Insights into Diabetes Prevalence in the US

Team:

Leyda Almodóvar

Neal Edgren

Chiara Mattamira

Shravan Patankar

Mentor:

Bailey Forster

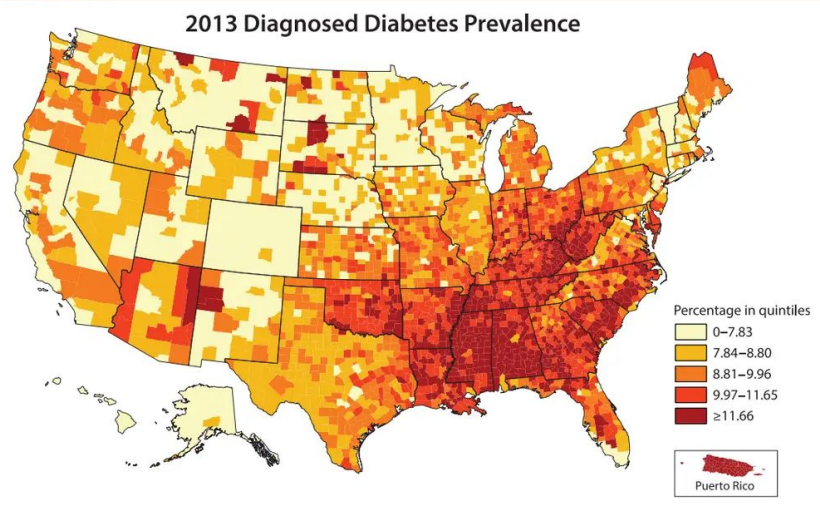
Diabetes Stats

11.6%

of American in 2021 had diabetes

Number of Americans diagnosed with diabetes every year

2.1m



% Diabetes is distributed highly unequally across counties

Our Project

Goal: Analyze high risk for diabetes based on demographic, socioeconomic, environmental, and health behaviors data

Motivation:

- Understand which populations are at risk at a local level
- Identify key risk predictors
- Provide insights to help make informed policy decisions

Features

Demographic → Age, race, sex

Socio-economic → Income, education, unemployment

Health Habits → Drinking, smoking, exercising

Environmental → Pollution, traffic, healthy food access

Access to Care → Vaccination, primary care, and insured rates

Data Info and Data Cleaning

- **Source:** County Health Rankings & Roadmaps (CHR&R)
- **Size:** 3200 rows and 88 columns
- **Removed**
  - rows with aggregated state data
  - counties with more than 35 null values
  - categorical data
  - features with more than 100 missing values
  - features that are redundant or obviously correlated
- **Imputed** missing values using Knn with n = 10

Demographic

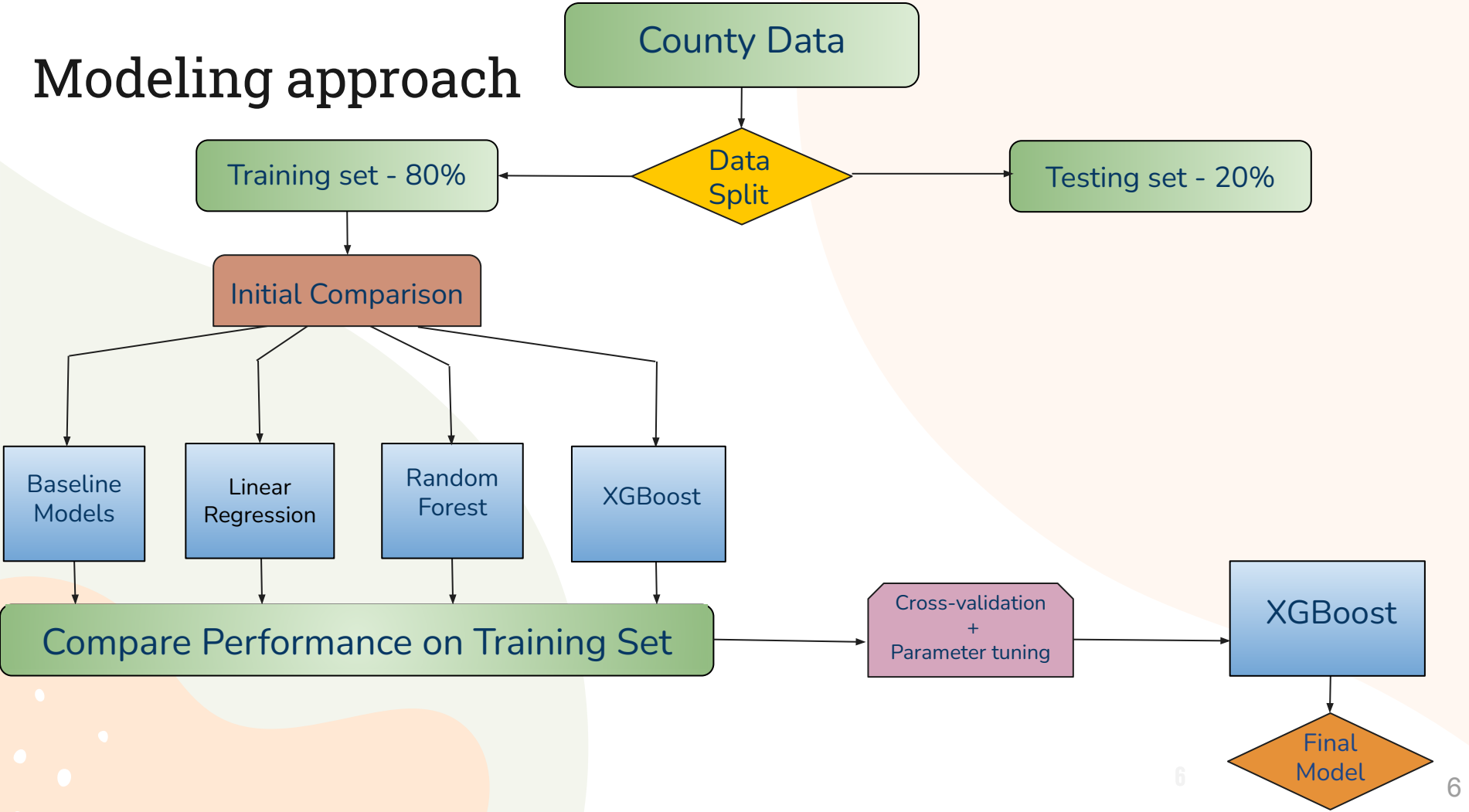
Socio-economic

Health Habits

Environmental

Access to Care

Modeling approach



Initial Performance Comparison

	RMSE for Training set	RMSE for Validation set
Mean model	2.23	2.31
Random sampling	3.16	3.26
SLR on % w/ Obesity	1.64	1.64
Linear regression	0.47	0.50
Random Forest	0.20	0.564
XGBoost	0.02	0.556

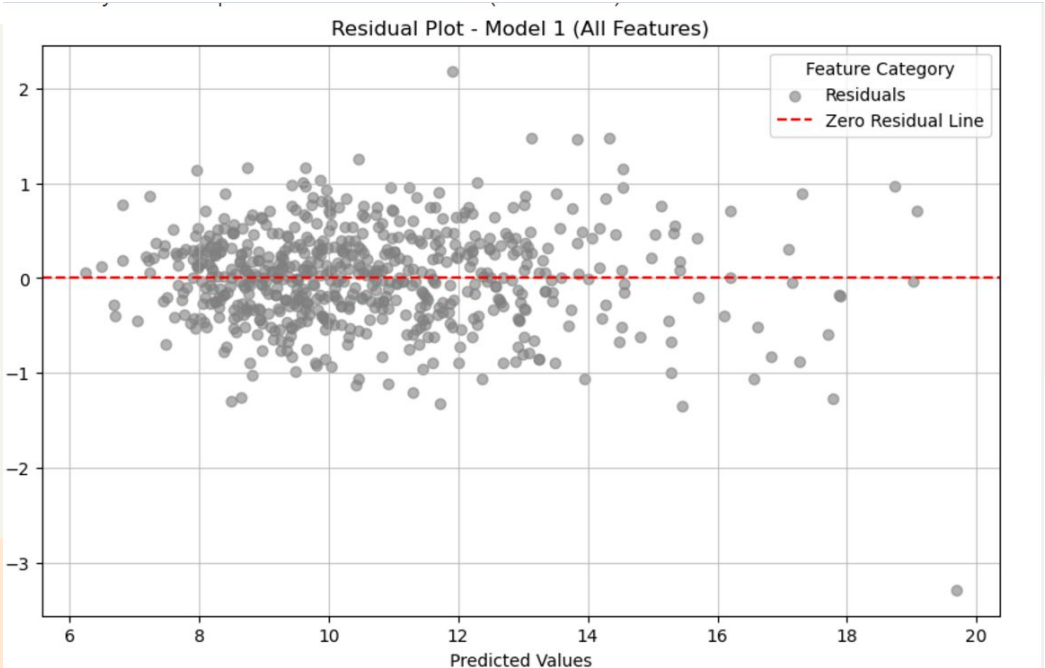
Model Evaluation

- Root Mean Square Error (RMSE)
  - magnifies large errors and ignores small ones
  - biased
- Mean Absolute Error (MAE)
  - treats all errors equally
  - unbiased
- Mean Absolute Percentage Error
  - more interpretable comparison

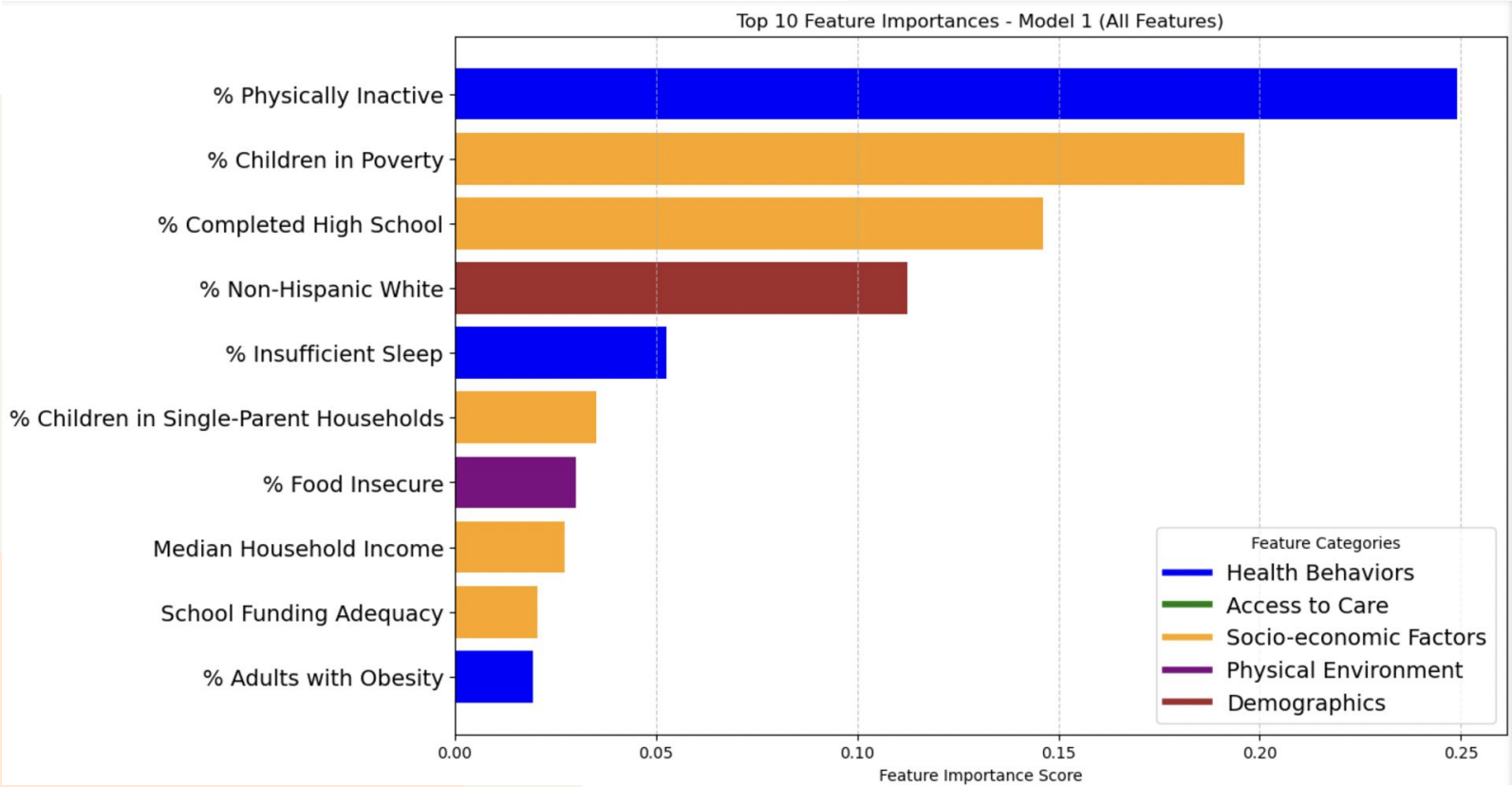
Tuned XGBoost Model	Validation Error on 80/20 split of training set	Test error after fitting to full training set
RMSE	0.51	0.49
MAE	0.39	0.37
MAPE	3.7%	3.5%

Sample Predictions:

	County	State	% Adults with Diabetes	Predicted	Residual
472	Lincoln	Arkansas	12.2	12.9	-0.7
273	Richmond City	Virginia	12.5	12.4	0.1
449	Marion	Georgia	13.3	13.0	0.3
391	Caribou	Idaho	8.4	8.1	0.3
577	Adair	Missouri	10.6	10.4	0.2



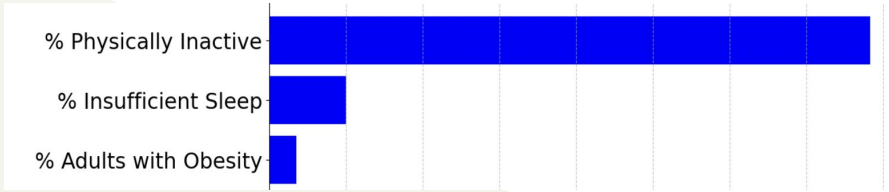
Modeling: Final Model Feature Importance



	Full Model	Health Behavior	Socio-economic	Demographic	Physical Environment	Access to care
RMSE (validation set)	0.31	0.52	0.72	1.02	1.34	1.74

Importance by Feature Group

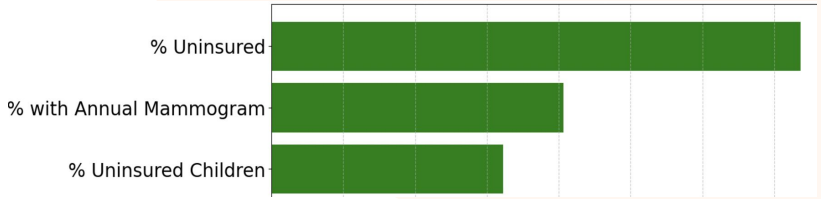
Health Behaviors



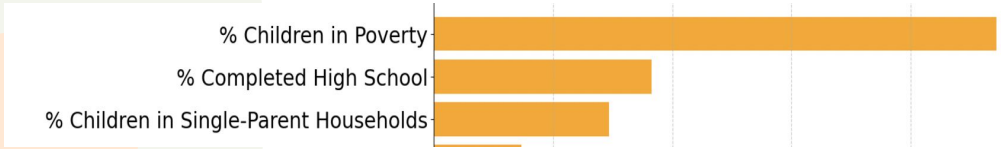
Physical environment

Demographics

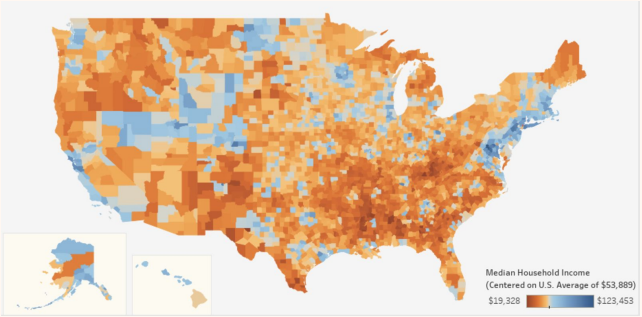
Access to care



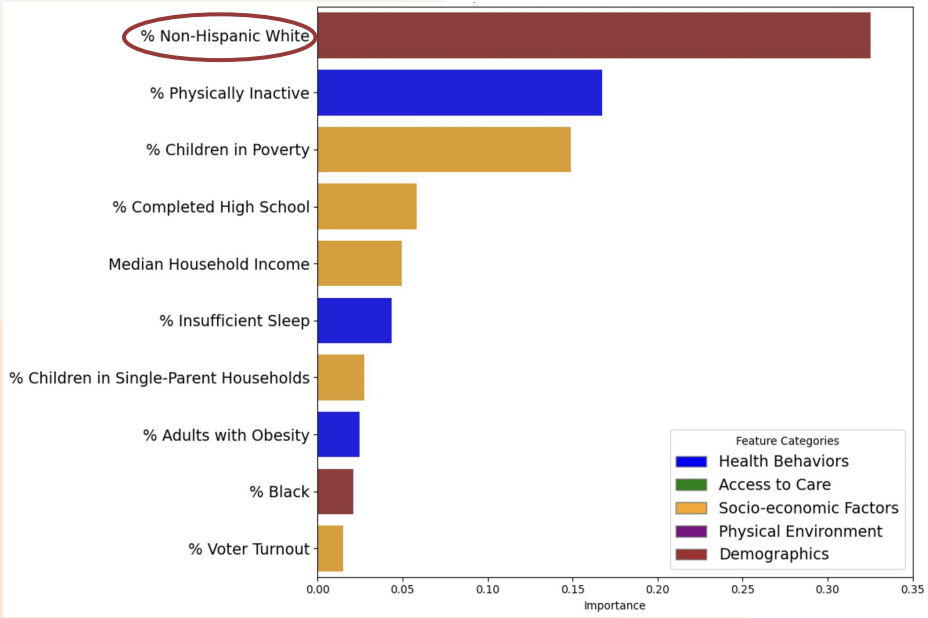
Socio-economic factors



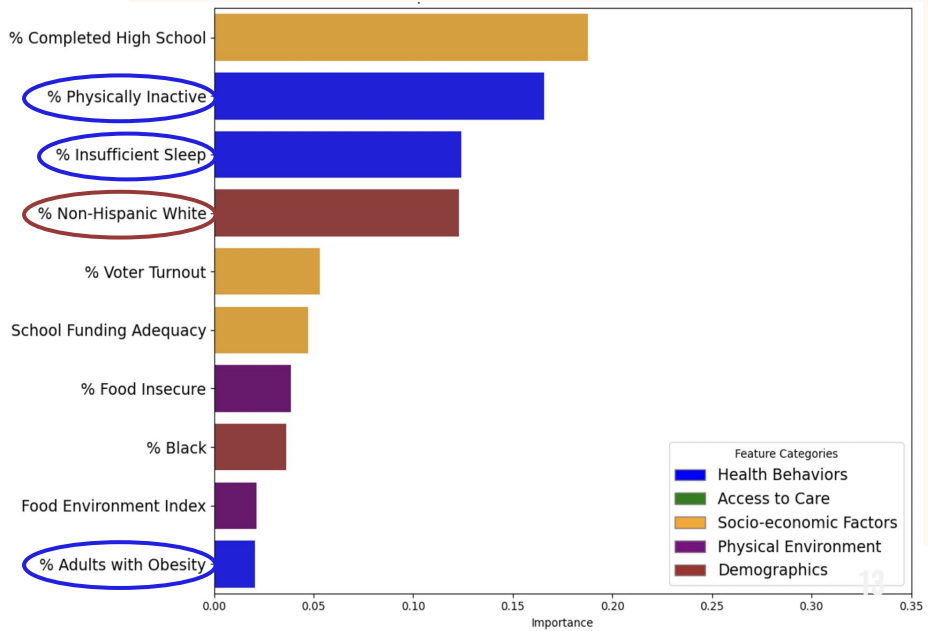
Low vs High Household Income



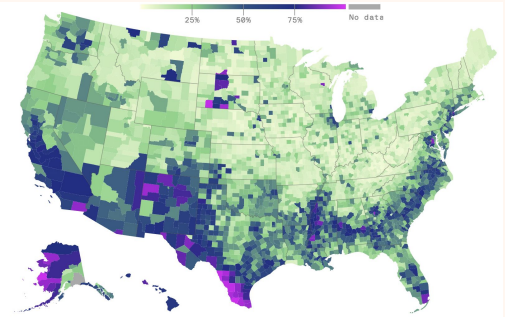
Below Median



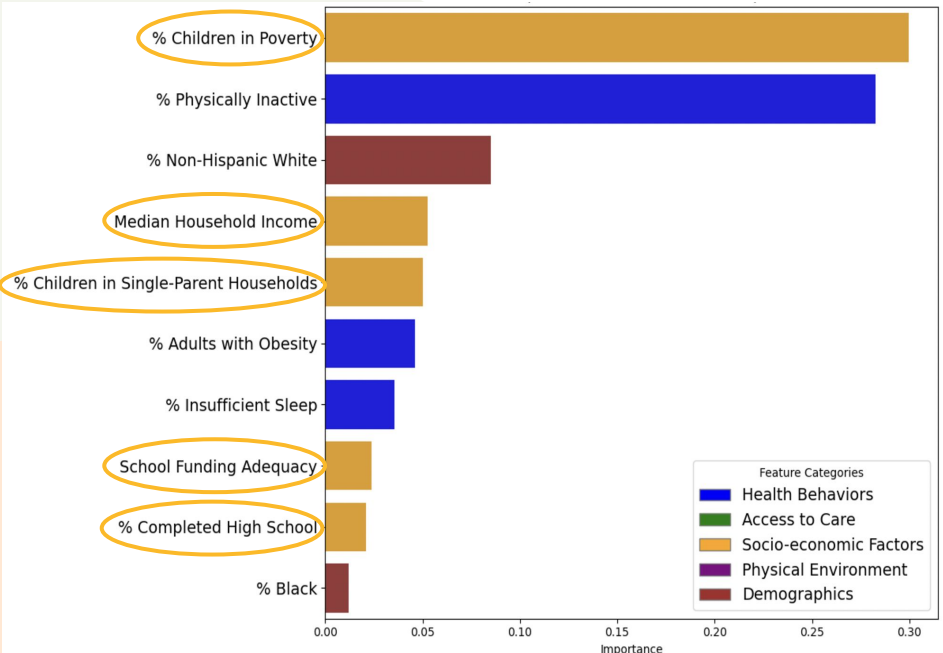
Above Median



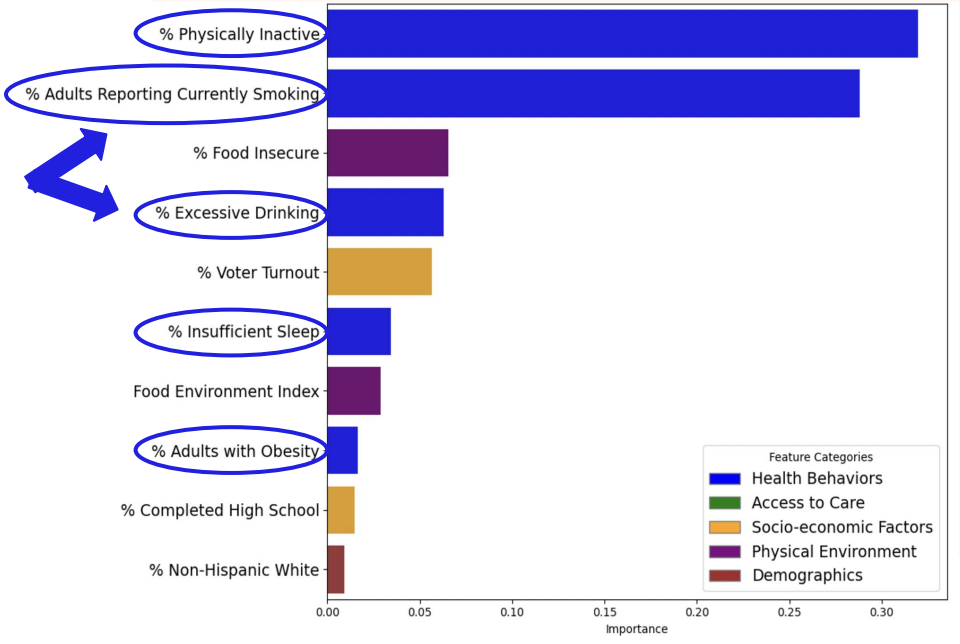
Split by % Non-Hispanic White



Below Median

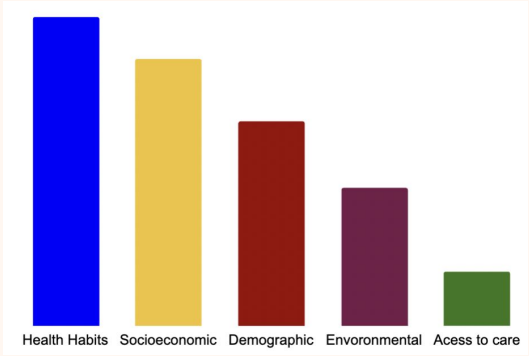


Above Median



Summary and Future Directions

- **Summary:**
  - Both **health behaviors** and **socio-economic** factors play a significant role in diabetes prevalence in the US.
  - Feature importance varies by **income** and **race**.



- **Future Directions:**
  - Focus on a specific state/geographical region
  - Inferential model
  - Further reduce features list

Acknowledgements

- **Our mentor** Bailey Forster
- **Erdös Institute**
  - Steven Gubkin
  - Alec Clott
  - Roman Holowinsky