A Regionally Generalized Machine Learning Framework Towards Census-Enabled Multi-Factor Non-Communicable Disease Analyses

Kevin Geng[1,^], Sudith Thota[1,*], Anish Kataria[2,#]
[1]Dublin High School, California, USA
[2]Princeton University, New Jersey, USA

ICAISE 2024

Introduction

In recent years, non-communicable diseases, diseases intransmissible through human interaction, have cemented themselves as a leading cause of death globally. These diseases are exacerbated by various social, economic and geographical factors within communities. Despite their resounding consequences on community health, the compilation of medical surveys is both costly and time consuming. This study uses a machine learning approach to predict the population proportion with Asthma, Cancer, Diabetes, High Blood Pressure, and Poor Mental Health within census tract populations in three states.

Data

US Environmental Justice Index (EJI)

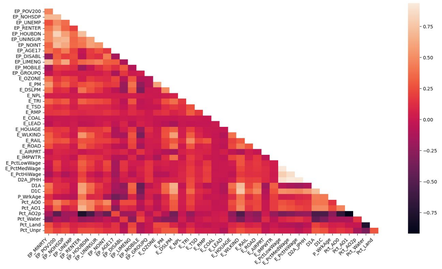
U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION

atsdr.cdc.gov/placeandhealth/eji/index

US Smart Location Database (SLD)

epa.gov/smartgrowth/smart-location-mapping

Features



Economic: 8 features
Environmental Justice: 12 features
Environment/Geographic: 8 features
Sociodemographic: 13 features

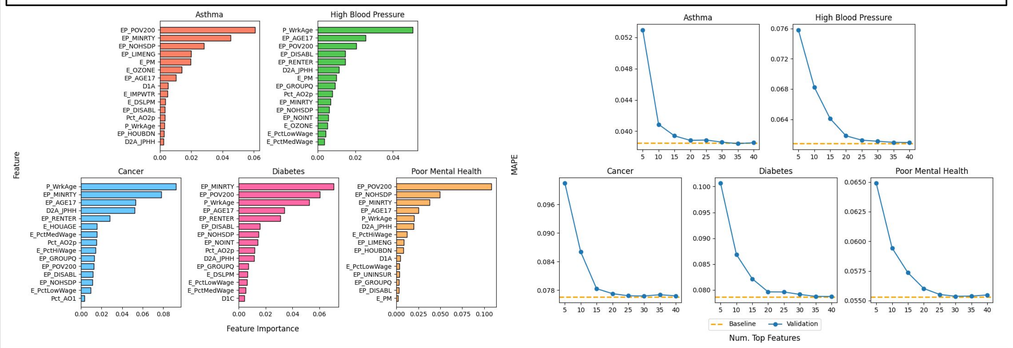
NCDs: 5 features

Comparative Model Analysis

Cross Validated 8-Model MAPE Comparison

	Cubist	Elastic Net	GBDT	k-NN	Lasso	MLP	Random Forest	Ridge
Asthma	**3.849%**	5.723%	4.472%	5.379%	6.091%	4.493%	4.526%	5.609%
Cancer	**7.656%**	9.900%	8.240%	12.025%	18.806%	7.767%	8.517%	9.869%
Diabetes	**7.866%**	10.859%	8.735%	10.635%	11.837%	8.115%	8.703%	10.699%
HBP^	**6.077%**	7.946%	6.699%	7.906%	8.323%	6.625%	6.949%	7.904%
PMC*	**5.531%**	7.408%	5.932%	7.763%	7.542%	5.600%	6.289%	7.169%

Feature Importances & Feature Subset Performance



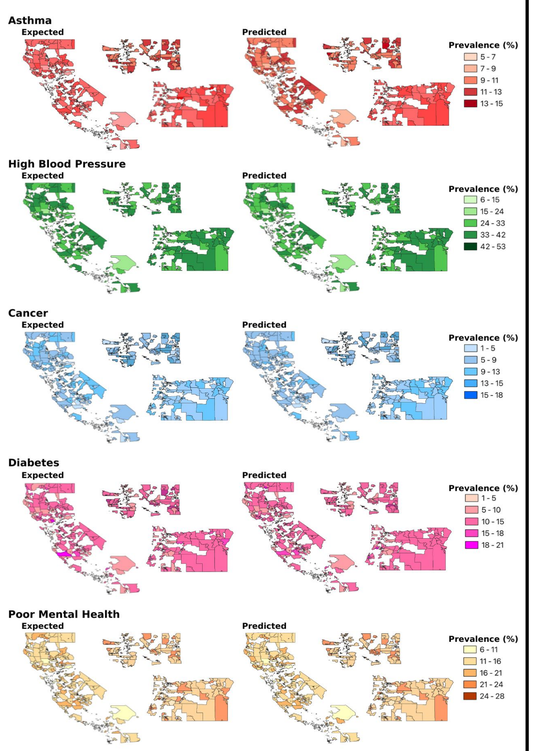
Test Set Validation (Top 20 vs. All Features

Evaluations conducted on test set, split using a 90/10 scheme (Train/CV, Test).

	All Features			Top 20 Features		
	R^2	RMSE	MAPE	R^2	RMSE	MAPE
Asthma	0.857	0.510	3.903%	0.871	0.480	3.849%
Cancer	0.826	2.086	6.094%	0.803	2.186	6.077%
Diabetes	0.915	0.545	7.690%	0.904	0.568	7.656%
HBP^	0.886	0.909	7.684%	0.863	1.017	7.866%
PMC*	0.902	0.971	5.495%	0.904	0.953	5.531%

^ High Blood Pressure * Poor Mental Health

Cubist Model Geospatial Results



Train/CV/Test Data includes 10,243 census tracts from California, Oregon, and Washington, preprocessed using both tract removal and *k-NN* feature imputation.

Feature importance was retrieved using permutation importances.

Key Findings

Through an evaluation of 8 different regression models on their predictive capabilities of disease prevalence, the Cubist model held the strongest results for every disease, achieving considerably low relative mean value errors of 3.849%, 7.656%, 7.866%, 6.077%, and 5.531% towards each of the non-communicable diseases, respectively. Out of the four input classifications split from 41 input features- Sociodemographic, Economic, Environment/Geographic, and Environmental Justice- Sociodemographic features consistently were the highest contributors. A lightweight model with comparable accuracy to the full feature set (≤0.1% difference) was found to require only the 20 most important features from every Cubist model. This study provides insight towards machine learning models' capabilities in enhancing the United States' holistic understanding of domestic epidemiology determinants.

Contact

^ kevin@ncdcare.org
* sudith@ncdcare.org
# anish@ncdcare.org