Predicting Cancer Mortality Rate

Problem Statement:

Cancer was the second leading cause of death, after heart disease, in the United States in 2020. Predicting cancer mortality rates by region and identifying important socioeconomic factors is important for policy makers and healthcare industries. The goal for this project is to analyze the effect of socioeconomic status on cancer mortality rate on the county level. This analysis includes data from clinicaltrials.gov, cancer.gov & census.gov to examine cancer trials, mortality, incidence and demographics between 2010 and 2016. The general hypothesis is that poorer and underprivileged regions would have higher per capita cancer incidence and death rates.

Data Cleaning

The complete dataset was obtained from data.world. The data was aggregated from a number of sources including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov. The dataset contains cancer deathrate data for 3047 counties in the United States with 32 explanatory variables. Three columns had missing values. Column 'PctSomeCol18_24' was dropped since it had more than 70% missing values. The other two columns were imputed with the median values in the data preprocess pipeline after splitting the data into train and test sets. Only two of the columns in the dataset were categorical: 'binnedInc' and 'Geography'. The 'binnedInc' column containing binned median income per capita was encoded using ordinal encoding in the preprocess pipeline. The 'Geography' column which contained the county names and corresponding states. The state names for each entry were extracted and States data was used to analyze the data by state.

Exploratory Data Analysis

The average cancer mortality rate per capita (100,000) is normally distributed ranging from 60 to 360 and it varied significantly by state.

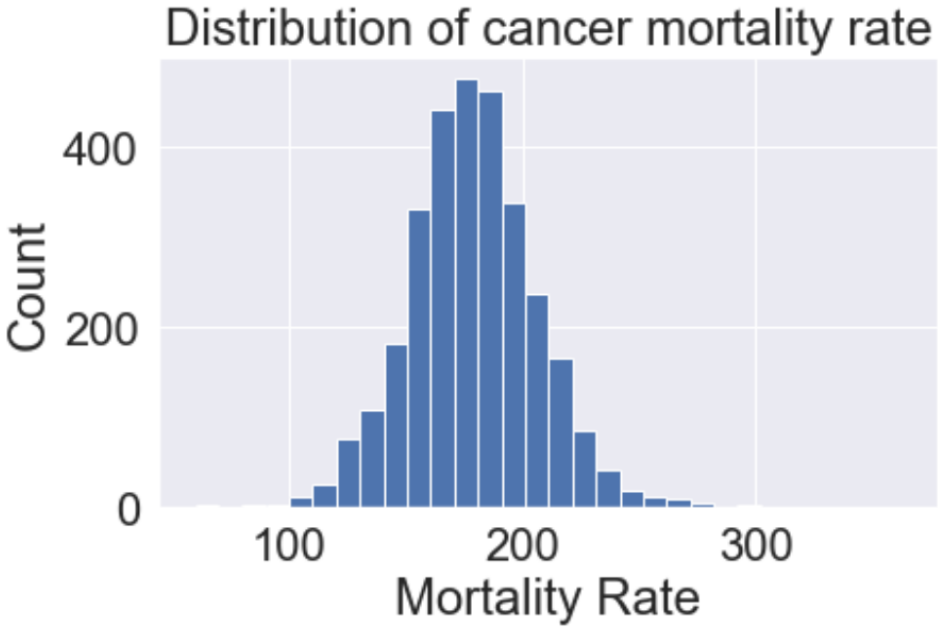

Fig 1. Distribution of cancer mortality rate

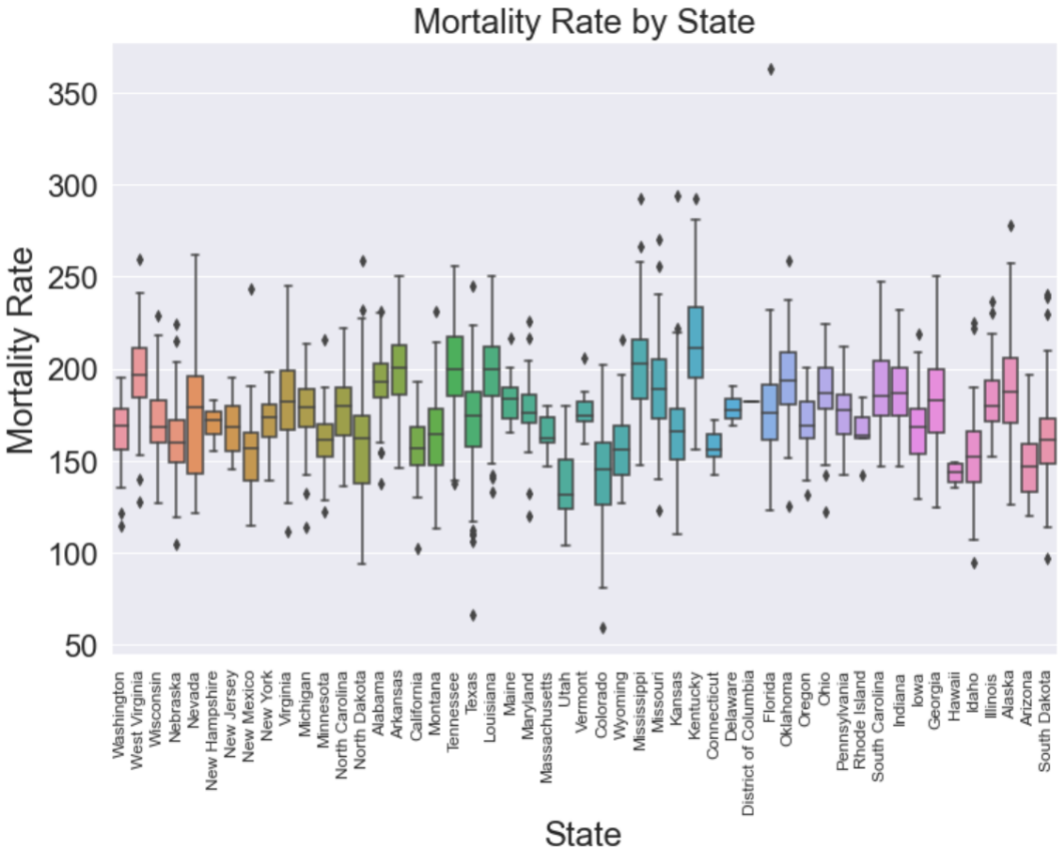

Fig 2. Cancer mortality rate by county

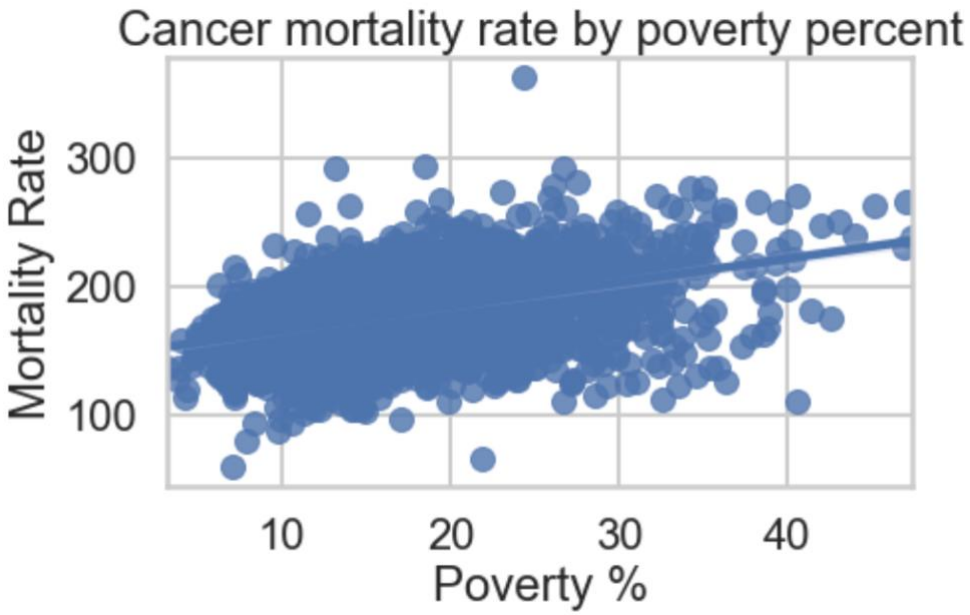

Fig. 3. Cancer mortality rate by poverty percentage.

There are several highly correlated independent variables. There are several highly correlated independent variables such as the avgDeathsPerYear ~ avgAnnCount, avgAnnCount ~ popEst2015, avgDeathsPerYear ~ popEst2015, medIncome ~ percentprivatecoverage, Percentmarried ~ pctMarriedHouseholds and Porvertypercent -pctpubliccoveragelone. This makes sense since most of these variables can be mathematically derived from one another.

Cancer mortality rate is not highly correlated with any variables. However, one of the most interesting findings is poorer cancer patients die at a higher rate than those who are wealthier. However, there is no correlation between cancer mortality and clinical trials per capita. It is possible that the quality of treatment or exposure to carcinogens is the reason why poorer counties have a higher mortality rate. Further analysis broken down by types of cancer and exposure to specific carcinogen is required to make conclusions.

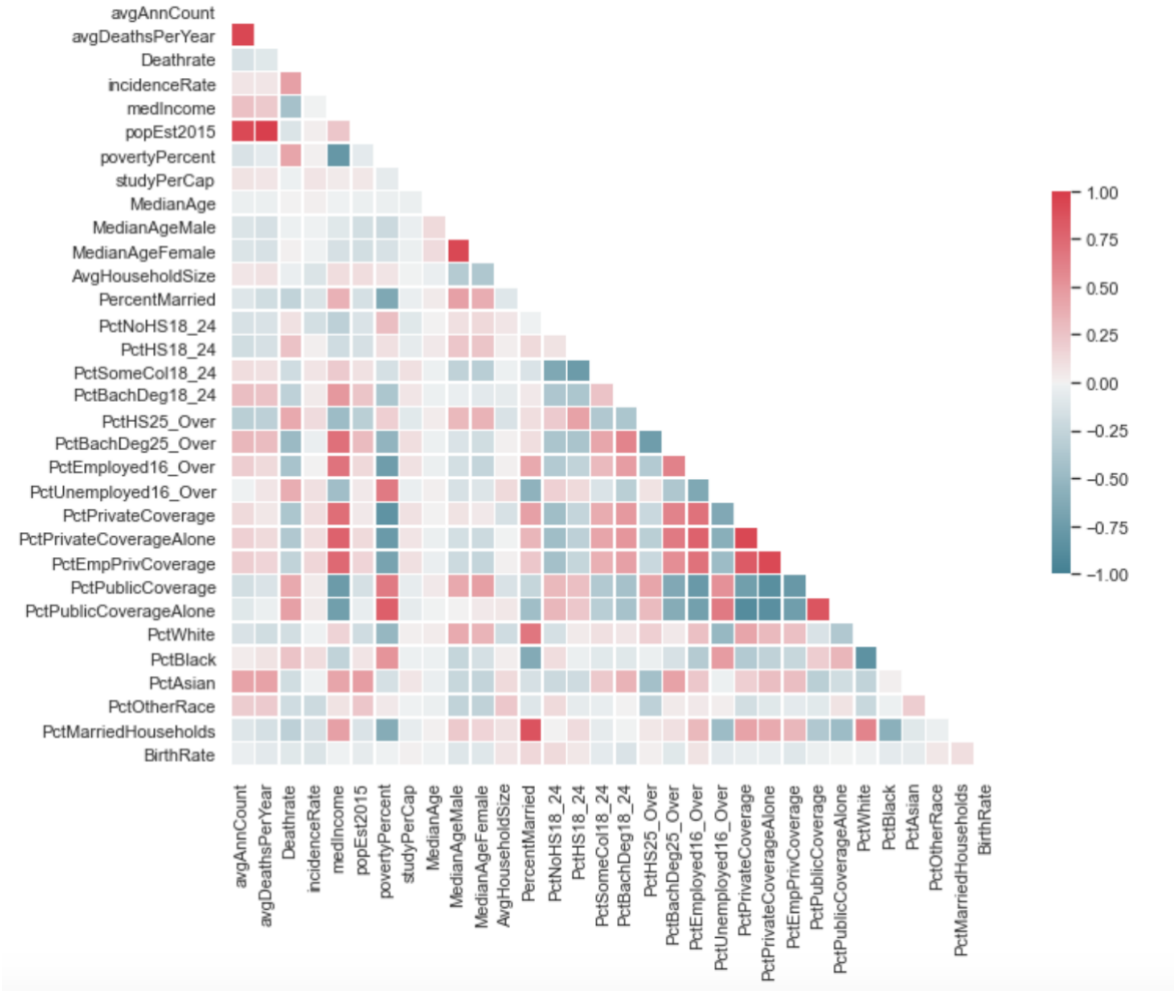

Fig 4. Correlation of data variables

Model Selection

General linear regression models and tree based models were trained on the dataset. Grid search for k-best features showed the optimum number of features to be used with the basic linear regression model is 26 features. Regularization did not improve the performance of the models. Ridge, Lasso and ElasticNet models performed similar to the base linear regression model even after hyperparameter tuning. This means that dropping or reducing the influence of some variables has little effect on the performance of the linear model. The low R^2 values of the linear model indicate a weak linear relationship between the target dependent variable and the independent variable. The linear models also showed slightly lover performance on the test data compared to the train dataset. The R^2 values are lower and the MSE and RMSE values are higher for the test dataset. This shows that the models were overfitting the training set.

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

Table 1. Comparison of performance of regression models. **A.** shows the calculated metrics for the models on the train dataset. **B.** shows the calculated metrics for the models on the test dataset

The tree based models performed better than the linear models. The Random Forest model was slightly better than the linear models on the test dataset with hyperparameter tuning. It appears there was no overfitting with this model as observed for the general linear models. The best

model was the XGBoost model. Hyperparameters were also tuned using grid search for this model. The model performed well on the training data and slightly more on the test data. The metrics are well within the acceptable range of a good predictive model with an R^2 value of 0.7 and fairly low mean absolute error and root mean squared error.

Although there was multicollinearity between the independent variables, I did not attempt to filter the variables since this would not have had added any significant improvement to the performance of the linear models and the tree based models are generally not affected by presence of highly correlated explanatory variables.

Best Model

The XGBoost regression model was the best model of all the models that were trained on the data. The performance metrics were significantly better in comparison. The hyperparameters are shown in table 3. The model performed similarly for both for the train and test data. Thus this model can be reliably used for predicting cancer mortality rate using the regional socioeconomic variables indicated in this model.

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

Table 2. Performance metrics of the XGB regression model.

XGB Parameter	learning _rate	Max_ depth	gamma	min_ child_weight	colsample _bytree	reg- alpha	n_ estimators
Best	0.1	3	0.1	5	0.5	50	1000

Table 3. Hyperparameters for the best XGB regression model.

The built-in feature importance function was used to score the features based on their weighted contribution to the model. The scores are shown in figure 5.

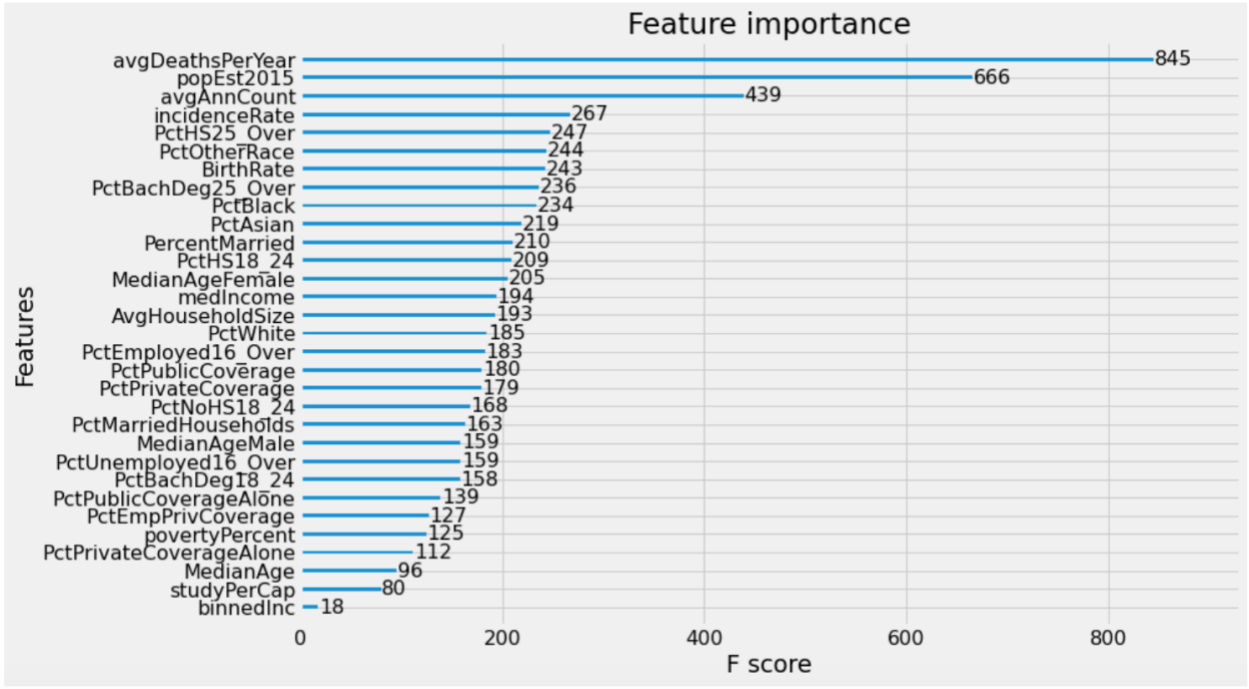

Fig. 5. Feature importance bast on the XGB model

Conclusion

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