The current state of reality is the average death rate from cancer for US counties is 179 cases per 100,000 population.

It is in society's interest to understand factors which can lead to cancer risks. Being able to better understand these factors can lead people to make better informed decision about where they live. And it can lead county governments to make better decisions to reduce their population's exposure.

Success for this project will be defined as identifying the ML model which yields the best evaluation metrics from predictions. Because this is a supervised regression problem, we will use MAE and MSE as evaluation metrics.

The data was aggregated from several sources, including census.gov, clinicaltrials.gov, and cancer.gov. This data has been put into the file cancer_reg.csv, which is found at https://data.world/nrippner/ols-regression-challenge.

We went through the steps of the data science workflow to execute the project. These steps are detailed below.

Data wrangling

- 1 column missing over half the values, so it was dropped
- 2 other columns with null values, imputed the mean
- Got rid of binnedInc column because it reflects the same info as medIncome column
- Created state column
- Checked for duplicated rows

EDA

- Some median age values were unreasonably high, so we dropped all rows with median age over 70 years
- Histogram shows deathrate is almost normally distributed.
- We apply a normal distribution as the deathrate
- We take a sample
- From the sample mean and std, we have 95% confidence all state means fall within the margin of error calculated from the z table and the t table.
- The incidenceRate is not normally distributed
- We look at boxplot of deathRate and medianIncome. We observe higher deathrate with lower incomes.
- We observe a linear regression plot with deathRate increasing along with incidenceRate
- The heatmap shows the correlations between the features and deathRate.
- Linear regression plot shows no relationship between MedianAge and deathRate.
- We get Pearson R values for all numerical features. The top positive values belong to PctPublicCoverageAlone, incidenceRate, and povertyPercent. The top negative values belong to PctBachDeg25_Over, medIncome, and PctEmployed16_Over.
- We observe histograms of all features, and many appear to be normally distributed.
- We observed histograms of grouped state date, but found nothing interesting.

Preprocessing

- We created a dummy regression model to get baseline mse and mae scores.
- We split the data into train and test sets
- We trained a linear regression model and found it had better metrics than the dummy model. Then we found this model best performs with 27 features.
- We found the random forest model performs best with 784 estimators and using the median strategy for imputing missing data.
- We employed Featuretools to generate extra features. This process generated 30 additional features.

Modeling

- We trained several regression models. All are scored on R^2. We compare MAE and MSE across the models. Hyperparameters tuning went through several iterations to fine tune the values.
- OLS model had MAE 13.8 and MSE 342.5
- Random forest model with max depth 50 and estimators 811 had MAE 14.1 and MSE 398.5
- We trained 2 gradient boosting models.
- The first was a decision tree regressor trained on the training data and also twice on the residuals. This model had MAE 17.1 and MSE 539.2
- The second was a gradient boosting regressor model with learning rate 0.2, max depth 2, estimators 5000. This model performed well with MAE 10.7 and MSE 231.3
- ElasticNet model had MAE 14.8 and MSE 407.8
- We used Beyesian optimization to optimize the hyperparameters for Light GBM regressor model. The best model had MAE 17.1 and MSE 532.9

The best model is the gradient booster regressor model with MSE 231. But this model takes a long time to train with 5000 estimators. We would have to understand any time constraints to determine this is indeed best. But even this model with 10 times fewer estimators still performed better than the alternative models in this exercise.

This model can be used to determine the expected cancer death rate for other US counties having similar info available. We can also look at the most important features and find strategies to affect the target death rate. With this dataset, education shows as being more important than many other features. Perhaps counties can better subsidize higher education to incentive people to continue their schooling.