Springboard Data Science

Capstone Two Project Report

By: Bini Teklehaimanot

# Problem statement

Are there significant geographic variations in healthcare access for blacks in the US? Do these variations correlate with factors such as the availability of primary care providers or income?

# Data Wrangling

I first set out to explore the Area Health Resources data and NCHS mortality data. The NCHS data did not have the necessary features to address the problem statement. The AHRF data was focused on race and gender distribution in different healthcare fields by geography. This is very useful data that I later determined was not necessary. However, it is data that I intend to use for future work. The two charts below demonstrate the gender and race distributions, hence the reason I want to use the data in the future. The decision to use a different dataset was determined during the EDA phase.

A graph of a number of people

AI-generated content may be incorrect.

# Exploratory Data Analysis

During the exploration data analysis phase, I found additional data that might support the problem I am trying to solve. The Center for Medicare and Medicaid Services provides total cases, total charges, and total days of care by zip code and by the provider, used as a health\_services dataframe in the project. Additionally, the CMS provided data on physician offices by location. By using publicly available zip code data, I added city and state features to health\_services.

The relationship between the three main features of health\_services is linear (example plot below). Some outliers exist; as total cases increase, we see greater variance. The variability is more significant as days of care increase; the charges are no longer close to the regression line.

Using the IRS average gross income (AGI) data, I discovered that a correlation of 0.130 suggests a very weak positive correlation between total income and total days of care by ZIP code. This means that income level has little to no effect on the number of days in care. The same holds true for cases and charges. There is very little correlation between care and income. It is important to note that there is a limitation on the current data set as it is on Medicare patients.

A graph showing a red line and a blue line

AI-generated content may be incorrect.

While there is little correlation between income and the features, we can see a high concentration at the low-income values. An example plot is given below. The plot is similar for total days of care and total charges.

A graph of a graph showing the difference between a total and a total case

AI-generated content may be incorrect.

The final dataset that proved to be helpful for the hypothesis was found from the CDC. This data has the same limitations as previously discussed. It focuses on Medicare and Medicaid patients in the 50 states of the US for 2023. This data contained 647 columns. I removed columns that appeared to hold similar information. I also dropped columns that have the same value across all rows. Finally, I merged the health\_services dataframe with CDC data, resulting in 40 columns. Many of the patterns discovered during the EDA did not lend to an evident trend.

# Model Preprocessing with feature engineering

Once the model was ready from the previous steps, I extracted the access column as y and the rest of the data as X to fit the data. Access, in the source data, was given as 1, 2, 3, 7, 8, or 9. 1 and 3 represent yes for one or more locations. I changed these to 1. All other values are related to no variations, which I set to 0. Since the model was built on a 50/50 split, I refitted the model with a more realistic 80/20 train-test split.

I found marital status, experiencing discrimination, awareness (vigilance) of discrimination, and pain limits activity were the top features affecting access. I examined various scenarios to enhance access for African Americans. Here's how I approached it. I used a range of improvements or degradation from -20% to +20% in increments of 5%. The scenarios are:

* Suppose marital support improves
* Suppose experienced discrimination decreases
* Suppose awareness of discrimination improves
* Suppose depression events reduce
* Suppose pain limits activity improves

The incremental changes are impacting access. An example plot given below is discrimination versus change in access. The model detected meaningful patterns! Vigilance and discrimination have the strongest effects. Pain limits activity and marital status influence access, but only in extreme cases. Depression events may require further investigation. It might be that it is underreported due to cultural stigma. The data indicates that approximately 74% of the population is white. Therefore, the model could be biased. Different algorithms were tried, and Random Forest was the one that best represented the problem I set out to solve.

A graph with a line and a dotted line

AI-generated content may be incorrect.

# Access findings and recommendation

**📍 Current Position: Black Race Access to Services**

Our analysis indicates that Black individuals (Race = 2) currently have an access rate of approximately 89%. This reveals a measurable disparity in access compared to other racial groups within the dataset, highlighting the need for targeted interventions.

**📈 Key Modeling Insights: How Access Could Improve**

The modeling simulations suggest that targeted improvements in specific social and health factors can significantly increase access for Black individuals. Firstly, a 20% decrease in experienced discrimination is associated with an increase in access. However, it's crucial to note that the effect plateaus after a certain point, indicating that other barriers still exist beyond discrimination alone. Secondly, enhancing overall wellness has the strongest effect on access, suggesting that health-related quality of life is a major determinant of accessibility. Reducing the impact of pain-limiting activities also contributes positively, reinforcing the connection between physical well-being and access. Lastly, higher vigilance, or awareness of discrimination, positively impacts access. This suggests that providing resources to navigate systemic challenges could help mitigate barriers in practical ways.

**🚀 Strategic Approach for Business Leadership**

To influence leadership decisions, the findings should be framed in terms of business impact, equity goals, and data-driven ROI. A data-backed business case should be presented, outlining the model’s quantified impact of reducing discrimination and improving wellness as a way to broaden access to key services. Highlight how higher access increases customer retention, engagement, and satisfaction among underrepresented groups. Operational feasibility should be addressed by identifying low-cost, high-impact initiatives that can address key barriers. For example, offering wellness-based incentives or funding awareness programs could have tangible benefits. Finally, align initiatives with Diversity, Equity, and Inclusion (DEI) and Environmental, Social, and Governance (ESG) goals. Many organizations have DEI commitments—use this model to showcase measurable progress toward racial equity goals. Incorporate findings into Corporate Social Responsibility (CSR) initiatives to strengthen leadership buy-in.

**📜 Policy Recommendations for Advocates**

Advocacy groups can use these findings to push for policy-level changes that improve access for underrepresented races. Firstly, expand access to healthcare and wellness support by targeting chronic pain, mobility, and preventive care. Advocate for employer-based incentives to improve health equity. Secondly, push for stronger anti-discrimination policies in healthcare, employment, and financial services to address systemic discrimination. Fund legal assistance and community education programs to help individuals navigate discrimination cases. Lastly, enhance community awareness and resource navigation by creating awareness campaigns about systemic barriers and how individuals can navigate them. Partner with Black-led advocacy organizations to develop practical guides for overcoming access challenges.

**🔮 Future Improvements: Which Scenarios Should Be Studied Further?**

Recommended areas for further research include wellness and physical health factors, as they are the strongest predictor and need further exploration into specific interventions. Vigilance and awareness training should be studied to understand how structured programs can empower individuals to navigate systemic challenges. Additionally, the intersection of race and economic status should be explored to understand how financial resources interact with access barriers in a racialized way. Less impactful areas, such as extreme discrimination reductions alone, should be de-prioritized. While important, the model suggests plateauing effects, meaning that other structural barriers still need to be addressed.

# Conclusion

This analysis provides actionable strategies for business leadership, policymakers, and advocates to drive measurable improvements in access for Black individuals. By targeting wellness, reducing systemic discrimination, and increasing awareness-based interventions, organizations can bridge the access gap and create a more equitable system. Next steps include presenting findings to key stakeholders, prioritizing wellness-based and awareness-driven interventions, and exploring partnerships with community organizations and policy advocates.

# Future scope of work

**Data Limitations & Deficiencies**

While this model provides valuable insights, certain data limitations may have impacted the depth of analysis. Specifically, the lack of granular access data, which is currently a binary view (Yes/No), fails to capture the quality, consistency, or frequency of healthcare access. Future work should focus on adding metrics like frequency of visits, wait times, or types of healthcare services accessed to provide a richer picture. Additionally, missing or limited socioeconomic data, beyond factors like total adjusted gross income, could offer more context. Factors such as education level, employment stability, or neighborhood healthcare availability would provide more context. Finally, potential bias in data collection, where survey responses or administrative data may have systematic underreporting from marginalized communities, could lead to overestimation of access rates compared to real-world conditions.

**What Additional Access Data Would Be Useful?**

To enhance the model’s predictive power, several additional data points would be valuable. Understanding the impact of transportation and distance to healthcare centers by incorporating data on physical proximity to providers would be beneficial. Examining digital access and telehealth usage would reveal whether these services are bridging the gap for those with historically lower access. More granular data on out-of-pocket costs and insurance barriers, such as copay amounts, denied claims, and prior authorization issues, could better explain access disparities.

**Surprising Insights for Business Executives**

Key takeaways that may surprise business leaders include the fact that discrimination reduction has a ceiling. While reducing discrimination initially increases access, further reductions do not yield additional improvements beyond a certain point. Furthermore, vigilance, or awareness of discrimination, is a game-changer. Employees or patients who are more aware of discrimination barriers tend to navigate them better, leading to higher access rates. Lastly, wellness plays a bigger role than expected. The strongest predictor of access was not income or insurance but overall wellness, suggesting that health initiatives may be more impactful than financial subsidies alone. To validate these insights, we can survey business teams and healthcare providers to see if the results align with their field experiences, compare findings against external research, and pilot small-scale interventions like wellness programs or discrimination awareness campaigns to measure real-world impact.

**How Could the Business Make Use of This Model?**

The business can utilize this model for several practical applications in decision-making. Scenario planning for policy decisions allows for simulating different healthcare access policies before implementation. Equity-focused strategy development enables prioritization of initiatives with the strongest impact on access, such as wellness programs over direct subsidies. Predictive modeling for outreach and engagement helps identify communities at the highest risk of poor access and proactively design interventions.

**Making the Model Accessible for Business Analysts**

To make the model more accessible, various deployment approaches can be considered. An interactive web-based dashboard would allow business users to input different policy changes and see real-time predictions on access impact. This could be built using tools like Streamlit, Dash, or Power BI with embedded machine learning models. Automated report generation, with a scheduled monthly or quarterly model run, would summarize access trends for leadership, including top factors affecting access and potential policy levers. Exposing the model as a self-service API for scenario testing would allow business teams to submit different parameter combinations and get predicted access rates, enabling non-technical users to test different scenarios independently.

**🎯 Conclusion: Next Steps for Business Adoption**

Short-term actions should include presenting findings and exploring pilot interventions based on key insights, such as wellness programs and discrimination awareness training. In the mid-term, deploying a self-service tool for scenario testing would empower business analysts to explore policy options independently. Long-term, integrating more detailed socioeconomic and behavioral data would refine access predictions further. With the right data and tools, this model can become a long-term asset for shaping equitable healthcare policies.

Data Source:

**Centers for Disease Control and Prevention (CDC)..** *National Health Interview Survey 2023: Sample Adult File – Codebook*. National Center for Health Statistics. Retrieved from <https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2023/adult-codebook.pdf>

**Centers for Disease Control and Prevention (CDC).** *National Health Interview Survey, 2023: Sample Adult File [Data set].* National Center for Health Statistics. Retrieved from <https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2023/adult23csv.zip>