### Statistical Machine Learning Techniques for Predicting Lung Cancer Prevalence

#### Adekunle Ajiboye, Fekadu Bayisa, Zhuoran Wang

#### Data Science Bootcamp Trainees

Erdős Institute

April 21, 2025

#### Introduction

-			D	
	loto '	CIODCO.	Kootcompo I	L KOUDOOC
	ald.	менсе	DODICATIO	I l'allees

æ

Lung cancer poses a major public health concern

- 一司

Lung cancer poses a major public health concern

**Motivation:** To learn the underlying factors contributing to lung cancer prevalence

3/17

- 個 ト - ヨ ト - ヨ ト

Lung cancer poses a major public health concern

**Motivation:** To learn the underlying factors contributing to lung cancer prevalence

**Goal:** To identify key features associated with lung cancer incidence

3

イロト イポト イヨト イヨト

Table: General overview of study features, with the underlying population being individuals aged 18 and over

Group	Features	Source
Demographics	% Male, % Female, % Black, % White, % Hispanic, %	US Census ACS
	Age ≥65	
Behavioral	Prevalence of Smoking, Binge Drinking, and Obesity	CDC PLACES
Socioeconomic	% Below Poverty, Social Deprivation Index (SDI)	US Census, Graham Center
Environmental	PM2.5 Air Quality ( $\mu g/m^3$ )	EPA Downscaler

- 3

4/17

人口区 人間区 人居区 人居区

Table: General overview of study features, with the underlying population being individuals aged 18 and over

Group	Features	Source
Demographics	% Male, % Female, % Black, % White, % Hispanic, %	US Census ACS
	Age ≥65	
Behavioral	Prevalence of Smoking, Binge Drinking, and Obesity	CDC PLACES
Socioeconomic	% Below Poverty, Social Deprivation Index (SDI)	US Census, Graham Center
Environmental	PM2.5 Air Quality ( $\mu g/m^3$ )	EPA Downscaler

**Feature engineering:** Age-based 6 features are created from the existing gender and race related features

< 回 ト く ヨ ト く ヨ ト

Table: General overview of study features, with the underlying population being individuals aged 18 and over

Group	Features	Source
Demographics	% Male, % Female, % Black, % White, % Hispanic, %	US Census ACS
	Age ≥65	
Behavioral	Prevalence of Smoking, Binge Drinking, and Obesity	CDC PLACES
Socioeconomic	% Below Poverty, Social Deprivation Index (SDI)	US Census, Graham Center
Environmental	PM2.5 Air Quality ( $\mu g/m^3$ )	EPA Downscaler

**Feature engineering:** Age-based 6 features are created from the existing gender and race related features

Unobserved or missing values are removed, as there is no clear rationale for imputation

#### Statistical machine learning techniques

Data Science Bootcamp (Trainees)

**Model formulation:** Let  $Y_i$  denote the number of lung cancer cases in a population of size  $N_i$ 

**Model formulation:** Let  $Y_i$  denote the number of lung cancer cases in a population of size  $N_i$ 

We model  $Y_i$  using Poisson distribution given by

$$P(Y_i = y_i) = \frac{(N_i \lambda_i)^{y_i} e^{-N_i \lambda_i}}{y_i!}, \ i = 1, 2, \cdots, n$$

Poisson generalized linear model: We model the incidence rate  $\lambda_i$  as  $\lambda_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ 

Poisson generalized linear model: We model the incidence rate  $\lambda_i$  as  $\lambda_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ 

Using the data  $y_1, y_2, \ldots, y_n$ , the objective function with elastic net regularization can be given by

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left[ y_i \log N_i + y_i \mathbf{x}_i^{\top} \boldsymbol{\beta} - N_i e^{\mathbf{x}_i^{\top} \boldsymbol{\beta}} - \log(y_i!) \right] + \alpha \left[ \gamma \|\boldsymbol{\beta}\|_1 + \frac{1-\gamma}{2} \|\boldsymbol{\beta}\|_2^2 \right], \quad \alpha \ge 0, \ \gamma \ge 0$$

Poisson generalized linear model: We model the incidence rate  $\lambda_i$  as  $\lambda_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ 

Using the data  $y_1, y_2, \ldots, y_n$ , the objective function with elastic net regularization can be given by

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left[ y_i \log N_i + y_i \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} - N_i e^{\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}} - \log(y_i!) \right] + \alpha \left[ \gamma \|\boldsymbol{\beta}\|_1 + \frac{1-\gamma}{2} \|\boldsymbol{\beta}\|_2^2 \right], \quad \alpha \ge 0, \ \gamma \ge 0$$

We use this regularization approach to select features that are associated with the lung cancer incidence rate

**XGBoost with Poisson Loss Function:** We model the rate  $\lambda_i$  using tree at *t*-th iteration as  $\log(\lambda_i) = f(\mathbf{x}_i, \eta_t)$ , which is parameterized by  $\eta_t$ 

< 4 → <

**XGBoost with Poisson Loss Function:** We model the rate  $\lambda_i$  using tree at *t*-th iteration as  $\log(\lambda_i) = f(\mathbf{x}_i, \eta_t)$ , which is parameterized by  $\eta_t$ 

The objective function in XGBoost with Poisson loss function can be given by

$$L(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{x}, \eta_t) = \sum_{i=1}^n \left( N_i e^{f(\mathbf{x}_i, \eta_t)} - y_i f(\mathbf{x}_i, \eta_t) \right)$$

**XGBoost with Poisson Loss Function:** We model the rate  $\lambda_i$  using tree at *t*-th iteration as  $\log(\lambda_i) = f(\mathbf{x}_i, \eta_t)$ , which is parameterized by  $\eta_t$ 

The objective function in XGBoost with Poisson loss function can be given by

$$L(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{x}, \eta_t) = \sum_{i=1}^n \left( N_i e^{f(\mathbf{x}_i, \eta_t)} - y_i f(\mathbf{x}_i, \eta_t) \right)$$

We use this approach to assess the importance of the selected variables in lung cancer incidence

#### Results of the study

Data Science Bootcamp (Trainees)

Image: A matrix of the second seco

3 ) 3

#### **Data splitting:** Training data (70%) and testing data (30%)

Data Science	Bootcamp (	Trainees
--------------	------------	----------

- ∢ /च 🕨 🕨

э

#### Hyperparameter selection:

Five-fold cross-validation on training data

#### Hyperparameter selection:

Five-fold cross-validation on training data

**Evaluation metric:** Mean Absolute Error (MAE)

#### Hyperparameter selection:

Five-fold cross-validation on training data

Evaluation metric: Mean Absolute Error (MAE)

Poisson GLM: Hyperparameter Tuning via Cross-Validation



3 K K 3 K April 21, 2025

#### Hyperparameter selection:

Five-fold cross-validation on training data

**Evaluation metric:** Mean Absolute Error (MAE)

#### Elastic Net: Selected features





#### Hyperparameter selection:

Five-fold cross-validation on training data

**Evaluation metric:** Mean Absolute Error (MAE)

#### Elastic Net: Selected features





Table 1: Selected Features Using Elastic Net Regularization

Features	Parameter estimates
Pct_BelowPoverty_18andOver	-1.339
Pct_Population_Male_65andOver	2.797
Pct_Population_Female_65andOver	2.743
Pct_Black_Female_65andOver	2.487
Pct_White_65andOver	0.136
Pct_White_Male_65andOver	-0.484
Pct_Hisp_65andOver	-0.885
Pct_Hisp_Female_65andOver	-1.819
BINGE_CrudePrev	-0.017
CSMOKING_CrudePrev	0.125
OBESITY_CrudePrev	0.039
Median_Household_Income	-0.087
ZCTA_pm2_5	-0.069
sdi_score	0.034
Pct_White_Male_Between18and65	1.086
Pct_White_Female_Between18and65	1.296
Pct_Black_Female_Between18and65	1.434
Pct_Hisp_Male_Between18and65	-0.328
Pct_Hisp_Female_Between18and65	-1.263

**Learning rate tuning:** It is tuned using cross-validation

**Learning rate tuning:** It is tuned using cross-validation

**Evaluation metric:** Mean Absolute Error (MAE)

**Learning rate tuning:** It is tuned using cross-validation

**Evaluation metric:** Mean Absolute Error (MAE)



**Learning rate tuning:** It is tuned using cross-validation

**Evaluation metric:** Mean Absolute Error (MAE)

**Gain index:** Feature with high gain improves the model performance



**Learning rate tuning:** It is tuned using cross-validation

**Evaluation metric:** Mean Absolute Error (MAE)

**Gain index:** Feature with high gain improves the model performance



Table 2: Gain-based Feature Importance

Feature	Importance
Median_Household_Income	26.08
ZCTA_pm2_5	6.787
OBESITY_CrudePrev	5.602
Pct_Population_Female_65andOver	4.88
Pct_Hisp_Female_Between18and65	4.859
CSMOKING_CrudePrev	2.414
BINGE_CrudePrev	1.775
Pct_Population_Male_65andOver	1.69
Pct_Hisp_Female_65andOver	1.47
sdi_score	1.31
Pct_White_Male_Between18and65	1.205
Pct_White_65andOver	1.193
Pct_White_Male_65andOver	1.19
Pct_Hisp_65andOver	1.166
Pct_Black_Female_Between18and65	1.122
Pct_White_Female_Between18and65	1.108
Pct_Black_Female_65andOver	1.061
Pct_Hisp_Male_Between18and65	0.992
Pct BelowPoverty 18andOver	0.399

# Poisson generalized linear model



Data Science Bootcamp (Trainees)

< ■> < ■> ■ April 21, 2025

# Poisson generalized linear model

Mean absolute error: 6.313



# Poisson generalized linear model

Mean absolute error: 6.313

XGBoost



(b) (a) (E) (b)

# Poisson generalized linear model

Mean absolute error: 6.313

XGBoost





Erdős Institute

# Poisson generalized linear model

Mean absolute error: 6.313

XGBoost

Mean absolute error: 5.963











Data Science Bootcamp (Trainees)

Erdős Institute

April 21, 2025

14 / 17

2

In general, we can conlude that older Black and White seniors are positively associated with higher lung cancer counts

In general, we can conlude that older Black and White seniors are positively associated with higher lung cancer counts

Poverty and Hispanic populations show negative associations

In general, we can conlude that older Black and White seniors are positively associated with higher lung cancer counts

Poverty and Hispanic populations show negative associations

Smoking is positively associated while Income and PM2.5 are modestly negatively associated

Obesity and Smoking are also significant predictors

Obesity and Smoking are also significant predictors

Age and Race demographics have moderate importance

Obesity and Smoking are also significant predictors

Age and Race demographics have moderate importance

XGBoost improved prediction accuracy by 5.54% compared to the Poisson Generalized linear model

That concludes our presentation

Thank you for your attention!

э