SL4HOI: Supervised Learning for Predicting Health Opportunity Index across States

Zhiyuan Song*
University of Virginia
Charlottesville, Virginia, United States
bfy8kq@virginia.edu

Zihan Mei*
University of Virginia
Charlottesville, Virginia, United States
zm4hy@virginia.edu

Liran Li*
University of Virginia
Charlottesville, Virginia, United States
zqj6pe@virginia.edu

N. Rich Nguyen
University of Virginia
Charlottesville, Virginia, United States
nn4pj@virginia.edu

ABSTRACT

Health disparity is a critical issue, with access to health opportunities unevenly distributed across populations. The Virginia Health Opportunity Index (HOI) provides a comprehensive measure of the social determinants of health (SDH), yet its complex computation limits its applicability beyond Virginia. This research aims to simplify the HOI calculation process and extend its utility to other states by developing a Supervised Learning (SL) model using readily available data from the American Community Survey (ACS). We acquire and process ACS data for Virginia, train and validate a Random Forest model to predict HOI, and test its applicability in North Carolina and California. Our model demonstrates robust performance, with positive correlations between predicted HOI and life expectancy and low p-value in all states tested. This study has implications for public health policy, enabling more accessible and generalizable tools for assessing health opportunities and facilitating targeted interventions to promote health equity.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Supervised learning by regression*; • **Applied computing** → **Health informatics**.

KEYWORDS

HOI, Supervised learning, Feature learning, Regression Tasks

ACM Reference Format:
Zhiyuan Song, Zihan Mei, Liran Li, and N. Rich Nguyen. 2024. SL4HOI: Supervised Learning for Predicting Health Opportunity Index across States. In *Proceedings of (KDD-UC '24).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

*Zhiyuan, Zihan, and Liran are undergraduate students. They contributed equally to this research. Professor N. Rich Nguyen is the research advisor.

KDD-UC '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Health is a fundamental human right and a critical component of overall well-being [1]. However, access to health opportunities is not evenly distributed across populations, leading to disparities in health outcomes [2]. To address this issue, it is essential to develop tools that can accurately assess and predict health opportunities at a granular level. The Virginia Health Opportunity Index (HOI) is one such tool that provides a comprehensive measure of the social determinants of health (SDH) across the state of Virginia [3, 4].

The HOI is a complex measure that incorporates over 30 variables, 13 indicators, and four profiles to create a single, composite index. While the HOI offers valuable insights into health opportunities, its computation relies on intricate data processing techniques, limiting its accessibility and applicability beyond Virginia [3].

This study aims to simplify the HOI calculation process and extend its utility to other states by developing a machine learning model that can predict HOI using readily available data from the American Community Survey (ACS) [5]. By leveraging machine learning techniques, we seek to create a more accessible and generalizable tool for assessing health opportunities information, enabling policymakers, public health professionals, and researchers to identify areas of need and develop targeted interventions to improve health outcomes [6, 7].

Our approach involves acquiring and processing ACS data for Virginia, training and validating a machine learning model to predict HOI and testing the model's applicability to other states (North Carolina and California). We evaluate the model's performance using various metrics and validate its usefulness by examining the correlation between predicted HOI and life expectancy data from the Institute for Health Metrics and Evaluation (IHME) [8, 9].

The development of a streamlined model for predicting HOI democratizes access to this index, facilitating its use in diverse contexts and contributing to the development of evidence-based public health policies and interventions [10]. By identifying areas with lower predicted HOI values, policymakers can allocate resources to address health disparities and promote health equity across communities.

2 RELATED WORK

The Virginia Health Opportunity Index (HOI) is a comprehensive tool developed by the Virginia Department of Health to

measure SDH and identify health disparities across the state of Virginia. [3, 11]. Ogojiaku et al.'s study, "The Health Opportunity Index: Understanding the Input to Disparate Health Outcomes in Vulnerable and High-Risk Census Tracts" [3], offers a detailed analysis of the HOI. Their research highlights how the HOI can pinpoint areas with significant health disparities, enabling targeted public health interventions. The study emphasizes the complexity of the HOI, noting that its calculation involves extensive data processing and integration of various socioeconomic and environmental factors. Other notable indices include:

- **Social Vulnerability Index (SVI):** An index used to evaluate community-level vulnerability and resilience, primarily in the context of disaster response and healthcare accessibility [12].
- **Area Deprivation Index (ADI):** A census-based measure gauging regional deprivation and its influence on health [13].

Both indices share a common goal with the HOI: to provide a detailed understanding of the SDH and to help design better public health strategies [3, 12, 13].

Predictive Models in Public Health. Machine learning models have been increasingly utilized in public health to predict outcomes and identify patterns within complex datasets [6]. Ensemble methods like Random Forests and XGBoost have shown high efficacy in capturing non-linear relationships within health data [14, 15]. These models are particularly useful for handling high-dimensional data and making accurate predictions.

Public Health Policy Applications. The application of predictive models in public health policy has significant implications for resource allocation, intervention planning, and prioritization of under-developed regions. Predictive analytics have been effectively used to guide public health strategies and improve health outcomes [16]. For instance, predictive models have been employed to allocate resources during the COVID-19 pandemic, ensuring that interventions are targeted at the most vulnerable populations [17].

Building on the foundational work of Ogojiaku et al. [3] and leveraging insights from machine learning and public health policy research, our study aims to enhance the applicability of the HOI.

3 METHODOLOGY

To ensure the reliability and robustness of our model in predicting the HOI and its applicability across various states, we implemented comprehensive testing methods. This included rigorous steps in data collection, data preprocessing, model training, and testing.

3.1 Data set

To reduce the need for extensive data pre-processing and enhance accessibility beyond Virginia, we chose to use raw data from the ACS, instead of using the processed data provided by the Virginia Department of Health [18, 19]. This shift enables us to develop a more general model for predicting HOI, making it usable for ordinary individuals and applicable across various states.

For this study, we specifically gathered data concerning Census tracts in Virginia for the period from 2013 to 2017, aligning with the time frame of the VDH_VA_HOI dataset. Out of 249 available tables, we meticulously selected 21 tables that are closely related to health conditions [20]. This selection includes data on employment status, poverty status, and health insurance coverage, among others.

Table 1: Summary of Datasets Used in the Study

Dataset Name	Description
VDH_VA_HOI	Health Opportunity Index Dataset for Virginia
ACS_VA	2013-2017 ACS Virginia data
ACS_CA	2013-2017 ACS California data
ACS_NC	2013-2017 ACS North Carolina data

A complete list of the tables utilized in our study is detailed in the appendix of this paper.

After acquiring Virginia data from the American Community Survey (ACS), we integrated the HOI values from the VDH_VA_HOI dataset into the ACS_VA dataset based on matching census tracts. This integration allowed us to train and validate regression learning models to predict the HOI accurately.

Using a similar approach, we acquired the 2013-2017 ACS data for California, referred to as the ACS_CA dataset. We ensured that the features in the ACS_CA dataset precisely matched those in the ACS_VA dataset to maintain consistency in the model application. Additionally, we acquired the 2013-2017 ACS data for North Carolina (ACS_NC) in the same manner to further test the generalizability of our model.

3.2 Data processing

The data acquired from ACS require processing before they are suitable for model training. However, our approach contrasts sharply with the extensive data transformations typically performed by the Virginia Department of Health, which involve numerous complex statistical methods [3].

- **Dealing with Missing Data:** Our dataset contains some columns and entries with missing values. To address this, we have implemented a succinct approach: where data is absent, we either remove these entries or impute the missing values with the median of the corresponding data.
- **Feature Selection:** To enhance the model's performance and interpretability, we selected features with a correlation coefficient of at least 0.25 with the HOI. This threshold ensures that only variables with a meaningful relationship to HOI are included, thereby reducing noise and focusing on the most impactful predictors.
- **Combining Features:** ACS provides an extensive breakdown of variables such as income levels and housing prices, often including dozens of columns for a single parameter. To streamline our analysis and reduce complexity, we have consolidated these classifications into three broad categories: low, medium, and high. This helps us to avoid overfitting when training our model. An example is illustrated in 1.
- **Scaling:** The data exhibit considerable variability in measurement scales. To mitigate potential issues during model training arising from this heterogeneity, we employ a Standard Scaler [21]. This scaling process normalizes the data, ensuring that all features contribute equally to the model's performance and improving the reliability of our predictions.
- **Principal Component Analysis (PCA):** Our dataset contained 54 features after previous data processing steps, which

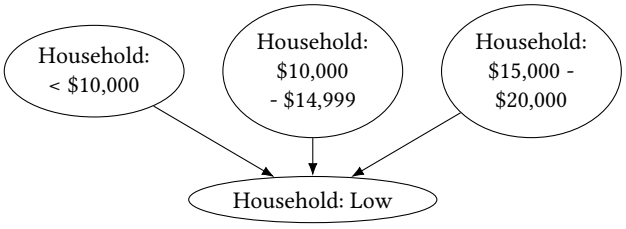

Figure 1: Example of column aggregation

is still highly complex. Consequently, PCA [22] was applied to reduce the dimensionality of the dataset and enhance model performance, since many of the features are highly unrelated. By transforming the original features into a set of uncorrelated components, PCA captures the most significant variance with fewer dimensions, simplifying the model and reducing computational complexity. This approach is consistent with previous studies on the HOI, "The Health Opportunity Index: Understanding the Input to Disparate Health Outcomes in Vulnerable and High-Risk Census Tracts" [3], which also utilized PCA for data reduction and simplification.

In this study, we chose a threshold of 99% variance for PCA. The choice of 99% variance ensures that almost all the information from the original dataset is retained, providing a balance between dimensionality reduction and information retention. PCA helps to reduce the dataset from 54 features to 35, simplifying the model and making it computationally more efficient. Additionally, fewer components reduce the risk of overfitting, leading to better generalization of new data.

The scree plot (Figure 2) demonstrates that the cumulative explained variance increases rapidly with the first few components and then plateaus. By the 35th component, 99% of the total variance is captured, indicating that these 35 components retain nearly all the information from the original dataset while reducing dimensionality [23].

The 3D PCA scatter plot comparison (Figure 3) shows the distribution of data points in the first three principal components for both the full set of components and the 35 components that capture 99% variance. The plots illustrate that reducing the number of components to 35 does not significantly alter the overall structure and spread of the data, confirming that the chosen components sufficiently represent the original dataset.

3.3 Training Model for Virginia

We trained regression models for predicting Virginia's HOI values based on ACS_VA and VDH_VA_HOI datasets. We chose to focus on tree-based models due to their efficiency and effectiveness in making predictions [24]. Specifically, we tried Decision Tree, Random Forest, and XGBoost models [25]. Linear regression was not considered ideal for this study because our study works in high dimensional settings [26]. The complexity and interactions between different SDH focus on non-linear, making tree-based models more suitable in our setting [27, 28].

To optimize the performance of our model, we employed a grid search technique [29] combined with 5-fold cross-validation [30]

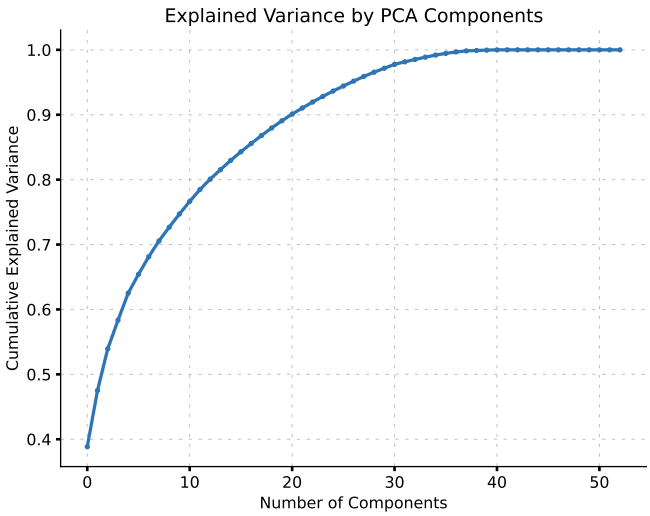

Figure 2: Scree plot: Explained Variance by PCA Components

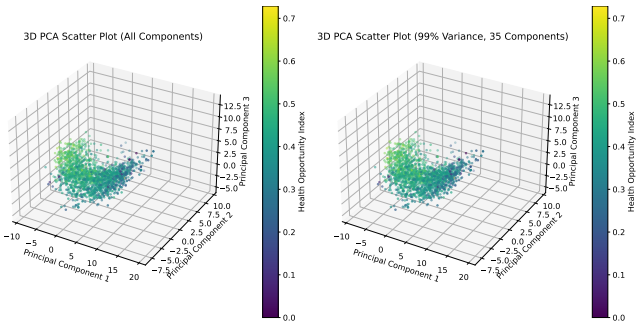

Figure 3: Comparison of 3D PCA Scatter Plots: All Components vs. 35 Components (99% Variance)

to conduct the fine-tuning process. This approach enables us to systematically explore a wide range of hyperparameter combinations to identify the set that yields the best performance. By utilizing cross-validation, we effectively minimize the risk of overfitting, ensuring that our model generalizes well to new, unseen data. This method not only enhances the accuracy of our predictions but also bolsters the reliability of the model across different datasets.

3.4 Validity Test using Life Expectancy

To show that the predicted HOI of our model indeed reflects the health conditions of an area, we utilize health condition data from the Institute for Health Metrics and Evaluation (IHME) [9], which provides the life expectancy of each county across the US [3]. One of the critical ways to measure the effectiveness of the HOI is by examining its relationship with life expectancy [3]. Life expectancy at birth is a comprehensive indicator that reflects the cumulative impact of various health determinants over a person's lifetime [3].

To align our model's output with IHME data, we aggregated the predicted HOI values from Census Tracts up to the county level by calculating the weighted average of our predicted HOI, using the total population as the weight factor. Then we test the correlations

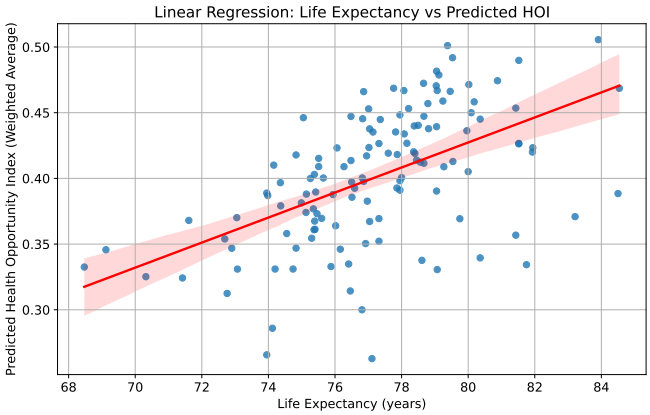

Figure 4: Correlations between the weighted average of predicted HOI and life expectancy for counties in Virginia

and p-values between our predictions and the life expectancy data from IHME using a simple linear regression model [31, 32].

3.5 Testing Applicability to Other States

To establish that our model can be effectively utilized beyond Virginia, we will evaluate its performance in two other states, following the methodology previously described. We choose the following two states for our analysis:

- **North Carolina**: North Carolina is chosen for its similarities to Virginia in terms of economic status, population demographics, and health policies [33]. We anticipate that our model will integrate seamlessly in this state, providing a robust test of its effectiveness in comparable settings.
- **California**: In contrast to North Carolina, California represents a diverse scenario with significant differences from Virginia, especially in terms of geographic location, economic structure, and demographic composition. Evaluating our model in California will shed light on its adaptability and reliability in more varied and challenging environments.

4 RESULTS

Following the experimental procedures outlined previously, this section will present the outcomes of our study. We will demonstrate the effectiveness of our model in accurately predicting the HOI. Our analysis will detail how the model performs across different states and discuss key insights that emerged from the correlation between our predicted HOI and life expectancy.

4.1 Model Selection

Based on the result of model training displayed in Table 2, the Random Forest model emerged as the most effective, achieving a Mean Squared Error (MSE) of 2.81×10^{-3} (HOI is measured under a scale of 0-1 [3]) and an R-square value of 0.622 [34, 35]. While these metrics are not perfect, they remain significant. It is essential to recognize that the HOI is inherently a relative measure, designed to assess comparative health conditions across different areas. Consequently, even with the variability introduced by using less processed data, the model's predictions remain practically valuable. The results underscore that the predicted HOI values are consistent and informative regarding the health status of an area.

Table 2: General Model Performance

Model	MSE	MAE	R^2
Decision Tree	4.42×10^{-3}	5.21×10^{-2}	0.405
Random Forests	2.81×10^{-3}	4.19×10^{-2}	0.622
XGBoost	2.90×10^{-3}	4.25×10^{-2}	0.610

Table 3: Statistics about the correlation between the weighted average of predicted HOI and life expectancy for counties in Virginia

correlation coef	p-value
0.534	6.92×10^{-11}

4.2 Application to Virginia

After we applied the life expectancy metric to validate Virginia's predicted HOI, we found that despite the moderate positive correlation coefficient of 0.534 between our predicted HOI and life expectancy, the p-value [36] of 6.92×10^{-11} confirms a statistically significant association ($p < 0.05$) [37], supporting the notion that improved health opportunities do contribute to higher life expectancy in Virginia. However, it's important to note that life expectancy [38] is influenced by a multitude of factors, including healthcare access, socioeconomic status, and public health policies [39–41]. Our model serves as a valuable tool in identifying areas where interventions can potentially enhance health outcomes.

From this analysis, it is clear that our Random Forest model, which predicts the HOI using raw data from the ACS, is both dependable and valuable. Despite the simplification in data processing, the model effectively identifies and quantifies health opportunities across different areas. This capability makes it an essential tool for uncovering underlying health issues within communities and provides actionable insights that can guide improvements in public health initiatives.

4.3 Application to other states

To evaluate the applicability and effectiveness of our model beyond Virginia, we conducted performance tests using life expectancy data in two other states: North Carolina and California. By following the established data processing steps and testing procedures, we ensured a consistent approach across different geographical contexts. The results from these tests are presented below:

4.3.1 North Carolina. Our analysis demonstrates a moderate positive linear relationship between the predicted Health Opportunity Index and life expectancy for North Carolina (NC), with a correlation coefficient of 0.559. P-value of 3.25×10^{-9} ($p < 0.05$) shows it is statistically significant which supports the hypothesis that our model is robust and can be effectively applied to regions beyond Virginia (at least for those states that share similarity with Virginia).

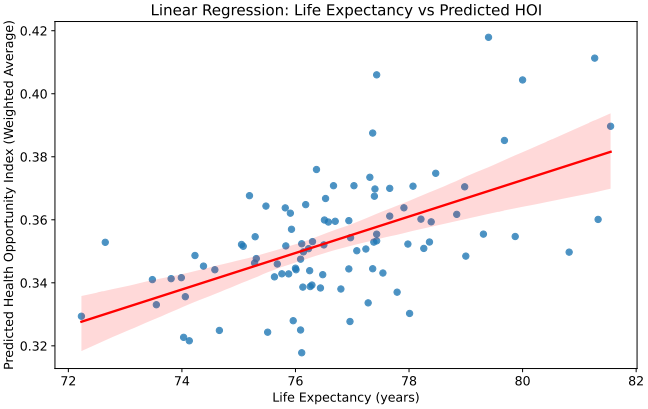

Figure 5: Correlations between the weighted average of predicted HOI and life expectancy for counties in North Carolina

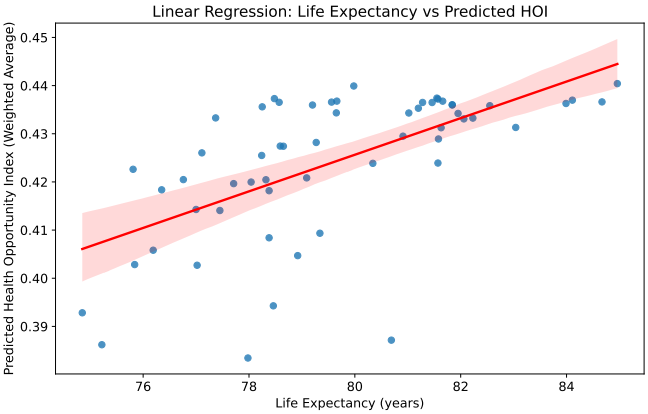

Figure 6: Correlations between the weighted average of predicted HOI and life expectancy for counties in California

4.3.2 California. The correlation coefficient for California stands at 0.615, underscoring the model's ability to generalize across diverse economic and cultural contexts. This robust correlation and P-value of 2.82×10^{-7} ($p < 0.05$) indicates that our model retains its predictive accuracy and relevance even in states that differ significantly from Virginia in terms of economic, cultural, and political characteristics. This means that our model can even be applied to a wider range of regions.

4.4 Outliers analysis

Contrary to our initial expectations, the correlation coefficients obtained from testing the model across different states revealed an interesting pattern. Specifically, California demonstrated the highest correlation coefficient between predicted HOI and life expectancy at 0.615, followed by North Carolina with 0.559, and Virginia with the lowest at 0.534. This outcome was unexpected, as we anticipated the highest correlation in Virginia, followed by North Carolina, due to their similarities, and the lowest in California, since we trained our model entirely based on Virginia's data.

Table 4: Statistics about the correlation between the weighted average of predicted HOI and life expectancy for counties in North Carolina

correlation coef	p-value
0.559	3.25×10^{-9}

Table 5: Statistics about the correlation between the weighted average of predicted HOI and life expectancy for counties in California

correlation coef	p-value
0.615	2.82×10^{-7}

To understand this phenomenon, we conducted a detailed analysis of the outliers in each state's data (Shown in Figures 7, 8, and 9). We used the Z-score method of determining the outliers, with a threshold of 1.5 (This means data points with an absolute value of residuals exceeding the standard deviation of the residuals scaled by 1.5 are defined as outliers). [42, 43]

Our investigation highlighted specific features common among these outliers that might influence the observed correlations:

- **Small size:** A significant number of outliers are found in exceptionally small counties, with some even encapsulated within another county. The limited size of these areas may lead to more fluctuations in HOI and life expectancy, skewing the data disproportionately. This, in turn, can lead to unexpected predictions in our model.
- **Outlying location:** Other outliers are located on the peripheries of their respective states. These locations may be subject to external influences, either from neighboring states or from environmental factors such as proximity to the ocean, which could impact health outcomes and corresponding HOI values.

This explains why California, despite its economic and cultural differences from Virginia, exhibited the highest correlation coefficient (0.615) among the states tested. Unlike Virginia, which has a complex county structure with many small counties, California has fewer counties with larger average sizes. This structural difference results in fewer outliers in California, enhancing the consistency of our model's predictions in this state.

This observation reveals that while our model was developed exclusively with Virginia data, it is capable of adapting effectively to other states, often with equal or superior performance. However, the results also highlight the model's sensitivity to local conditions, such as county size and structure, which can significantly impact its predictive accuracy.

5 DISCUSSION

The discussion section delves into the broader implications of our findings and explores the practical applications, limitations, and future directions for our study. This comprehensive examination is crucial for understanding the potential impact and areas for enhancement of our work.

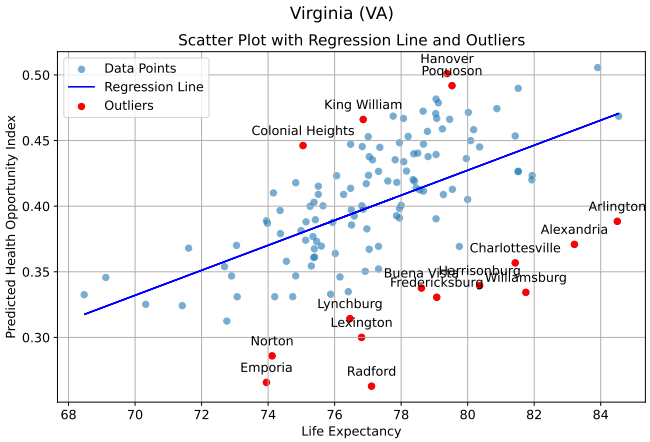

Figure 7: Outlier analysis for Virginia

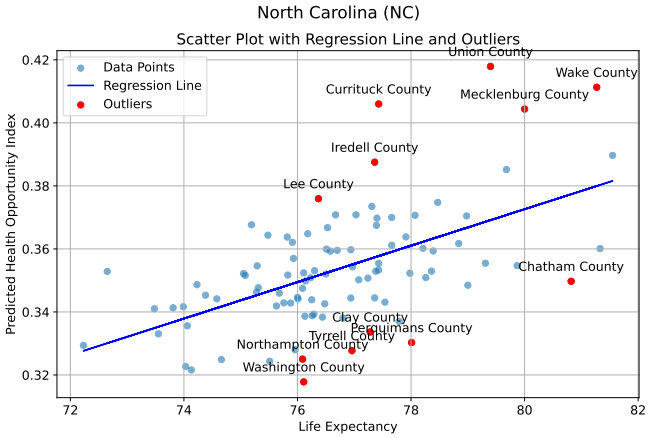

Figure 8: Outlier analysis for North Carolina

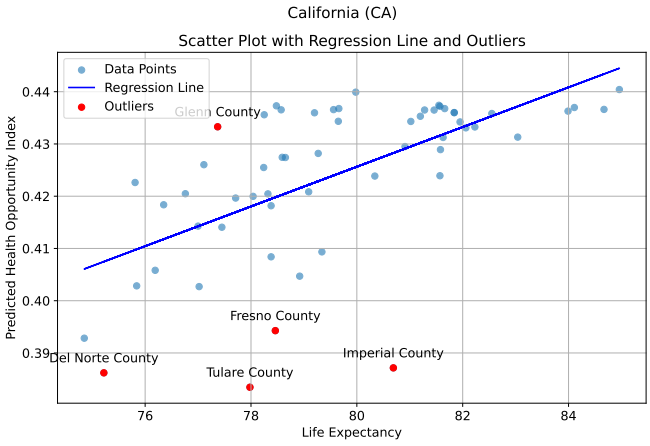

Figure 9: Outlier analysis for California

5.1 Application

The real-world applications of using our model to predict the HOI extend beyond Virginia and have broader implications for public health research and practice. Leveraging readily available data from sources like the ACS can streamline the prediction process for HOI, facilitating the development of nationwide evidence-based policy recommendations aimed at improving overall public health outcomes [10]. By targeting interventions towards areas identified as most disadvantaged, communities facing health disparities can receive the necessary support and resources to enhance their well-being [3]. Using existing ACS data also minimizes the need for extensive data collection efforts, reducing costs and time associated with model development.

5.2 Limitation and Future Work

We may need access to more years of data to improve the accuracy and relevance of our predictions. Our current study relies on 2013-2017 ACS data. Collecting and analyzing more recent data, including post-pandemic datasets, will help refine and validate the HOI prediction model, ensuring it remains relevant and accurate in current contexts. Future work should focus on acquiring and analyzing updated ACS data to refine and validate the HOI prediction model. In this case, our model can be integrated into public health surveillance systems to continuously monitor changes in HOI over time, enabling early detection of emerging health disparities and allowing for timely interventions to mitigate their impact on population health, especially considering the potential changes in health opportunities due to events like the COVID-19 pandemic [44].

Incorporating deep learning models may further enhance the accuracy and predictive power of the HOI model [45]. While our current approach using random forests has shown promise, deep learning techniques could capture more complex patterns within the data, potentially leading to more precise predictions.

We may also need to consider more metrics for evaluating the model's performance. Currently, we use life expectancy as a validation metric, but additional metrics such as Potential Years of Life Lost (PYLL) and other health outcome indicators could provide a more comprehensive assessment of the model's effectiveness [46]. These additional metrics would offer deeper insights into the public health implications of the predicted HOI values and help validate the model's utility in different contexts.

6 CONCLUSION

Our findings highlight the effectiveness of the Random Forest model in predicting HOI, achieving a moderate positive correlation with life expectancy data. The successful application of our model to North Carolina and California demonstrates its robustness and adaptability. The model's ability to generalize across different states underscores its potential as a valuable tool for public health analysis and policy-making despite its sensitivity to local geographic and demographic characteristics [10]. By continuously refining the model and expanding its applicability, we can contribute to the development of evidence-based public health policies and the promotion of health equity across states.

REFERENCES

[1] Anita Pereira. Live and let live: Healthcare is a fundamental human right. *Conn. Pub. Int. LJ*, 3:481, 2003.

[2] Michael Marmot and R Bell. Fair society, healthy lives., 2009.

[3] Chinonso N. Ogojiaku, JC Allen, Rexford Anson-Dwamena, Kierra S. Barnett, Olorunfemi Adetona, Wansoo Im, and Darryl B. Hood. The health opportunity index: Understanding the input to disparate health outcomes in vulnerable and high-risk census tracts. *International Journal of Environmental Research and Public Health*, 17(16), 2020.

[4] Michael Marmot and Richard Wilkinson. *Social determinants of health*. Oup Oxford, 2005.

[5] Edward Herman. The american community survey: an introduction to the basics. *Government Information Quarterly*, 25(3):504–519, 2008.

[6] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[7] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, , and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[8] Christopher JL Murray, Jerry Abraham, Mohammed K Ali, Miriam Alvarado, Charles Atkinson, Larry M Baddour, David H Bartels, Emelia J Benjamin, Kavi Bhalla, Gretchen Birbeck, et al. The state of us health, 1990-2010: burden of diseases, injuries, and risk factors. *Jama*, 310(6):591–606, 2013.

[9] Rita Rubin. Profile: Institute for health metrics and evaluation, wa, usa. *The Lancet*, 389(10068):493, 2017.

[10] Dirk T Ubbink, Gordon H Guyatt, and Hester Vermeulen. Framework of policy recommendations for implementation of evidence-based practice: a systematic scoping review. *BMJ open*, 3(1):e001881, 2013.

[11] Virginia Department of Health. The virginia health opportunity index.

[12] Jasmine Cassy Mah, Jodie Lynn Penwarden, Henrique Pott, Olga Theou, and Melissa Kathryn Andrew. Social vulnerability indices: a scoping review. *BMC public health*, 23(1):1253, 2023.

[13] Andrew J Knighton, Lucy Savitz, Tom Belnap, Brad Stephenson, and James VanDerslice. Introduction of an area deprivation index measuring patient socioeconomic status in an integrated health system: implications for population health. *EGEMs*, 4(3), 2016.

[14] Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar, R Sujatha, Jyotir Moy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai, and Ohyun Jo. Covid-19 patient health prediction using boosted random forest algorithm. *Frontiers in public health*, 8:357, 2020.

[15] Wei Dong, Yimiao Huang, Barry Lehane, and Guowei Ma. Xgboost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. *Automation in Construction*, 114:103155, 2020.

[16] Robert M Kaplan and John P Anderson. A general health policy model: update and applications. *Health services research*, 23(2):203, 1988.

[17] Emma S McBryde, Michael T Meehan, Oyelola A Adegboye, Adeshina I Adekunle, Jamie M Caldwell, Anton Pak, Diana P Rojas, Bridget M Williams, and James M Trauer. Role of modelling in covid-19 policy development. *Paediatric respiratory reviews*, 35:57–60, 2020.

[18] National Research Council et al. *Using the American Community Survey: benefits and challenges*. 2007.

[19] Mark Mather, Kerri L Rivers, and Linda A Jacobsen. The american community survey. *Population Bulletin*, 60(3), 2005.

[20] American Community Survey. Social explorer.

[21] Md Manjurul Ahsan, MA Parvez Mahmud, Pritom Kumar Saha, Kishor Datta Gupta, and Zahed Siddique. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3):52, 2021.

[22] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[23] J Brown. Choosing the right number of components or factors in pca and efa. *JALT Testing & Evaluation SIG Newsletter*, 13(2), 2009.

[24] Hongge Chen, Huan Zhang, Si Si, Yang Li, Duane Boning, and Cho-Jui Hsieh. Robustness verification of tree-based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[25] Liangyuan Hu and Lihua Li. Using tree-based machine learning for health studies: Literature review and case series. *International journal of environmental research and public health*, 19(23):16080, 2022.

[26] Noora Shrestha. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2):39–42, 2020.

[27] Ramal Moonesinghe, Eleanor Fleming, Benedict I Truman, and Hazel D Dean. Linear and non-linear associations of gonorrhea diagnosis rates with social determinants of health. *International journal of environmental research and public health*, 9(9):3149–3165, 2012.

[28] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[29] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

[30] Michael W Browne. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132, 2000.

[31] Kelly H Zou, Kemal Tuncali, and Stuart G Silverman. Correlation and simple linear regression. *Radiology*, 227(3):617–628, 2003.

[32] Sybil L Crawford. Correlation and regression. *Circulation*, 114(19):2083–2088, 2006.

[33] DA Herbert Jr, Jack Bacheler, Sean Malone, and Dan Mott. Thrips control options in virginia/north carolina: overviews, insights and updates. In *Proc. Beltwide Cotton Conf., New Orleans, LA*, pages 9–12, 2007.

[34] Tianfeng Chai, Roland R Draxler, et al. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, 7(1):1525–1534, 2014.

[35] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623, 2021.

[36] Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

[37] Nicole M White, Thirunavukarasu Balasubramaniam, Richi Nayak, and Adrian G Barnett. An observational analysis of the trope "a p-value of< 0.05 was considered statistically significant" and other cut-and-paste statistical methods. *PLoS One*, 17(3):e0264360, 2022.

[38] Max Roser, Esteban Ortiz-Ospina, and Hannah Ritchie. Life expectancy. *Our world in data*, 2013.

[39] Wim JA van den Heuvel and Marinela Olaroiu. How important are health care expenditures for life expectancy? a comparative, european analysis. *Journal of the American Medical Directors Association*, 18(3):276–e9, 2017.

[40] John Mirowsky and Catherine E Ross. Socioeconomic status and subjective life expectancy. *Social Psychology Quarterly*, pages 133–151, 2000.

[41] Atheendar S Venkataramani, Rourke O'Brien, and Alexander C Tsai. Declining life expectancy in the united states: the need for social policy as health policy. *JaMa*, 325(7):621–622, 2021.

[42] Vaibhav Aggarwal, Vaibhav Gupta, Prayag Singh, Kiran Sharma, and Neetu Sharma. Detection of spatial outlier by using improved z-score test. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 788–790. IEEE, 2019.

[43] Abdulmalik Shehu Yaro, Filip Maly, and Pavel Prazak. Outlier detection in time-series receive signal strength observation using z-score method with s n scale estimator for indoor localization. *Applied Sciences*, 13(6):3900, 2023.

[44] Marco Ciotti, Massimo Ciccozzi, Alessandro Terrinoni, Wen-Can Jiang, Cheng-Bin Wang, and Sergio Bernardini. The covid-19 pandemic. *Critical reviews in clinical laboratory sciences*, 57(6):365–388, 2020.

[45] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[46] G Wiesner and EK Bittner. Life expectancy, potential years of life lost (pyll), and avoidable mortality in an east/west comparison. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 47:266–278, 2004.

A REPRODUCIBILITY

To reproduce our results, please visit the GitHub repository created for this project. The repository details how data was converted into a usable format for the models, how to use our code to train the models, and how to directly implement them. The repository contains a Jupyter Notebook that provides a comprehensive, step-by-step analysis from data acquisition to the final results.

- **Data Preparation:** Instructions on how to convert raw data into a format suitable for model training.
- **Model Training:** Guidelines on using our code to train the models, including hyperparameter settings and training procedures.
- **Implementation:** Steps to implement the trained models and reproduce the results presented in the paper.

You can find the repository at the following URL: https://github.com/LiranLizqj6pe/SL4HOI-Supervised-Learning-for-Predicting-Health-Opportunity-Index-across-States.git

B LIST OF FEATURES

B.1 Raw Features Selected

The following features were selected from Social Explorer Tables: ACS 2017 (5-Year Estimates):

- **A00002:** Population Density (Per Sq. Mile)
- **A00001:** Total Population
- **A10003:** Average Household Size
- **A10003B:** Average Household Size of Renter-Occupied Housing Units
- **A12004:** School Enrollment for the Population 3 Years and Over
- **A12003:** School Dropout Rate for Population 16 to 19 Years
- **A17005:** Unemployment Rate for Civilian Population in Labor Force 16 Years and Over
- **A14001:** Household Income (In 2017 Inflation Adjusted Dollars)
- **A14010:** Median Family Income (In 2017 Inflation Adjusted Dollars)
- **A14028:** Gini Index of Income Inequality
- **A10038:** Monthly Housing Cost (Renter-Occupied Housing Units)
- **A17002:** Employment Status for Total Population 16 Years and Over
- **A10003B:** Age of Householder (Renter-Occupied Housing Units)
- **A10039B:** Monthly Housing Costs as a Percentage of Household Income in the Past 12 Months (Renter-Occupied Housing Units)
- **A13002:** Poverty Status in 2017 of Families by Family Type by Presence of Children Under 18 Years
- **A13003A:** Poverty Status in 2017 for Children Under 18
- **A13003B:** Poverty Status in 2017 for Population Age 18 to 64
- **A13003C:** Poverty Status in 2017 for Population Age 65 and Over
- **A13004:** Ratio of Income in 2017 to Poverty Level
- **A20001:** Health Insurance

C LIST OF OUTLIERS

C.1 Virginia

- Alexandria
- Arlington County
- Buena Vista
- Charlottesville
- Colonial Heights
- Emporia
- Fredericksburg
- Hanover County
- Harrisonburg
- King William County
- Lexington
- Lynchburg
- Norton
- Poquoson
- Radford
- Williamsburg

C.2 California

- Del Norte County
- Fresno County
- Glenn County
- Imperial County
- Tulare County

C.3 North Carolina

- Chatham County
- Clay County
- Currituck County
- Iredell County
- Lee County
- Mecklenburg County
- Northampton County
- Perquimans County
- Tyrrell County
- Union County
- Wake County
- Washington County