

Section 2. Medical science

<https://doi.org/10.29013/YSJ-22-1.2-8-17>

Yue Wang,

*Place of study: A high school junior at George School
City, country: Pennsylvania, United States*

A STUDY OF US EXPENDITURES ON CANCER TREATMENT WITH DATA ANALYSIS AND MACHINE LEARNING

Abstract. Cancer is the second leading cause of death around the world, causing cancer cost to be an important social issue in the United States. News reports show that American cancer patients spent more than \$21 billion on their care in 2019. (US News [2]) In this research, data analysis has been done based on the national expenditure on cancer treatment from 2010 to 2020 through the use of Python language and available third party libraries. Also, a machine learning classification model has been trained, developed and tested to help predict the cost of cancer treatment in the next few years. Among four different machine learning regression algorithms that are applied (i.e linear regression, lasso regression, random forest regression, and gradient boosting regression), gradient boosting regression is the best fit for the model, aiming to produce the most accurate prediction to inform people and government officials.

Keywords: cancer cost, correlation, machine learning, linear regression, lasso regression, random forest regression, gradient boosting regression.

Introduction

As the second leading cause of death, cancer has a high risk of taking people's lives away, which brings about questions regarding the cost of cancer treatment. A news report recently reveals that "American cancer patients spent more than \$21 billion on their care in 2019." (US News [2]) This figure emphasizes how much US citizens have spent on cancer treatment and to what extent some of them must have suffered from the high cost brought by different kinds of cancer. In addition, this figure is increasing every year – \$190.2 billion in 2015 and \$208.9 billion in 2020, an increase of 10 percent that is due to aging and growth of the US population. (National Cancer Institute [5]) To stress the heavy cost of dealing

with cancer, U.S. Bureau of Labor Statistics provided data contrasting between American people's average monthly income pre-tax and their average cost of cancer treatments per month. It is obvious to see from the comparison the inequality between how much people earned and how much they spent, in which income is about \$3600, while cost is about \$20000. (Karen Selby [3]) Aside from the out-of-pocket medical costs, expenditure spent on commuting also accounts for the large amount of spending. For example, among the approximate \$21 billion of cancer cost in 2019, \$16.22 billion was spent on out-of-pocket medical costs and \$4.87 billion on traveling expenses. (US News [2]) The high costs of cancer treatment exert especially great pressure on

the poor, those who are uninsured or underinsured, and blue-collar workers who may lose wages as a result of health issues. (Karen Selby [3]) According to the survey, 23% of US citizens aged 19–64 were “underinsured” in 2018 since their out-of-pocket health care costs were equal to 10 percent or more of their yearly income, meaning that insurance cannot cover their expenses on cancer treatment. (Karen Selby [3]) On the other hand, blue-collar workers are facing serious situations too, not only because they are less likely to have employer-based insurance coverage than white-collar workers, but also due to the fact that their annual mean wages weigh much less than the monthly cost of some cancer drugs, causing the cost of cancer to be unaffordable for them. Under such circumstances, people are expecting ways to lower the financial burden caused by cancer treatment, which not only means to reduce cancer patients’ out-of-pocket costs, but also to address the long-term financial impact (‘The Cost of Cancer [5]).

In this research paper, with the data collected in *Data.World* (xprizeai-health [1]) on the estimation of the national expenditures for cancer care from 2010 to 2020 under different assumptions of cancer incidence and survival trends, data analysis is going to be completed through the use of Python language and available third party libraries. Furthermore, it is possible to predict the cost of cancer treatment in the US in the next few years based on previous analysis through the use of machine learning algorithms.

Data analysis

The data used in this study is from *Data.World* (xprizeai-health [1]), which is the world’s largest collaborative data community. This dataset is an estimation of the national expenditures for cancer care from 2010 to 2020 in billion dollars under different assumptions of cancer incidence and survival trends, including 1258 entries of different kinds of cancer cost. The descriptions of column features and their corresponding values are shown below in Table 1:

Table 1.– Description of column features

Type	AllSites, bladder, brain, breast, cervix, colorectal, esophagus, head_neck, kidney, leukemia, lung, lymphoma, melanoma, ovary, pancreas, prostate, stomach, uterus, other
Year	Numerical type, including 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, and 2020
Sex	Both sexes, females, males
Age	All ages
Survival	Different combinations – (incidence, survival at constant rate), (incident follows recent trend, survival constant), (survival follows recent trend, incidence constant), (incidence, survival follow recent trends)
Cost_increase	Annual cost increase, numerical type, including 0%, 2%, and 5%
Cost_total	Total costs, numerical type
Cost_initial	Initial year after diagnosis cost, numerical type
Cost_continue	Continuing phase cost, numerical type
Cost_last	Last year of life cost, numerical type

Table 2.– Below presents 5 sample rows from the dataset:

	Type	Year	Sex	Age	Survival	Cost_increase	Cost_total	Cost_initial	Cost_continue	Cost_last
1	2	3	4	5	6	7	8	9	10	11
0	AllSites	2010	Both Sexes	All ages	Incidence, Survival at constant rate	0%	124565.6	40463.5	46642.8	37459.2

1	2	3	4	5	6	7	8	9	10	11
1	AllSites	2010	Both Sexes	All ages	Incidence follows recent trend, Survival constant	0%	122420.8	38552.7	46671.9	37196.3
2	AllSites	2010	Both Sexes	All ages	Survival follows recent trend, Incidence constant	0%	125397.7	40463.5	47136.3	37797.9
3	AllSites	2010	Both Sexes	All ages	Incidence, Survival follow recent trends	0%	123236.3	38552.7	47155.7	37527.8
4	AllSites	2010	Both Sexes	All ages	Incidence, Survival follow recent trends	2%	123236.3	38552.7	47155.7	37527.8

Table 2: Sample data

The next process is to learn the dataset by analyzing the data and understanding the relationships between these different columns through the use of Python Data Analysis (Pandas) and Python Data Visualization Libraries (Matplotlib and Seaborn).

First, through the use of countplot in Python, which shows the counts of observations in each cat-

egorical bin using bars (Geeksfor Geeks [4]), I make a plot about different types of cancer sites shown in Figure 1 below. In this plot, each bar represents the number of counts of different types of cancer. According to the same height of bars, we can conclude that data is very evenly distributed across different cancer sites in this dataset.

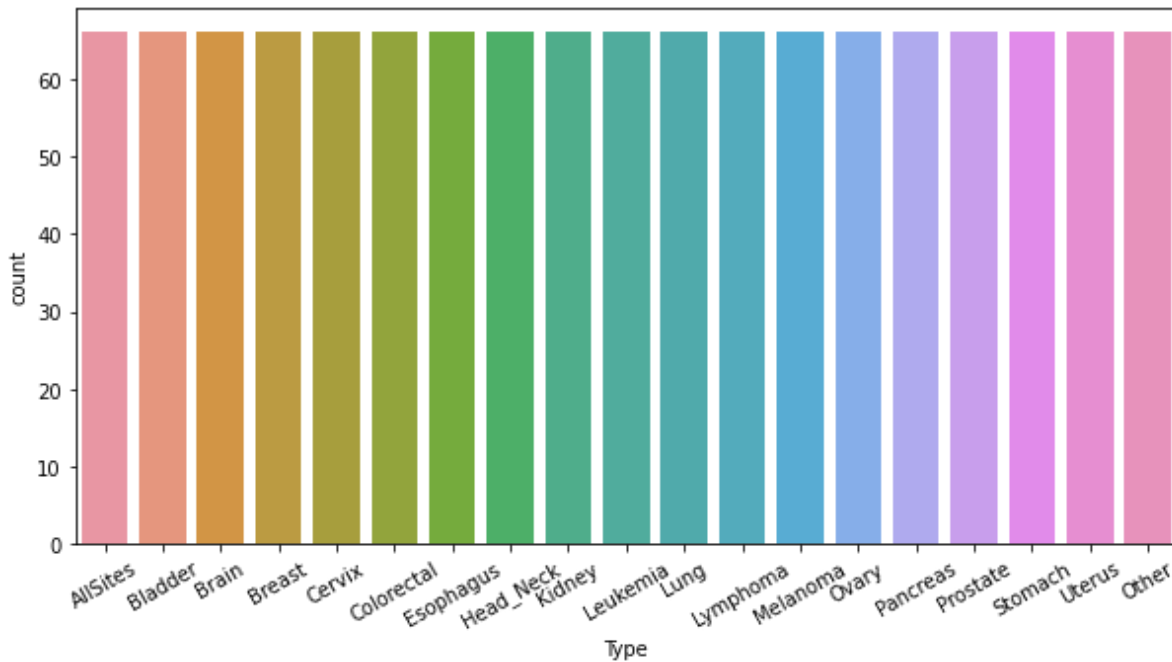


Figure 1. Plot for Type

The same measurement can be used for creating the plot of different sexes shown in Figure 2 below. In this plot, each bar represents different types of sexes – females, males, and both sexes considered as a whole. Evidently, the bar of “both sexes” is the

highest, and the bar of “females” is the second highest, which is higher than the bar of “males”, meaning that the population of females is larger than that of males included in this dataset.

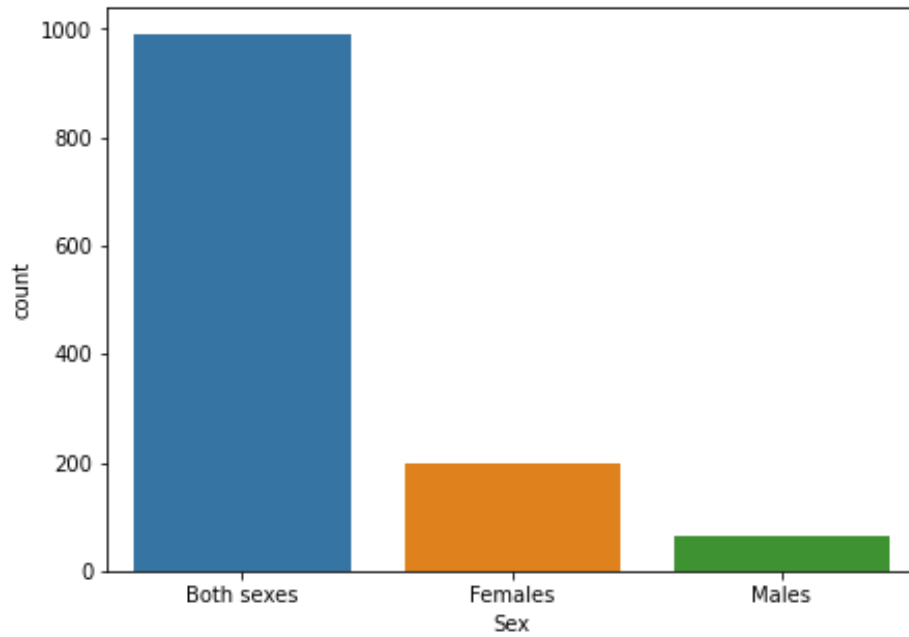


Figure 2. Plot for Sex

Also, the plot of survival and incidence was created as shown in Figure 3 below. In this plot, each bar represents different combinations of survivals – incidence and survival at constant rate, incidence follows recent trend and survival constant, incidence follows recent trend and survival constant, survival follows recent trend and incidence constant, survival follows recent trend and incidence constant, survival follows recent trends and incidence constant, survival follows recent trends and incidence follows recent trends.

recent trend and incidence constant, and incidence and survival follows recent trends. We can make the conclusion that since the first three heights are the same, the counts for them are the same, which are much lower than the fourth one.

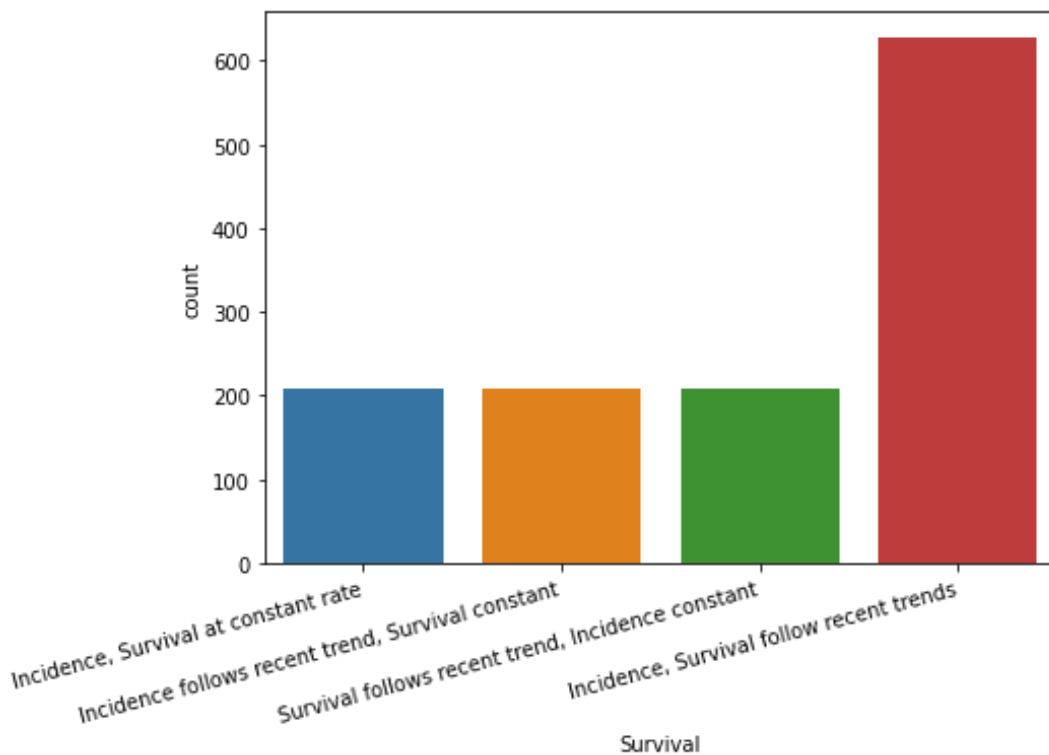


Figure 3. Plot for Survival

Next, I plot the histograms to investigate the distribution of numeric columns. As shown in Figure 4, we can draw some conclusions. For example, the data about years is evenly distributed since the number of counts for specific years is the same. For dif-

ferent percentages of increasing cost, the frequencies of 2% and 5% are roughly the same, which are much smaller than that of 0%. On the other hand, the rest of four distributions about costs are rightly skewed with some outliers.

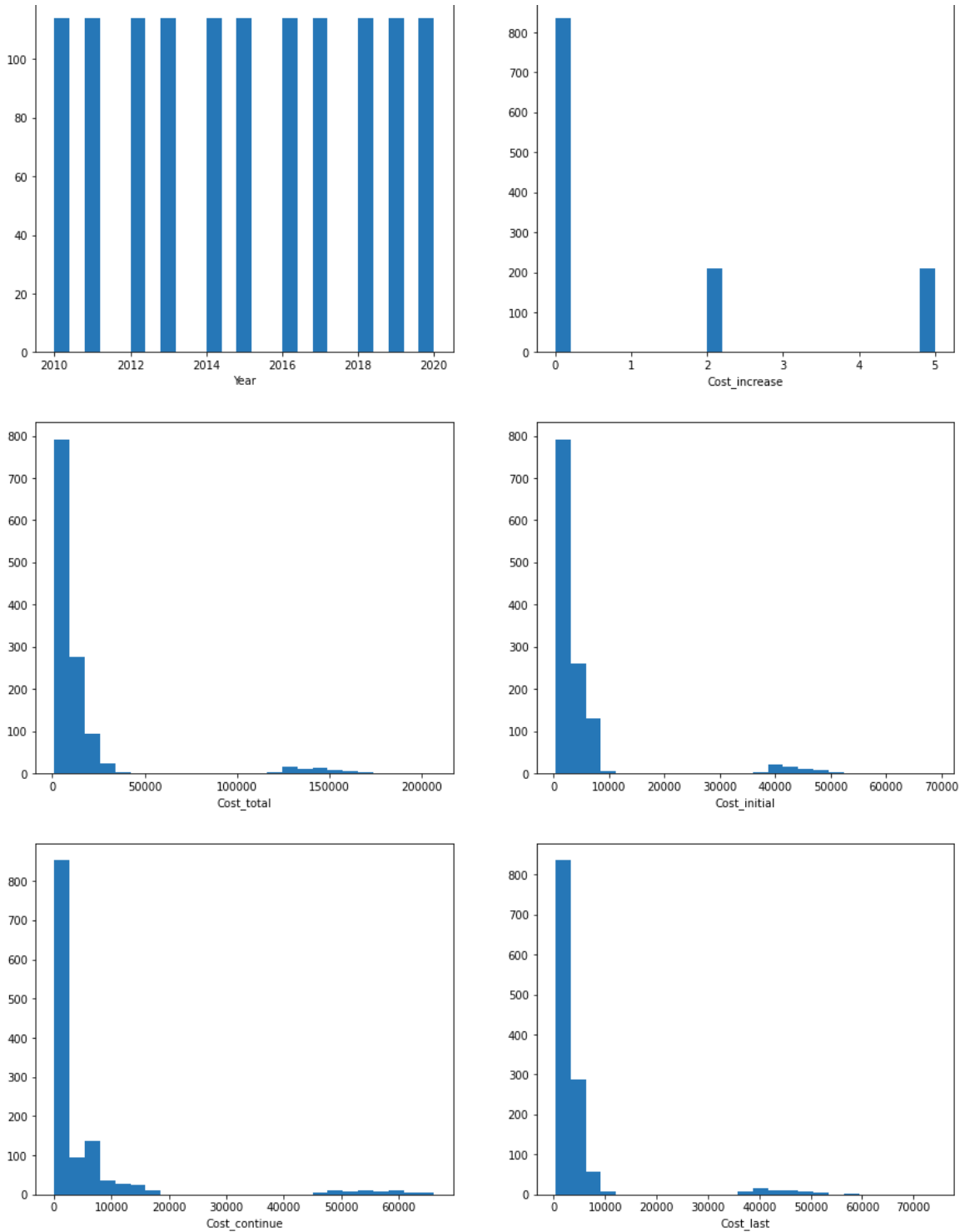


Figure 4. Distribution plots for each feature column

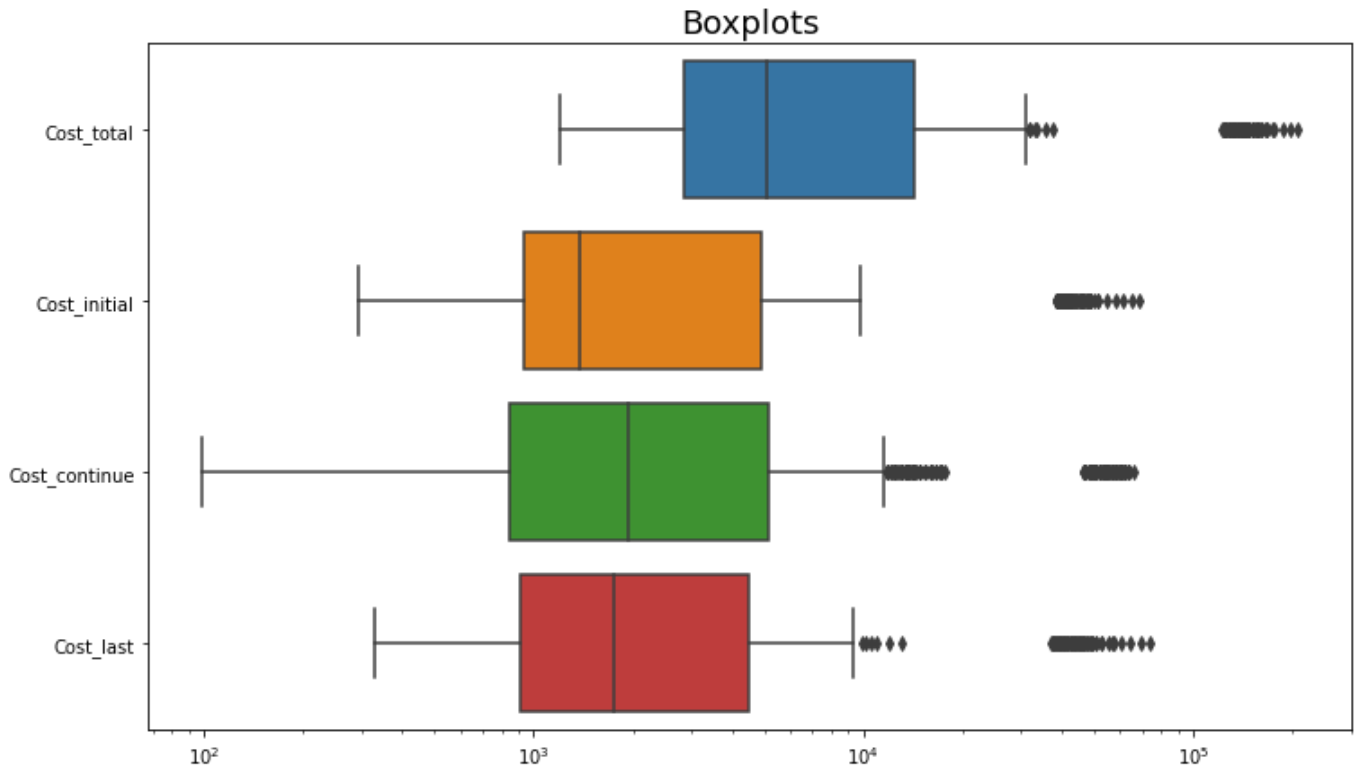


Figure 5. Boxplots for Costs

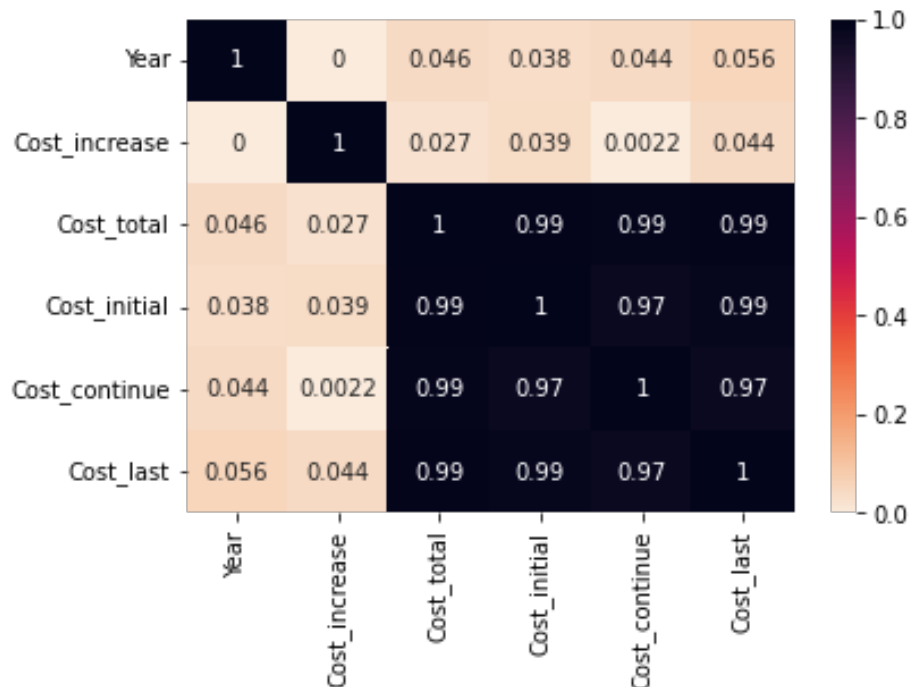


Figure 6. Correlation heatmap

To further explore these outliers, boxplot was made for specific kinds of cancer costs, in which we can have a clear idea about the values of minimum,

Q_1 , median, Q_3 , and maximum as well as the distribution of these outliers (Figure 5).

Correlation heatmap is also produced to visualize the correlation matrix. A correlation number of 1 or close to 1 indicates that two columns are highly correlated, while a correlation number of 0 or close to 0 shows that two columns hardly correlate with each other. In accordance with the figure below, year and annual cost increase serves as two factors that are not much influencing. On the contrary, the total costs, the costs of initial year after diagnosis, continuing phase costs, and costs of last year of life have strong correlation with each other, which is rather understandable since these four factors represent various costs during different stages in the cancer treatment process.

Machine learning

This section describes the approach to develop the machine learning regression model for prediction of cancer costs.

Preprocessing

Preprocessing is an important procedure before feeding the data into the model, which is also an integral step since the output of the model can be directly influenced by the quality of the data.

The specific step of preprocessing is below:

1) One hot encoding the sex column. According to Table 1, some column features are categorical variables, and some are numerical variables. Sex is a one of these categorical variables, which needs to be converted to numerical variables for the machine learning algorithm to understand. In this way, one hot encoding serves as a good way to prepare data

for an algorithm. (Dinesh Yadav [6]) For the sex column, after the conversion, all entries of “Both_sex” become “1” and those of “Male” and “Female” become “0”.

2) One hot encoding the survival column. According to Table 1, survival is also a categorical variable. Through one hot encoding, it is separated into four different columns – “Incidence, survival at constant rate” to “Both_constant”, “Incidence follows recent trend, survival constant” to “ConstSurv_TrendIncid”, “Survival follows recent trend, incidence constant” to “ConstIncid_TrendSurv”, and “Incidence, survival follows recent trends” to “Both_trend.” Whether being “1” or “0” depends on what kind of survival features a specific individual has.

3) Converting the “Cost_increase” column. Since the original “Cost_increase” column is object type due to the percentage sign, which is difficult to be used in an algorithm, 0%, 2%, and 5% are changed into 0, 1, and 2 respectively.

4) Frequency encoding the type column. According to table 1, there are 10 different types of cancers exhibiting in the type column, which result in too many unique values. Therefore, I attempt to group those values by frequency since I don’t want 10 new columns. By mapping values to dataframe and dropping the original column, a new column “Type_freq_encode” is created.

After all four steps, the 10 new sample rows from the dataset are below shown in Table 3:

Table 3.– New Sample Data

	Year	Cost_increase	Cost_total	Cost_initial	Cost_continue	Cost_last	Both_sex	Sex_F	Sex_M	ConstSurv_TrendIncid	Both_constant	Both_trend	ConstIncid_TrendSurv	Type_freq_encode
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	2010	0	124565.6	40463.5	46642.8	37459.2	1	0	0	0	1	0	0	0.052632
1	2010	0	122420.8	38552.7	46671.9	37196.3	1	0	0	1	0	0	0	0.052632
2	2010	0	125397.7	40463.5	47136.3	37797.9	1	0	0	0	0	0	1	0.052632
3	2010	0	123236.3	38552.7	47155.7	37527.8	1	0	0	0	0	1	0	0.052632

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	2010	1	123236.3	38552.7	47155.7	37527.8	1	0	0	0	0	1	0	0.052632
5	2010	2	123236.3	38552.7	47155.7	37527.8	1	0	0	0	0	1	0	0.052632
6	2010	0	3980.7	978.7	1895.8	1106.3	1	0	0	0	1	0	0	0.052632
7	2010	0	3885.2	923.3	1872.3	1089.7	1	0	0	1	0	0	0	0.052632
8	2010	0	3987.7	978.7	1900.2	1108.8	1	0	0	0	0	0	1	0.052632
9	2010	0	3891.9	923.3	1876.5	1092.2	1	0	0	0	0	1	0	0.052632

We also obtain the frequency for different types of cancer, which are identical for every cancer, so we can drop this column.

Regression models

For this research, I apply four different machine learning regression algorithms, including linear regression, lasso regression, random forest regression, and gradient boosting regression. During each run, I first apply the train set to train the model, then use the model on test data to make predictions which is to test the model’s performance.

First, linear regression is the base model that can be used to perform basic regression tasks. Mean absolute error is a typical metric used to evaluate a regression model, which with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set. Based on our data, the mean absolute error is about 0.0366985. Besides, r square is a necessary statistical measure of

how close the data are to the regression line (Minitab, 2013), which leads to the result of about 0.999, meaning that the model fits the data well.

According to the previous correlation heatmap shown in **Figure 6**, four cost factors are highly correlated with each other. As a result, we can apply lasso regression to address the multicollinearity, which is a good solution to reduce the magnitude of the coefficients of the model while keeping other features the same (Andrea Perlato [10]). Based on our data, the mean absolute error is about 0.55354.

A random forest is a supervised machine learning algorithm used for classification and regression that is constructed from decision tree algorithms. (Afroz Chakure [6]) It is a bagging technique, which operates by constructing a multitude of decision trees at training time and outputting the mode of classes or mean prediction of trees. The figure below shows how a random forest works.

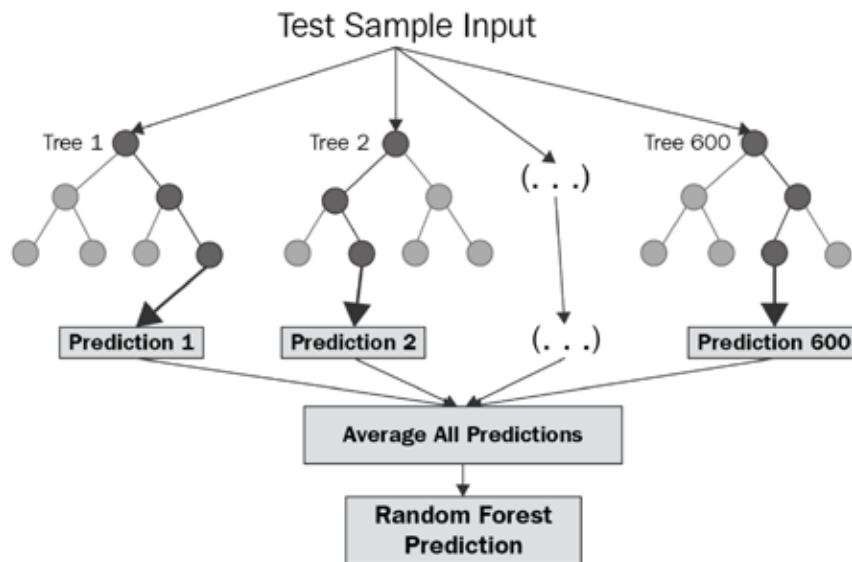


Figure 7. Random Forest Structure

According to the random forest regression, the R square value is about 0.99567.

Also, gradient boosting is also a machine learning algorithm used for classification and regression that is constructed from decision tree algorithms. Different from random forest, gradient boosting is a boosting technique, which works by building weaker prediction models sequentially where each model tries to predict the error left by the previous one. According to the gradient boosting regression, the R square value is about 0.99896.

Since the target variables are very skewed as the values of R square are rather close to 1, we should address the skewness before applying any regression model. Therefore, log transformation is an appropriate way that can be used to remove skewness from the predictor. (Dario [9]) By log transforming y , the skew coefficient changed from 3.8372 to 0.9349.

Besides, to address the multicollinearity problem, since Cost_continue, Cost_last, and Cost_initial all are very correlated with each other as the correlation coefficient is about 0.99, we can drop them and only keep Cost_initial in consideration.

By using linear regression and lasso regression model again for new transformed data, the values of

R square are about 0.53246 and 0.52546 respectively. We also compare the R squares inferred from both random forest regression and gradient boosting regression, which are about 0.934734 and 0.93939285 respectively. Therefore, gradient boosting regression is the most appropriate model since its R square value approaches 1 the most, meaning that the model fits the data best.

Conclusion

In this research, four different machine learning regression algorithms have been applied to develop and train the cancer cost prediction model, which are linear regression, lasso regression, random forest regression, and gradient boosting regression. Since several factors are highly correlated with each other, we use log transformation to reduce the skewness. Based on our data, gradient boosting regression is the best machine learning algorithm since its R square is higher than that of other three algorithms, which is about 0.93939285, meaning that it fits the data well. Therefore, we can use gradient boosting regression to predict the cost of cancer treatment in the United States in the future. And we also hope to improve the dataset in order to produce the model with higher accuracy.

References:

1. Watson IBM. "Expenditures for Cancer Care – Dataset by Xprize Ai-Health." Data.world, 19 July 2017. URL: <https://data.world/xprizeai-health/expenditures-for-cancer-care/workspace/project-summary?agentid=xprizeai-health&datasetid=expenditures-for-cancer-care>
2. "Cancer Costs U. S. Patients \$21 Billion a Year." US News. URL: <https://www.usnews.com/news/health-news/articles/2021-10-26/cancer-costs-us-patients-21-billion-a-year>
3. Selby Karen. "Americans Can't Keep Up with the High Cost of Cancer Treatment." Mesothelioma Center – Vital Services for Cancer Patients & Families, 20 Aug. 2021. URL: <https://www.asbestos.com/featured-stories/high-cost-of-cancer-treatment>
4. "Financial Burden of Cancer Care." Financial Burden of Cancer Care, 20 July 2021. URL: https://progressreport.cancer.gov/after/economic_burden.
5. "The American Cancer Society Cancer Action NetworkSM (ACS CAN). Is Making Cancer-and the Affordability of Cancer Care-a Top Priority for Public Officials and Candidates at the Federal, State and Local Levels". The Costs of Cancer, Oct. 2020. URL: <https://www.fightcancer.org/sites/default/files/National%20Documents/Costs-of-Cancer-2020-10222020.pdf/> Accessed: 16 Dec. 2021.

6. Yadav D. Categorical encoding using label-encoding and one-hot-encoder. Medium. (2019, December 9). Retrieved January – 22, 2022. URL: from <https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd>
7. Editor M. B. (n.d.). Regression analysis: How do I interpret R-squared and assess the goodness-of-fit? Minitab Blog. Retrieved January 22, 2022. URL: from <https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>
8. Chakure A. Random Forest and its implementation. (2020, November 6). Medium. Retrieved January 22, 2022. URL: from <https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f>
9. Radečić D. Top 3 methods for handling skewed data. (2020, January 4). Medium. Retrieved January 22, 2022. URL: from <https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45>
10. Parleto A. Deal multicollinearity with lasso regression. (2020). Andrea Perlatto. Retrieved January 22, 2022. URL: from <https://www.andreaperlato.com/mlpost/deal-multicollinearity-with-lasso-regression>