Data Imputation and Social Determinants of Health

Bay-esian Watch:
Stephanie Balarezo, Carmin Berberich, Cara Dunnavant, Megan Schaeb

Socially Determined

Dr. Alan Lattimer	Chief Analytics Officer, Socially Determined
Mike Ackermann	PhD Student, Virginia Tech
Quiyana Murphy	PhD Student, Virginia Tech

07 December 2022

Executive Summary

Our team, Bay-esian Watch, worked with Socially Determined this semester to improve risk score models surrounding the five social determinants of health (SDoH): economic strain, food insecurity, housing instability, transportation barriers, and health literacy challenges. The data used in their models can be missing or uncertain. We have explored imputation methods to predict an individual's missing or uncertain data field with increased accuracy and improve Socially Determined's economic risk model.

We began with imputing wealth score data field bins for individuals in the provided Infutor data. Literature reviews were performed to find correlations between other fields in the Infutor data that could decrease the margin of error. We discovered strong relationships between wealth score and estimated household income, homeowner status, and education level. Three classification methods were then implemented utilizing the correlated fields to predict an individual's wealth score bin.

The Pareto distribution was identified as a national distribution for wealth and was used to sample a wealth value. The distribution is specific to each individual based on demographics such as state of residence, age, race, and education level. After simulating Pareto distributions, and thorough discussion with Socially Determined, the determination of the final Pareto distribution was not achievable due to a lack of real-world data for comparison. The beginning of the imputation process was then repeated within the educational sector of the SDoH; however, we concluded with the literature review due to time constraints and have provided our recommendations moving forward with the exploration of the field.

1 Problem Statement

Socially Determined envisions a future where the social determinants of health risk factors are identified and quantified, potential interventions are scored and prioritized, and community interventions can be taken using these analytics. They currently utilize predictive models to assess an individual's risk score, but are often missing data or do not know its validity with certainty. Uncertain data needed to be conditioned on a background distribution to gain insights into the ground truth. Missing data needed to be imputed with the most likely values for fields within the dataset. Our team's main focus was to address these issues and update their models by leveraging data correlations, additional data sources, and literature review. The improvement of their predictive risk score models could assist Socially Determined by informing their clients on how to reduce the severity of adverse health outcomes and help achieve health equity.

We have received expert assistance from the following individuals throughout this project:

Mike Ackermann and Quiyana Murphy were project mentors and met with us weekly throughout the semester to guide us through our project. They both were extremely helpful when we encountered challenges and assisted us with many technical components as well as our research and writing. They demonstrated their expertise in distributions, data analysis, and research skills.

Dr. Alan Lattimer helped us understand the goals and aspirations of Socially Determined through his enthusiasm and optimism around public health and it inspired a passion for our project. He also took the time to invite us to the Socially Determined Headquarters and explained the products produced by the company.

2 Ethical Considerations

An area where our project has important ethical considerations is model limitations due to bias in our dataset, which is sourced from either Infutor or American Community Survey (ACS), which contains demographics data from the U.S. census. Infutor data consists only of individuals who have a substantial online presence (e.g. using social media regularly, online shopping, etc.) and own credit cards. Additionally, the specific Infutor data we were given by Socially Determined is sampled from two considerably wealthy counties in the DMV area: Fairfax, VA, and Washington, D.C. Our project was built around data that was not representative of the entire U.S. There is nonresponse bias regarding the ACS data—some may choose not to participate—and it does not account for undocumented immigrants or those without homes. In order to deal with these considerations, Socially Determined behaves ethically by acknowledging the uncertainty in their models and the bias in their data. Furthermore, we researched distributions during our literature review that will apply to the entire country for various variables surrounding the social determinants of health, in hopes to increase representation.

3 Literature Review

To familiarize ourselves with different aspects of our project, we reviewed literature on the social determinants of health, data imputation, and Bayesian networks.

3.1 "Addressing Health Equity and Social Determinants of Health Through Healthy People 2030"

To delve deeper into the topic of our project, we researched and reviewed the SDoH and their part in building healthier communities. "Addressing Health Equity and Social Determinants of Health Through Healthy People 2030" is a peer-reviewed article from the *Journal of Public Health Management and Practice* that provides background information on health equity and strategies to achieve it in the U.S. It also emphasizes that SDoH are "mostly responsible for health inequities" since they "influence a wide range of health and well-being outcomes, functioning, and quality-of-life outcomes and risks" (Gómez et al., 2021, p. S251). This idea reinforces the mission of Socially Determined as they work to combat health inequity and gauge the impacts of SDoH in communities.

3.2 "Comparison of Performance of Data Imputation Methods for Numeric Dataset"

We investigated multiple sources to familiarize ourselves with data imputation practices. A research paper titled "Comparison of Performance of Data Imputation Methods for Numeric Dataset" contrasted seven data imputation methods: mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian linear regression, linear regression, non-Bayesian, and random sample on five different datasets. The researchers evaluated each method's performance using the normalized root mean square error (RMSE) method (Jadhav et al., 2019, p. 913-933). This source was helpful as we considered different data imputation methods to use as well as different error metrics, such as RMSE, for method evaluation.

3.3 "Bayesian Network Data Imputation with Application to Survival Tree Analysis"

To better understand Bayesian networks, we reviewed the journal article "Bayesian network data imputation with application to survival tree analysis". This article explained the structure of Bayesian networks and the use of posterior distributions as a method of data imputation (Rancoita et al., 2016, p. 373-387). Moreover, discussed methods of model evaluation, such as utilizing mean squared error (MSE), and the possibility of bias among missing data. With information regarding missing data, Bayesian networks, data imputation, and model evaluation, this article should serve as a useful reference as we work to improve Socially Determined's individual-level economic risk model.

4 Project Criteria

Our team had originally established Data Exploration and Selection of Data Imputation Methods as the two most substantive components of our project. However, after some discussion with our coach and sponsor, we decided to replace Data Exploration with Evaluation of Data Imputation Methods as a substantive component.

4.1 Selection of Data Imputation Methods

- *Language.* Code should be written in Python, specified by the client.
- *Number of Imputation Methods.* We aim to develop 3 to 5 imputation methods.

- *Dynamic.* The models should be able to intake any variable amount of data. A model is dynamic if it can handle the largest and smallest datasets that Socially Determined has utilized in the past.
- *Code Format.* The code should be readable and replicable. On a scale of 1 (easy) to 5 (hard), a panel of Socially Determined analysts would score the code legibility as a 1 or 2.

4.2 Evaluation of Data Imputation Methods

- *Accuracy.* Performance of the data imputation method should have at most 10% error when compared to a known full dataset.
- *Sample Size.* The sample size when testing a data imputation method should be 16,000 individuals for the economic social determinant of health because 16,000 out of 48,000 individual records have no missing economic data. Sample sizes will vary as we get into other determinants of health, but should always be the number of records with no missing data specific to the variable.
- *Speed.* The model evaluation method should take no longer than 0.005 seconds per 10,000 individuals in the testing set when run on a laptop with an Intel Core i7-1065G7 processor.

5 Selected Solutions

After considering our project criteria and discussion with our sponsor we chose to move forward with the following selected solutions.

5.1 Classification Methods Selected

Three supervised machine learning methods–decision tree classification, random forest classification, k-nearest neighbor (kNN) classification–were investigated in hopes of filling in gaps in Socially Determined's wealth score data. All three algorithms are common techniques and fairly simple to implement, making them useful tools for classification problems.

A decision tree "works on a set of decisions derived from the data and its behavior" (Roy, 2020). The model uses the attributes of a training dataset to form "nodes," or points where decisions need to be made, such that every path through nodes results in a distinct classification output corresponding to unique values in the dependent variable (*What is a Decision Tree?*, n.d.). There are multiple algorithms employed to decide when to split a node and how many sub-nodes to split into (Chauhan, 2022). Each node requires a binary response, either the condition presented at the node is "true" or it is "false"; that true or false response determines which node will be encountered next. When predicting with the model, each observation will traverse through the nodes to determine in which category it is best suited.

Random forest classification, "consists of a large number of individual decision trees that operate as an ensemble" (Yiu, 2019). Each of the individual trees within the random forest will spit out its own class prediction and then the class predicted most frequently becomes the overall prediction for the model (Yiu, 2019). The algorithm uses two different methods to ensure there is

little to no correlation between each decision tree; bagging allows each individual tree to randomly sample from the dataset with replacement so that there are distinct trees, and feature randomness is when “each tree in a random forest can pick only from a random subset of features" when splitting a node, causing more variation and diversification while lowering the correlation amongst trees (Yiu, 2019). The most important parameter with the random forest is the number of trees included, and a rule of thumb is to use between 50-400 (Ellis, n.d.).

The kNN algorithm works by creating groups among the individuals in the training data based on similar characteristics in the predictor data fields. An individual’s group can then be predicted by determining which of the groups the individual is most similar to. The k value in the algorithm determines how many neighbors will be checked to classify the individual’s group; different values of k can lead to overfitting or underfitting (*What is the K-nearest neighbors algorithm?*, n.d.). The optimal k value is usually the square root of the total number of samples (Band, 2020).

5.2 Distribution Method Selected

Socially Determined currently imputes quantitative wealth values with the median of an individual’s wealth score bin, but we aimed to improve this by instead sampling from a background distribution within the bounds of the wealth score bin. We identified the Pareto distribution as the background distribution for U.S. net worth. We then sought to parameterize Pareto in order to determine the correct shape for an individual’s wealth distribution based on demographics such as state, age, race, and education level.

The Pareto distribution is named after an Italian engineer who created the distribution to reflect the distribution of wealth in a society. One of the main components of this distribution is that it follows the 80/20 rule; in societal wealth terms, this could mean that 20% of society owns 80% of the land/resources (*Pareto distribution,* n.d.). The probability density function for the Pareto distribution is as follows:

$$p(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}},$$

where α represents the shape of the distribution and x_0 represents the scale (*Pareto distribution,* n.d.).

6 Results

Below we have described the exploration of the correlations between the variables discussed in the literature review. These variables were then used as predictors in multiple classification methods in an effort to predict wealth score. Lastly, the process of conditioning the Pareto distribution is outlined.

6.1 Relationship Exploration for Wealth Score

Throughout our literature review, we discovered relationships between wealth score and EHI, and wealth score and homeowner status (Hira et al, 2012; Holzhauer, 2021). We investigated these correlations within the Infutor data in addition to wealth score and education.

6.1.1 Wealth Score and EHI

Within the Infutor data, which contains observations on 44,988 individuals, EHI and wealth score are categorical variables with ordered bins; bin "1" for EHI and bin "A" for wealth score represent the lowest value ranges, while bin "12" for EHI and bin "M" for wealth score represent the highest ranges.[1] Figure 1 below shows the proportion of individuals with a certain wealth score given their EHI score. For example, the bar corresponding to a wealth score of "M" and an EHI score of "12" (the far right bar in Figure 1) shows that approximately 70% of individuals in the dataset with EHI scores of "12" have a wealth score of "M". The plot reflects that individuals with low EHI scores often have low wealth scores, and vice versa—those with high EHI scores tend to have high wealth scores. The direct relationship between EHI and wealth is less defined in the middle of the plot (middle range of the wealth score bins) but is still observable. Figure 1 supports the findings from the literature review that wealth score and household income are correlated (Hira et al, 2012). A chi-squared test resulted in an extremely small p-value (displayed as 0.0 in Python), which confirmed that EHI is significantly correlated with wealth score within the Infutor data.

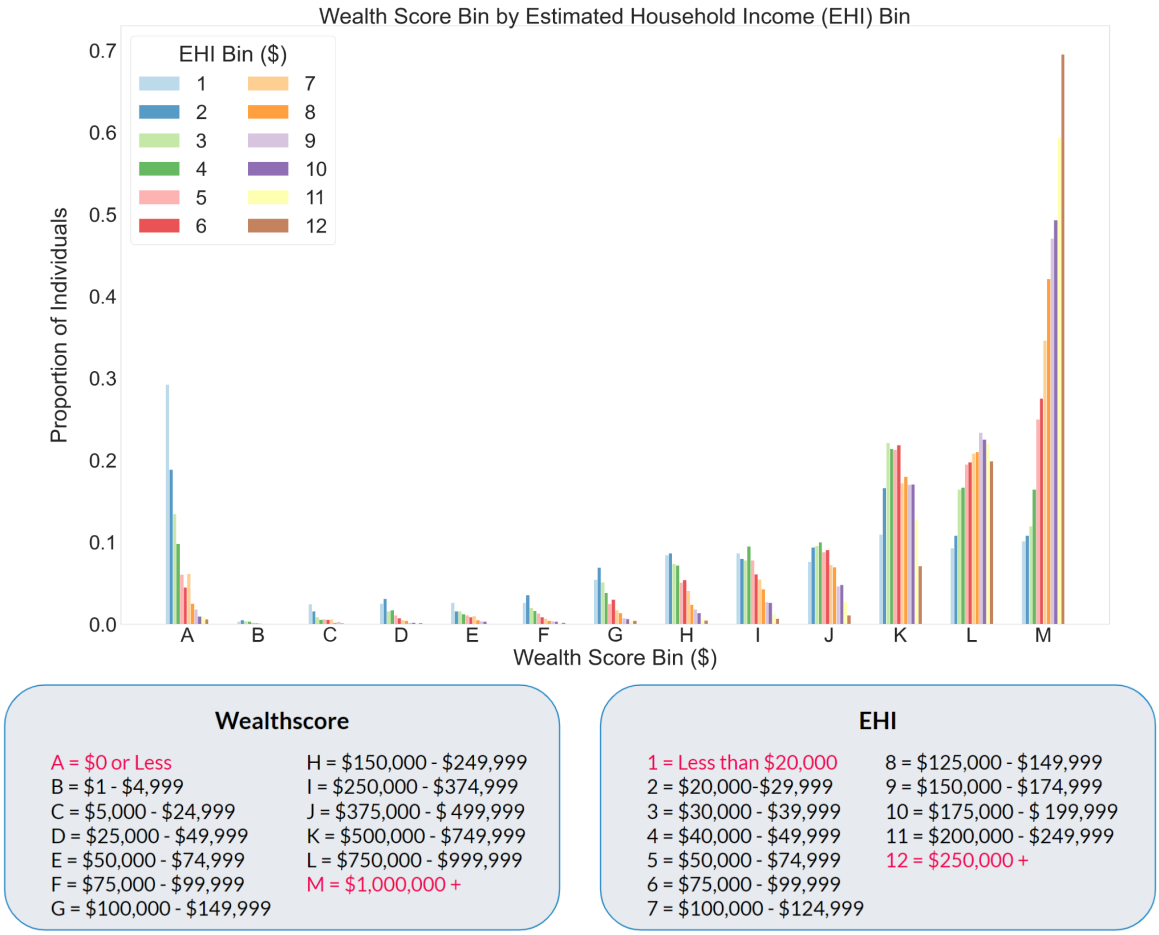

Figure 1: Wealth Score Bin Given EHI Bin visualized using data from Infutor

[1] The EHI bins in the Infutor dataset are represented by alphabetic characters "A" through "L". In Section 2.2.1, the EHI bins are redefined using the numeric characters "1" to "12" ("A" became "1", "B" became "2", and so on) to distinguish EHI from wealth score, which uses alphabetic bins "A" through "M".

6.1.2 Wealth Score and Homeowner Status

Figure 2 shows a logarithmic count of how many people either rent or own their home in each wealth score bin. The logarithm of the original count was taken in order to clearly view the differences in the proportions of renters and homeowners. It can be observed that as wealth scores increase, the less likely individuals are to rent a home. It is important to note that the lowest wealth score bin, "A," has the highest count of renters compared to any other bin, whereas the highest wealth score bin, "M," seemingly has no renters. Figure 2 supports the findings from the literature review that homeowners will have a higher median net worth and validated the idea that homeowner status is significant to one's wealth score (Holzhauer, 2021). A chi-squared test resulted in an extremely small p-value (displayed as 0.0 in Python), which confirmed that homeownership is significantly correlated with wealth score within the Infutor data.

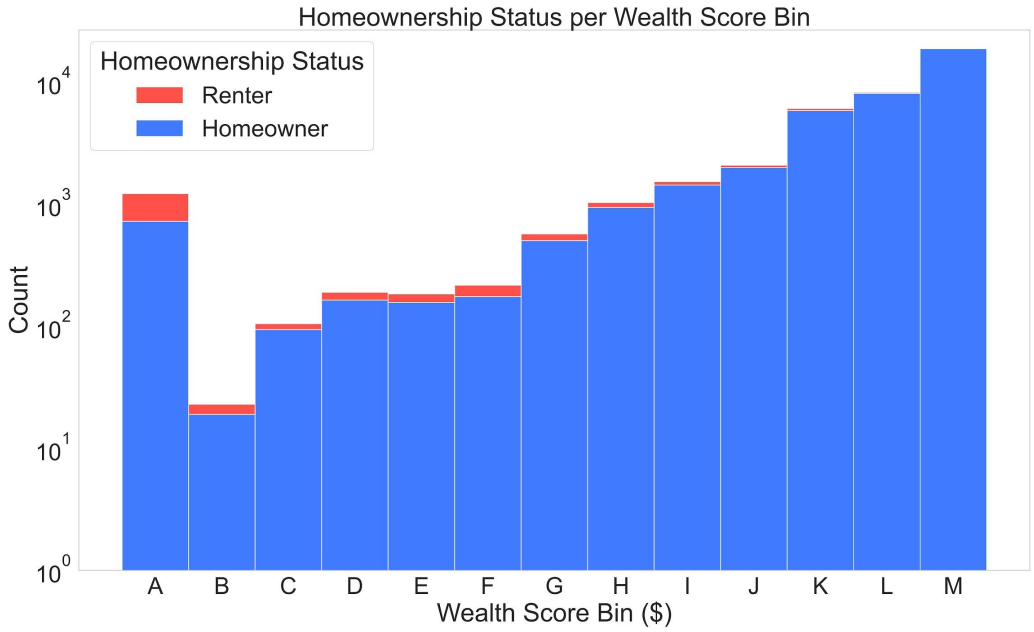

Figure 2: Wealth Score Bin Given Homeownership visualized using data from Infutor

6.1.3 Wealth Score and Educational Attainment

Figure 3 displays the logarithmic count of individuals in each 'wealthscr' bin with regard to their highest completed education level; the logarithm was taken for the same reason as with homeowner status. Interestingly, the lowest wealth score bin, "A," has a high count of individuals who completed graduate school. We assumed that individuals could be in this bin due to student loan debts, failure of finding a successful job after graduation, low wages from assistantships or research, etc. A chi-squared test resulted in an extremely small p-value (displayed as 0.0 in Python), which confirmed that educational attainment is significantly correlated with wealth score within the Infutor data.

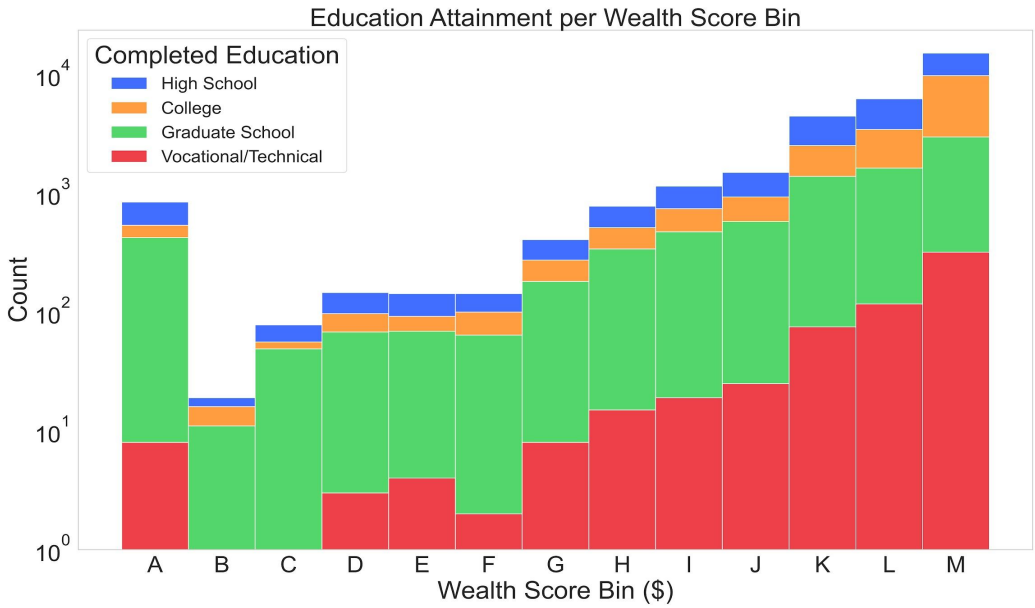

Figure 3: Wealth Score Bin Given Education Level visualized using data from Infutor

6.2 Classification Models

To predict an individual's wealth score, two types of classification models were constructed using EHI, homeowner status, and educational attainment as predictor variables.

6.2.1 Probabilities of Missing Data Based on Missing Data Fields

Before implementing any wealth score classification techniques, the probability of having missing data in one data field given that there is missing data in another field was explored. Table 1 presents these probabilities for wealth score and the three data fields investigated in Section 2.2. The rows represent what data field was missing while the columns represent how likely it was to also be missing that variable. For row i and column j in Table 1, let $\mathbf{A}_j$ be the set of individuals with missing data in the data field specified by column j and let $\mathbf{A}_i$ be the set of individuals with missing data in the data field specified by row i. Then, the value in Table 1 located at (i, j) is

$$(\text{Table 1})_{i,j} = \frac{\#\left(A_i \cap A_j\right)}{\#\left(A_i\right)}.$$

For instance, if there is a missing value for one's wealth score, there is a 54% chance that their EHI is also unknown. An example of the calculation can be seen below.

$$\frac{\#\ of\ missing\ 'ehi'\ values\ when\ 'wealthscr'\ was\ missing}{total\ \#\ of\ missing\ 'wealthscr'\ observations} = \frac{6}{11} = 0.5455$$

We repeated this calculation to find the probabilities pertaining to EHI, homeowner status, and education. The wealth score field has low probabilities of missing data due to the fact that only 11 individuals are missing values, whereas 728 individuals are missing EHI, 4,582 are missing homeowner status, and 13,949 are missing education. Despite a relatively low number of missing

data values, the wealth score field's importance to the economic risk model ensured that wealth score imputation was a top priority in this project.

Table 1: Missing Data Probability Matrix calculated with data from Infutor

	Wealth Score	EHI	Homeowner Status	Education
Wealth Score	1	0.5455	0.636	0.636
EHI	0.0082	1	0.918	0.988
Homeowner Status	0.0015	0.1458	1	0.938
Education	0.0005	0.0515	0.308	1

6.2.2 Classification Model Implementation

From the relationships seen between wealth score and EHI, homeowner status, and education, a decision tree classification model, a random forest model, and a kNN model were fit using different combinations of predictors to impute wealth scores. The same process was used with all three algorithms. The first step was to create a full dataset, so any row that contained missing data in either the 'wealthscr' column or any of the predictor columns was dropped from the set. After removing the rows with missing data, the remaining 30,747 individuals were broken up into training and testing sets, with 80% of the data being used for training and the other 20% for testing. From the training data, three different combinations of predictors (all three predictors, just EHI and homeownership, and just EHI and education) were transformed into dummy variables and used to fit the three classification algorithms. Each decision tree, random forest, and kNN model was run 15 times using different training and testing sets for each trial.

All three models were built using the Scikit-Learn library in Python, which is commonly abbreviated as 'sklearn.' The decision tree algorithm from sklearn.tree did not require any parameters. The random forest algorithm from sklearn.ensemble was run with 100 trees in the forest (n_estimators=100) since our model did not consist of too many features and still aligned with the rule of thumb previously mentioned. The kNN algorithm from sklearn.neighbors was run with k=175 (the square root of 30,747, the total number of non-missing observations). Other k values were explored, but we chose to follow kNN recommendations and continue with the square root of the total number of non-missing observations.

6.2.3 Classification Model Results

The average accuracy, rounded to three significant figures, of each classification method for the 15 trials was calculated by the following equation, where **A** is the number of correct predictions in a trial and **B** is the total number of predictions in that trial. At the suggestion of our sponsor, a correct prediction includes any prediction in its actual bin or in a bin directly next to the actual bin. The exception to this one-bin margin of error on both sides is with bin "A" and bin "M," where the one-bin margin is only on one side.

$$Accuracy = \frac{A}{B}.$$

The average accuracy of each predictor combination for the decision tree model, random forest model, and kNN model are shown in Table 2.

Table 2: Average Accuracy for Decision Tree Classifier Models, Random Forest Classifier Models and kNN Classifier Models

Predictors Included	Decision Trees	Random Forest	kNN
ehi, educationcd, homeownercd	0.718	0.720	0.721
ehi, homeownercd	0.707	0.707	0.710
ehi, educationcd	0.712	0.714	0.720

Including the margin of error in the accuracy calculation increased the average accuracy by at least 40% for each combination of predictors. The decision tree, random forest, and kNN models that use all three predictors provide the best results out of the different predictor combinations for each classification technique. From this analysis, kNN classification looks to have a slightly higher accuracy compared to the decision tree and random forest models, but the accuracy values for all methods are extremely close.

The limitation of any of these decision tree, random forest, or kNN models is that the predictor data must be available (i.e. not missing from the dataset). Additionally, since the training and testing sets are selected randomly, the average accuracy score will change slightly each time the model is run. In conclusion, decision tree, random forest, and kNN are all viable supervised learning methods to impute wealth score data, although kNN does provide the highest accuracy when including the margin of error for all three predictor combinations; therefore, kNN may be the best method for wealth score imputation.

6.3 Conditioning the Pareto Distribution on Wealth Score

To sample a wealth score bin and value for an individual the Pareto distribution was explored.

6.3.1 Pareto Parameters

The Pareto distribution has two parameters: x_0, representing scale, and α, representing shape. There is no way to determine the correct value of x_0 from the provided set of Infutor data. The parameter α was calculated using randomly generated x_0 values, and then the simulated distributions were compared against the histogram of individual wealth scores from publicly available data. We used the following equation, where T represents total median wealth which is detailed further in the report, to simulate α values from known median wealth values and a sample of randomly selected x_0 values in the range [0, 1] (*Pareto distribution,* n.d.).

$$\alpha = log_{(\frac{T}{x_0})}\sqrt{2}$$

The more specific the conditions are to fit the Pareto, the more accurate the wealth samples could be for individuals, resulting in a more accurate computation of risk score for the individuals in

our dataset. We were able to construct code that could potentially use median wealth values for individuals' data from the U.S. Census Bureau's Survey of Income and Program Participation (SIPP) based on state, age, race, and education level in the specifications for the Pareto distribution. This code is intended to be utilized by those at Socially Determined who have access to the described demographic data. We utilized the State Wealth Asset Ownership table for median wealth based on states (*State-Level Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020*) and the Wealth Asset Ownership table for median wealth based on age, race, and education level (*Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020*). An example of state median wealth data is shown below in Table 3, and the data for the rest of the demographics (i.e. age, race, and education level) are presented similarly.

Table 3: Sample Data for State Median Wealth from SIPP (*State-Level Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020*); (B) denotes a missing median net worth due to lack of data

State	**Median Net Worth**
Alabama	88,910
Alaska	(B)
Arizona	149,300
Arkansas	78,100
California	247,500

The overall median wealth logic is implemented by calculating a proportion for each demographic for an individual with the formula

$$p = \frac{|M-A|}{A},$$

where M is the median wealth of the demographic and A is the average wealth of the demographic. In this formula, the median wealth comes from the "Median Net Worth" column corresponding to each demographic and the row that matches the individual's category within that demographic. The average wealth is calculated by averaging the "Median Net Worth" of the demographic. Once a proportion is calculated for each demographic, the total median wealth, T, is calculated with the following formula:

$$T = \frac{S\,(p_1 + p_2 + p_3)}{3},$$

where S is state median wealth and each p_i for $i = 1, 2, 3$ corresponds to the proportions calculated from the individual's age, race, and education level's median net worths. Regarding the Pareto distribution parameters, since it is impossible to calculate α without x_0 and vice versa, we had to randomly generate x_0 values and calculate the corresponding α values.

6.3.2 Pareto Sensitivity

We tested the sensitivity of the α values to different extremes of wealth values to be wary of a potential impact. As seen in our code, a slight inverse relationship was discovered between imputed wealth values and the value of α. However, we determined this relationship would not have a significant impact on the distribution.

After checking the sensitivity of α to extreme wealth values, we simulated Pareto distributions for an individual with the following demographics: white, from the state of Virginia, in the age range of 55-64, and possessing a Bachelor's degree. Figure 4 visualizes these Pareto distributions with parameters calculated from the median wealth value for an individual with these demographics and five randomly generated x_0 values in the range [0, 1].

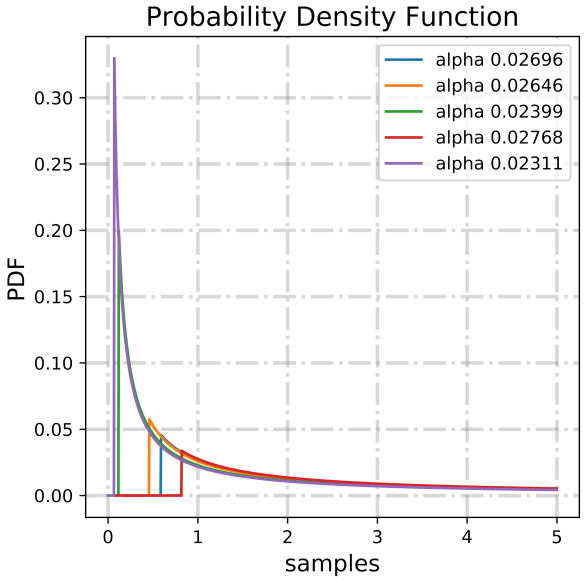

Figure 4: Various Pareto Distributions for an Individual (who is white, from the state of Virginia, in the age range of 55-64, and possessing a Bachelor's degree)

6.4 Relationship Exploration for Education

Throughout our literature review, we discovered relationships between education and EHI, homeowner, and gender (*Homeownership by Education,* 2022; *Median Household Income by Education*, 2022; Parker, 2021). We investigated these correlations within the Infutor data.

6.4.1 Education and EHI

Figure 5 below shows the proportion of individuals with a certain educational attainment level given their EHI score. For example, the bar corresponding to "High School" and an EHI bin of "1" (the far left bar in Figure 5) shows that just over 50% of individuals in the dataset with EHI scores of "1" concluded their education at the high school level. The plot reflects that individuals with low EHI scores have often attained lower levels of education and vice versa. The direct relationship between EHI and education is visible primarily at the "High School" and "Graduate School" levels. While there is a relatively high proportion of individuals with an EHI in the highest bin ("12") who finished their education at the college level, the EHI score proportions do not appear to vary substantially. No distinct relationship can be detected for the vocational/ technical school level. Figure 5 supports the findings from the literature review that income and educational attainment are correlated (*Median Household Income by Education*, 2022). A chi-squared test resulted in an extremely small p-value (displayed as 0.0 in Python), which confirmed that EHI is significantly correlated with educational attainment within the Infutor data.

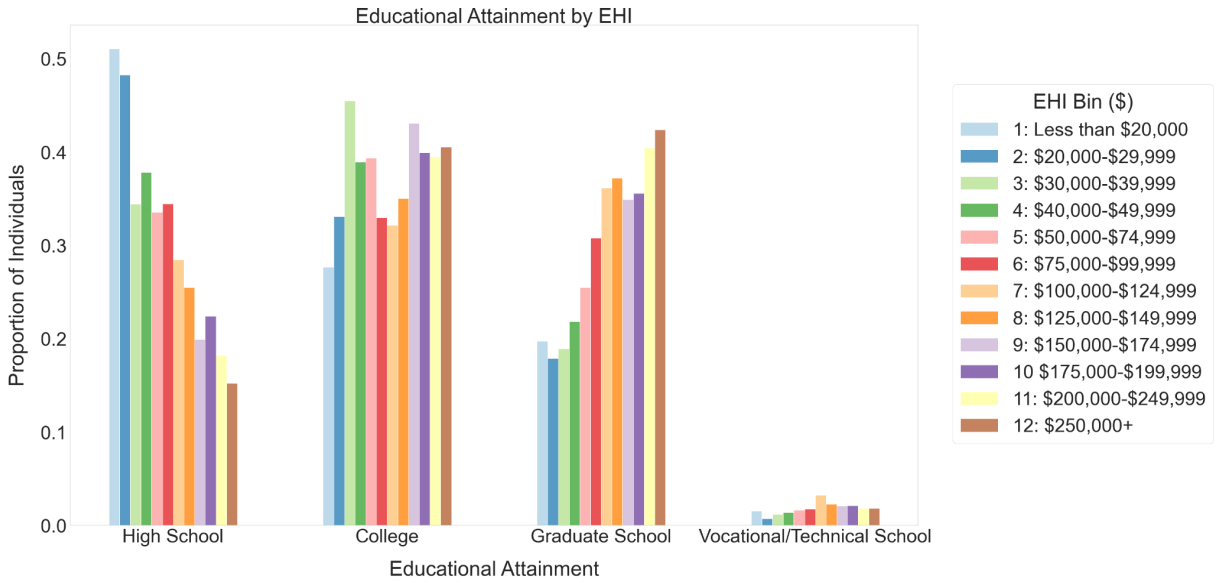

Figure 5: Educational Attainment by EHI Bin visualized using Infutor data

6.4.2 Education and Homeowner Status

Figure 6 displays the probability of being a homeowner or renter given an individual's level of completed education. We observed that those who attended graduate school are approximately twice as likely to be a homeowner than a renter. Those with lower levels of education, high school and college, have a combined 80% probability of being a renter compared to the 60% probability of being a homeowner. Figure 6 supports the findings from the literature review that homeowners are more likely to be degree-holders with higher levels of education (*Homeownership by Education,* 2022). A chi-squared test resulted in a p-value of 4.56e-25, which confirmed that homeownership is significantly correlated with education attainment within the Infutor data.

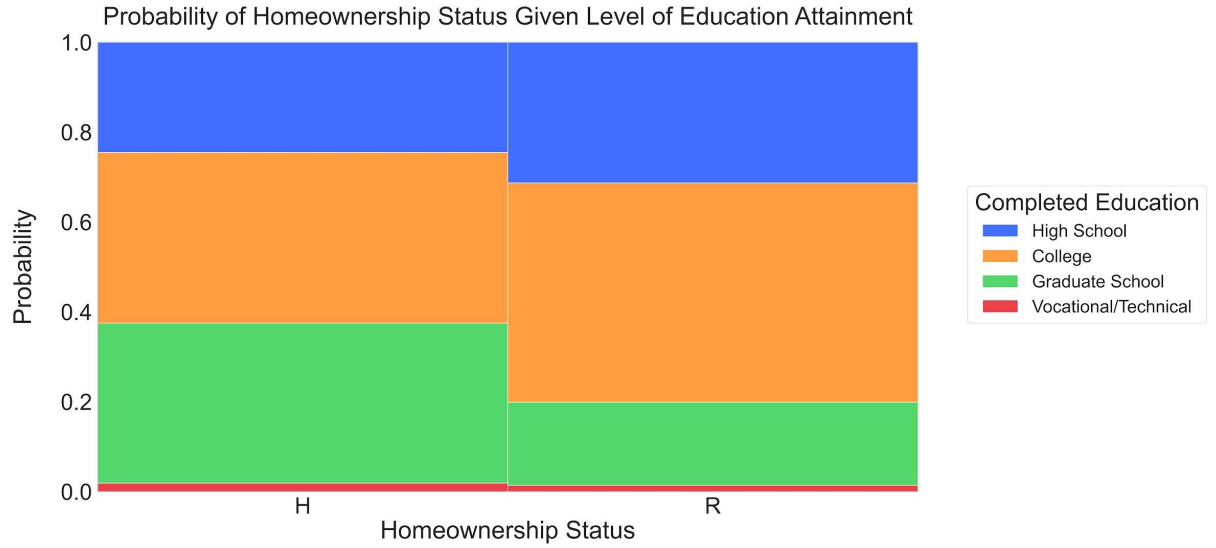

Figure 6: Probability of Homeownership Status Given Level of Education Attainment visualized using Infutor data

6.4.3 Education and Gender

Figure 7 shows the frequency of gender in each category of education. We noted that females in our dataset are more likely to have attained education in all four levels. This observation confirms the idea from the literature review that women are more likely to be enrolled and complete higher education than men (Parker, 2021). A chi-squared test resulted in a p-value of 7.50e-08, which confirmed that gender is significantly correlated with education attainment within the Infutor data.

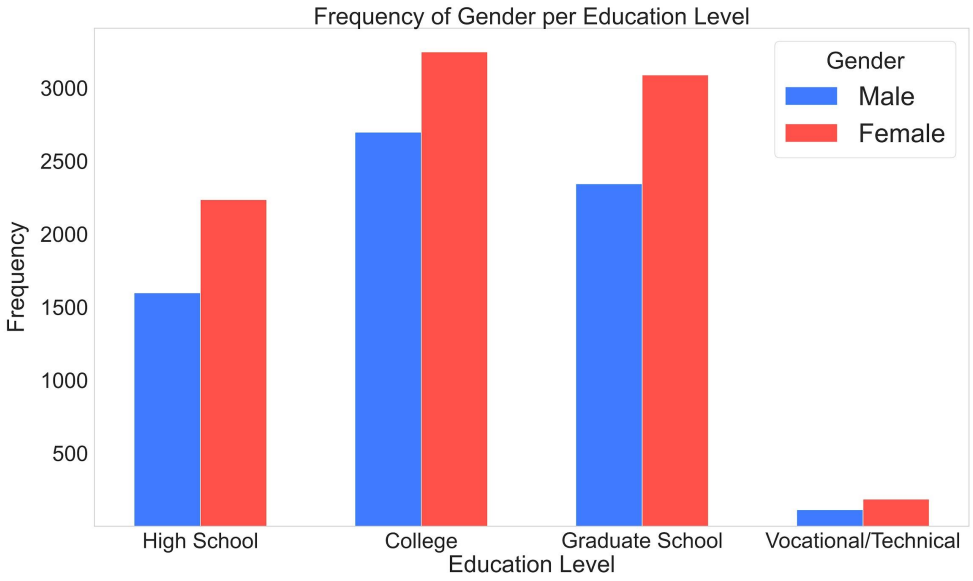

Figure 7: Frequency of Gender per Education Level visualized using Infutor data

7 Limitations

The biggest limitation we've encountered has been the lack of data publicly available concerning wealth, hindering our ability to fit a background distribution. After identifying the Pareto distribution as the optimal background distribution for wealth in the United States, we sought to parameterize this distribution for individuals utilizing certain demographics. For example, distinct Pareto distributions could be fit for each individual based on state of residence, race, age, and education level. The more specific the conditions are to fit Pareto, the more accurate the wealth samples could be for individuals, resulting in a more accurate computation of risk score for the individuals in our dataset. After creating different sets of Pareto parameters using our developed methods, which are detailed in the Results section, we were not able to continue because we did not have known parameters with which to compare them. Our sponsor agreed that continuing further with the wealth model was outside the realm of data science and better suited for a social scientist to continue researching.

8 Interpretation of Results

As discussed earlier, our results show that many attributes of an individual in a risk score model can be predicted using information from other data fields. For example, one's wealth score can be predicted using estimated household income (EHI), homeowner status, and educational

attainment. Supervised learning techniques such as decision trees, random forest, and kNN are all viable ways of predicting one's wealth score bin, with kNN having a slightly higher accuracy than the other two. Our results also imply that educational attainment can be predicted using EHI, homeowner status, and gender, although classification techniques were not yet run to predict this field. The information in this report can be applied to update risk score models pertaining to the SDoH if the process can be replicated on unbiased data to validate our results and someone with access to the necessary population data can use our research to determine the correct parameters for the wealth Pareto distribution. All of these results and the processes have been detailed in reports specifically created for Socially Determined. The code written and datasets used have also been uploaded to their private GitHub repository for future reference if they wish to apply these methods to other data fields.

9 Team Roles

Each team member played an important role in the production of our Capstone project. Members worked to their strengths and provided versatility when other members were in need of help. Outlined below are the contributions of each individual member.

Stephanie Balarezo

Technical Contributions: Stephanie coded the simulation of the Pareto distribution parameters, wrote code to make a Pareto distribution unique for individuals based on their demographics, and coded to test the sensitivity of Pareto distribution to wealth values.

Nontechnical Contributions: Stephanie led the team meetings, delivered the Elevator Pitch, contributed to weekly reports for the client, contributed to Tech Memos, delivered the demo for the Tools and Techniques Workshop, contributed to the Wealth Score report requested by the client, and will deliver a portion of the final presentation.

Carmin Berberich

Technical Contributions: Carmin collaborated with Cara to create the kNN model for wealth score, created the random forest model for wealth score, the NaN probability matrix for wealth score and its predictors, and plotted with Seaborn to research relationships within the following variables: homeowner status, wealth score, educational attainment, and gender.

Nontechnical Contributions: Carmin contributed to weekly reports for the client, contributed to Tech Memos, delivered a major portion of the Midterm Presentation, reviewed literature for the wealth score and educational attainment fields, contributed to the Wealth Score Report requested by the client, and will deliver portions of the final presentation

Cara Dunnavant

Technical Contributions: Cara collaborated with Carmin to create the kNN model for wealth score, coded the simulation of the Pareto distribution parameters, and accessed the

datasets needed to create a unique Pareto distribution based on the individual's demographics.

Nontechnical Contributions: Cara contributed to weekly reports for the client, contributed to Tech Memos, delivered the PowerPoint portion of the Tools and Techniques Workshop, contributed to the Wealth Score Report requested by the client, and will deliver portions of the final presentation.

Megan Schaeb

Technical Contributions: Megan created decision trees for the wealth score variable and plotted with Seaborn to find relationships between the wealth score and EHI, and educational attainment and EHI.

Nontechnical Contributions: Megan contributed to weekly reports for the client, contributed to Tech Memos, delivered a major portion of the Midterm Presentation, contributed to the Wealth Score Report requested by the client, reviewed literature for the wealth score and educational attainment fields, and will deliver parts of the final presentation.

10 Conclusions and Future Work

At this point we could not find any general U.S. wealth data on the SIPP (population data) besides the median wealth for state, age, race, and education level. Since we could not build our distribution using our Infutor data (sample data), we had no way to compare our simulated distributions to real-world data and determine the best pair of parameters. We concluded this to be a good stopping point in terms of the wealth score variable exploration. In the future when more information is available, and a correct Pareto distribution is decided based on the best pairs of parameters, the next steps would be to set the bounds of the bins against the distribution and sample a random value from said bin in the distribution to be assigned as an individual's wealth amount.

However, Socially Determined has acknowledged that the sample of Infutor data used to conduct the analysis in this report contains undercoverage sampling bias. Those in the provided Infutor dataset were from some of the wealthiest counties in America and they all had credit card information which indicates that the responses were skewed. In the future, the processes described herein should be replicated on unbiased data to ensure the validity of the results.

In terms of the educational field exploration, we stopped our process after finding correlations with the field. Moving forward, Socially Determined should implement the classification methods we described using the found correlations as predictor variables.

If our team were to recreate this project, we would explore more categorical imputation and evaluation methods because time was initially used to explore numerical imputation methods when we should have been focusing on categorical methods.

11 Bibliography

Band, A. (2020, May 23). *How to find the optimal value of K in KNN?* Towards Data Science. Retrieved from https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb

Chauhan, N. S. (2022, February 9). *Decision Tree Algorithm, Explained*. KDnuggets. Retrieved from https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Ellis, C. (n.d.). *Number of trees in random forests*. Crunching the Data. Retrieved December 4, 2022, from https://crunchingthedata.com/number-of-trees-in-random-forests/

Gómez, C. A., Kleinman, D. V., Pronk, N., Gordon, G. L. W., Ochiai, E., Blakey, C., ... & Brewer, K. H. (2021). Practice Full Report: Addressing Health Equity and Social Determinants of Health Through Healthy People 2030. *Journal of Public Health Management and Practice*, 27(6), S252.

Hira, T. K., Sabri, M. F., & Loibl, C. (2012). Financial Socialization's impact on investment orientation and household net worth. *International Journal of Consumer Studies*, *37*(1), 29–35. https://doi.org/10.1111/ijcs.12003

Holzhauer, B. (2021, August 24). *Here's the average net worth of homeowners and renters*. CNBC. Retrieved from https://www.cnbc.com/select/average-net-worth-homeowners-renters/

Homeownership by Education: Degree-Holding Owners Surge as Those Without High School Drop 30%. (2022, April 6). Point2 Homes. Retrieved December 4, 2022, from https://www.point2homes.com/news/us-real-estate-news/homeownership-by-education-us.html

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10), 913-933.

Median household income by education U.S. 2021. (2022, October 11). Statista. Retrieved December 4, 2022, from https://www.statista.com/statistics/233301/median-household-income-in-the-united-states-by-education/

Pareto distribution (n.d.). Wikipedia. Retrieved November 22, 2022, from en.wikipedia.org/wiki/Pareto_distribution.

Parker, K. (2021, November 8). *What's behind the growing gap between men and women in college completion?* Pew Research Center. Retrieved December 4, 2022, from https://www.pewresearch.org/fact-tank/2021/11/08/whats-behind-the-growing-gap-between-men-and-women-in-college-completion/

Rancoita, P. M., Zaffalon, M., Zucca, E., Bertoni, F., & De Campos, C. P. (2016). Bayesian network data imputation with application to survival tree analysis. *Computational Statistics & Data Analysis*, 93, 373-387.

Roy, A. (2020, November 6). A Dive Into Decision Trees. Medium. Retrieved from https://towardsdatascience.com/a-dive-into-decision-trees-a128923c9298

State-Level Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020 (2022, August 31). United States Census Bureau. Retrieved from https://www.census.gov/data/tables/2020/demo/wealth/state-wealth-asset-ownership.html

Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020 (2022, August 31). United States Census Bureau. Retrieved from https://www.census.gov/data/tables/2020/demo/wealth/wealth-asset-ownership.html

What is a Decision Tree? (n.d.). Master's in Data Science. Retrieved December 2, 2022, from https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/

What is the K-nearest neighbors algorithm? (n.d.). IBM. Retrieved from https://www.ibm.com/topics/knn#:~:text=The%20k%20value%20in%20the,as%20its%20single%20nearest%20neighbor.

Yiu, T. (2019, June 12). *Understanding Random Forest*. Medium. Retrieved December 6, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

"We have neither given nor received unauthorized assistance on this assignment."