Socioeconomic and demographic disparities in fitness center visits

Giorgos Ioannou*, George Pallis*, Marios Dikaiakos*, Christos Nicolaides*

* University of Cyprus, Nicosia, Cyprus
Email: {ioannou.george, pallis.george, dikaiakos.marios, nicolaides.christos}@ucy.ac.cy

***Abstract*—Socioeconomic and demographic disparities in physical activity participation are strong predictors of obesity prevalence, underscoring the critical need to understand and address unequal participation in fitness centers across communities. In this work, we integrate large-scale visit data, which include visits to fitness centers and other types of locations, with census-derived socioeconomic and demographic attributes, to investigate patterns related to physical activity. Specifically, we analyze over 50 million daily visits from more than 15 million users and construct socioeconomic profiles for more than 150,000 neighborhoods, including median household income, educational attainment, age distributions, and gender ratios. We develop machine learning models that predict the proportion of residents visiting fitness centers, and show that demographic and socioeconomic features alone explain over half of the observed variation in visit rates (R^2 ranging from 0.536 to 0.572 across three months). Explainability analyses via linear regression and SHapley Additive exPlanations (SHAP) of LightGBM models reveal that income and education are the strongest predictors of fitness center visits, while a higher female population proportion is positively associated with fitness center visits. We further show that residents aged 18-24, and those with higher education, are more likely to visit local fitness facilities, whereas income has no effect on choosing facilities closer to home. Our findings offer crucial insights for designing targeted interventions and policies to alleviate physical activity disparities and promote equitable access to fitness facilities.**

***Index Terms*—behavioral modeling, explainable AI, predictive analytics, data analytics**

I. Introduction

Our world faces a global inactivity crisis [1], with over 30% of the population failing to meet physical activity recommendations [2], facing increased health risks. Despite this widespread crisis, physical inactivity does not impact all segments of society equally. Socioeconomic disparities play a significant role, with lower income associated with lower physical activity [3], while gender influences the preferences between various forms of activity [4]. Notably, inequality in the distribution of physical activity within countries is a stronger predictor of obesity than average activity levels [5]. Effectively addressing activity inequality requires large-scale, innovative, and technology-driven methods that go beyond participant-based methods, to develop targeted policies.

Recent technological advancements enabled the collection of large-scale mobility data, instrumental in studying population behavior and policy outcomes [6]. Typically, activity is quantified through steps measured from smartphone accelerometers. Incentive-based applications using this approach [7] have increased step counts, though these are strongly influenced by temperature and rainfall [8]. Alternatively, exercise sessions can be identified from visit data, comprising timestamped visits to fitness centers and other types of locations such as restaurants. For example, analysis of over 20 million individuals revealed that exercisers share distinct visit patterns over non-exercisers, one example being a higher visit rate to coffee shops [9].

This study advances physical activity research by integrating large-scale visit data, comprising more than 15 million users making more than 50 million daily visits, with U.S. Census attributes at the census block group (CBG) level. Each CBG typically contains 600-3,000 residents. We construct comprehensive CBG profiles, that include median household income, education attainment, age distributions, and gender ratios. Using machine learning (ML), we develop models that reliably predict the proportion of residents visiting fitness centers, solely based on their home CBG profile. Attribute influence is assessed using ordinary least squares (OLS) for linear effects and SHapley Additive exPlanations (SHAP) to uncover nonlinear interactions. Moreover, we apply logistic regression to identify how each attribute influences whether or not an individual visits a fitness center at their home CBG.

To our knowledge, this is the first large-scale integration of mobility and census-derived socioeconomic data to examine physical activity disparities. We propose the relevant research questions, along with our contributions for tackling them.

1) "What proportion of visits from various income and education groups are allocated to fitness centers, full-service restaurants, and fast-food restaurants?" We analyze more than 50 million daily visits, incorporate socioeconomic profiles from each individual's home CBG, and calculate visit proportions across location categories.
2) "How accurately can CBG's socioeconomic and demographic profiles predict the proportion of their residents who visit fitness centers?" We develop multiple ML regression models using data from over 150,000 CBGs, with analyses repeated across three months (October-December 2019), to predict fitness center visitor rates based solely on the profile of the CBG.
3) "What is the relative importance of CBG attributes (income, education, age, gender) in predicting fitness center visit rates, and how do these factors interact?" We quantify feature importance through OLS analysis,

and use SHAP on LightGBM to construct partial dependence plots that reveal complex interactions between the demographic and socioeconomic variables.

4) "How do socioeconomic and demographic factors influence the likelihood of visiting fitness centers within one's home CBG?" We apply logistic regression with more than 500,000 fitness center visitors to determine which factors influence local facility usage patterns.

II. Related Work

Understanding how demographic and socioeconomic factors shape physical activity is vital for designing effective policies [9], [10]. Traditionally, such insights relied on self-reported surveys, which consistently show that age negatively correlates with activity, whereas education, income, and male sex are positively associated [10]. However, these methods are subject to recall and social desirability biases [11], and are generally limited to a small number of participants [12], hindering population-level assessments. Large-scale mobility data overcome these limitations and enable unprecedented insight into population behavior. People follow reproducible mobility patterns [13], a finding which has supported precise epidemic simulations [6], [14]. During COVID-19, mobility networks showed that 49% of nursing home cases stemmed from shared staff transmitting the virus between facilities [15], a vulnerability later highlighted by the CDC [16].

Building upon these foundations, mobility data has been used to explore physical activity and inform public health policy. Analysis of steps from 46 countries revealed that inequality in how activity is distributed within populations is a better predictor of obesity prevalence over average activity levels [5], underscoring the need for equity-focused interventions. Moreover, analysis of visit patterns from over 20 million individuals revealed that exercisers show distinct mobility behaviors compared to non-exercisers, including a preference for healthy food venues over fast-food restaurants [9]. These differences provide insights into patterns associated with physical activity participation, offering targets for policy design.

Despite the advantages of mobility data, the lack of demographic attributes poses challenges in interpreting findings, which are typically addressed through integration with census-derived socioeconomic information. During COVID-19, this integration showed that individuals in lower-income areas were less able to reduce their movements and had higher likelihood of being infected [6], while parks in wealthier areas saw smaller declines in visits [17]. These findings demonstrate how combining mobility and socioeconomic data uncovers disparities and guides targeted policy interventions.

Quantifying the impact of demographic and socioeconomic determinants on activity requires explainable ML models. Ordinary least squares (OLS) regression is widely used for its interpretable coefficients that estimate variable effects while controlling for confounders. OLS has assessed intervention effectiveness in increasing fitness center visits [18] and evaluated peer effects on exercise habits [19]. When linear models fail to capture complex interactions, black-box ML approaches offer superior performance but lack interpretability. SHAP [20] address this and is the leading method for quantifying feature's contribution to model predictions. In healthcare, SHAP identified predictors of arrhythmias [21] and type 2 diabetes [22]. By making complex models transparent, SHAP links predictive power with actionable insights for policymakers.

III. Datasets

A. Visit data

We utilize Veraset's visit dataset[1], which comprises records from over 20 million anonymized U.S. devices, recording visited locations with timestamps. Data are collected via opt-in consent through mobile applications and Software Development Kits (SDKs), ensuring compliance with privacy laws such as the CCPA[2]. Previous research utilized Veraset and SafeGraph[3] visit datasets, with SafeGraph being the majority owner of Veraset, to assess COVID-19 impacts and inform public policy [6], [9]. The visit datasets were shown to be unbiased, with the sampling highly correlated to the true census populations [6], [23]. Each visit contains an anonymized device ID and a location category. The most visited categories are full-service restaurants, fast-food restaurants, and fitness centers. Within the fitness center category, some of the most visited locations include Planet Fitness, CrossFit, Fit4Mom, and Club Pilates, with each targeting different demographic segments. Veraset identified the home CBG for more than 15 million individuals by analyzing nighttime location patterns over six-week windows preceding each analysis date[4].

B. U.S. census

We obtain demographic and socioeconomic data from the U.S. Census Bureau's 5-year American Community Survey (ACS) for 2016–2020[5]. Each CBG is mapped to its corresponding median household income. Moreover, for every CBG, we compute the proportion of residents according to educational attainment, age group, and gender.

C. University Locations

To identify the CBG of all U.S. colleges and universities, we first obtain institutional coordinates from an open dataset, which maps the names with their corresponding longitude and latitude[6]. Subsequently, for each extracted longitude-latitude, we use the API from the Federal Communications Commission[7], which accepts coordinates and returns their CBG. Finally, we map each institution's name to its corresponding CBG, enabling integration with the visit and census datasets.

IV. Results

Each subsection below corresponds to the analysis of a research question, alongside its contribution, from section I.

[1] https://www.veraset.com/datasets/visits
[2] https://www.veraset.com/insights/ensuring-data-security
[3] https://www.safegraph.com/
[4] https://www.veraset.com/datasets/home-work
[5] https://www.census.gov/programs-surveys/acs
[6] https://public.opendatasoft.com/explore/dataset/
[7] https://geo.fcc.gov/api/census/

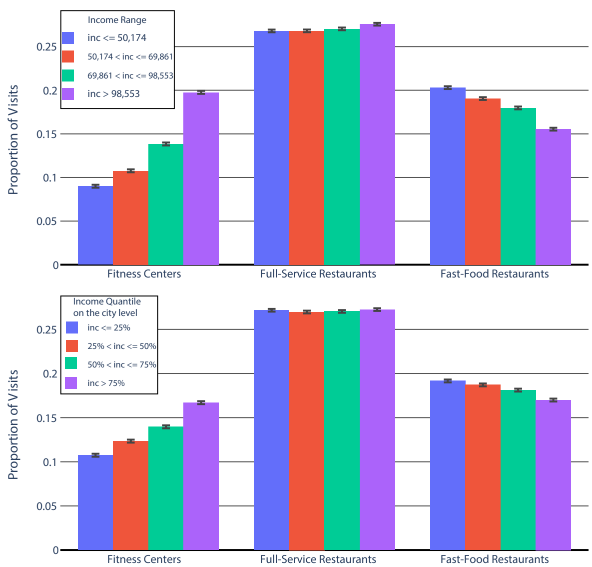

Fig. 1: Proportion of visits to location categories by home CBG income level. Top panel: users grouped by national income quartiles. Bottom panel: users grouped by city-level income quartiles. Bars represent the 95% confidence intervals.

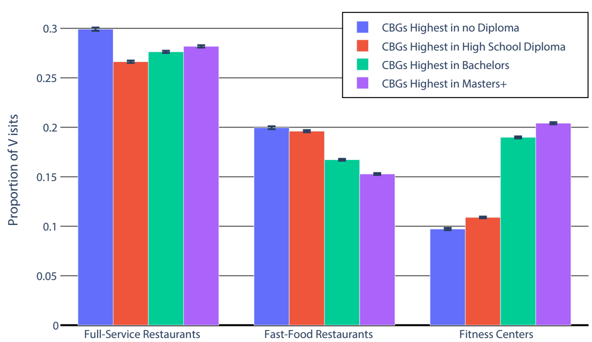

Fig. 2: Proportion of visits to location categories by home CBG educational attainment. Groups represent the 20,000 CBGs with the highest proportions of residents in each educational category. Bars represent 95% confidence intervals.

A. Location visits

We compare the proportion of visits to fitness centers, full-service restaurants, and fast-food restaurants, based on the income and educational attainment of the home CBG of users.

The income-based results are shown in Figure 1. Daily analysis throughout the three-month study period revealed identical patterns, with the figure displaying results from one such day to illustrate these consistent findings. The top panel partitions users into four equally-sized groups, with each user assigned the median household income of their home CBG. The bottom panel accounts for income segregation that exists at the city level. We compare the median income of each CBG against the income distribution within the corresponding city, and assign each CBG to a quartile based on local economic status. The patterns observed in both approaches are consistent. Residents of higher-income CBGs dedicate a greater proportion of their visits to fitness centers and a smaller proportion to fast-food restaurants. At the national level, these differences are more pronounced due to income segregation across different areas [24]. No significant differences exist in full-service restaurant visits. However, there may exist variations in the characteristics of full-service restaurants visited, as people from lower-income CBGs tend to visit smaller and more crowded locations within the same category [6].

The education-based results are shown in Figure 2. Unlike income, education per CBG is represented as the proportion of residents across attainment levels rather than a single median value. Therefore, to form comparison groups, we selected the 20,000 CBGs with the highest share of residents in each category (no high school diploma, high school diploma, bachelor's, and master's+). Results show that residents from CBGs with higher educational levels have higher visit rates to fitness centers and lower rates to fast-food restaurants. Full-service restaurant visits follow a mixed pattern, rising from high school completion onward, yet peaking among residents of CBGs with predominantly no high school diploma.

B. Regression Models

We analyzed visit data from the first four weeks of October, November, and December 2019 to identify fitness center visitors, individuals with at least one visit per period. For each CBG, we computed the proportion of fitness center visitors and trained ML models using only census-derived socioeconomic and demographic variables, along with state-level effects. The target variable was the proportion of fitness center visitors in each CBG. Table I shows three example entries from the dataset we constructed and used for analysis. Table II shows results across the three months. We evaluated 15 ML regression models using 10-fold cross-validation, reporting Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and average R^2. For clarity, we present only the seven top-performing models. The baseline is a naive regressor that always predicts the monthly average proportion of fitness center visitors among CBGs.

The ML models show significant predictive capability, with R^2 values exceeding 0.5 across all periods and models. This indicates that socioeconomic and demographic variables alone explain more than half of the variance in fitness center visit patterns at the CBG level. Consistent temporal results indicate stable relationships between these factors and facility use across the studied months, despite seasonal variation in exercise behavior. LightGBM achieved the best performance across all metrics and periods.

C. Feature Importance

To quantify the influence of each demographic and socioeconomic factor, we apply linear regression and use SHAP to LightGBM, the best-performing model.

Linear regression's competitive performance against non-linear models (R^2 within 0.02 vs. LightGBM) and substantial MAE improvement over baseline (0.150 to 0.097 in November) shows that effects on fitness center participation are largely linear, supporting its use for prediction and interpretation. We apply OLS to quantify the expected change in the

TABLE I: Three sample entries (n=163,995) used in the regression for November, 2019. The dependent variable is the rate of individuals that visited a fitness center at least once. Numbers rounded to 2 decimals for clarity.

CBG	Median Income	No diploma	High School	Bachelors	Masters+	Male	Female	18-24	25-44	45-64	65+	US state	Rate
10010201001	39167	0.23	0.43	0.2	0.14	0.42	0.58	0.16	0.23	0.39	0.22	al	0.38
10010201002	70699	0.17	0.49	0.22	0.12	0.53	0.47	0.13	0.35	0.34	0.18	al	0.47
10010202001	39750	0.12	0.71	0.14	0.02	0.38	0.62	0.08	0.43	0.22	0.27	al	0.27

TABLE II: Performance comparison of regression models predicting CBG-level fitness center visitor rates. Models utilize demographic and socioeconomic variables as predictors (see Table I). Evaluation metrics are averaged over 10-fold cross-validation. Baseline model always predicts the monthly average visit rate among CBGs (0.41, 0.42, 0.40 for respective months).

	October, 2019				**November, 2019**				**December, 2019**			
Model	**MAE**	**MSE**	**RMSE**	**R2**	**MAE**	**MSE**	**RMSE**	**R2**	**MAE**	**MSE**	**RMSE**	**R2**
LightGBM	0.096	0.015	0.122	0.549	0.095	0.014	0.120	0.572	0.094	0.014	0.119	0.536
Extra Trees	0.097	0.015	0.123	0.549	0.096	0.015	0.122	0.560	0.096	0.015	0.121	0.522
Gradient Boosting	0.098	0.015	0.124	0.530	0.097	0.015	0.123	0.551	0.097	0.015	0.122	0.517
Linear	0.098	0.015	0.124	0.530	0.097	0.015	0.123	0.551	0.097	0.015	0.122	0.515
Bayesian Ridge	0.098	0.015	0.124	0.530	0.097	0.015	0.123	0.551	0.097	0.015	0.122	0.515
Ridge	0.098	0.015	0.124	0.530	0.097	0.015	0.123	0.551	0.097	0.015	0.122	0.515
Random Forest	0.098	0.015	0.124	0.528	0.097	0.015	0.123	0.551	0.097	0.015	0.122	0.515
Baseline	0.148	0.033	0.181	0	0.150	0.034	0.184	0	0.142	0.031	0.175	0

proportion of fitness center visitors associated with a one-unit increase in each predictor variable, providing direct interpretation and statistical significance. To avoid multicollinearity, we remove one reference category from each categorical group (no high school diploma, age 18–24, male). Coefficients reflect the relative effect of each included category compared to the respective reference one. Table III shows the results, omitting state-fixed effects for clarity.

All demographic and socioeconomic variables are statistically significant ($p<0.05$), indicating robust relationships between them and fitness center participation. The consistency of signs and magnitudes across the three months demonstrate temporal stability across the studied periods. Income shows that every \$10,000 rise in median household income increases fitness center visitation by a proportion of 0.0132 (1.32%). In education, up to high school diploma completion is slightly negatively associated, but a 10% increase in bachelor's degree holders corresponds to a 2.5% higher visitation rate relative to areas with greater proportions of residents without a high school diploma. Similarly, master's+ show strong positive correlations, though slightly lower than bachelor's degree impacts. A 10% higher female proportion raises visitation by 0.718%, contrasting prior step-count studies where males were more active [10], suggesting women may prefer structured exercise environments. Age effects show modest but still significant effects. Adults aged 45-64 demonstrate the highest fitness center visit rates, followed by the 25-44 group. These patterns differ from general physical activity research measuring step counts, where younger individuals typically record higher daily steps [10]. This difference might occur because visiting fitness centers requires financial investment.

We utilize SHAP values [20] to interpret the predictions of LightGBM, the top-performing model. Intuitively, SHAP measures how much each feature shifts a prediction from its baseline by averaging the change in model output with

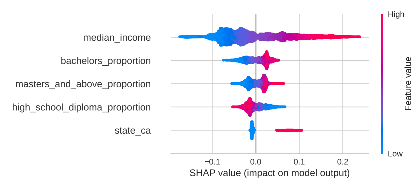

Fig. 3: Feature importance showing the top 5 features. Each dot represents a CBG, with color indicating feature values, and horizontal position showing impact direction and magnitude.

and without that feature across all feature subsets. Figure 3 shows the five most important features. Income ranks first followed by education level. Higher income and proportion of bachelor's or master's+ increase visitation, whereas higher proportions with only high school education is important towards the opposite direction. The fifth most important feature is whether the CBG is in California.

Figure 4 shows the partial dependence relationships between income and the other socioeconomic and demographic indicators for predicting fitness center visit rates per CBG in November, 2019. The plots for October and December contain identical patterns. Most features follow a consistent positive or negative effect. Higher proportions of residents without diplomas or with only high school education correlate with lower incomes (blue dots) and negative SHAP values, reducing predicted visitation. In contrast, more bachelor's or above show the opposite patterns. Female proportion shows a modest positive association. The 18-24 age group is negatively associated, except in CBGs where they exceed 75%, where fitness visit rates increase substantially. We found that these CBGs predominantly correspond to universities (87 of 162 CBGs with more than 90% young adults). Residents aged 45–64 are positively associated with visitation, especially in the highest-income CBGs. For the 25–44 group, visitation rises until they make up 40% of residents, then declines.

TABLE III: OLS regression results for predicting the proportion of fitness center visitors per CBG.

Variable	October 2019 Coef.	October 2019 Lower CI (95%)	October 2019 Upper CI (95%)	November 2019 Coef.	November 2019 Lower CI (95%)	November 2019 Upper CI (95%)	December 2019 Coef.	December 2019 Lower CI (95%)	December 2019 Upper CI (95%)
Constant	0.063	0.048	0.077	0.070	0.056	0.085	0	-0.015	0.014
Median Income	1.324×10^{-6}	1.300×10^{-6}	1.350×10^{-6}	1.323×10^{-6}	1.300×10^{-6}	1.350×10^{-6}	1.334×10^{-6}	1.310×10^{-6}	1.360×10^{-6}
High School Prop.	-0.0206	-0.027	-0.015	-0.0232	-0.029	-0.017	-0.0086	-0.014	-0.003
Bachelor's Prop.	0.2515	0.246	0.257	0.2624	0.257	0.268	0.2572	0.251	0.263
Master's+ Prop.	0.2048	0.198	0.212	0.2211	0.214	0.228	0.1957	0.189	0.203
Female Prop.	0.0718	0.063	0.081	0.0739	0.065	0.083	0.0705	0.062	0.079
Age 25-44 Prop.	0.0161	0.008	0.024	0.0179	0.010	0.026	0.0567	0.049	0.064
Age 45-64 Prop.	0.0220	0.014	0.030	0.0282	0.020	0.037	0.0784	0.0707	0.087
Age 65+ Prop.	-0.0312	-0.039	-0.024	-0.0252	-0.033	-0.018	0.0168	0.009	0.024

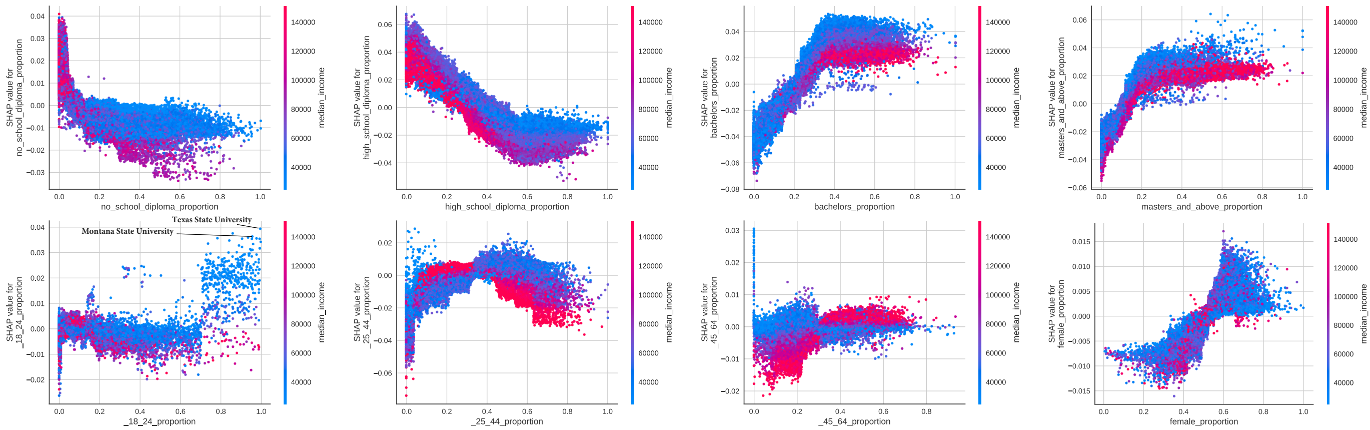

Fig. 4: Partial dependence plots showing the relationship of income with demographic and socioeconomic factors. Each dot is a CBG from the training dataset of the ML model. The SHAP value indicates how important the variable on the x-axis is for the prediction. A negative value means it is important towards predicting less proportion of visits to fitness centers.

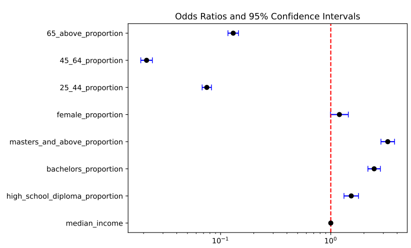

Fig. 5: Impact of socioeconomic and demographic variables for visiting a fitness center in the home CBG.

D. Spatial Proximity Barrier

We analyzed visits within and outside users' residential CBGs and their association with demographic and socioeconomic factors using logistic regression predicting home-CBG visits. Residents without a high school diploma, males, and the 18–24 age group were set as reference categories. Figure 5 shows that young adults (18–24) are the most likely to visit a fitness center within their home CBG, whereas adults aged 45–64 are more inclined to travel outside. Higher educational attainment increases the likelihood of home-CBG visits, while income has no association and gender effects are minimal.

V. Discussion

Previous research identified disparities in physical activity, typically measured by step counts, which are sensitive to extreme temperatures [8]. Safety concerns, such as neighborhood crime or inadequate lighting, further discourage park use, particularly in disadvantaged areas [25]. Our study expands this body of work by focusing on fitness center visits instead of steps, enabling the quantification of socioeconomic and demographic disparities in a domain requiring financial investment and therefore subject to stronger inequities. Identifying individuals with at least one visit per period effectively measures access rather than activity frequency. While our datasets lack complete population coverage and sub-CBG granularity, strong predictive accuracy and consistent statistical significance across time periods support the robustness of our findings. From a policy perspective, our results support targeted interventions, such as subsidized community fitness programs [26], improved accessibility through strategic placement in underserved neighborhoods [27], and public health agencies creating affordable fitness opportunities [28].

Consistent with prior findings, income and education show positive associations with participation. However, fitness center data also reveal distinctions from step-count measures. While earlier studies linked female populations with lower activity [10], we find a positive relationship, suggesting structured environments like Fit4Mom or Pilates may enhance female engagement. Addressing gender disparities thus requires ensuring equitable access to female-oriented facilities. Moreover, previous research linked younger age with higher activity levels, but fitness facility usage only revealed this pattern in university areas, likely due to free gym access. Targeted subsidies for low-income young adults could therefore alleviate age and income-related disparities.

Finally, our primary objective was to identify and quantify socioeconomic and demographic drivers rather than maximize predictive accuracy, as understanding which factors influence participation is more actionable for policy. Moreover, our results provide correlations and not causations. Achieving higher predictive accuracy, and potentially causations, would require integrating additional datasets capturing local gym density, pricing, accessibility, and more factors. Nevertheless, our results show that neighborhood-level socioeconomic and demographic factors alone can explain over half the variation in fitness center usage. Our large-scale correlational findings provide valuable guidance for policy design, which is essential for reducing future epidemics of inactivity [10].

VI. Conclusion

In this work, we integrated U.S. census socioeconomic and demographic indicators with over 50 million daily visits and identified disparities in fitness center participation. We found that income and education are positively associated with fitness center visits and negatively associated with fast-food restaurants. We further developed ML models that explained more than half of the variance in fitness center participation, solely based on CBG-level socioeconomic and demographic profiles. We analyzed linear regression coefficients and computed SHAP values for LightGBM, quantifying each feature's contribution to predictions. We further constructed partial dependence plots based on SHAP values to illustrate how pairs of features jointly influence predictions. Moreover, we used logistic regression to examine which factors influence the likelihood of residents visiting a fitness center within their home CBG. Our results reveal significant disparities in fitness center participation. Our framework supports data-driven policy interventions, such as optimizing new facility locations to serve undeserved socioeconomic groups.

Acknowledgment

Code available at: https://github.com/GeoIoan98. We thank Veraset for the dataset. This project was funded by European Union's Horizon Europe research and innovation program under Grant Agreement No: 101135826.

References

[1] P. C. Hallal, L. B. Andersen, F. C. Bull, R. Guthold, W. Haskell, and U. Ekelund, "Global physical activity levels: surveillance progress, pitfalls, and prospects," *The lancet*, vol. 380, no. 9838, pp. 247–257, 2012.

[2] F. Bull, J. Willumsen, G. Stevens, and T. Strain, "Global levels of physical inactivity in adults: off track for 2030," 2024.

[3] K. Shuval, Q. Li, K. P. Gabriel, and R. Tchernis, "Income, physical activity, sedentary behavior, and the 'weekend warrior'among us adults," *Preventive medicine*, vol. 103, pp. 91–97, 2017.

[4] K. Molanorouzi, S. Khoo, and T. Morris, "Motives for adult participation in physical activity: type of activity, age, and gender," *BMC public health*, vol. 15, pp. 1–12, 2015.

[5] T. Althoff, R. Sosič, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec, "Large-scale physical activity data reveal worldwide activity inequality," *Nature*, vol. 547, no. 7663, pp. 336–339, 2017.

[6] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of covid-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[7] A. Derlyatka, O. Fomenko, F. Eck, E. Khmelev, and M. T. Elliott, "Bright spots, physical activity investments that work: Sweatcoin: a steps generated virtual currency for sustained physical activity behaviour change," *British Journal of Sports Medicine*, vol. 53, no. 18, pp. 1195–1196, 2019.

[8] F. Pontin, N. Lomax, G. Clarke, and M. A. Morris, "Socio-demographic determinants of physical activity and app usage from smartphone data," *Social Science & Medicine*, vol. 284, p. 114235, 2021.

[9] G. Ioannou, G. Pallis, M. Dikaiakos, and C. Nicolaides, "Utilizing large-scale human mobility data to identify determinants of physical activity," *Scientific Reports*, vol. 15, no. 1, p. 3420, 2025.

[10] A. E. Bauman, R. S. Reis, J. F. Sallis, J. C. Wells, R. J. Loos, and B. W. Martin, "Correlates of physical activity: why are some people physically active and others not?," *The lancet*, vol. 380, no. 9838, pp. 258–271, 2012.

[11] L. G. Sylvia, E. E. Bernstein, J. L. Hubbard, L. Keating, and E. J. Anderson, "A practical guide to measuring physical activity," *Journal of the Academy of Nutrition and Dietetics*, vol. 114, no. 2, p. 199, 2013.

[12] J. L. Hicks, T. Althoff, R. Sosic, P. Kuhar, B. Bostjancic, A. C. King, J. Leskovec, and S. L. Delp, "Best practices for analyzing large-scale health data from wearables and smartphone apps," *NPJ digital medicine*, vol. 2, no. 1, p. 45, 2019.

[13] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[14] C. Nicolaides, D. Avraam, L. Cueto-Felgueroso, M. C. González, and R. Juanes, "Hand-hygiene mitigation strategies against global disease spreading through the air transportation network," *Risk Analysis*, vol. 40, no. 4, pp. 723–740, 2020.

[15] M. K. Chen, J. A. Chevalier, and E. F. Long, "Nursing home staff networks and covid-19," *Proceedings of the National Academy of Sciences*, vol. 118, no. 1, p. e2015455118, 2021.

[16] T. M. McMichael, "Covid-19 in a long-term care facility—king county, washington, february 27–march 9, 2020," *MMWR. Morbidity and mortality weekly report*, vol. 69, 2020.

[17] H. Zhao, B. J. Mailloux, E. M. Cook, and P. J. Culligan, "Change of urban park usage as a response to the covid-19 global pandemic," *Scientific Reports*, vol. 13, no. 1, p. 19324, 2023.

[18] K. L. Milkman, D. Gromet, H. Ho, J. S. Kay, T. W. Lee, P. Pandiloski, Y. Park, A. Rai, M. Bazerman, J. Beshears, *et al.*, "Megastudies improve the impact of applied behavioural science," *Nature*, vol. 600, no. 7889, pp. 478–483, 2021.

[19] S. Aral and C. Nicolaides, "Exercise contagion in a global social network," *Nature communications*, vol. 8, no. 1, p. 14753, 2017.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[21] T. Y. Lee, S. H. Ha, and H. M. Lee, "Predicting mental health using lifelog data: Application of median resampling and data augmentation techniques," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 793–797, IEEE, 2024.

[22] M. Lugner, A. Rawshani, E. Helleryd, and B. Eliasson, "Identifying top ten predictors of type 2 diabetes through machine learning analysis of uk biobank data," *Scientific reports*, vol. 14, no. 1, p. 2102, 2024.

[23] R. F. Squire, "What about bias in the safegraph dataset?." https://medium.com/safegraph/what-about-bias-in-the-safegraph-dataset-b95b7208893c, 2019.

[24] E. Moro, D. Calacci, X. Dong, and A. Pentland, "Mobility patterns are associated with experienced income segregation in large us cities," *Nature communications*, vol. 12, no. 1, p. 4633, 2021.

[25] A. L. Bedimo-Rung, A. J. Mowen, and D. A. Cohen, "The significance of parks to physical activity and public health: a conceptual model," *American journal of preventive medicine*, vol. 28, no. 2, pp. 159–168, 2005.

[26] C. Barragan, "Social relationships and the importance of community-based fitness programs (cbfp)," *Journal of Women & Aging*, vol. 33, no. 4, pp. 428–441, 2021.

[27] A. S. Mendoza-Vasconez, S. Linke, M. Muñoz, D. Pekmezi, C. Ainsworth, M. Cano, V. Williams, B. H. Marcus, and B. A. Larsen, "Promoting physical activity among underserved populations," *Translational Journal of the American College of Sports Medicine*, vol. 1, no. 14, pp. 125–132, 2016.

[28] H. Speake, R. J. Copeland, S. H. Till, J. D. Breckon, S. Haake, and O. Hart, "Embedding physical activity in the heart of the nhs: the need for a whole-system approach," *Sports Medicine*, vol. 46, pp. 939–946, 2016.