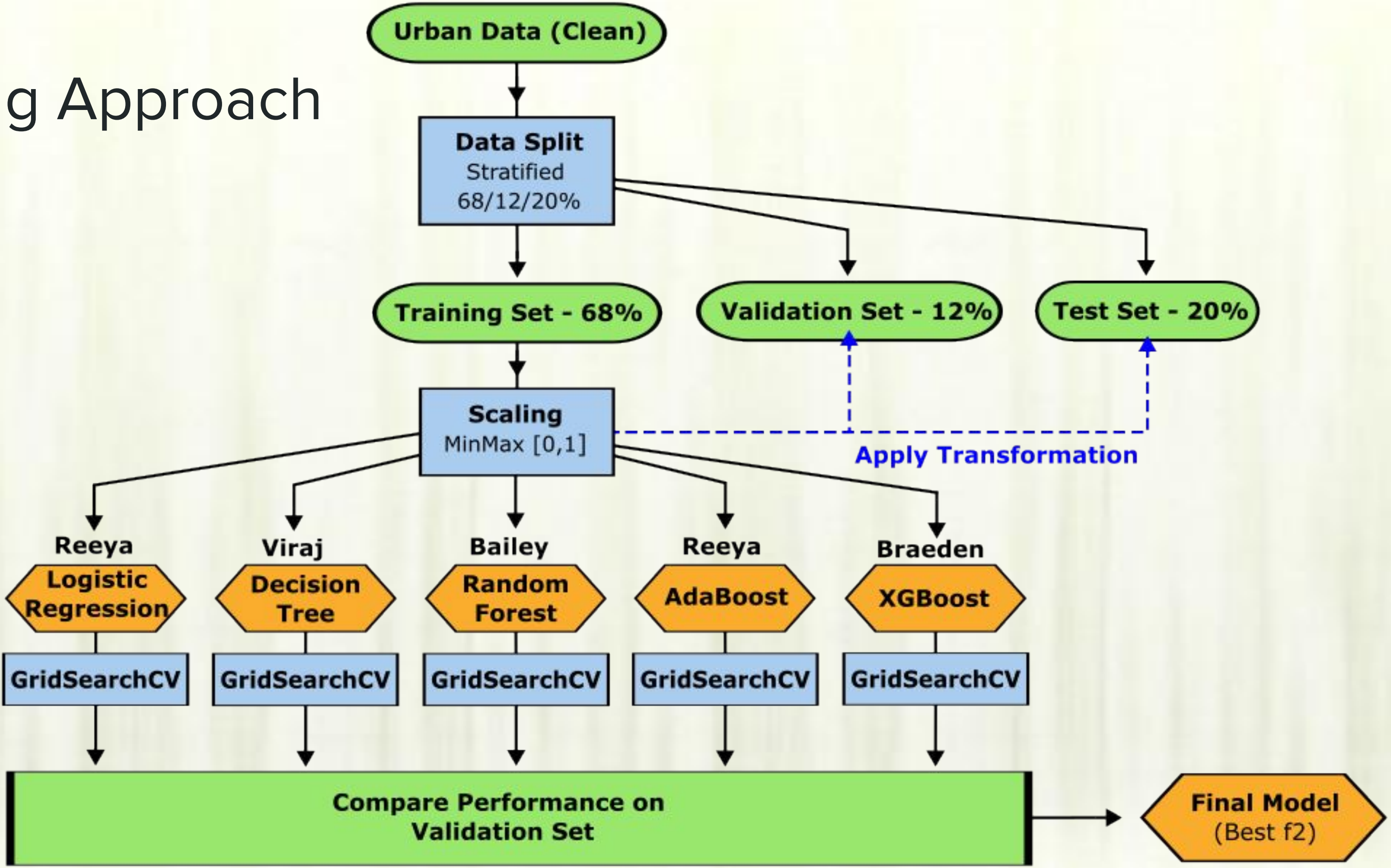
Use when the goal is to get a balanced performance on FP & FN

Use when there is more emphasis on reducing FP

Use when there is more emphasis on reducing FN

https://twitter.com/prithivida/status/1496100101877641216

Modeling Approach



Urban Data (Clean)

Data Split
Stratified
68/12/20%

Training Set - 68%

Validation Set - 12%

Test Set - 20%

Scaling
MinMax [0,1]

Apply Transformation

Reeya
Logistic Regression
GridSearchCV

Viraj
Decision Tree
GridSearchCV

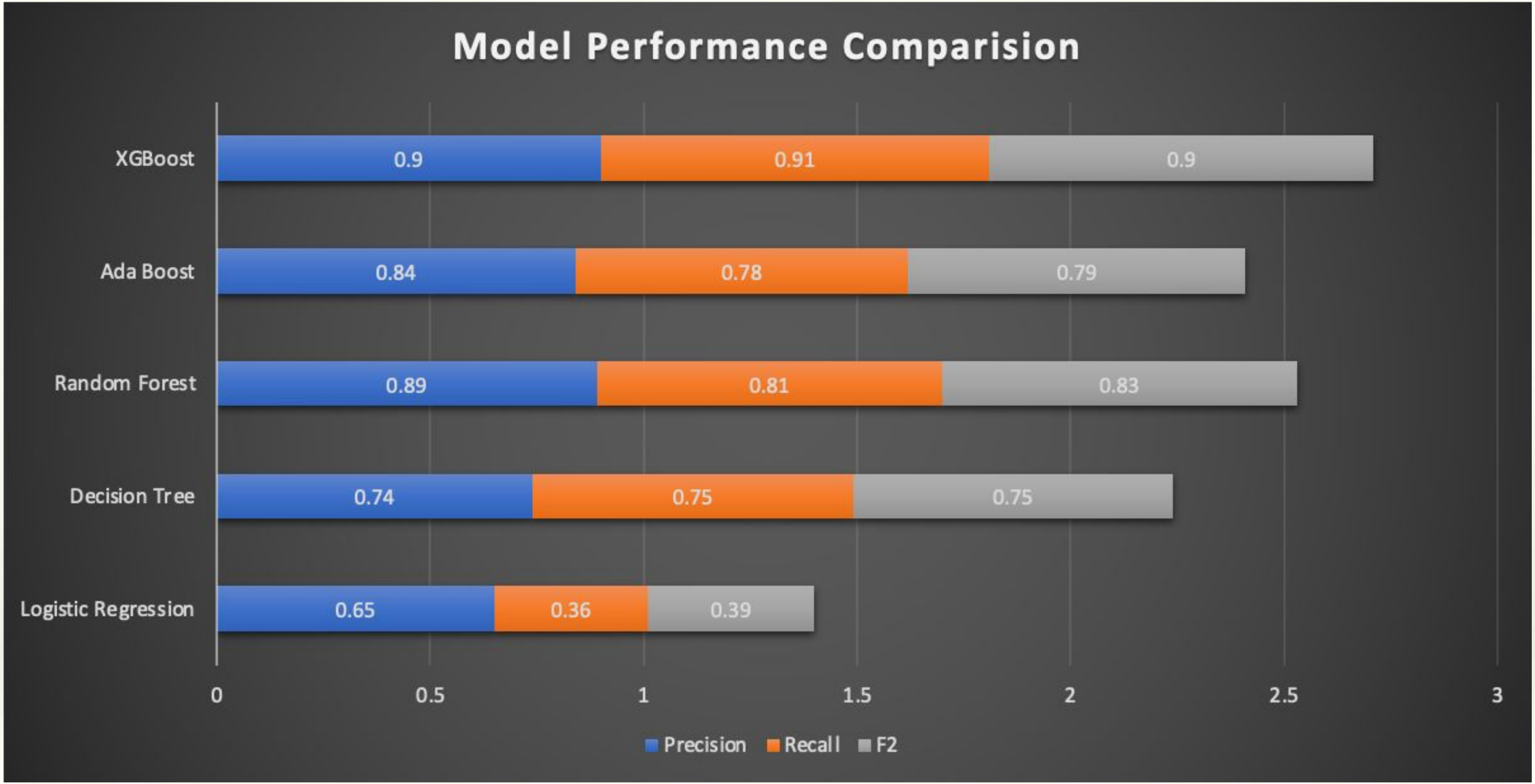
Bailey
Random Forest
GridSearchCV

Reeya
AdaBoost
GridSearchCV

Braeden
XGBoost
GridSearchCV

Compare Performance on Validation Set

Final Model
(Best f2)

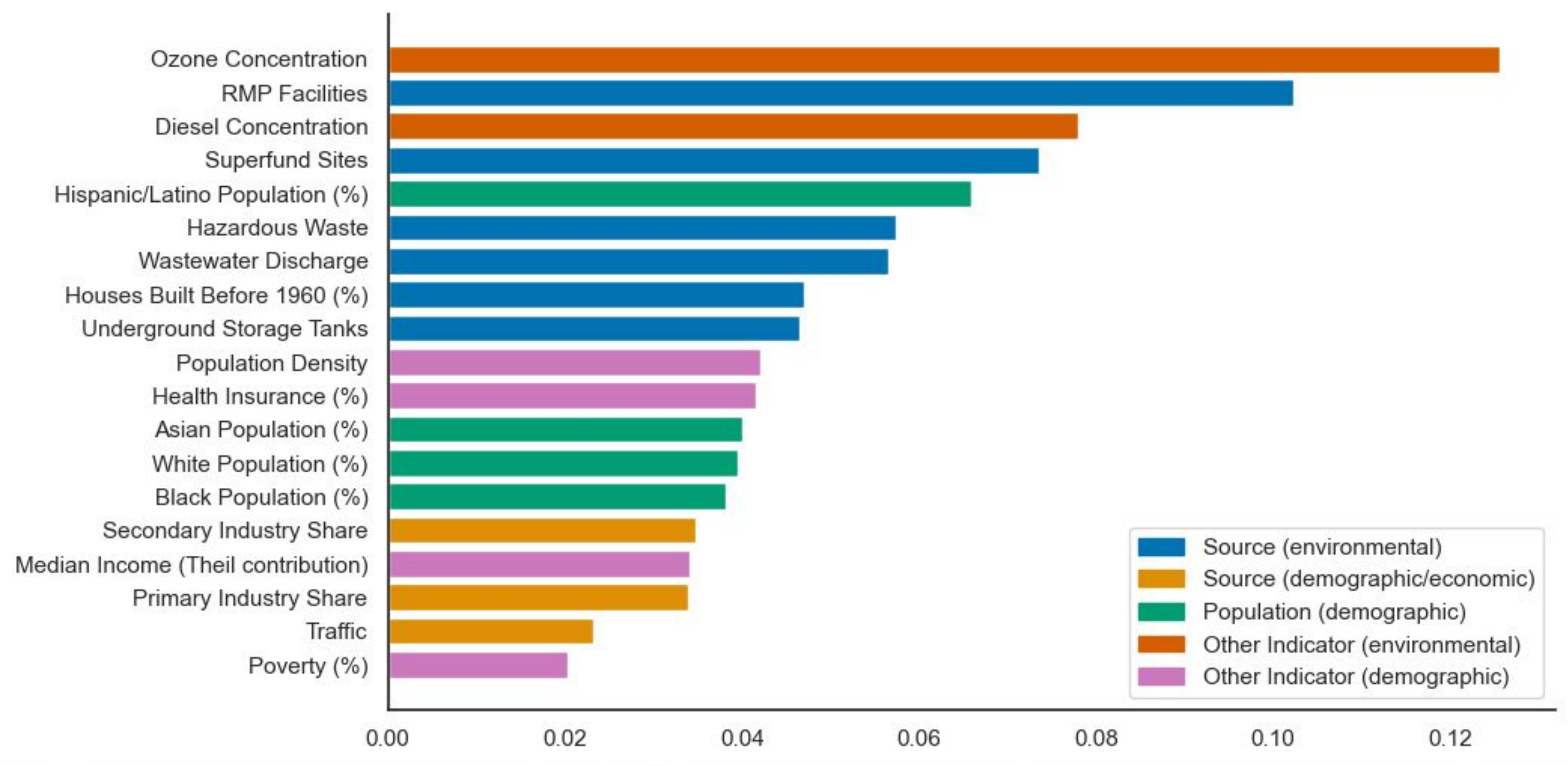

Model Performance Comparision
XGBoost
0.9
0.91
0.9
Ada Boost
0.84
0.78
0.79
Random Forest
0.89
0.81
0.83
Decision Tree
0.74
0.75
0.75
Logistic Regression
0.65
0.36
0.39
0
0.5
1
1.5
2
2.5
3
Precision
Recall
F2

Modeling: Inference and Interpretation

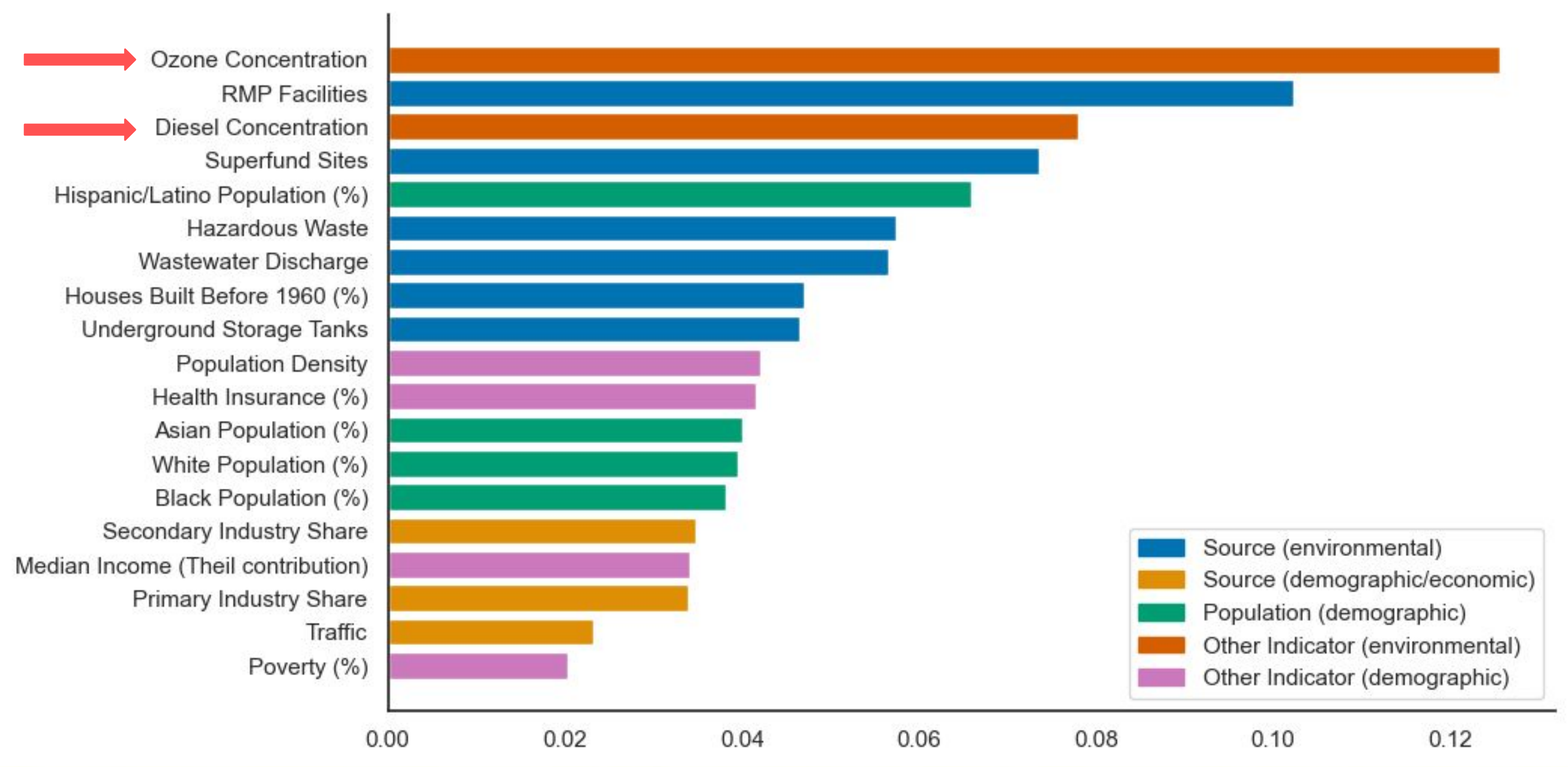
Modeling: XGBoost Final Model Evaluation

	Validation	Test	Baseline
Accuracy	93%	93%	34%
f2 Score	**90%**	**89%**	**72%**
Recall	91%	89%	100%
Precision	90%	91%	34%
Area under PR-curve	97%	97%	34%

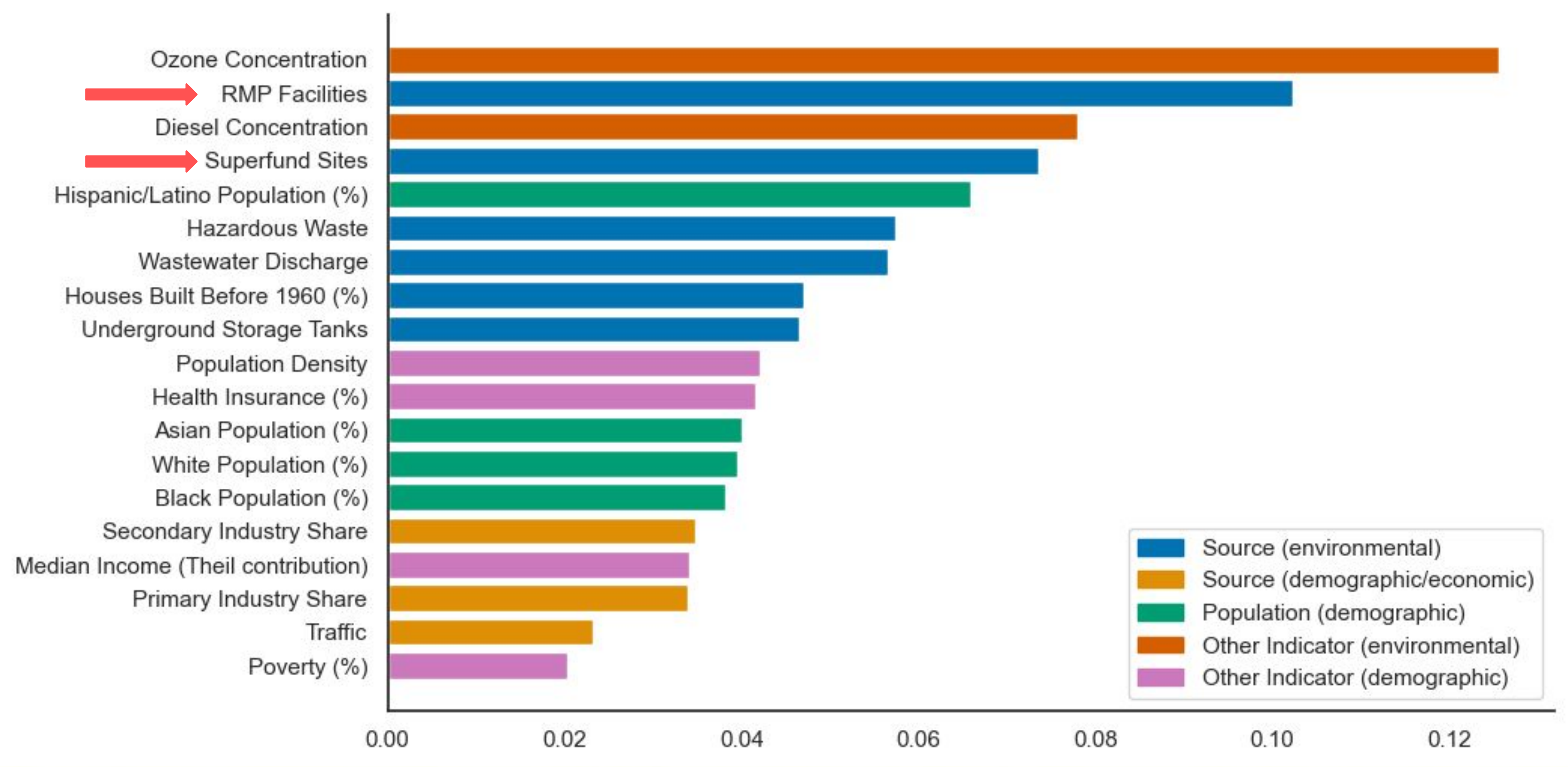
Modeling: XGBoost Final Model Feature Importance



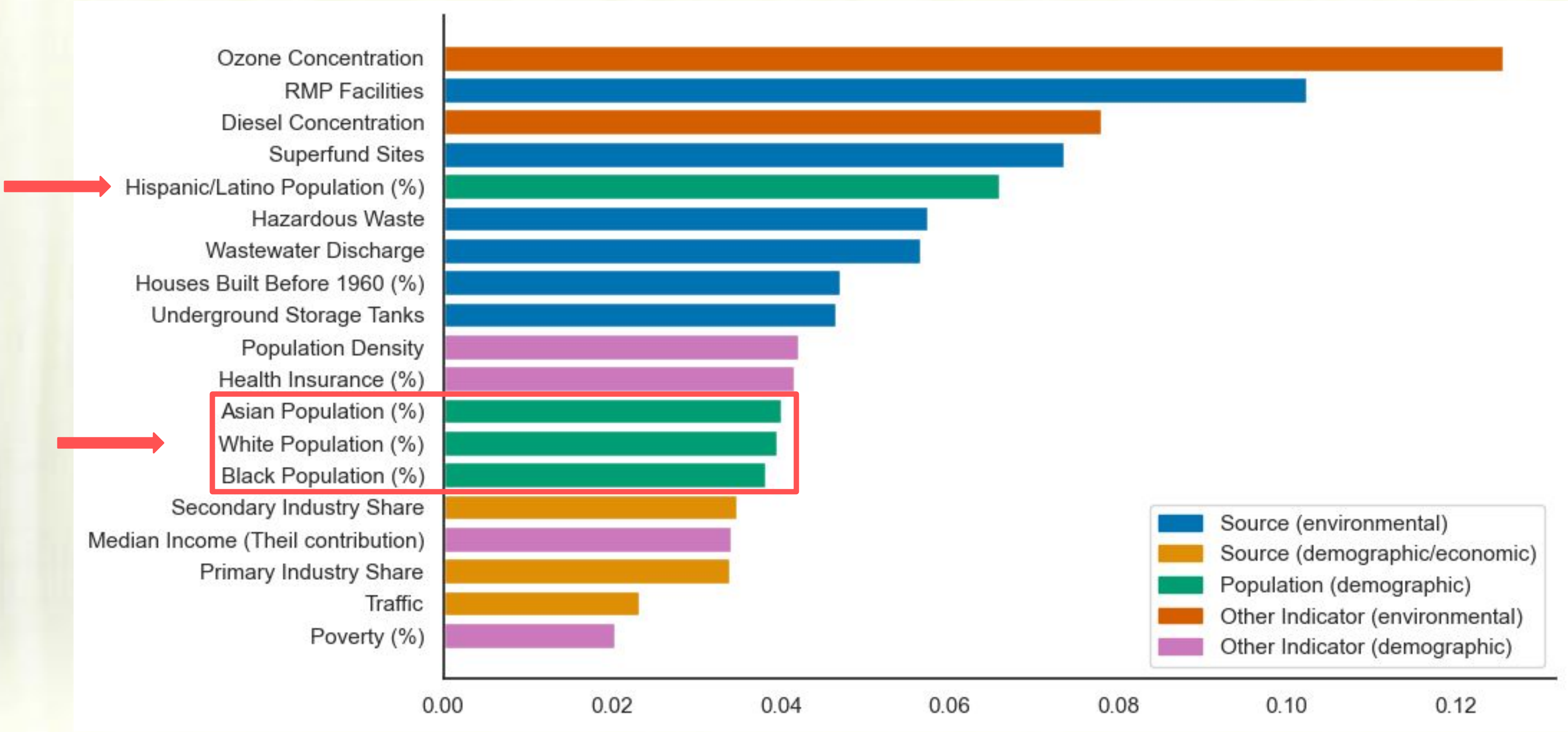
High PM2.5 risk is associated with high Ozone/Diesel risk



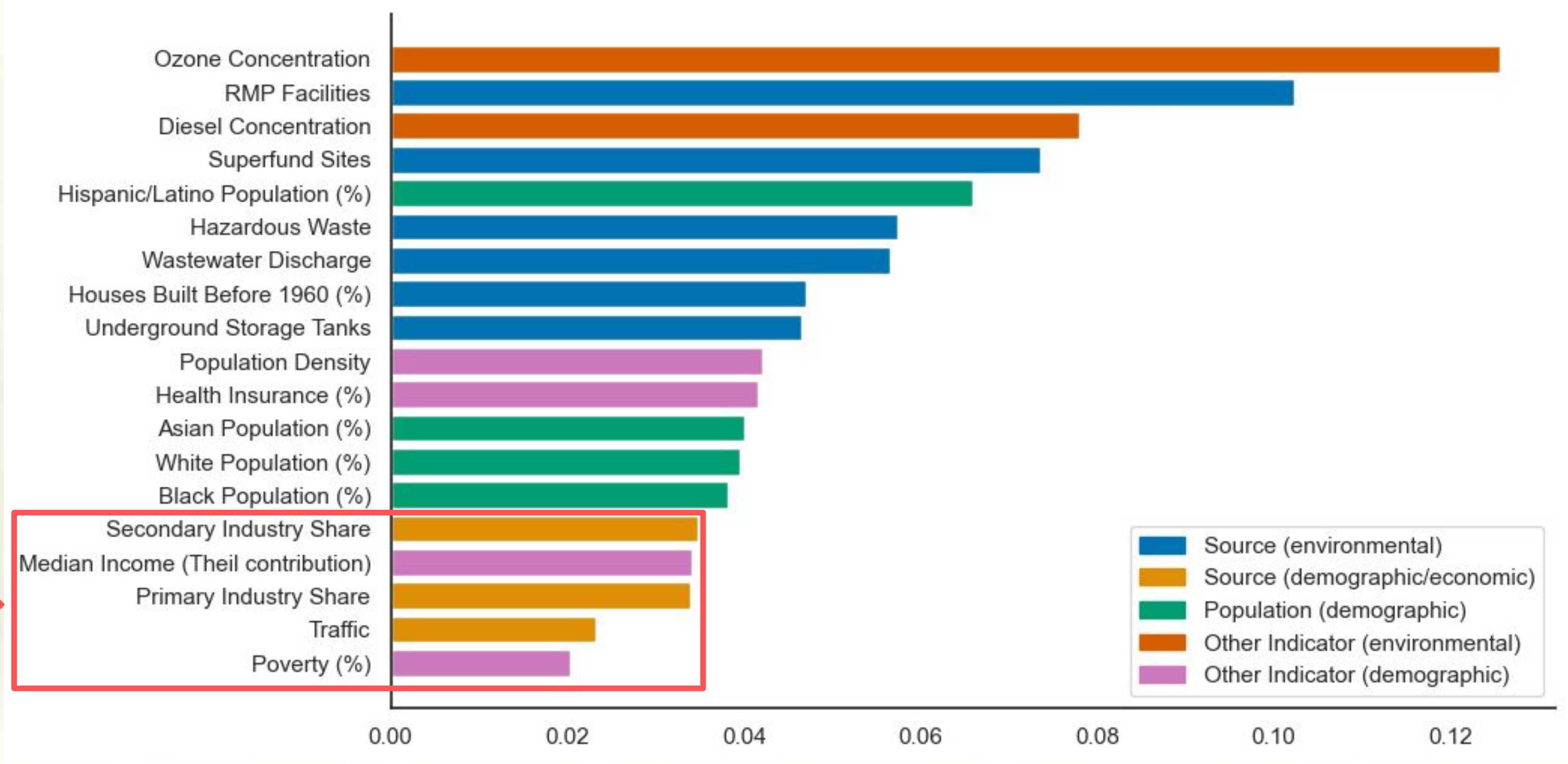
RMP and Superfund sites are the biggest sources of risk



Hispanic/Latino populations experience outsized risk



Economic factors are less predictive at a highly local scale



Diving Deeper: Predictive Comparison of Feature Groups

	Full Model	Without Ozone/Diesel	PM2.5 Sources	Demographic	Baseline
Accuracy	93%	84%	76%	67%	34%
f2 Score	89%	78%	73%	61%	72%
Recall	89%	78%	77%	64%	100%
Precision	91%	76%	61%	52%	34%
Area under PR-curve	97%	87%	75%	60%	34%

Diving Deeper: Predictive Comparison of Feature Groups

	Full Model	Without Ozone/Diesel	PM2.5 Sources	Demographic	Baseline
Accuracy	93%	84%	76%	67%	34%
f2 Score	89%	78%	73%	61%	72%
Recall	89%	78%	77%	64%	100%
Precision	91%	76%	61%	52%	34%
Area under PR-curve	97%	87%	75%	60%	34%

Diving Deeper: Predictive Comparison of Feature Groups

	Full Model	Without Ozone/Diesel	PM2.5 Sources	Demographic	Baseline
Accuracy	93%	84%	76%	67%	34%
f2 Score	89%	78%	73%	61%	72%
Recall	89%	78%	77%	64%	100%
Precision	91%	76%	61%	52%	34%
Area under PR-curve	97%	87%	75%	60%	34%

Diving Deeper: Predictive Comparison of Feature Groups

	Full Model	Without Ozone/Diesel	PM2.5 Sources	Demographic	Baseline
Accuracy	93%	84%	76%	67%	34%
f2 Score	89%	78%	73%	61%	72%
Recall	89%	78%	77%	64%	100%
Precision	91%	76%	61%	52%	34%
Area under PR-curve	97%	87%	75%	60%	34%

Wrapping Up

Summary and Future Directions

- Results:
  - Binary classifier with **93% accuracy** and **89% f2-score**
  - **New insights** into causes and distribution of PM2.5 risk
- Future directions:
  - Separate classification models for **target populations**
    - E.g. control for areas with high hispanic/latino populations
    - Features: **PM2.5 sources** and **health outcomes**
  - Multinomial model: low, medium, high risk based on WHO, EPA, US standards
  - Rural model: how does feature importance change?

Acknowledgements

- Thanks: Roman Holowinsky, Matt Osborne, Alec Clott, Erdös Institute.
- Thank you to our project mentor, Kenny Salau, for his support from the beginning to the completion of the project.
- Dataset Sources:
    - U.S. Environmental Protection Agency (EPA), 2023. EJScreen Technical Documentation
    - United States Census Bureau