A STORY THROUGH DATA

Predicting Health Disparity in the United States

ELINA RANKOVA

DATA SCIENTIST
BROOKLYN, NY

Presentation Agenda

- BUSINESS OBJECTIVE
- DATA UNDERSTANDING
- MODELING
- RESULTS
- LOOKING TO THE FUTURE
- CONTACT INFO

Business Objective

In recent years more companies and institutions are specializing in public health are leaning on data and technology to inform business decisions.

This project aims to identify overall health disparity across the United States by defining a Health Disparity Index to help direct business opportunities and close much needed care gaps.

Data Understanding

About the Data

CDC PLACES and SDOH data spanning 2017-2021
8287 rows from the SDOH data and 780890 from the PLACES

Includes health measure and categories for counties in the US
SDOH, Health Outcomes, Prevention, Health Risk Behaviors, Health Status, Disabilities

Columns with too much missingness are dropped
Columns not present in SDOH were dropped as well

MIsspelled columns were combined with the misspelled column dropped
`Geolocatioin`/`Geolocation` & `MeasureID`/`MeasureId`

Proper pre-split transformations applied
`State` → categorical, `LocationID` → object

Features with either duplicated or unnecessary information were dropped

Data Understanding

Feature Preview

Category
measure group

Location ID
county unique identifier

TotalPopulation
population in the county

State
two letter abbreviation

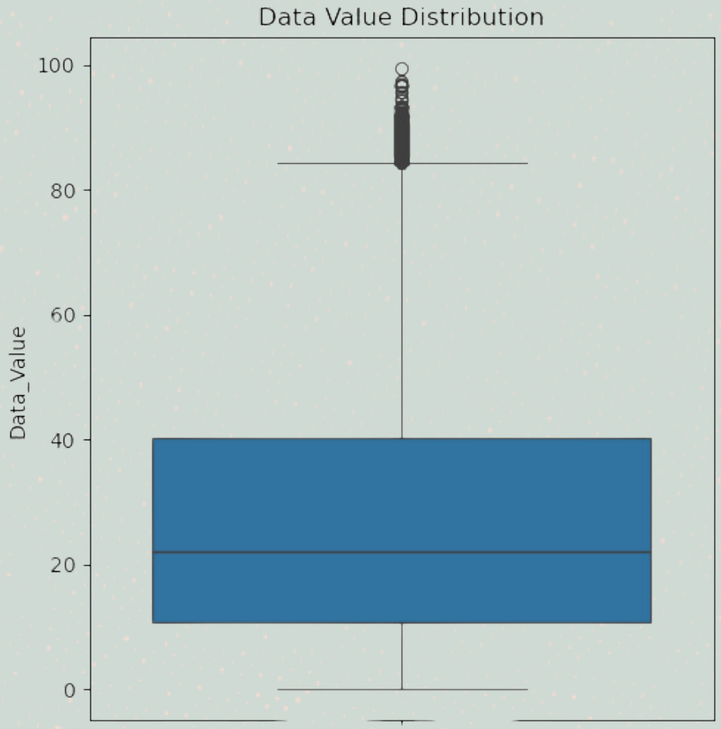
Short Answer Text
short text of measure

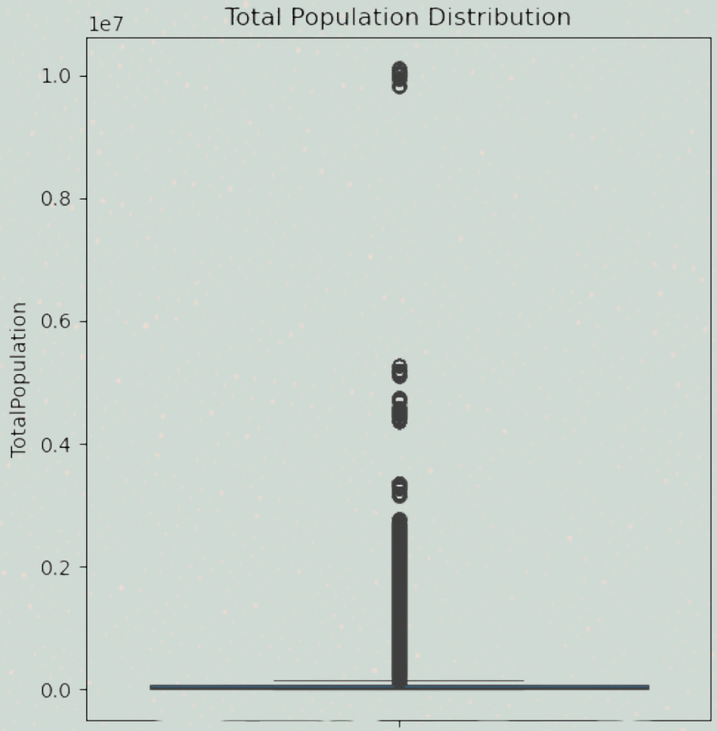
Location Name
county name

Data Value
percentage of measure depending on data value type

Data Value Type
percentage aggregation type

Data Understanding





Statistical Analysis

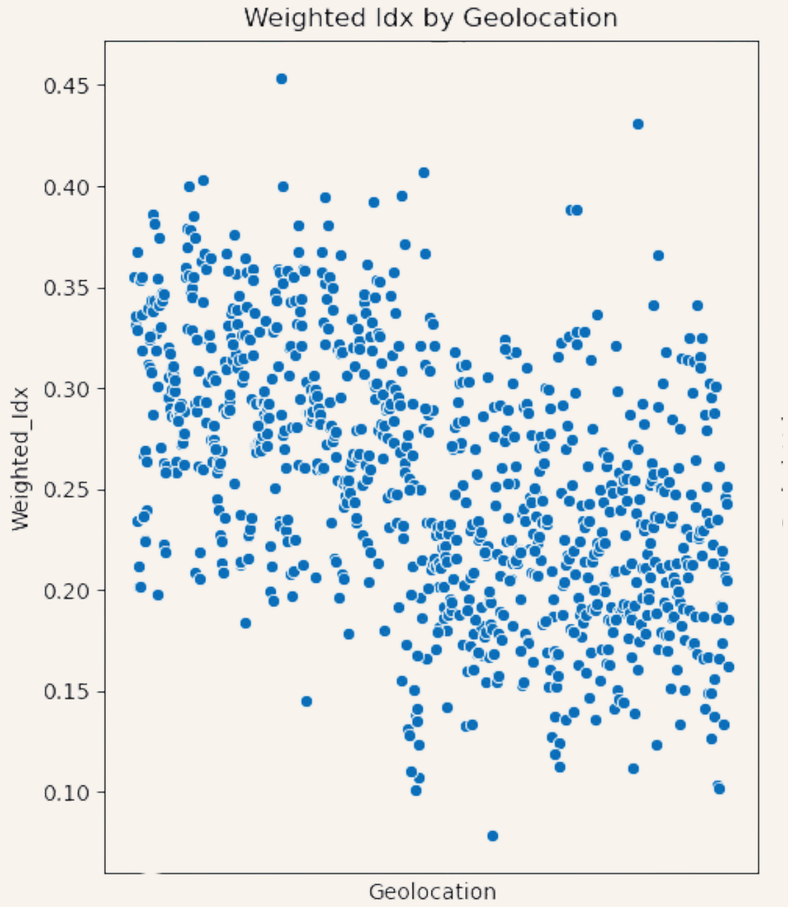
Outliers were dropped using IQR
Data_Value and TotalPopulation

Data_Value was normalized with new Scaled_Value feature created
`RobustScaler` used to keep consistent with IQR analysis

Data Understanding

Health Disparity Index

Scaled data values aggregated by Geolocation and Data Value Type and weighted by the population.

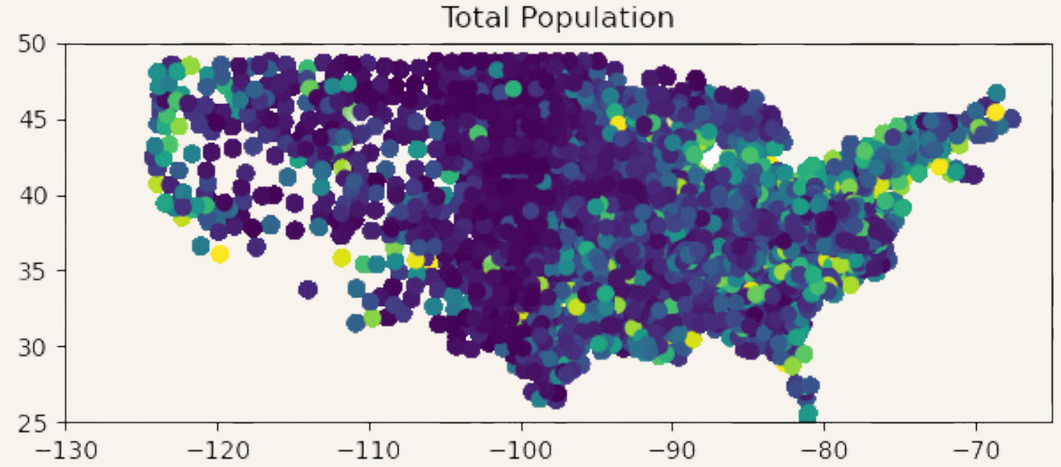


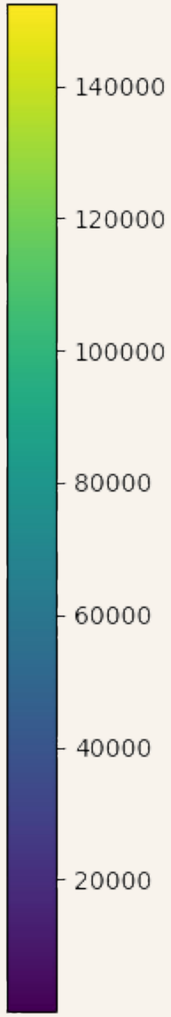
All aggregation was done on a record level so that each scaled value was aggregated based on the record's health measure and data value type.

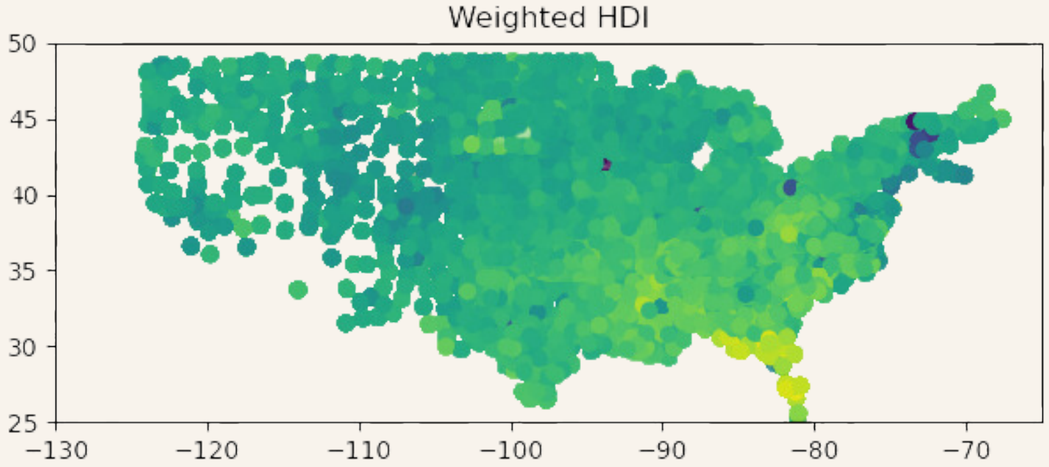
When population weight is added, we see geographical distinction

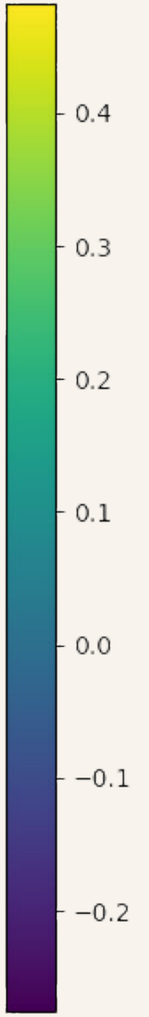
Data Understanding

Health Disparity Index









Modeling

Base

METRIC	TRAIN VALUE	TEST VALUE
RMSE	0.035	3.76E+09
R-SQUARED	0.703	-3.35E+21
MAE	0.031	9.20E+07

Cross validation and parameter tuning is conduced through a custom class offering the option of `RandomizedSearchCV` or `gp_minimize` for paramter tuning

Modeling

Lasso

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
LASSO	0.047	0.047	0.481	0.480	0.039	0.039

Modeling

Ridge

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
LASSO	0.047	0.047	0.481	0.480	0.039	0.039
RIDGE	0.035	0.037	0.703	0.679	0.031	0.033

Modeling

Ridge

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
~~LASSO~~	~~0.047~~	~~0.047~~	~~0.431~~	~~0.480~~	~~0.039~~	~~0.039~~
RIDGE	0.035	0.037	0.703	0.679	0.031	0.033

Modeling

RandomForest

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
RIDGE	0.035	0.037	0.703	0.679	0.031	0.033
RANDOM FOREST	0.049	0.0497	0.425	0.412	0.042	0.042

Modeling

RandomForest

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
RIDGE	0.036	0.036	0.680	0.680	0.033	0.033
NEURAL NETWORK						0.034

Modeling

VotingRegressor

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
RIDGE	0.035	0.037	0.703	0.679	0.031	0.033
VOTING	0.037	0.038	0.675	0.657	0.033	0.033

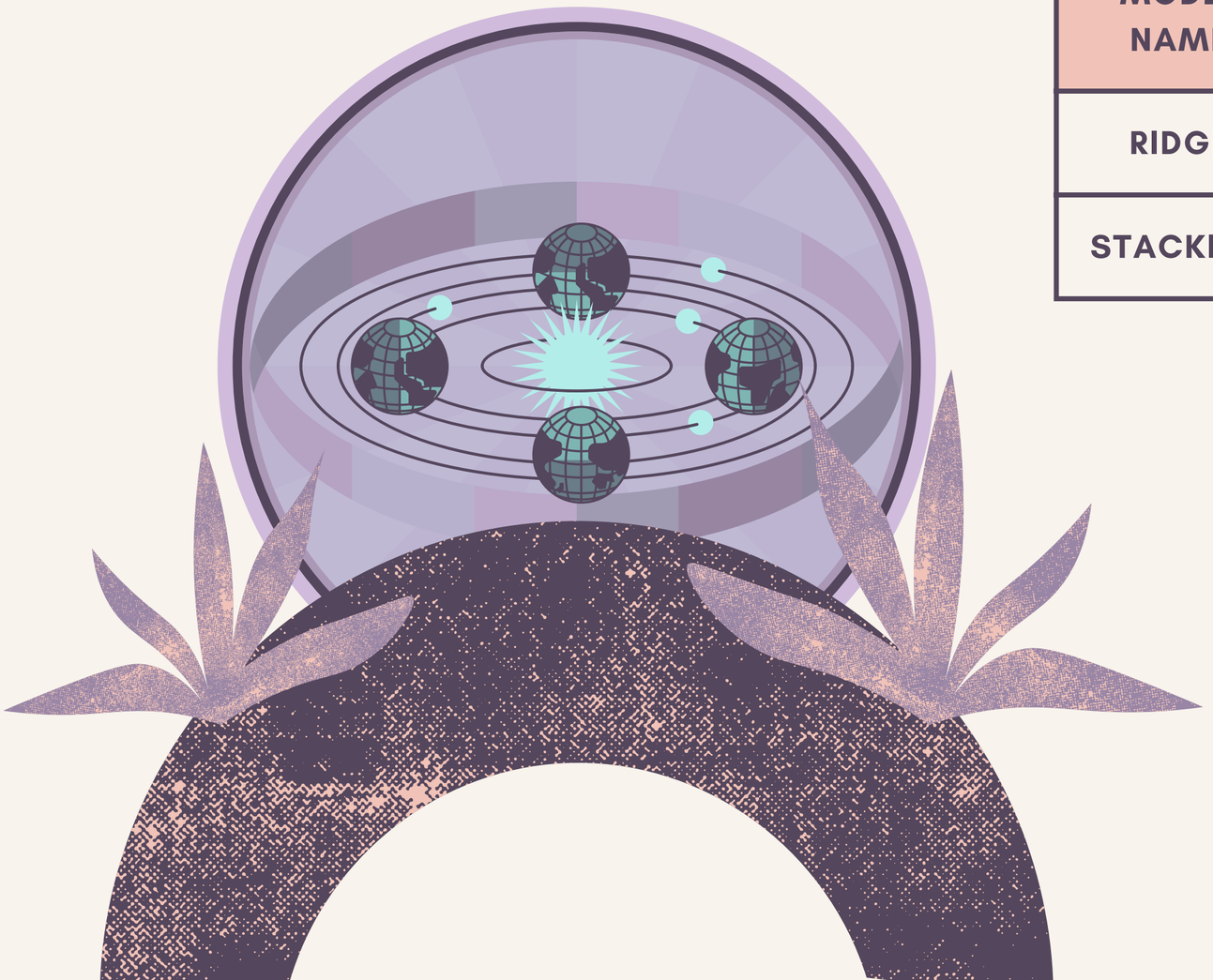
Modeling

VotingRegressor

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
RIDGE	0.035	0.037	0.703	0.679	0.031	0.033
VOTING	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.033

Modeling

StackingRegressor

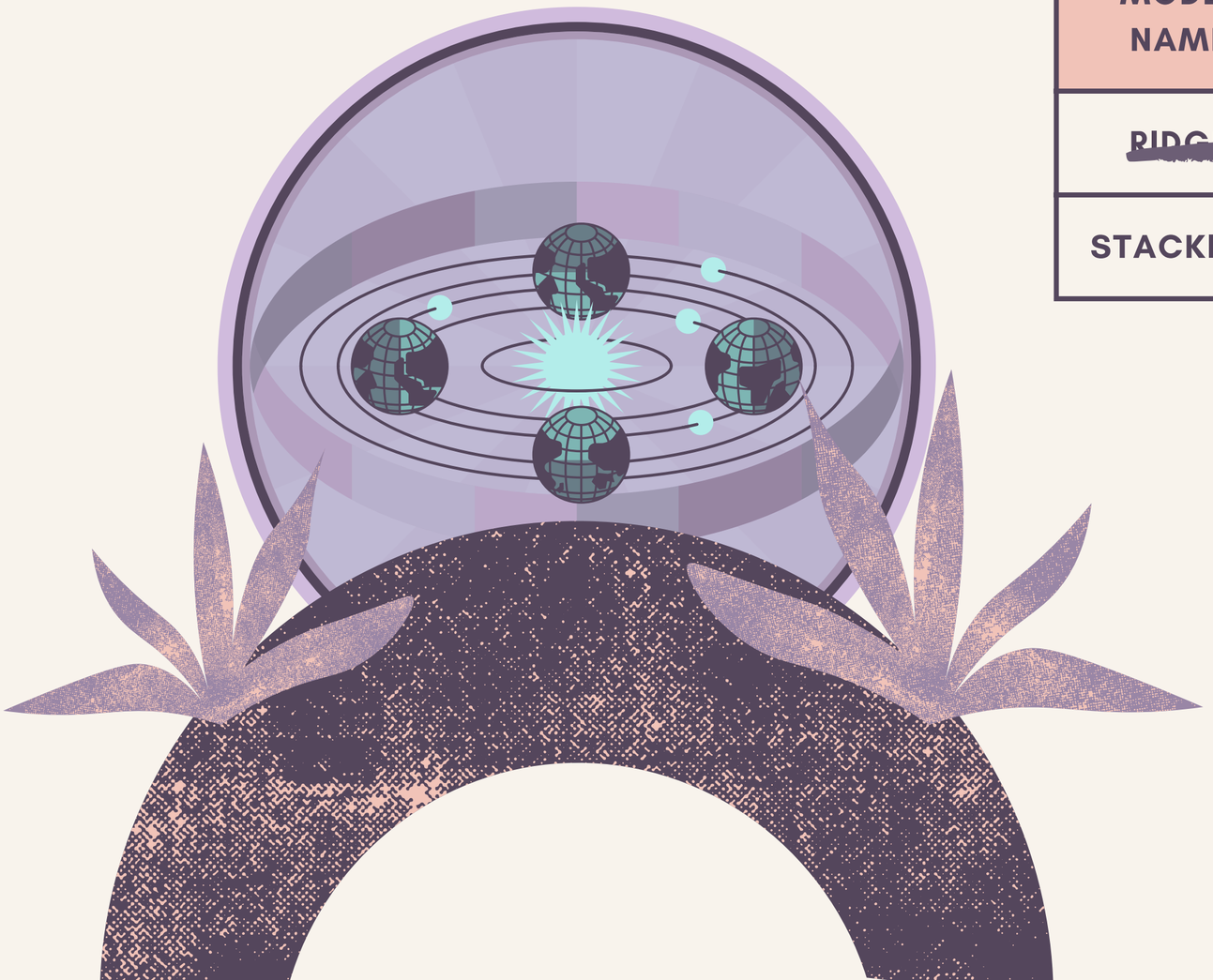


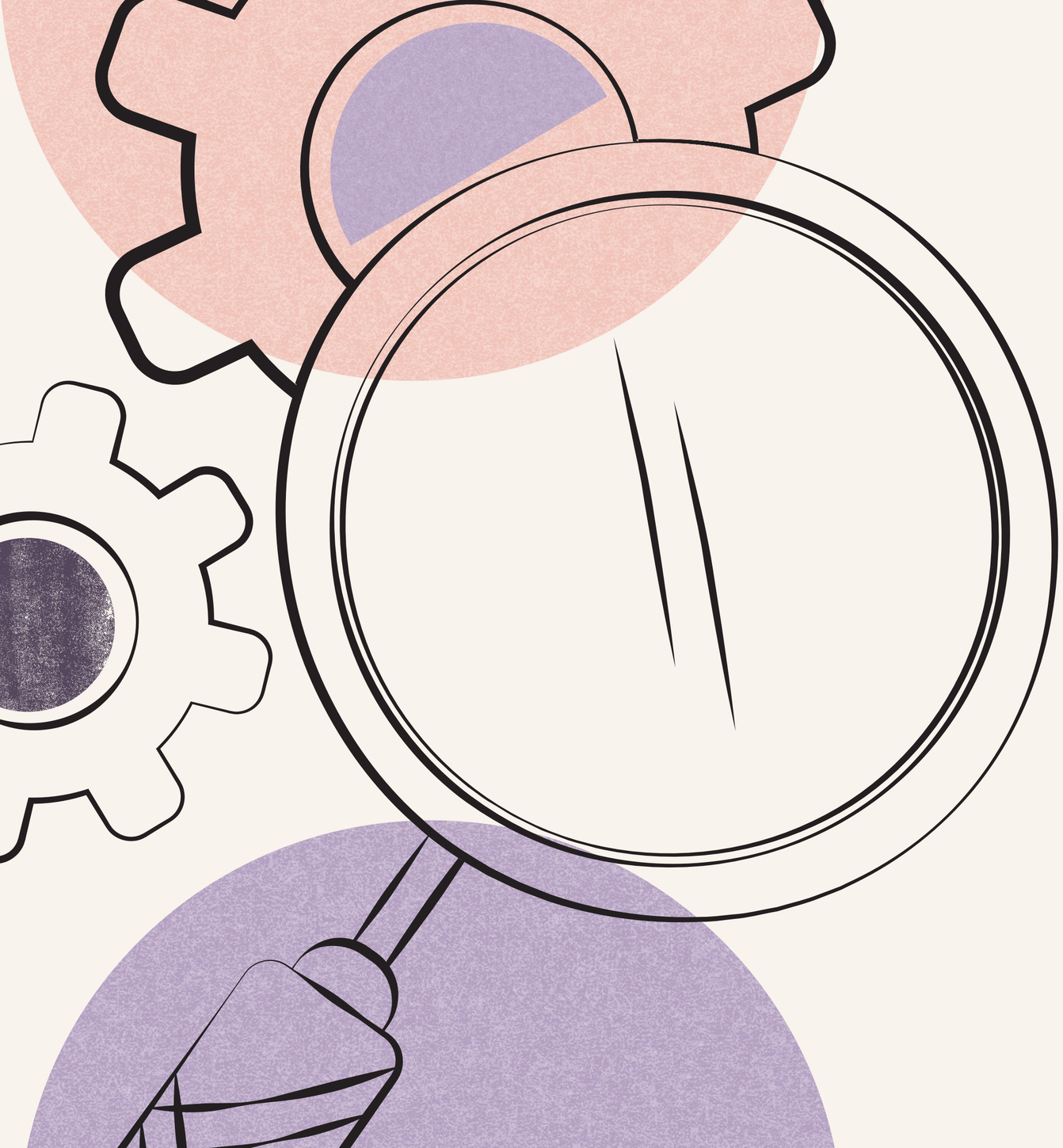
MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
RIDGE	0.035	0.037	0.703	0.679	0.031	0.033
STACKING	0.035	0.037	0.703	0.680	0.031	0.033

Modeling

StackingRegressor

MODEL NAME	TRAIN SCORE	TEST SCORE	TRAIN R2	TEST R2	TRAIN MAE	TEST MAE
RIDGE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.033
STACKING	0.035	0.037	0.703	0.680	0.031	0.033





Results

Holdout Set

Base	Final
RMSE: 3.76e+09	**RMSE:** 0.036
R-Squared: -3.35e+21	**R-Squared:** 0.698
MAE: 9.20e+07	**MAE:** 0.031

Looking Ahead

1. *Computational resources*
2. *Deeper geolocation analysis*
3. *SDOH measures by year*
4. *Upscaled deployment*

Thank you!

CONTACT WITH ANY BUSINESS INQUIRIES

ELINARANKOVA@GMAIL.COM

GITHUB | BLOG | LINKEDIN