Predicting Cancer Mortality Rate

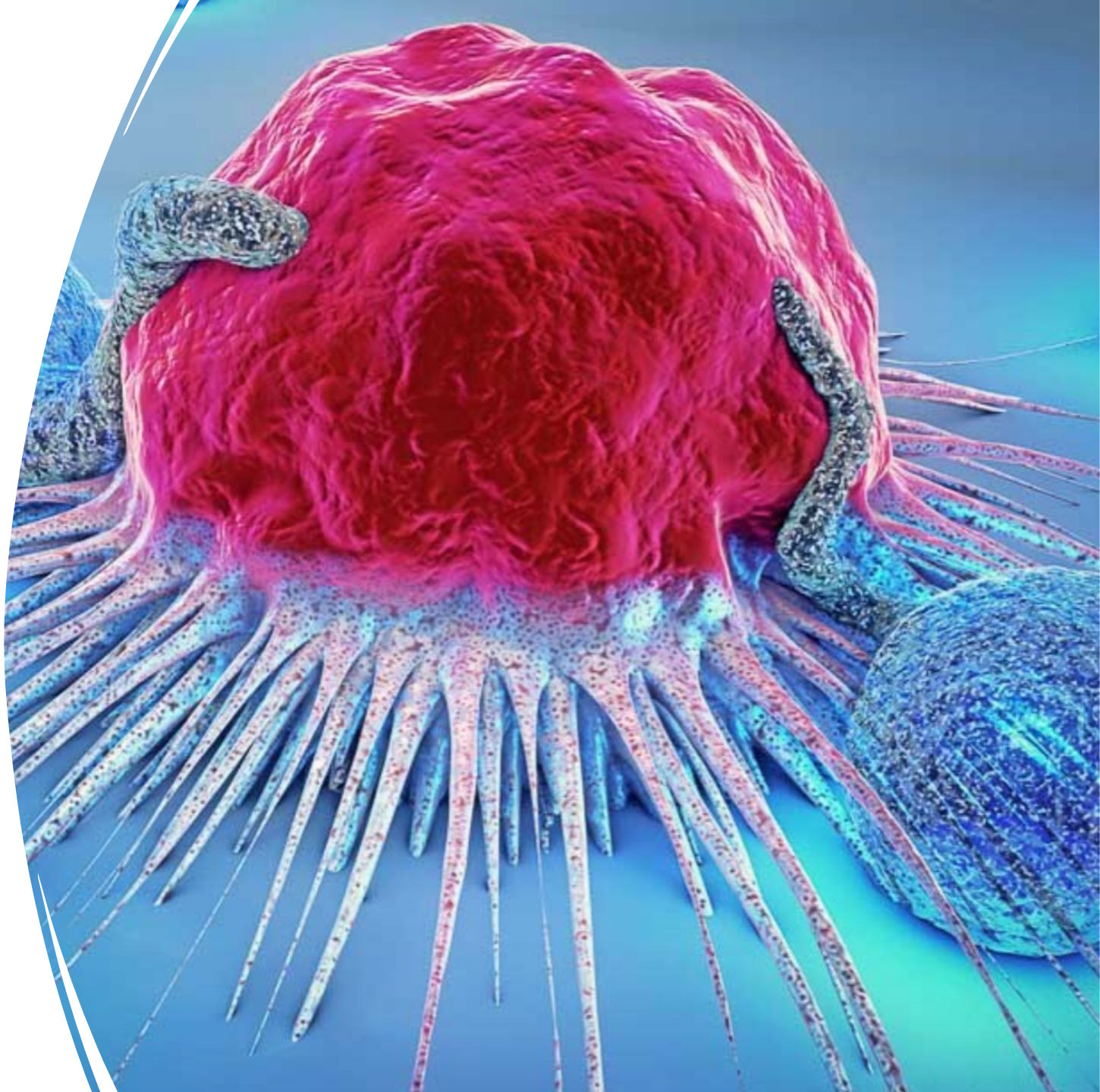
Linah Rusere

Mentor: Kenneth Gil-Pasquel

Springboard

Problem Statement

- Cancer is the second leading cause of death, after heart disease. Can we predict regional cancer mortality rates by analyzing the socioeconomic factors?



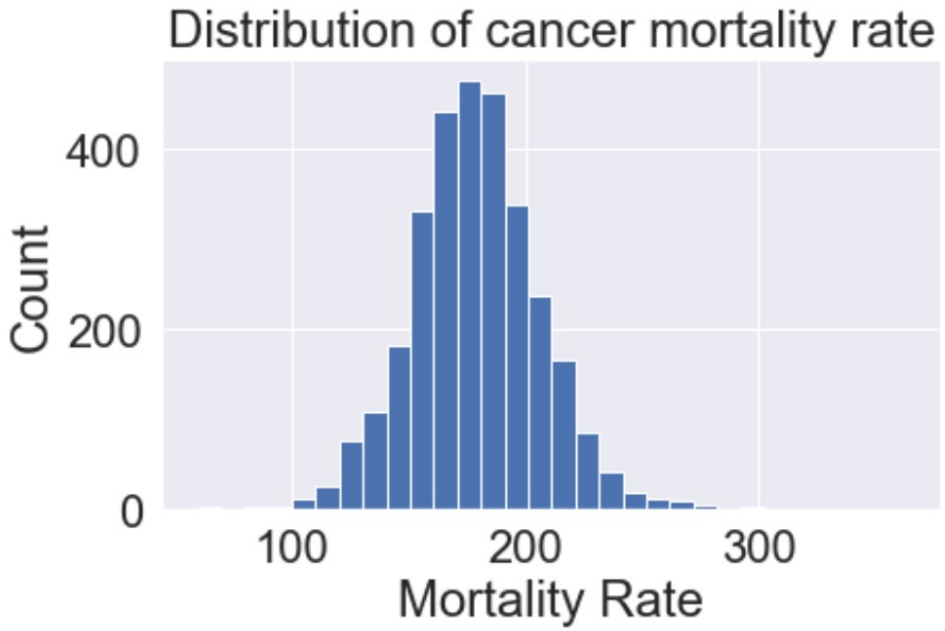
Stakeholders

- The General Public
- Healthcare Providers
- Policy Makers

Data Sources

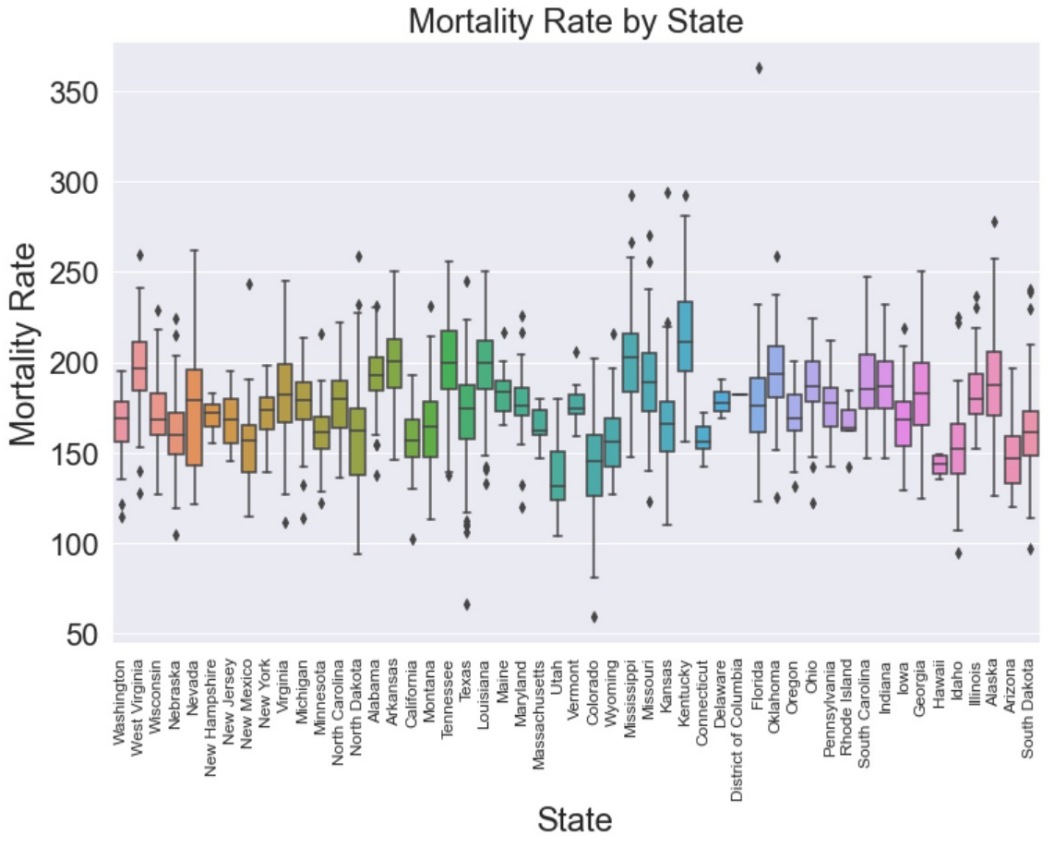
- Data.world
  - The dataset was aggregated from a number of sources including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov.

Distribution of cancer mortality rate

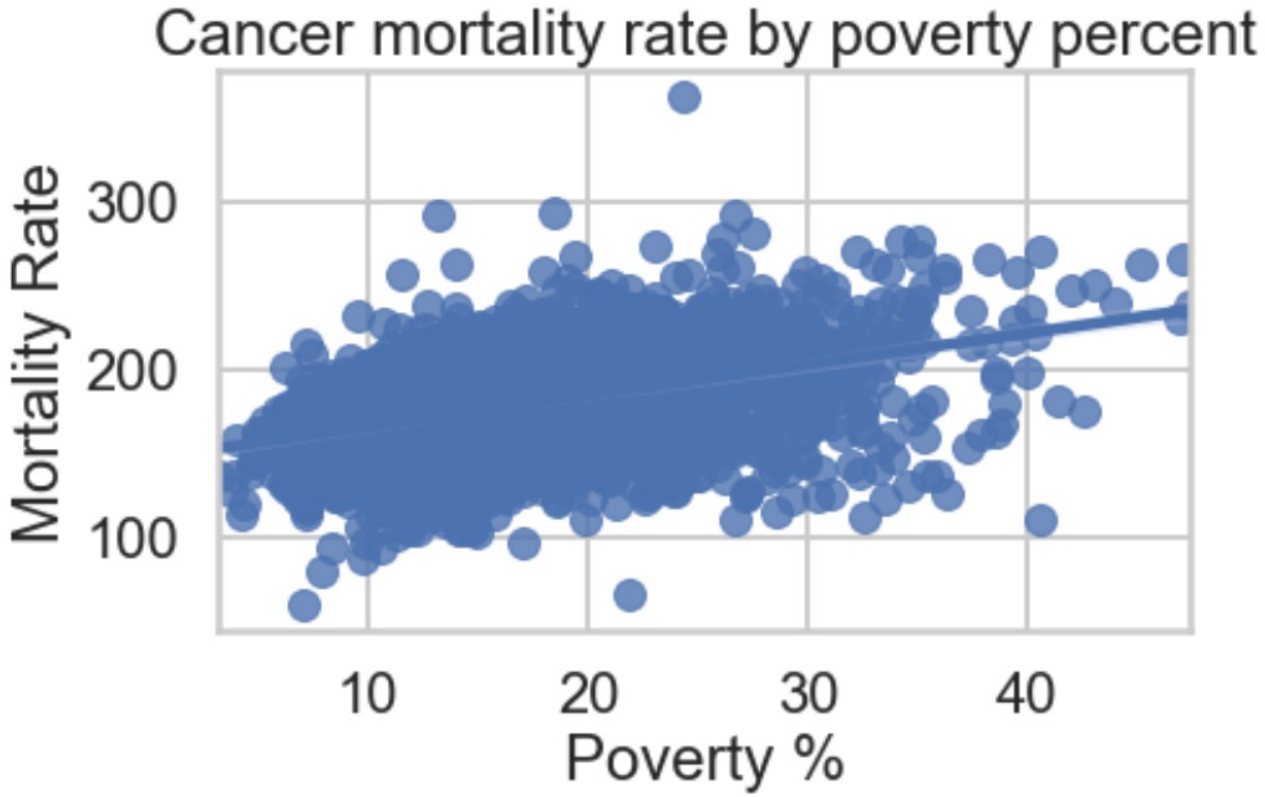


The average cancer mortality rate per capita (100,000) is normally distributed

Average mortality rate by state

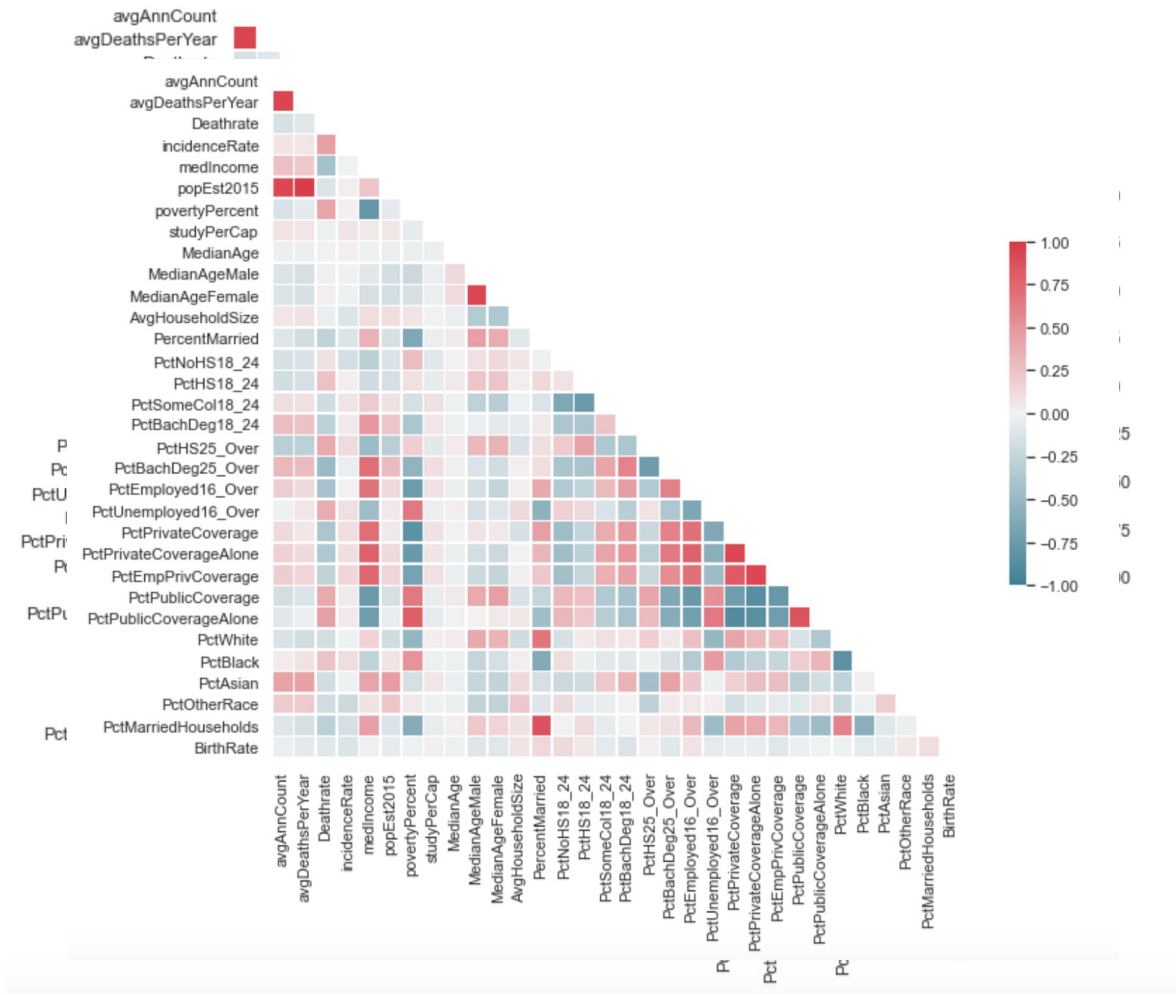


Cancer mortality rate vs poverty percentage



- Poorer cancer patients die at a higher rate than those who are wealthier.

Correlation of data variables



- There are several highly correlated independent variables
- Cancer mortality rate is not strongly correlated with any variables

Training and Testing Model Metrics

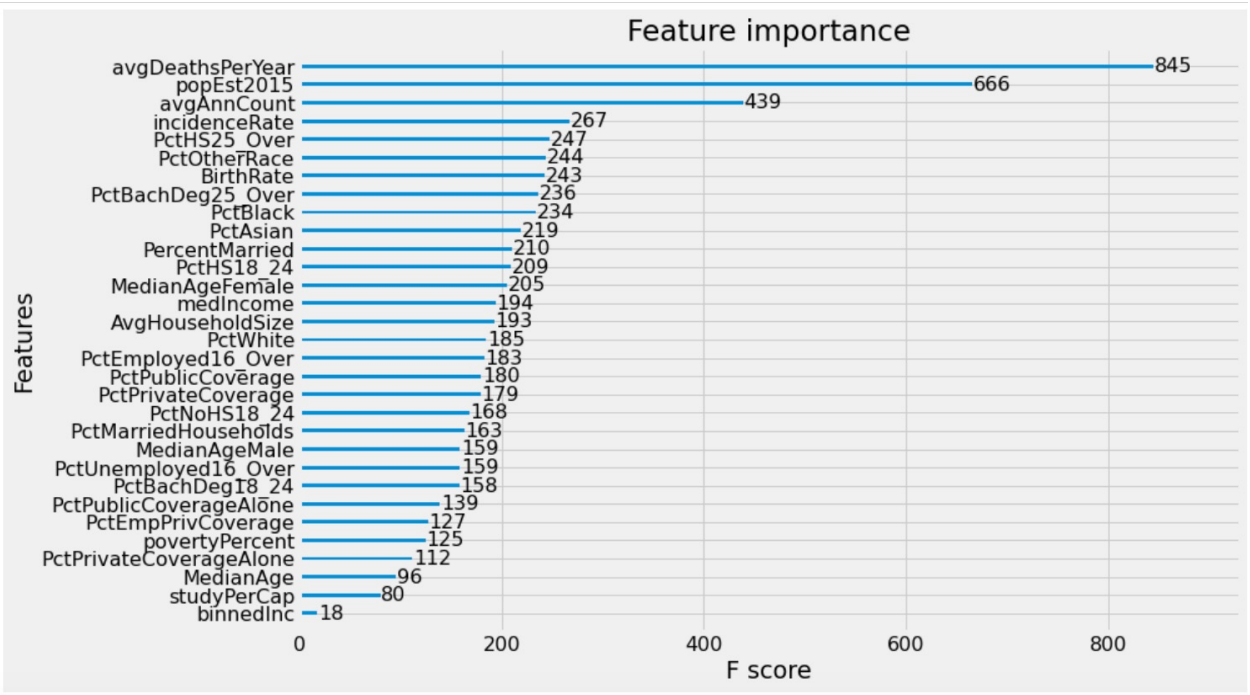
A.

TRAIN	Linear Regression	Ridge	Lasso	ElasticNet	XGB	Random Forest
R^2	0.52	0.52	0.52	0.52	0.68	0.54
MSE	342.57	356.46	356.06	356.06	237.02	348
RMSE	18.51	18.87	18.87	18.87	15.40	18.65

B.

TEST	Linear Regression	Ridge	Lasso	ElasticNet	XGB	Random Forest
R^2	0.46	0.46	0.47	0.47	0.70	0.53
MSE	438.44	436.53	432.26	432.26	245.39	378.45
RMSE	20.93	21.37	20.79	20.79	15.66	19.45
MAE	15.66	15.64	15.57	15.57	11.08	14.33

XGB Feature Importance



Most Important features

- Average Deaths caused by cancer Per Year
- Population
- Cancer Incident rate

XGBoost Best Model Metrics

XGB	R^2	MSE	RMSE	MAE
train	0.68	237.02	15.40	11.35
test	0.70	245.39	15.66	11.08

- Performance metrics of the XGB regression model.

Conclusion

- The optimized linear models revealed a weak linear relationship between the dependent and explanatory variables.
- XGBoost showed the best predictive power compared to the other models.

- Predicting cancer mortality rate using a basic linear models of socioeconomical variables is not as accurate as using the non-linear tree-based regression models.