ATU™ ARKANSAS TECH UNIVERSITY

April 5, 2024

Analyzing the Impact of Socioeconomic Factors on Cancer Clinical Trails Accessibility in the U.S. Using Machine Learning

Presenters:
Krysta Ray
Hiromi Honda

Understanding Cancer Clinical Trials

- Clinical trials test new ways to find prevent and treat cancer.
- Trail placement and selection is meant to be randomized to prevent bias.
- Many socioeconomic and geographical factors are believed to be a barrier to trail access
- Predictive modeling may reveal solutions by:
  - Uncovering patterns and insights to trial access
  - Identifying the factors that lead to disparities in access
  - Using insights to propose polices to improve healthcare equity

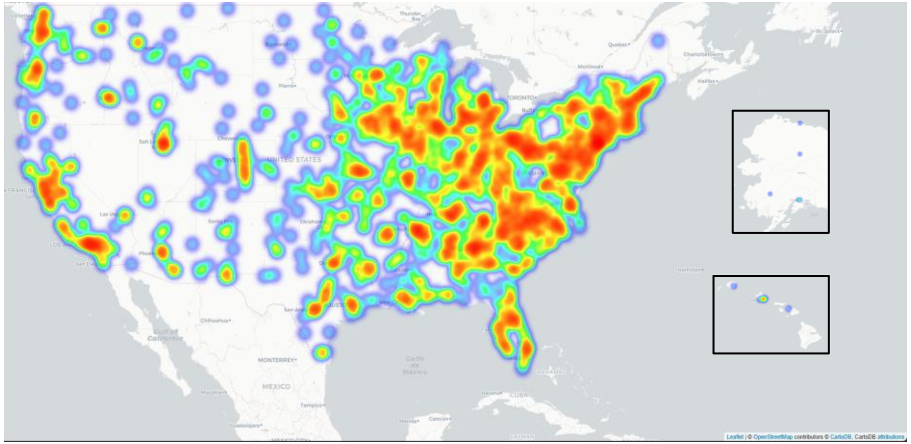


Image 1. Map of U.S. showcasing clinical trials offered using Leaflet

Previous Trial Participation and Access Studies

- Few machine learning studies but many statistical studies
- Patients with lower income (<50k) were 29% less likely to participate
- Higher populated areas were found to conduct more trials
- 5-year relative survival for all cancers combined is 14% lower among residents of poorer counties

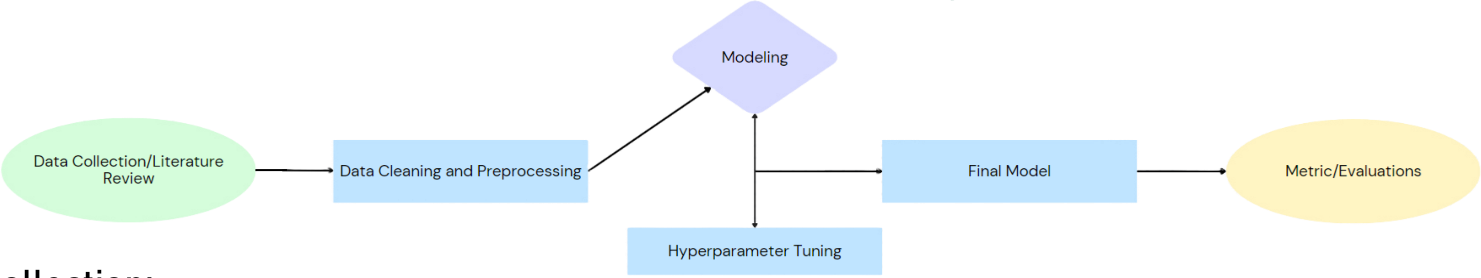
<5%

of adult cancer patients enroll in cancer clinical trials

70%

of adult cancer patients are willing to participate

Methodology


Data Collection/Literature Review
Data Cleaning and Preprocessing
Modeling
Hyperparameter Tuning
Final Model
Metric/Evaluations

1. Data Collection:

- *Data collected from secondary data set created by Noah Ripper (US Census Bureau, ClinicalTrials.gov, and other sources for counties in the US from 2010-2015)*

2. Data Preprocessing:

- *Clean and prepare data for analysis*
- *Remove duplicates and handle missing values*
- *Convert data types as needed*
- *Group data by county instead of zip code*

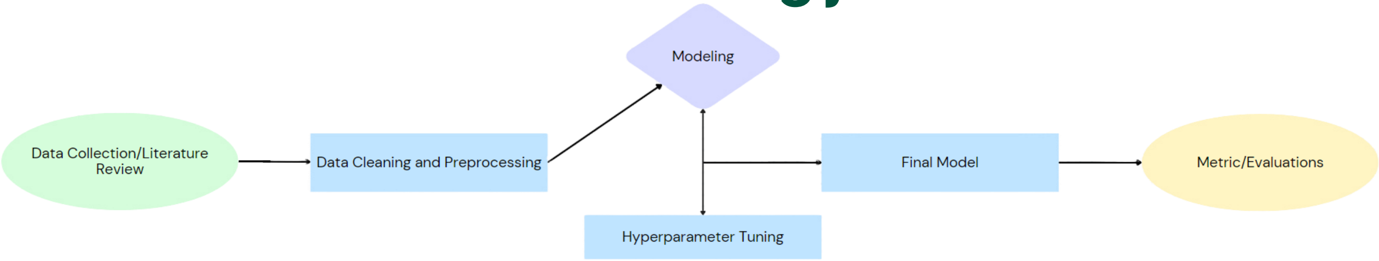
3. Model Development:

- Correlation Analysis
  - *Determine target and features*
- Regression Models
- Determine best performing model
  - *Gradient Boosting*
  - *Random Forest*
  - *Linear Regression*
  - *K Neighbors*

4. Hyperparameter Tuning

- Selecting parameters for optimal results from the model

Methodology *contd.*


Data Collection/Literature Review
Data Cleaning and Preprocessing
Modeling
Hyperparameter Tuning
Final Model
Metric/Evaluations

4. Model Evaluation:
   - *Evaluate model performance using metrics: RMSE, MSE, MAE, R2 score*
   - *Iterate through modeling process until best accuracy is received*
5. Results and Insights:
   - *Analyze and interpret model predictions*
   - *Identify socioeconomic patterns that contribute to trail access and participation*
   - *Provide actionable insights for the healthcare industry and policymakers*

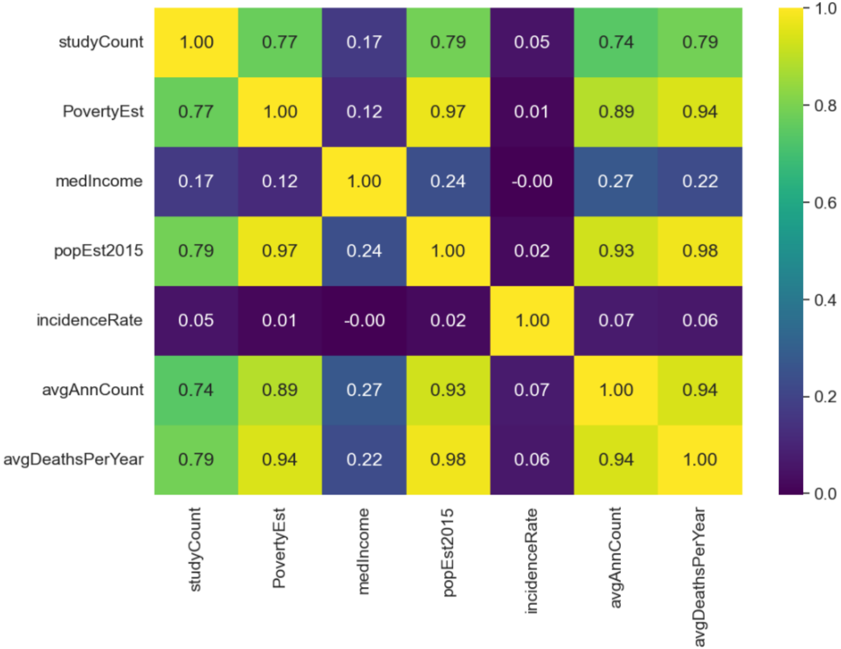
Data Collection and Preprocessing

- A secondary data set created by Noah Ripper includes trail data from the US Census Bureau, ClinicalTrials.gov, and other sources for counties in the US from 2010-2015
- Four goals were identified for data preprocessing:
  - Initial reading
  - Type conversion
  - Missing data handling
  - Erroneous data handling
- Rows were listed by zip code and grouped by county. The sum of studies in each zip code was added to each county
- Encoding any string values using a label encoder

Initial Data Insights

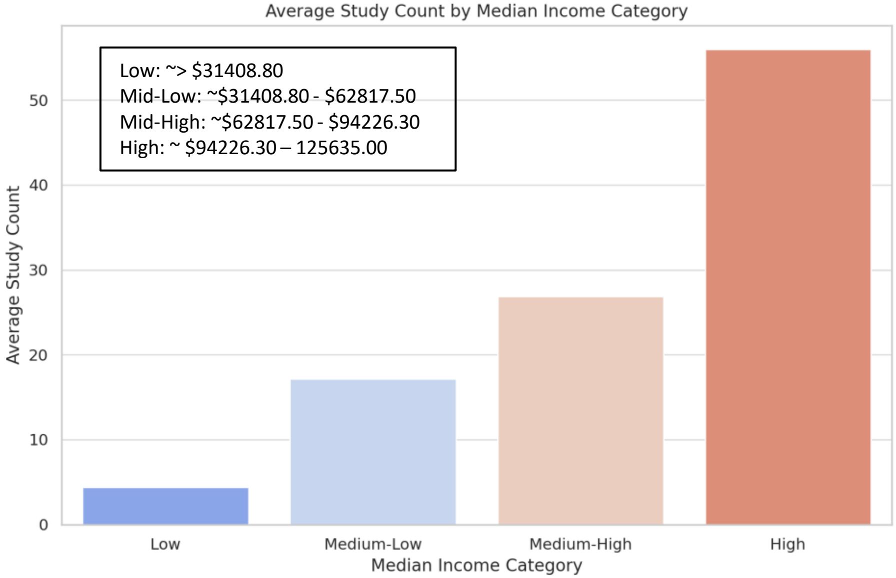
- Socioeconomic factors seemed to correlate highly with the number of studies offered.
- Using the correlation and feature importance analysis, we selected the features that affected the target (most to least important).
  - popEst2015
  - PovertyEst
  - incidenceRate
  - avgDeathsPerYear
  - medIncome

Correlation Analysis

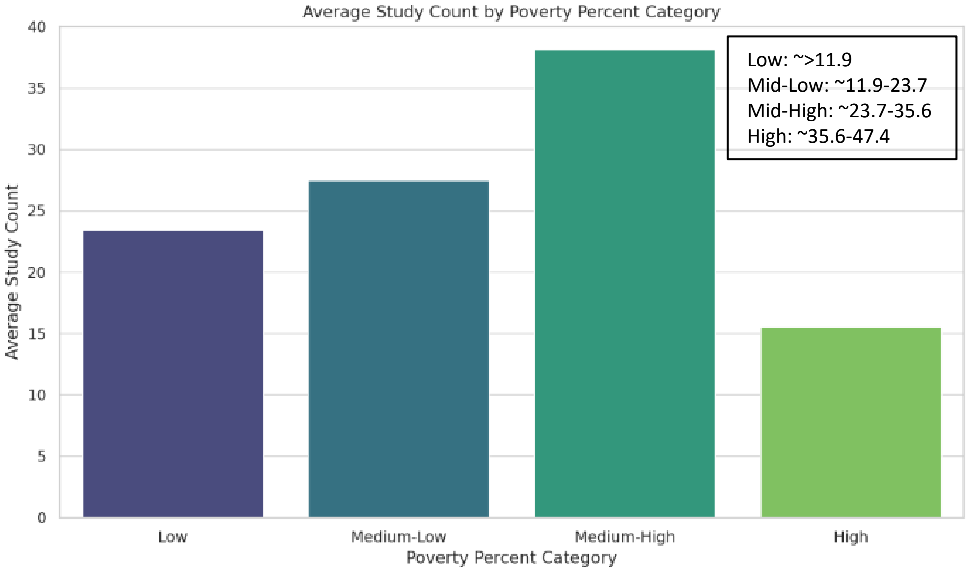


	studyCount	PovertyEst	medIncome	popEst2015	incidenceRate	avgAnnCount	avgDeathsPerYear
studyCount	1.00	0.77	0.17	0.79	0.05	0.74	0.79
PovertyEst	0.77	1.00	0.12	0.97	0.01	0.89	0.94
medIncome	0.17	0.12	1.00	0.24	-0.00	0.27	0.22
popEst2015	0.79	0.97	0.24	1.00	0.02	0.93	0.98
incidenceRate	0.05	0.01	-0.00	0.02	1.00	0.07	0.06
avgAnnCount	0.74	0.89	0.27	0.93	0.07	1.00	0.94
avgDeathsPerYear	0.79	0.94	0.22	0.98	0.06	0.94	1.00

Initial Data Insights


Average Study Count by Median Income Category

Low: ~> $31408.80
Mid-Low: ~$31408.80 - $62817.50
Mid-High: ~$62817.50 - $94226.30
High: ~ $94226.30 – 125635.00


Average Study Count by Poverty Percent Category

Low: ~>11.9
Mid-Low: ~11.9-23.7
Mid-High: ~23.7-35.6
High: ~35.6-47.4

Model Comparison

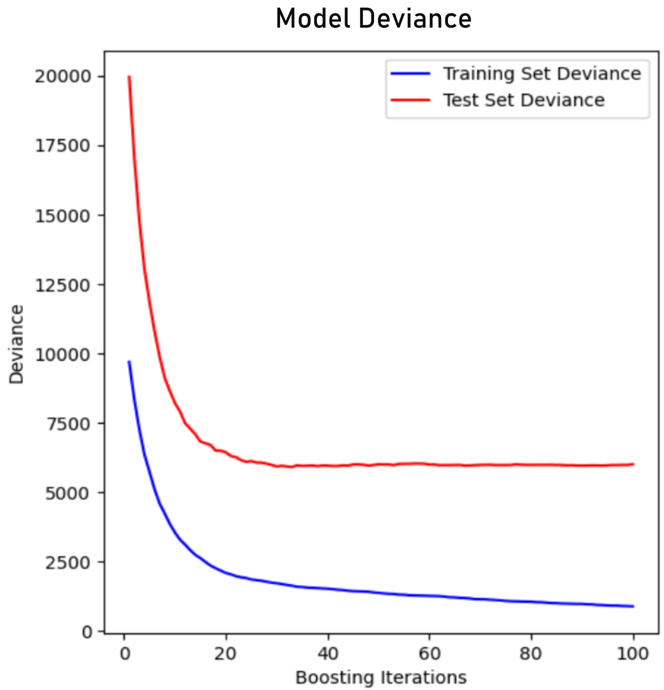
Comparison of Regression Models

- Four different types of regression models were used to predict where trials would be.
- Gradient boosting regression had better overall performance compared to the others.
- Preliminary model accuracy was 74.47%

Metric	Gradient Boosting	Random Forest	Linear Regression	K Neighbors
R2 Score	0.7447	0.7279	0.6716	0.6225
Root Mean-Squared Error	77.53	80.05	87.93	94.28
Mean Squared Error	6010.86	6407.50	7732.33	8887.83
Mean Accuracy Error	21.27	21.42	24.71	26.04

Gradient Boosting Regression

- Deviance is the goodness-of-fit statistic for a statistical model
- A model can be improved by plotting the deviance and determining whether the model is underfitting or overfitting
  - Overfitting: When the model gives accurate predictions for training data but not for test data
  - Underfitting: When the model does not recognize the relationship between the dependent and independent variables
- Slight Overfitting but close to the ideal fit



Prospective: Next Steps

Advancing Cancer Clinical Trial Access in the US

- Currently, we are working to improve the accuracy of the model. It has improved to ~78% so far.
- Our goal is to have an accuracy of at least 90% by the end of this research.
- We plan to achieve more precise and actionable insights with our model to guide interventions and improve equitable access to care

- Machine learning models are continually being improved and enhanced!
- Hyperparameter tuning will play a big part in improving our model's performance
- Re-analyze data for any missed outliers
- By identifying and understanding the factors that lead to disparities in clinical trial access, we can propose targeted interventions and policies to improve healthcare equity and cancer treatment outcomes across the United States.

Acknowledgements

- Dr. Robin Ghosh, Hiromi Honda, Musfikur Rahaman
- Dept. Of Engineering & Computing Sciences
- Arkansas Tech University, Russellville, Ar-72801

References

- Islami F, Guerra CE, Minihan A, Yabroff KR, Fedewa SA, Sloan K, Wiedt TL, Thomson B, Siegel RL, Nargis N, Winn RA, Lacasse L, Makaroff L, Daniels EC, Patel AV, Cance WG, Jemal A. American Cancer Society's report on the status of cancer disparities in the United States, 2021. CA Cancer J Clin. 2022 Mar;72(2):112-143. doi: 10.3322/caac.21703. Epub 2021 Dec 8. PMID: 34878180.

- National Cancer Institute. (2024, March 8). *How do clinical trials work?*. National Cancer Institute. https://www.cancer.gov/research/participate/clinical-trials/how-trials-work

- Price, K. N., Lyons, A. B., Hamzavi, I. H., Hsiao, J. L., & Shi, V. Y. (2020a). Facilitating clinical trials participation of low socioeconomic status patients. *Dermatology*, *237*(5), 843–846. https://doi.org/10.1159/000511889

- Unger JM, Cook E, Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. Am Soc Clin Oncol Educ Book. 2016;35:185-98. doi: 10.1200/EDBK_156686. PMID: 27249699; PMCID: PMC5495113.

- Unger JM, Gralow JR, Albain KS, Ramsey SD, Hershman DL. Patient Income Level and Cancer Clinical Trial Participation: A Prospective Survey Study. *JAMA Oncol.* 2016;2(1):137–139. doi:10.1001/jamaoncol.2015.3924

- Data set created by Noah Ripper: Cancer Trials - dataset by nrippner | data.world

Questions?

ATU™
ARKANSAS TECH
UNIVERSITY

Thank you!

ARKANSAS TECH UNIVERSITY