EXPLORING HEART DISEASE IN CALIFORNIA

Senior Project Internal Documentation

According to the Centers for Disease Control, heart disease remains the leading cause of death. This project aims to visually explore heart disease metrics in California.

Created by Jarod Morris

Advised by Dr. Lubomir Stanchev

Table of Contents

Motivation:

According to the Centers for Disease Control, heart disease remains the leading cause of death in California. My project is an analytic tool that seeks to quantify and evaluate the degree of scientific and clinical research efforts aimed at reducing the heart disease burden. This metrics and analytic platform seek to assess the level of team science collaboration involved in the research and interventions (clinical trials) in the heart disease space among leading publicly funded medical academic institutions (in California).

Data Sources:

California Research Institutions –
This project focuses on the publication data from the list of teaching colleges in California from Wikipedia: https://en.wikipedia.org/wiki/Category:Teaching_hospitals_in_California

Heart disease Related Terms -
The list of terms used to query publication and funding data relevant to heart disease:
https://www.webmd.com/heart-disease/heart-disease-glossary-terms

Publications -
Queried publications from 2015-2021 PubMed related to Heart Disease Terms for each affiliation:
https://pubmed.ncbi.nlm.nih.gov/

Example Query:
(Angina OR Angiogram OR cardiac catheterization OR Anticoagulant OR Atherosclerosis OR hardening of the arteries OR Beta-blocker OR Calcium-channel blocker OR Catheter OR Coronary artery disease OR Dyspnea OR Electrocardiogram OR ECG OR EKG OR Heart attack OR myocardial infarction OR Heart-lung bypass machine OR cardiopulmonary OR Heart surgery OR Ischemia OR Off-pump heart surgery OR Plaque)
AND Harbor–UCLA Medical Center[affiliation]
AND (2015:2021[pdat])

Project Funding Data –

Funding information from 2015-2021 gathered from NIH Reporter based on heart disease related publication / projects. Data includes Project Total Cost, Fiscal Year, County, Project Terms:

https://reporter.nih.gov/

Mortality Data –

Queried Population, Death Count, Cause of Death from 2015-2019 for mortality in California:

https://wonder.cdc.gov/Deaths-by-Underlying-Cause.html

2015-2019 Data by cause of death in each County:

https://data.chhs.ca.gov/dataset/death-profiles-by-county

Impact Factor by Journal –

JCR Clarivate calculated impact factor by journal:

https://impactfactorforjournal.com/journal-impact-factor-list-2019/

Water Quality –

Tables for lead sample amount by test site in California from 2015-2021:

https://ordspub.epa.gov/ords/sfdw/f?p=108:35::::::

Air Quality –

Tables for air quality data by county in California from 2015-2021:

https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

ZipCode by County –

Scraped websites XML to build table for California ZipCode to County information:

https://www.zip-codes.com/state/ca.asp

Code is in scripts/scrapeZipCodes.py

List of Affiliations per Publication –

Developed Python script using xpath to grab xml affiliation data from PubMed and create table using pandas with pmid, author name and affiliations. Generated 150k rows

Code is in scripts/affiliatonParser.py

Cleaning Data and Staging:

Python – Created python scripts to process tables:

- addAffColumn.py - to add associated affiliation as column to pub-med tables downloaded.
- mergeAQI.py – merge AQI tables for each year into one table.
- mergeWater.py – merge water content tables for each year into one table.

Code is in /scripts

DataBase – Provisioned an AWS – RDS – Database using SQL Server Express Edition. This database will hold the tables & views created from the data sources.

Microsoft Access – Create Initial Database and schema for tables. Imported all csv and excel tables into Access Database for easy import into SQL server. Set appropriate types for attributes across datasets.

SQL Table/View Creation:

Tables & Relevant Columns:

deaths_county_14-19 – Year, County Name, Strata, Cause of Death, Count

cdcWonder14-19 – Ten Year Age Group, Heart Disease Type, Hispanic Origin, Race, Gender, Deaths, Population & Year

zipCode – Zipcode, County Name

affPublicPrivate – Institution, Type

aqiCounty – County Name, Year, Days with [High/Med/Low] AQI

waterCounty - Zipcode, Year, Sample Measure (mg L)

affCounty – County Name, Affiliation

tableJCR_2019 – Impact Factor, Journal Name (ISO_Abbrev)

tablePublications – Journal Name, Publication Year, Affiliation, PMID

HeartDiseaseTerms – Term Name, Term Description

tableAffiliations – PMID, Affiliation

predictionModel – Year, County Name, Publication Count, Project Count, Total Cost, Average Lead Content, Bad Air Days, Death Count (heart disease), Death Count (all causes), LR, EBM, XGBoost, RFR

featureTable – feature index, LR, EBM, XGBoost, RFR

Views:

jcrJoin – Publication, Impact Factor

Design: Inner Join tableJCR_2019 on tablePublications, using attribute Journal Name

researchFunding – Project Terms, Title, Start-End Date, Organization, City, County

Design: Inner join on list of ZipCodes to County, Filter tableFunding for Projects which contain Heart Disease in the project Terms.

popDeath – Year, County Name, Cause of Death, Death Count (all causes)

Design: Grab total death count per county, per year for heart disease

mortalityView – Year, County Name, Strata, Cause of Death, Death Count (heart disease), Death Count (all causes)

Design: Inner join deaths_county_14-19 on popDeath, group by all columns

mortalityMapView – Year, County Name, Strata, Cause of Death, Death Count (heart disease), Death Count (all causes)

Design: Filter out total population and gender strata from mortalityView

InterCount – Affiliation, Inter Publication Count

Design: with statement to combine multiple views from table Affiliations

Infrastructure Build Out:

Tableau Desktop–

- Connected to AWS-RDS and imported tables and views. Created Tableau relationships between data.
- Creation of dashboards, connecting sheets of related information.

ReactJS -

- Created a website to host Tableau visualizations. Each page embeds a tableau dashboard. The website files are in /website.
- The website feature routes and uses a navigation bar.
- You can view the JavaScript files on a private repository: https://github.com/jzmorris/heartdiseaseviz
- The website is hosted using netlify, alongside a custom domain name https://www.heartdiseaseviz.com

Tableau Data Visualization:

Dashboard 1: Heart Disease Mortality by Demographics

Sheets: Mortality by Race, Mortality by Ethnicity, Mortality across the Lifespan, Mortality by Gender

Tables used: cdcWonder14-19

Global Filter by Year, Heart Disease Group, Race & Gender

Dashboard 2: Mortality Trends by County

Sheets: Top 10 Counties by Percentage of Total Deaths

Tables used: mortalityMapView

Global Filter on Race & Year

Dashboard 3: Research Output by Teaching Hospital

Sheets: Top 10 Publication Count, Public/Private Publication Count

Tables used: tablePublications, affPublicPrivate

Filter for top 10 by Public or Private institution.

Dashboard 4: Research Output by Impact Factor

Sheets: Top 10 Journals, Top 10 Institutions, Inter Publication Count

Tables used: tablePublications, jcrView, InterCount

Filter by Publication Year and Impact Factor

Dashboard 5: Federal Research Funding

Sheets: Top 10 Federally Funded Institutions, Trendlines for Research Funding, Publication Count & Mortality

Tables used: researchFunding

Filter by Top 10 Federally Funded Institutions by Fiscal Year

Dashboard 6: Comparing Maps by County

Sheets: Maps for Research Funding, Publication Count & Mortality by County

Tables used: researchFunding, affCounty, tablePublications, mortalityMapView

Filter each Map by Year

Dashboard 7: Prediction Map

Sheets: California Map, Death Percentage Prediction Lines

Tables used: predictionModel

Filter by 3 select Counties using selector on Map.

Dashboard 8: Explaining the Models

Sheets: Linear Regression, Random Forests, XGBoost & EBM Feature Importance’s

Tables used: featureTable

Prediction Model:

For the second part of the project, we want to predict the percentage of deaths related to heart disease by county in California.

I have created a set of prediction using the following models (Explainable Boosting Machine, Linear Regression, Random Forests and XGBoost)

The data frame used to train our predication models will consist of data from 2015 - 2019 for all counties in California. There is an 80/20 split on the sets, using 2015-2018 as our training years and 2019 as the testing year.

The training set consists of the following attributes Year, County, Air Quality Metric, and Average Lead Content in Water.

The output variable used to train these models is the percentage of deaths related to heart disease.

Code is in scripts/models.py

What Models Are Used in our Predications?

This process of combining the output of multiple individual models (also known as weak learners) is called **Ensemble Learning**.

Explainable Boosting Machine (EBM):

EBM is a glassbox model, designed to have accuracy comparable to state-of-the-art machine learning methods like Random Forest and Boosted Trees, while being highly intelligible and explainable.

A generalized additive model (GAM) is a white box model that is more flexible than logistic regression, but still interpretable. EBM uses GA^2M, which is a GAM with more complicated interpretations of the predictions (Frankowski).

EBM learns each feature function using modern machine learning techniques such as bagging and gradient boosting.

Because EBM is an additive model, each feature contributes to predictions in a modular way that makes it easy to reason about the contribution of each feature to the prediction.

Linear Regression:

Using the SKlearn Linear Regression Model. Linear Regression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation (SciKit).

Coefficients for the linear regression model represent the weight of an attribute on the output variable.

eXtreme Gradient Boosting (XGBoost):

XGBoost implements the gradient boosting decision tree algorithm.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models (Brownlee).

In this case we will be using the regression predictive model in order to predict the percentage of heart disease in a specific county.

XGBoost was developed for speed and performance. It comes with built-in cross validation.

Random Forests:

A decision tree is simply a series of sequential decisions made to reach a specific result (Sharma).

Random Forest is a tree-based machine learning algorithm that leverages the power of multiple decision trees for making decisions.

What features did we use to train our model?

Attributes collected per County in California:

- Death Count (heart disease)
- Death Count (all Causes)
- Year
- Project Count
- Total Cost
- Publication Count
- Combined Number of Bad Air Days
- Avg Lead Content in Water

How do we determine how accurate is our model?

Coeefficient of Determination (R2 Score) –

Defined as the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s) (Maini).

$$R2 = 1 - SS_{res} / SS_{tot}$$

SS_{res} is the sum of squares of the residual errors.
SS_{tot} is the total sum of the errors.

How well did the various Models perform?

XGBoost:

This model had an r2 score 0.844 and an accuracy of 96.5% on the testing set.

It can be referred that 84% of the changeability of the dependent output attribute can be explained by the model while the remaining 16% of the variability is still unaccounted for.

R^2 indicates the proportion of data points which lie within the line created by the regression equation. A higher value of R^2 is desirable as it indicates better results (Miani).

Feature Importance:

Based on the feature importance's calculated by the sklearn api, we can see that the Death Count, Year and Population have the greatest importance when determining the output variable.

The Average lead content is a counties water was about twice as importance as the bad Aqi metric.

Publication, Project costs and project counts were the least relevant attributes in the model.

Explainable Boosting Machine:

This model had an r2 score of 0.7236 and an accuracy of 98.04% using the testing set.

Feature Importance:

Based on our mean absolute score we can view the weights of each individual feature. For the EBM model we can see that the combined number of bad air days importance is higher than in the xgBoost model. The EBM Model found the Year and total Population to contribute most heavily to the output percentage.

Linear Regression Model (per County model):

This model had an accuracy of 99.79%. The r2 score was not applicable to this model since a separate model was created for each county. The dataset for each county is limited to just one year for each county in the testing set.

Feature Coefficients:

The Death count (all causes) had the greatest weight when fitting a linear model to our data points. The publication count was the last notable feature to have a large impact towards the fitting of the model.

Random Forests Regression:

This model had an r2 score of 0.862 and an accuracy of 98.3% on the testing set.

Feature Importance:

The Population and total count had the highest importance in the decision trees that were generated by the random forest algorithm. We can see that the number of combined bad Air days had slightly more weight than the average lead content of the water in a county. The Year had the least visible weight.

References:

Brownlee, Jason. “A Gentle Introduction to XGBoost for Applied Machine Learning.” *Machine Learning Mastery*, 16 Feb. 2021, https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.

Frankowski, Dan. “A Gentle Introduction to GA2Ms, a White Box Model.” *Blog*, 3 June 2019, https://blog.fiddler.ai/2019/06/a-gentle-introduction-to-ga2ms-a-white-box-model/.

Maini, Ekta. “Python - Coefficient of Determination-R2 Score.” *GeeksforGeeks*, 16 July 2020, https://www.geeksforgeeks.org/python-coefficient-of-determination-r2-score/.

Sharma, Abhishek. “Decision Tree vs. Random Forest - Which Algorithm Should You Use?” *Analytics Vidhya*, 12 May 2020, https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/.

“Sklearn.linear_model.Linearregression.” *Scikit*, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.