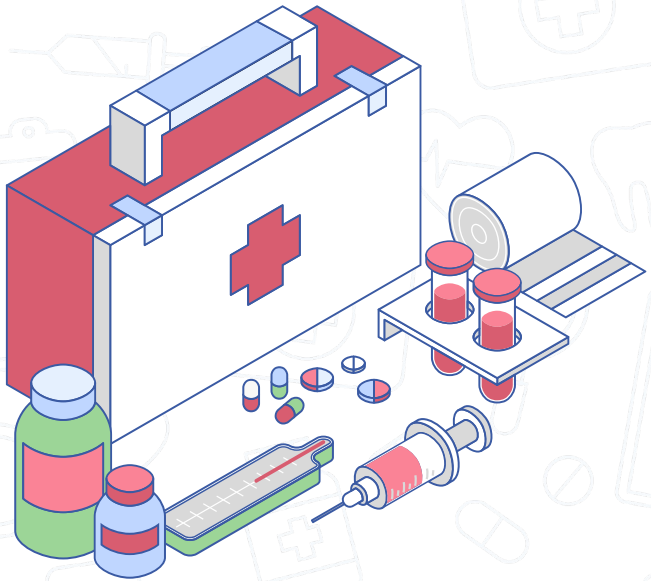


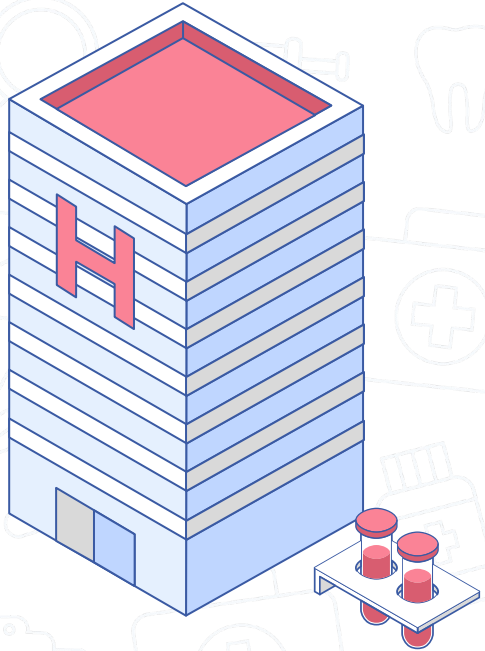
# PREDICTING PREVALENCE OF CANCER CASES

## TEAM: CRAMMA

Crystal, Ram, Emma

# TABLE OF CONTENTS





01

# CRAMMA

# OUR TEAM

## CRYSTAL

21’ MSBA
20’ BA

## RAM

22’ MA
21’ BA

## EMMA

22’ MA
21’ BA

# WHO WE ARE

Domain knowledge:

- Healthcare Market Research
- Healthcare Analytics
- Healthcare Product Development

Mission:

- Build cancer models for organizations in the healthcare industry to help cancer patients
- Provide governments and health providers with information that will help them predict onset of cancer in certain populations

# INITIAL QUESTIONS

- Is number of cancer cases seen in a region strongly predicted by the total population present, if not what other factor(s) could be causing the more frequent cancer cases seen.
- Do other factors help determine if cancer is more likely to develop among certain populations

# INITIAL HYPOTHESIS

- Our initial hypothesis was that income, insurance coverage, and age all would be important in predicting the number of cancer cases seen in a region. We assumed this because many times the ability of individuals to prevent and fight cancer is dependent on their financial capabilities and whether they have insurance as these can help them be healthier with regular checkups and afford chemotherapy.

02

# DATA PREPARATION

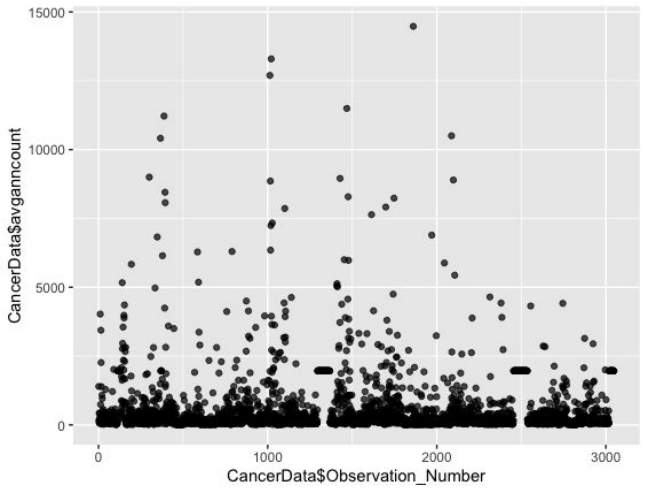
# DATA SOURCE

- The link for the data we used is https://data.world/nrippner/ols-regression-challenge. This data was collected by the National Cancer Institute. Since the data was attributed by the National Cancer Institute, a respectable and well-established organization, and used by the government, US Department of Health & Human Services, we are confident that this source is credible. After initially looking at this data, we thought choosing this data source would be beneficial to the project because it did not contain any missing values and there weren't any repeated categories.

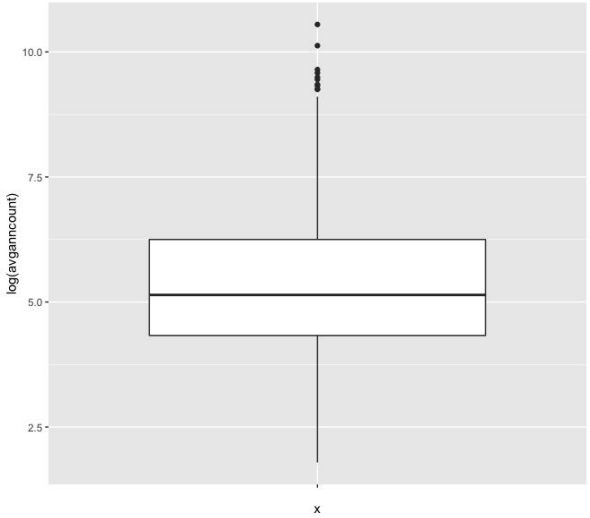
# DESCRIPTIVE STATISTICS

| | Average number of reported cancer cases (avganncount) | Median Income (medincome) |
|---|---|---|
| **Mean** | 606.3385 | $47,063.28 |
| **Standard Deviation** | 1416.356 | $12,040.09 |
| **Maximum** | 38150 | $125,635 |
| **Minimum** | 6 | $22,640 |
| **Median** | 171 | $45,207 |
| **Variance** | 2006065 | $144,963,787 |

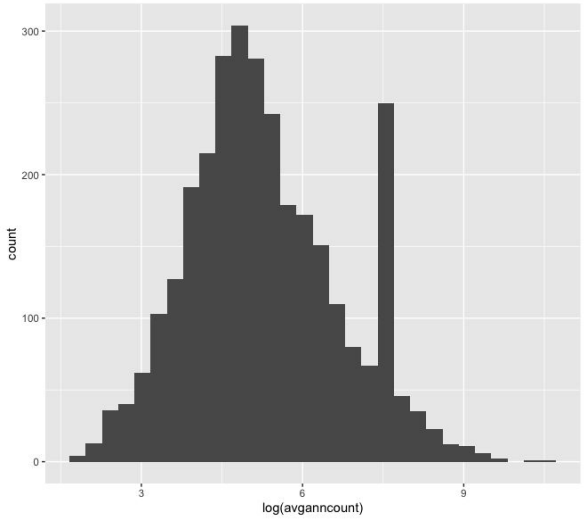
# DATA VISUALIZATION




Scatter Plot




Boxplot




Histogram

# TARGET VARIABLE

- Average Annual Number of Cancer Cases or “avganncount”

- We chose this as our target variable because we hypothesized that the remaining variables in the dataset may affect or help explain the number of cancer cases that are prevalent in a county population.

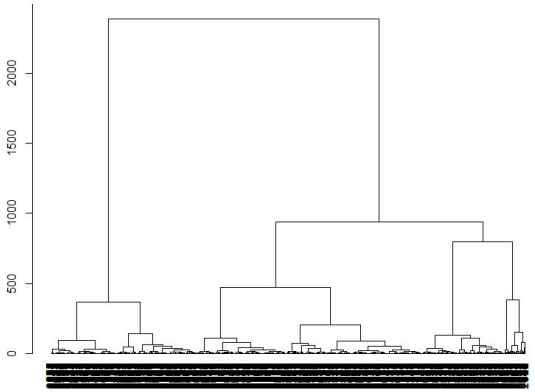
- To make classification models easier to interpret we converted the average number of cancer cases into a categorical variable that placed each value into a bin of either “low”, “medium”, or “high”

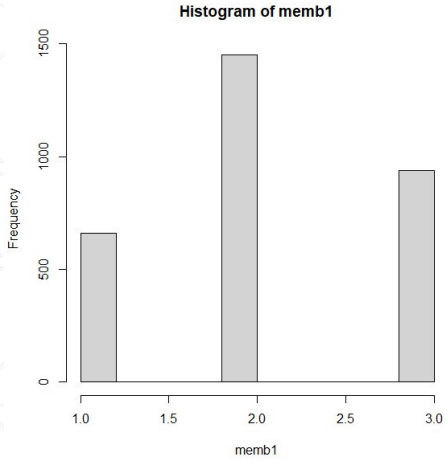
# PREDICTIVE VARIABLES

| | |
|---|---|
| **medianIncome** | Median income per county |
| **popEst2015** | Population of county |
| **MedianAge** | Median age of county residents |
| **PctPrivateCoverage** | Percent of county residents with private health coverage |
| **PctPublicCoverage** | Percent of county residents with government-provided health coverage |
| **PctMarriedHouseholds** | Percent of married households |
| **avgDeathsPerYear** | Mean number of reported mortalities due to cancer |

# CLUSTERING

- Identified 3 clusters based on the dataset to group similar data points
- Included this in our dataset to find out if these groups could provide new insights



03

# MODEL BUILDING

Prediction, Classification

# PREDICTION

- Backward selection method with polynomial term of average deaths
- Better when compared with Regression Tree (RMSE: 640.54)
- Percentage of private coverage is more significant than percentage of people covered by public insurance

```
Call:
lm(formula = CancerData$avganncount ~ avgdeathsperyear + popest2015 +
    pctprivatecoverage + pctpubliccoverage + I(avgdeathsperyear^2),
    data = CancerData)

Residuals:
    Min      1Q  Median      3Q     Max
-6486.1  -196.1  -112.4   -13.6  1946.7

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.746e+02  1.259e+02  -6.945 4.62e-12 ***
avgdeathsperyear     1.940e+00  8.869e-02  21.870  < 2e-16 ***
popest2015           7.288e-04  1.500e-04   4.858 1.25e-06 ***
pctprivatecoverage   1.356e+01  1.167e+00  11.615  < 2e-16 ***
pctpubliccoverage    4.310e+00  1.646e+00   2.619  0.00885 **
I(avgdeathsperyear^2) 7.806e-05  1.645e-05   4.746 2.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 466.4 on 3038 degrees of freedom
Multiple R-squared:  0.8299,    Adjusted R-squared:  0.8296
F-statistic:  2964 on 5 and 3038 DF,  p-value: < 2.2e-16
```

**RMSE: 465.93**

# CLASSIFICATION

- The parameters for bagging tree as a default were set to the following:
  - mfinal = 100
  - maxdepth=5
  - minsplit= 2
- After running a grid search, R studio found the best parameters to be:
  - mfinal = 3
  - maxdepth=4
  - minsplit= 2

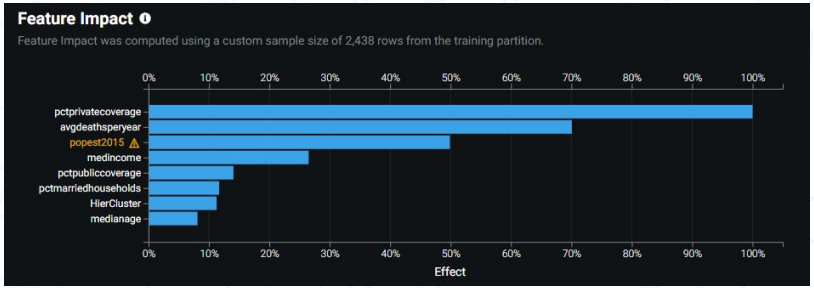
| | ACCURACY SCORE | FI |
|---|---|---|
| Classification Tree | 0.884 | 0.943 |
| Bagging Tree | 0.928 | 0.959 |
| Random Forest | 0.893 | 0.95 |
| Boosted Tree | 0.913 | 0.954 |

# CRAMMA VS. DATAROBOT

# PREDICTION

| | CRAMMA | DATAROBOT |
|---|---|---|
| | MULTIVARIATE LINEAR REGRESSION | RANDOMFOREST REGRESSOR: M43, BP64 |
| RMSE (HOLDOUT) | 465.92 | 418.02 |

- DataRobot's RandomForest model is much more robust compared to Cramma's Multivariate Linear Regression model
- Feature impact aligns with the Regression output from Cramma




DATAROBOT

# CLASSIFICATION

| | CRAMMA | DATAROBOT |
|---|---|---|
| | BAGGING TREE | EXTREME GRADIENT BOOSTED TREES CLASSIFIER: M39 BP25 |
| ACCURACY (HOLDOUT) | 0.93 | 0.9 |

| | FI | RECALL | PRECISION |
|---|---|---|---|
| HIGH | 0.86 | 0.76 | 0.98 |
| MEDIUM | 0.93 | 0.96 | 0.91 |
| LOW | 0.88 | 0.93 | 0.84 |

- The DataRobot results are very similar to the classification model that we have created using R. The main difference is DataRobot has broken it down by each class.
- Since DataRobot has an accuracy of .9 and our bagging tree has an accuracy of .93, this difference could indicate that the grid search we conducted via R packages might have been slightly more accurate than DataRobot.

**CRAMMA**

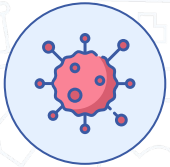
# 04

# CONCLUSION

# ACTIONABLE BUSINESS INSIGHTS

## INSIGHTS

Private insurance coverage plays a larger role in predicting cancer cases, indicating a high business potential for cancer treatment companies. Positive association between median income and average cancer cases, indicate need for government response



## SURPRISES

Analysis showed income and insurance coverage as most important variables, surprisingly not age, despite age being listed as highest risk factor in developing cancer (National Cancer Institute)

## USAGE

Models can be used to predict the average cancer cases expected in a region, for insurance companies to determine treatment budget, as well as pharmaceutical companies to estimate treatment manufacturing

# REFLECTION

## ROADBLOCKS

Finding the right data for this analysis was quite difficult

## COLLINEARITY

Average cancer mortalities may be explained by the other predictors just as they explain the average cancer cases

## NORMALIZATION

Valued interpretability more than accuracy in our model to explain to stakeholders, resulting in poor model performance

## TARGET LEAKAGE

Using cancer mortalities as a predictor for the number of cancer cases is using information that we would typically not have (looking into the future)

“The great thing about predictions is that you can be wrong.”

—CHRIS WIGGINS

# THANKS!

Do you have any questions?

*CONTACT US*
chariga@brandeis.edu
bram99@brandeis.edu
ebrown321@brandeis.edu