Final Report

BACKGROUND CONTEXT

The entirety of our team has experience in the healthcare industry which led us to choosing this topic. Crystal interned at a healthcare market research firm last summer. Ram was a medical research assistant and currently a health analytics intern. Emma was a product development intern at a healthcare startup based in Boston two summers ago. Our team wanted to focus on the various factors that may be correlated or help predict the number of the cancer cases we can expect to see in a region. This is important information to know and a useful model to have as it helps organizations that offer help for cancer individuals to be located in certain regions and it gives individuals a better estimation as to if they have a higher chance of developing cancer in a certain region.

INITIAL THOUGHTS

Cancer has been a lingering condition for years now and research has only led to the discovery of some causes of cancer, but not all causes. If we knew all the causes of cancer it would be preventable, but individuals still develop cancer despite being completely healthy and staying away from certain causes. In order to understand if there are any other factors that may cause cancer or help determine if cancer is more likely to develop, we decided to look at various variables among certain populations to see which ones, if any, help determine the number of cancer cases we would expect to see in a certain population.

Our initial hypothesis was that income and insurance coverage 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.

We also wanted to see if the number of cancer cases seen in a region is strongly predicted by the total population present, if this comes out to not be the case then there is some other factor that must be causing the more frequent cancer cases seen.

DATA SOURCES

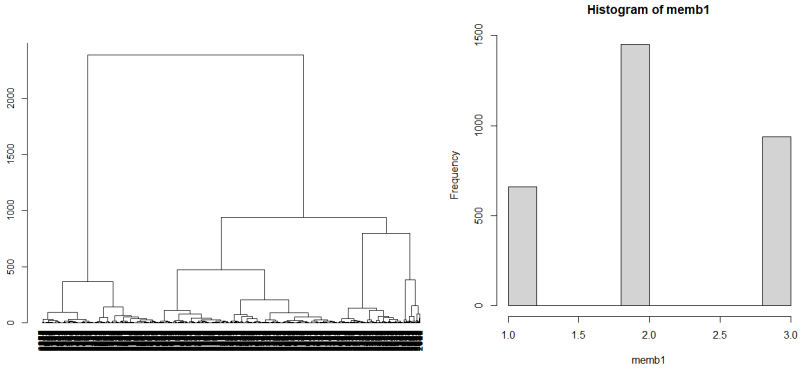
We used data from the data world website. The data we used was collected by the National Cancer Institute. Since the data was attributed by the National Cancer Institute, a respectable and well-established organization, and used by the government, US Department of Health & Human Services, we are confident 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.

DATA PREPARATION

The dataset is not too messy, but overall the data were in the right format and types.

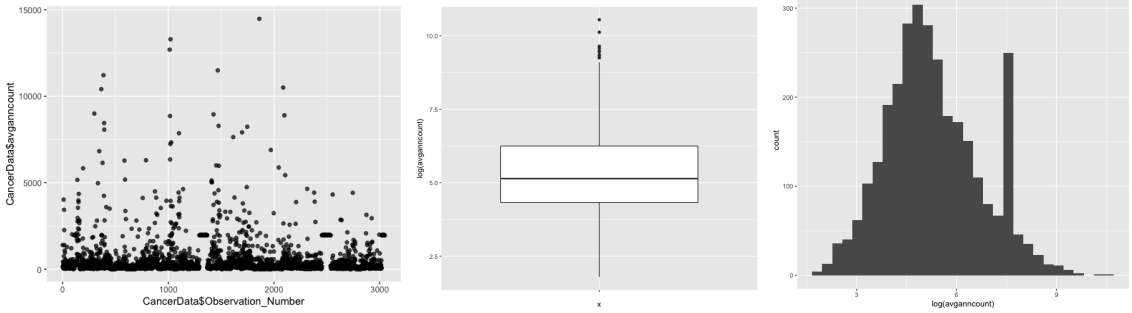
Clustering

Since our dataset did not have any classes that group multiple data points together, we used clustering to see if there are any similarities between groups of data points. From using hierarchical clustering with the euclidean and ward.D method, we found that our cancer dataset includes 3 different groups (see the distribution below).



We included the clusters into our analysis, to see if we can find any insights about the clusters in our models.

DESCRIPTIVE STATISTICS/VISUALIZATIONS



	Mean	Standard Deviation	Maximum	Minimum	Median	Variance
Average Number of Cancer Cases (avganncount)	606.33	1416.356	38150	6	171	2006065
Median Income (medincome)	$47,063.28	$12,040.09	$125,635	$22,640	$45,207	$144,963,787

Above we created a few visualizations to better understand the spread of our target variable, average number of cancer cases. In the scatterplot, the value seems to be condensed under 2500, so to gain a better understanding of the spread of the different counties we created a box plot and histogram but used the log of the average number of cancer cases. After doing so we can see that there are distinctly about 5 outliers according to the box plot and a few according to the small bar on the histogram. We calculated the basic descriptive statistics of the target variable and one of the predictive variables we expected to be quite significant in helping explain the target variable.

TARGET VARIABLE AND JUSTIFICATION

We chose our target variable to be the average annual number of cancer cases seen or avganncount. We chose this as our target variable because we infer that the remaining variables in the dataset may affect the number of cancer cases that are prevalent in a population.

The spread of the number of cancer cases seen across the populations is quite spread out as we can see in the histogram shown earlier, which typically determines if a variable would be a good target variable.

To make classification models easier to interpret using this continuous variable 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”

INTERPRETATION OF RESULTS

Prediction

```
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
```

The multivariate linear regression is our best model with an RMSE of 465.93, when compared with our other prediction model, Regression tree with an RMSE of 640.54. We filtered out the insignificant predictors using the backward selection method and were left with 5 variables: avgdeathsperyear, popest2015, pctprivatecoverage, pctpubliccoverage, I(avgdeathsperyear^2). We could interpret this model in a way that the 5 predictors play a large role in predicting the target variable, average annual cancer cases. An interesting insight that we found is that the

percentage of private coverage is more significant than the percentage of people covered by public insurance, which could indicate that we need further research for this.

Classification

	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

After comparing the various classification models we've created, the best one that we determined is the Bagging Tree with an accuracy score of 0.928 and F1 score of 0.959.
Mfinal = 3
Control = 4
Minsplit = 2
F1 = 0.959

In simple terms, a bagging tree model draws random samples out of the training sample (at the beginning of the code) and replaces that sample to get the different types of models. Although the HCluster is usually used as a predictor and not a target, we tried to use it as a target. Unfortunately, the HCluster took in a bunch of different factors that we were also using as predictors which gave us a result of F1 = 1. We knew that this was not actually the case. After finding out these results, we decided to create 3 bins ("high","medium", and "low") using the avganncount variable from our data set. We used the rest of the dataset variables (except for avganncount) as the set of predictors. The parameters for bagging tree as a default were set to the following:

mfinal = 100
maxdepth=5
minsplit= 2

After setting these parameters we got the following results:

Bagging Tree accuracy= 0.8917145 and F1= 0.9507295

After running a grid search, R studio found the best parameters to be:

mfinal = 3
maxdepth=4
minsplit= 2

We then used the randomforest() function using these new parameters and the resulting variable importance plot was as follows:

Bagging Tree accuracy= 0.9278394 and F1= 0.958525

The “mfinal” is the number of times that the bagging tree will run before finding the optimal model. As you can see above, the default was running the bagging tree 100 times but after running our grid search, we found that the bagging tree only needed to be run 3 times to find the optimal model. The “maxdepth” is just the maximum depth of the bagging tree. Whatever number you set this parameter to, it will make the tree stop when the depth is set equal to the maxdepth value. As you can see above, the default was having a maxdepth value of 5 but after running our grid search we concluded that the optimal maxdepth is 4. The minsplit is the minimum number of observations that are contained in a node for a split to try to occur. Both the default value in our code and the result after running a grid search was 2. Since the result was 2, this would be considered a terminal node.

COMPARISON WITH DATAROBOT

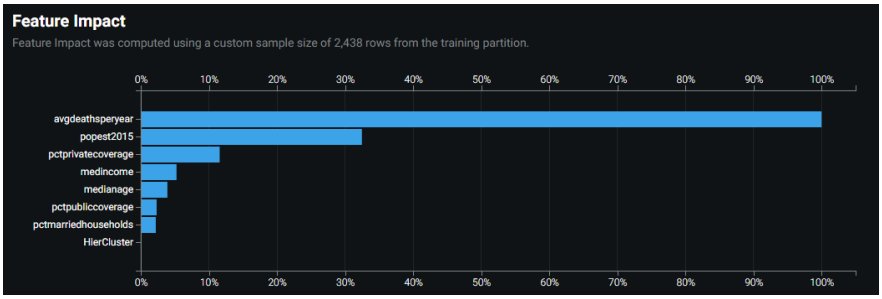
Classification

	R Model (Bagging Tree)	DataRobot (eXtreme Gradient Boosted Trees Classifier: M39 BP25)
Accuracy (Holdout)	0.93	0.90

Results from DataRobot are very similar to the classification model that we have created using R. Although the confusion matrix is difficult to compare since DataRobot has it broken down by each class,

	F1	Recall	Precision
High	0.86	0.76	0.98
Medium	0.93	0.96	0.91
Low	0.88	0.93	0.84

overall the accuracy is quite similar. This XG Boosted Tree model has an accuracy of 0.9, while the best model we have was actually Bagging (which is not available on DataRobot) with an accuracy of 0.93. Our second best model is Boosted with an accuracy of 0.91. This shows that both DataRobot and our model are very similar but the difference could indicate that the grid search we conducted via R packages might have been more accurate than DataRobot.

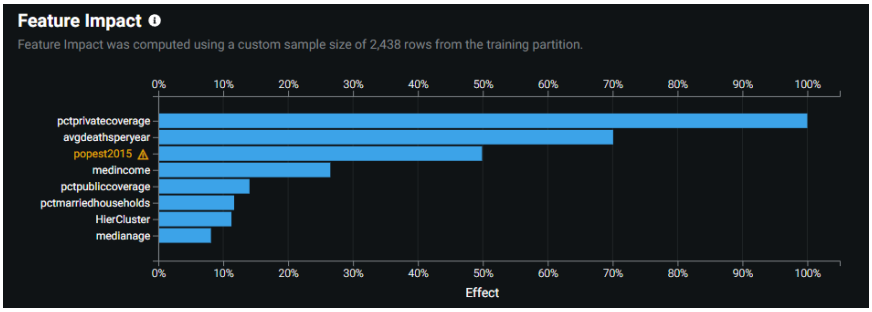


Looking at the Feature Impact chart from DataRobot, it is very similar to the Variable Importance Plot that we have created for the Decision tree in Report 3, with average death counts per year being the highest in importance followed by estimated population count.

Prediction

	R Model (Multivariate Linear Regression)	DataRobot (RandomForest Regressor: M43, BP64)
RMSE (Holdout)	465.92	418.02

Compared with our best Regression model, which is a linear regression model with an RMSE of 465.92, DataRobot's RandomForest Regression model performs much better with an accuracy of 418.02. This makes sense considering that RandomForest is an ensemble model that is much more robust and accurate compared to a linear regression.



Interestingly, when we take a look at the Feature Impact chart from DataRobot, we can see that the highest impact includes the percentage of private coverage followed by average death counts per year which we did not find in the classification model. It seems that in determining the average cancer count per year, the percentage of population with private coverage will impact the prediction. For further analysis, we can conduct more research to understand more about the relationship between private insurance and cancer cases.

Based on the results of this comparison, it is difficult to determine whether we should rely on DataRobot or our own models because it depends on the situation. Between prediction and classification, it is important to recognize that prediction will return a specified number, however classification will only return the classes (high, medium, low). This needs to be considered when applying the models in a real world setting, such that estimated ranges could use the classification model, whereas prediction would be used in applications where it is important to find an exact number.

CONCLUSION

Insights

It seems from the variables we deemed most significant as predictors that people may be acting careless about their health and the risk of developing cancer as they assume they are financially stable. In order to fix this we could advise the government and any non-profit organizations to make sure a proper education system is established that warns individuals of the various causes of cancer from established research over the years. As a result of this new and proper education we can expect individuals to be more cautious and see the rate of cancer cases decrease, along with the number of cancer related deaths.

Based on the random forest model we can advise chemotherapy companies to focus their marketing efforts toward more wealthy neighborhoods or regions as the model showed that median income was an important variable in predicting the number of cancer cases. Also these companies can focus their marketing efforts towards individuals with either private or public insurance coverage given variable importance plots marked both as models that could be important in determining the number of cancer cases. An explanation for this could be that people who know they are financially secure and have insurance coverage are less likely to be as healthy or cautious with their health. As sad as this reality is, chemotherapy is also very expensive, so those who are wealthier and/or have coverage are more likely they will purchase chemo treatments.

We can also suggest funeral homes use the classification and regression tree model, to predict and classify the number of average cancer cases or even the deaths to determine what characteristics of an area would have a high cancer death count, and hence be profitable.

Surprises

Our analysis showed that income and insurance coverage as the most important variables in helping predict the prevalence of cancer, but surprisingly not age, despite age being listed as the highest risk factor in developing cancer (National Cancer Institute). We expect that a reason for this could be due to target leakage or collinearity, which we will go into further detail later at the end of the report.

Usage

The model created can be used to help predict what regions are the most likely to have a high prevalence of cancer cases. This information is helpful for the government as they can not only make sure those regions have access to cancer treatments, but also allows them to work to take any preventative measures there may be to avoid developing cancer. This information can be used from a business standpoint by cancer treatment companies as they will be able to tell what regions will need their products more and choose to be located in those regions.

REFLECTIONS

Despite our model performing better than datarobot we expect there was some target leakage and collinearity in our model that prevented it from performing even better.

We used the average number of cancer mortalities as a predictive variable, but these cancer mortalities can most likely be predicted by the other variables we have used in the model. This is a situation of collinearity.

There is also a case of target leakage because by using cancer mortalities as a predictive variable for explaining the number of cancer cases, we are in a sense looking into the future to use information that we would not have. We cannot use cancer related deaths to predict how many cancer cases there will be. We did not catch this until after finishing our analysis because surprisingly the leakage did not result in a perfect accuracy model.

We can take this as a learning experience and make sure in our future models that we do not use a variable such as average cancer mortalities to predict the prevalence of cancer cases in a county population.

We also missed normalizing the dataset before incorporating that into the model, which could take into account the poor model performance, since we value a more interpretable model than

the accuracy. However, we do need to take this into account especially if we'd like to deploy this model for scale. As for outliers, we had removed a little bit from our data, but we found some more outliers that we might have missed through DataRobot. Although so, it seemed that DataRobot still went along and built the model for us, resulting in quite a high accuracy, so we think that the outliers could have been a minor issue, but we will definitely keep this in mind for our next model.

Even though our classification models were quite high, we still understand that these models can be wrong, even with a high accuracy since as Chris Wiggins' quoted in our presentation, predictions can be wrong. Thus, indicating that every model being used to predict or classify will always need to be taken with a grain of salt.