Journal of High School Science

Predicting affordability of a family paying medical and healthcare expenses – development and validation of a predictive model

Gao L

Submitted: April 8, 2023, Revised: version 1, October 1, 2023
Accepted: October 14, 2023

Abstract

The rising costs of medical and healthcare have led to unaffordable medical and healthcare expenses. Inability to afford proper and needed healthcare can have adverse effects on children's health as they grow, which may affect the development of society as today's children will eventually become key contributors of society. Our goal is to develop a practical and reliable model to predict whether families can afford their medical and healthcare expenses by identifying and evaluating the most significant factors, which can serve to improve healthcare affordability in the US. We use the 2021 National Survey of Children's Health dataset to develop the desired predictive model. Methods such as missing value imputation, dependent variable dichotomization, feature standardization, one-hot encoding, correlation, multi-collinearity test, logistic regression model, oversampling, regularization, and cross-validation are employed to identify the factors that have significant impacts on healthcare affordability. Model validation metrics , such as the ROC curve, AUROC score, and Wald testing were used to evaluate the performance of the predictive model. Our logistic regression model was developed with thirteen independent variables and one dependent variable. Among the thirteen predictor variables, twelve were statistically significant at 0.05 significance level. The model achieved an out-of-sample AUROC score of 0.91 and exhibited a robust ROC curve, indicative of accurate predictions. The developed logistic regression model can help in predicting whether a family can afford its medical and healthcare expenses and can be applied by insurance companies and welfare institutions to determine appropriate insurance premiums and levels of financial assistance tailored to different families, thus contributing to the overall improvement of medical and healthcare affordability.

Keywords

Medical and healthcare expenses, Logistic regression, Predictive model, National Survey of Children's Health, Correlation, Standardization, ROC curve, Machine learning, Insurance, Healthcare affordability

---

Leyi Gao, William P. Clements High School, 4200 Elkins Rd, Sugar Land, TX, USA, 77479.
Jasminegao222@gmail.com

1. Introduction

On Sept. 13, 2022, the U.S. Census Bureau published the 11.6 percent official poverty rate for 2021. This percentage represents 37.9 million individuals living in poverty. This figure does not include families with low incomes or poor financial status that are slightly above the poverty line (1). These families often allocate their entire income towards basic living expenses, leaving scant resources for healthcare, especially given the rising costs of medical insurance. The unaffordability of necessary healthcare can have a detrimental impact on the health of their children since access to health services is important for diagnosing potential health problems before they progress to advanced stages. Children can benefit significantly from healthcare services, and the inability to afford medical and healthcare expenses can lead to worsened health conditions, higher disability rates, low immunity, and long-term negative effects on children who are still in the developmental stage.

In a recent survey, it was found that 18% of Americans stated that "they would not be able to afford the healthcare they need" (2), meaning that about one in five Americans may not have access to necessary healthcare. If this situation continues to deteriorate, a significant number of children will be unable to receive proper healthcare, potentially affecting the overall welfare of society. While the government has made efforts to address this issue through programs like the U.S. Patient Protection and Affordable Care Act, some researchers argue that the extent of coverage can vary significantly depending on the type of insurance. For instance, a study conducted by JAMA Pediatrics found that children from publicly insured families (covered by Medicaid or CHIP) "experienced greater access to preventive medical and dental care" compared to children from privately insured families (88% to 83% for preventive medical and 77% to 73% for dental care). (3) In addition, families with private insurance "had the highest out-of-pocket costs (77% [75%-78%])" compared to those with public insurance —"Medicaid (26% [23%-28%]; $P < .001$) and CHIP (38% [35%-40%]; $P < .001$)". (3)

As children will eventually become a significant portion of the population, this current imbalance in insurance assistance for public health becomes an unresolved and increasingly pressing issue. Therefore, the objective of our study was to provide guidance to insurance companies and social assistance programs, enabling them to better support and identify families struggling to afford healthcare and facilitating them to adjust healthcare insurance plans to optimize resource allocation. In this study we aimed to predict whether a family would encounter problems in paying for medical and healthcare expenses based on various factors, including demographic status, physical and mental health conditions, health insurance coverage, and employment status, among others.

Literature search

Prior to our research, other scholars had presented other ideologies and methods to address the problem of unaffordable healthcare costs. For example, in 2017, Emanuel et al. proposed a method to measure the affordability of medical costs (4). In 2014, Pollitz et al. examined some circumstances and consequences of unpaid medical bills (5). In 2016, Hamel et al. analyzed the characteristics

of people with difficulty in paying for medical bills and the role of health Insurance (6). In 2022, Claxton et al. investigated how affordability of healthcare varies by income (7), while Schwartz et al. analyzed the affordability threshold of emergency department visits for families with private insurance (8). In 2016, Nyman and Trenz estimated the extent to which the Affordable Care Act helped increase affordability (9). In 2018, Cree et al. identified points to reach children in poverty with mental, behavioral, and developmental disorders (MBDD) (10). All these studies addressed the issue of unaffordable healthcare costs. However, there has been little research into how to develop a predictive model to assess whether families with diverse characteristics can afford medical and healthcare expenses. For example, Emanuel et al. proposed the Affordability Index, a ratio that divides the average cost of an employer-sponsored family insurance (ESI) policy by median household income, linking family income with insurance premiums (4). This Affordability Index can serve as a good way to illustrate the increasing financial burden of health insurance premium costs on average families receiving ESI. The advantages of the Index allow it to have several potential applications such as helping pertinent organizations and States to identify industries where health care costs vary disproportionately or to encourage health care organizations to control their prices increases that outpace the growth in household income. Despite its advantages and potential applications, the Affordability Index lacks the consideration of other potential factors affecting unaffordable health care costs and can only calculate a general ratio for the average population of the United States receiving ESI. This places limits on its measures and specificity. For example, it "may not reflect the financial burden of health care for those with Medicaid or Medicare coverage or the uninsured", and the cost of ESI only captures the direct costs (premiums) without considering the indirect financial burdens such as the time and travel costs associated with the health treatments. Comparing the Affordability Index with our predictive model, the Affordability Index provides a more basic approach that lacks certain variances when attempting to help improve the issue of unaffordable healthcare expenses, while our model focuses on a more precise and specific approach that allows to predict whether each family can afford health expenses according to their specific conditions, improving the lack of specificity of the Affordability Index (4).

Hamel et al. approached the issue of health care expenses by conducting a survey and using the data from the survey to find the relationship between problems paying medical bills with different demographic characteristics by logistic regression analysis, similar to our approach to the problem. They evaluated different features of those reported to have problems paying for their medical bills including household income, insurance status, deductibles, disability status, circumstances leading to problems paying medical bills, financial status, cost of the bills, age, race, sex, etc. Although their study mentions the use of logistic regression analysis, there is only one sentence, and it does not specify how they conducted this analysis, whether they used techniques such as correlation to pre-select factors to avoid the model being affected by non-significant predictors. Furthermore, the method does not mention whether they

performed logistic regression analysis after dividing the dataset into two parts for training and testing, and whether the model could avoid the overfitting and low precision challenges. Therefore, there could be potential errors in the results due to lack of machine learning and verification of the model. As we developed the model, we included the training and testing processes of the model, which could improve the precision of research results compared to their model (6).

In addition to the research conducted by Emanuel et al. And Hamel et al., Cree et al. attempted to enhance health care access and reduce costs. In their study, they analyzed data from the 2016 National Survey of Children's Health (NSCH) survey, a similar source to ours, on poverty levels, the use of publish assistance benefits, the use of health care services, and mental, behavioral, and developmental disorders (MBDDs) in children, in order to identify factors affecting children in poverty. This study identified that the lower the family income, the less access to health care services, and the higher the proportion of children diagnosed with MBDD. The study has potential implications for promoting public assistance programs, which "might offer collaboration opportunities for public health professionals and pediatricians to provide information, implement co-located screening programs or services, or facilitate connection to care". Nonetheless, the study does have some limitations: the data used is cross-sectional so it cannot "ascertain temporal associations or causality", the "sampling weights used to calculate nationally representative estimates" cannot completely avoid nonresponse bias, and there is potential for subjective parental reports on indicators in data utilized. When comparing with our study, Cree et al. focused on analyzing different factors affecting MBDD in children rather than factors affecting the affordability of health care. Therefore, the study cannot predict affordability. Moreover, unlike our study which employed a logistic regression model, Cree et al. did not use any statistical models in the analysis (10).

The significance of our research is that we built a machine learning model that can be used to predict whether a family can afford medical and health care expenses, filling the gaps in previous research on this topic. The result of this study could provide insights to insurance companies in determining their insurance premiums and support governmental or non-governmental welfare institutions in determining the various level of financial assistance that can be provided to different families based on their respective economic conditions and specific circumstances. Therefore, our research could help more people in society obtain affordable healthcare services and contribute to the improvement of the community.

2. Materials and Methods

2.1 Dataset

In this study, the dataset of 2021 National Survey of Children's Health (hereinafter referred to as 2021NSCH dataset) was utilized to identify the antecedents of healthcare expenditure affordability, such as factors related to the health condition of children, health insurance coverage, costs of medical and health care expenses, affordability of basic needs, employment status, special healthcare needs in Households, etc. The 2021NSCH dataset comprises 50,892 observations and 461 variables for families with children between 0

and 17 years old. NSCH is a national survey in the United States that provides rich data on various aspects of children's health and well-being, such as general and detailed health status, health care coverage and quality, demographic information, family structure, community, school, and social context.

2.2 Missing value imputation

It is common for real-world datasets to contain missing values variables. A missing value appears in the 2021NSCH dataset when there is a logical skip or no valid response. As most statistical models are unable to handle missing values present in the dataset, a convenient approach is to simply drop data containing missing values. However, depending on the quantity of missing values, such an approach may result in deleting too many sample points, which might introduce significant new selection bias and also carry the risk of losing valuable information that the classifier needs to learn model parameters. To avoid removing excessive sample points, another approach is to use the mode or mean to replace the missing values.

In this study, for the dependent variable, all missing values (15,086/29.6% of the observations) were dropped because replacing missing values of the dependent variable could introduce bias and make the prediction results less accurate. After deletion, 35806 observations remained.

For independent variables, any variables that had more than 25% of the observations with missing values were excluded from the analysis, resulting in a reduction of the number of variables from 461 to 237. Out of the remaining 237 variables, 192 of them had less than 25% of the observations with missing values. For these variables, missing values were replaced with the mode or median that were calculated based on the remaining observations (more than 75% of 35,806) with values, depending on whether they are quantitative or categorical variables. Specifically, missing values for categorical variables such as race, gender, and health conditions were replaced with the mode, while missing values for quantitative variables such as age, height, and weight were replaced with the median.

2.3 Dependent variable dichotomization

"A dichotomous variable, also known as a binary variable, is a categorical variable that can only take one of two values, usually represented as a Boolean (true or false) or an integer variable (0 or 1), where false or 0 typically indicates that the attribute is absent, and true or 1 indicates that it is present" (11).

In the 2021NSCH dataset, factors such as affordability for medical or health care expenses, health insurance coverage status, and employment status should be binary variables; they have two values but are not in binary type. Therefore, we adjusted the data of those variables to make them binary. For example, K3Q25 (Problems Paying for Medical or Health Care during the past 12 months) has an original value of 1 or 2. We replaced 2 with 0 to let 0 represent "No problem paying medical bills", and kept 1 to represent "Yes, had problems paying medical bills" to make K3Q25 binary.

2.4 Feature standardization for quantitative data

Different features of quantitative data usually have remarkably different value ranges. The technique of standardization ensures that different features carry equal weight, and the effects of different independent variables are comparable. First, the mean value (μ) and standard deviation (σ) are calculated for each variable. Then, each data point with respect to that feature is replaced by x^* calculated as:

$$x^* = \frac{x - \mu}{\sigma}.$$

In this study, the feature standardization process was done in SPSS (version 26, IBM), a commonly used program for statistical analysis in data science. We applied the feature standardization technique to quantitative variables such as age, weight, number of family members, times of moving to new addresses, poverty ratio, years in US, etc., to bring all features down to a common scale without distorting the differences in the range of their values. The rescaled data had a mean of 0 and a standard deviation of 1.

2.5 One-hot Encoding for nominal variables

A nominal variable is a kind of categorical variable whose levels are simply labels and do not contain any meaningful order. Although we want these categories to be equally weighted, problems often arise if we feed features directly into the model. The way to solve this problem is to create a binary dummy feature (variable) for each category of the nominal variable, a technique known as one-hot encoding. This step of creating dummy variables is repeated *n-1* times for each nominal variable with *n* categories of values (12). In this study, a total of 23 nominal factors required one-hot encoding, such as the race of the selected child, marital status, relation with child, insurance types, family structure, etc. After one-hot encoding in SPSS, a total of 83 dummy variables were created. For instance, the nominal factor of "Place Usually Goes Sick", has eight categories (1 Doctor's Office, 2 Hospital Emergency Room, 3 Hospital Outpatient Department, 4 Clinic or Health Center, 5 Retail Store Clinic or "Minute Clinic", 6 School (Nurse's Office, Athletic Trainer's Office), 7 Some other place, 8 Urgent Care Center). Therefore, a total of seven dummy variables were created for each value in the "Place Usually Goes Sick" variable, as shown in Table 1.

Table 1. One-hot encoding for “Place Usually Goes Sick”

		Dummy						
Categories of Place Usually Goes Sick	Place Usually Goes Sick	Doctor's Office	Hospital Emergency Room	Hospital Outpatient Department	Clinic or Health Center	Retail Store Clinic or "Minute Clinic"	School (Nurse's Office, etc.)	Some other place
Doctor's Office	1	1	0	0	0	0	0	0
Hospital Emergency Room	2	0	1	0	0	0	0	0
Hospital Outpatient Department	3	0	0	1	0	0	0	0
Clinic or Health Center	4	0	0	0	1	0	0	0
Retail Store Clinic or "Minute Clinic"	5	0	0	0	0	1	0	0
School (Nurse's Office, Athletic Trainer's Office.)	6	0	0	0	0	0	1	0
Some other place	7	0	0	0	0	0	0	1
Urgent Care Center	8	0	0	0	0	0	0	0

2.6 Correlation

Correlation is any statistical relationship, causal or not, between two randomly selected factors, normally referring to the degree of linear correlation between two variables. A correlation of +1 indicates that the variables are perfectly positively correlated, with a linear relationship 1 to 1. Thus, correlation provides both direction and strength. A negative correlation indicates that one variable decreases while the other variable increases. Values close to -1 or +1 represent a stronger relationship, while a correlation close to zero indicates a weak linear relationship between the two selected variables (13).

In this study, in order to identify which factors might be potential antecedents for the variable of interest (affordability for medical and health care expense), among 237 variables (after missing value deletion and imputation), we ran Pearson Correlations in SPSS for each possible pair of independent and dependent variable to determine whether there was a significant correlation between the two variables. We set a significance level of 0.05; the same significance level for the logistic regression model since correlation serves to select factors for the model. Significant factors with higher correlation coefficients and higher relevance to the affordability of healthcare were selected for further analysis by the logistic regression model. Regarding high relevance to the affordability of healthcare, according to *Health-Care Utilization as a Proxy in Disability Determination*, “access to health care is tied to the affordability of health insurance” and factors such as cost, family poverty level, and overall health condition of family members are common relevant factors for the population that cannot afford health expenses (14). In addition, the research also suggested that “the lack of health insurance has been identified as an important driver of health-care disparities" (14). A summary of the selected variables is shown in Figure 1.

Figure 1. Independent variables selected based on correlation coefficient and significance level

2.7 Multi-collinearity diagnostic test

Multicollinearity in regression analysis occurs when two or more predictor variables are highly correlated to each other, resulting in the inability to provide unique or independent information in the regression model. If the degree of correlation between variables is high enough, it can cause problems in fitting and interpreting the regression model (15).

In this study, we performed the multi-collinearity diagnostic test by checking the Variance Inflation Factor (VIF) on selected predictor variables in SPSS to ensure that the correlation and strength of correlation between these selected variables did not significantly overlap, thereby making the prediction model output unreliable.

A value of 1 indicates that there is no correlation between a given predictor variable and any other predictor variables in the model (15). A value between 1 and 5 indicates a moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention (15). A value greater than 5 indicates a potentially severe correlation between a given predictor variable and other predictor variables in the model, which means that the coefficient estimates and p-values in the regression output are likely unreliable (15).

Table 2. Multi-collinearity diagnostic test (Variance Inflation Factor)

Coefficients[a]								
		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
	(Constant)	0.094	0.002		61.994	0.000		
	ZK3Q21B	-0.056	0.002	-0.180	-32.566	0.000	0.669	1.494
	ZFPL_I1	-0.022	0.002	-0.069	-14.250	0.000	0.868	1.152
	ZTOTCSHCN	0.006	0.002	0.018	3.645	0.000	0.845	1.183
	ZACE1	0.077	0.002	0.247	47.731	0.000	0.765	1.308
	CUTHOURS	0.086	0.007	0.061	12.403	0.000	0.836	1.197
	AVOIDCHG	0.116	0.005	0.107	22.101	0.000	0.880	1.136
	ZA1_PHYSHEALTH	-0.006	0.002	-0.019	-3.425	0.001	0.642	1.559
	ZA1_MENTHEALTH	-0.002	0.002	-0.006	-1.018	0.309	0.645	1.550
	ZK3Q20	-0.009	0.002	-0.030	-5.032	0.000	0.559	1.789
	ZK3Q22	-0.009	0.002	-0.029	-4.862	0.000	0.577	1.733
	ZK2Q01	-0.007	0.002	-0.022	-4.311	0.000	0.773	1.293
	ZHOWMUCH	0.040	0.002	0.129	24.641	0.000	0.752	1.330
	ZATHOMEHC	0.004	0.002	0.012	2.346	0.019	0.788	1.269
a. Dependent Variable: y-variable								

According to the Table 2 above, none of the selected variables has a VIF value greater than 5, which indicated that multicollinearity was not be a problem in the regression model.

2.8 Logistic Regression

Predicted risks are calculated using logistic regression models. Logistic regression is a type of statistical model used to estimate the probability of an event occurring based on a set of independent variables that may be discrete, continuous, dichotomous, or a combination of these. Typically, it accepts a dichotomous dependent variable with a binary parameter, and one or more categorical or continuous independent variables.

The logistic regression model is represented by the formula

$$\ln\left(\frac{h_w(x^{i})}{1-h_w(x^{i})}\right)=w_0+w_1x_1+\ldots+w_nx_n$$

In the logistic regression, $h_w(x^{i})$ represents the probability of the sample being classified as the positive class, which represents unable to afford medical and healthcare bill, and each feature x_i has its specific weight w_i, where w_0 is the intercept while w_1 through w_n are the coefficients of the independent variables.

The task is to find a set of parameters $w_0,\ldots,w_n$ such that the cross-entropy cost function between the output $h_w(x^i)$ and the actual values y^i is minimized. (16)

$$J(w) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i \log\left(\mathrm{h_w}(x^i)\right) + (1-y^i)\log(1-\left(\mathrm{h_w}(x^i)\right)\right]$$

where, $h_w(x^i)$ is the predicted probability given the i-th input instance x^i, y^i is the true class label and m is the total number of input samples

In this study, we ran logistic regression model in Python via PyCharm (Community Edition 2022.1.3, https://www.jetbrains.com/edu-products/download/other-PCE.html).

2.8.1 Oversampling

In the data we used, the ratio between the minority class (coded as 1), where people reported having problems paying for medical and healthcare expenses, and the majority class (coded as 0), where people reported not having problems paying for medical and healthcare expenses, was highly unbalanced. Logistic regression models are sensitive to class imbalance; when the majority class dominates the dataset, the model tends to be biased towards predicting the majority class more frequently, leading to a high number of false negatives, misclassifying the minority instances as the majority class. (17)

To address the above issue, oversampling is employed as a preventive measure. In this approach, samples from the minority class are randomly selected, and these selected instances are duplicated to augment the dataset. This technique ensures a more balanced representation between the minority and majority classes, mitigating the adverse effects of class imbalance on the training process. (17)

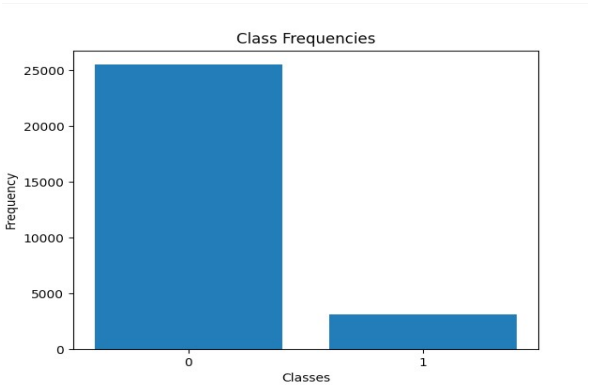

Figure 2. Training set class frequencies before oversampling

As discussed in section 2.2, we excluded 15,086 observations with missing values for the dependent variable, this left us with 35,806 remaining observations in our analysis dataset. To perform machine learning, we divided this dataset into two parts at an 80:20 ratio for training and testing, respectively. Before oversampling, as shown in Figure 2, the training set (80%/28,645 observations) consisted of 3,130 positive observations and

25,515 negative observations. After oversampling was performed, both the positive and negative classes were adjusted to 25515 observations.

2.8.2 Regularization and cross-validation

One of the primary concerns in training a machine learning model is to avoid overfitting. Overfitting occurs when the model attempts to capture the noise specific to the training set, which does not genuinely represent the true properties of the general data. As a result, the model's accuracy decreases because it essentially memorizes all the features of the training data without effectively learning the underlying patterns (18).

In order to secure better performance of our model, we attempted to reduce overfitting by performing regularization, a technique that penalizes the coefficients of variables in the logistic regression model by adding the coefficients of variables with a hyperparameter to the cross-entropy cost function (18). Each time a coefficient of a variable is updated to be significantly large, the value of the cost function was increased by the regularization term we added, which prompted the model to tune a large coefficient – penalizing and updating it to a smaller value (18). For regularization, we specifically chose to use an elastic-net regression, a variation of regularization that combines Lasso regression and Ridge Regression. Lasso regression, or L1 norm, involves adding a regularization term to "penalize absolute values of the coefficients" while setting less important or irrelevant variables in a regression to zero (19). On the other hand, Ridge Regression, or L2 norm, adds a regularization term to "penalize the size (square of the magnitude) of the regression coefficients" by making their values lower but never reaching zero (19).

By adding the regularization term (20):

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i \log\left(\mathrm{h_w}(x^i)\right) + (1-y^i)\log\left(1-\left(\mathrm{h_w}(x^i)\right)\right)\right] + \lambda\,[\alpha\sum_{\mathrm{j}=1}^{\mathrm{n}}|\mathrm{w_j}| + (1-\alpha)\sum_{\mathrm{j}=1}^{\mathrm{n}}\mathrm{w}_j^2]$$

α is a parameter to control the ratio of L1 and L2 penalties. For α=0 and 1, the ridge regularization and the Lasso regularization are calculated respectively. λ is the regularization strength and n is the number of the features

As a combination of both regularization methods above, elastic-net regression "allows a balance of both penalties, which can result in better performance than a model with either one or the other penalty on some problems" (21). It provides two hyperparameters, α and λ, that control the balance between the L1 and L2 penalties and the overall strength of regularization, respectively.

The default hyperparameters set the L1 penalty ratio and the L2 penalty ratio each to 50%, but they are not guaranteed to be the best ratio distribution for our model's prediction. Thus, we performed hyperparameter tuning through grid searching and 5-fold cross-validation over the α between 0 and 1 with a 0.1 step size and λ values from $1X10^{-5}$ to 100 to find to find the best parameters for our dataset (21).

The 5-fold cross-validation divides the training data into five equal parts, or 'folds', and

conducts five separate experiments to assess the model's performance with different regularization parameters. In each experiment, four folds are used for training, and one is reserved for validation, cycling through all the folds so that each is used once for validation. The set of regularization parameters that gives the best average performance across all experiments is selected, providing a robust method for tuning hyperparameters, and helping to prevent overfitting or underfitting (22). We chose 5-fold cross-validation because the total number of training cases (28,645) fell between 10,000 and 100,000. This choice struck a balance between reducing variation by increasing the number of folds while ensuring an adequate number of observations in each subset to train the model.

2.9 Model validation metrics

"A confusion matrix is a chart or table that summarizes the performance of a classification model or algorithm for machine learning processes" (23). It displays the number of True Positive, False Positive, True Negative, and False Negative by comparing the predicted outputs of a classifier model and the actual outputs.

True-positive rate (TPR), also known as sensitivity, of a model reflects the percentage of times that a classifier or model correctly predicts the positive class. It can be calculated by the equation TPR=TP/(TP+FN). In this equation, TP represents true positive outcomes, which are predicted outputs that match actual positive outputs in the dataset. FN represents false negative outcomes, which are predicted outputs that do not match the actual negative outputs in the dataset (23).

False-positive rate (FPR) of a model represents the percentage of times that a model incorrectly predicts the positive class (a Type I error). It can be calculated by the equation: FPR=FP/(FP+TN). FP represents false positive outcomes, which are predicted outputs that do not match the actual positive outputs in the data set. TN represents true negative outcomes, which are predicted outputs that match the actual negative outputs in the data set (23). The specificity of the model is 1 – FPR, and it reflects the percentage of times that a model correctly predicts the negative class.

A Receiver Operator Characteristics (ROC) curve is a graph that shows the performance of a classification model at all classification thresholds, plotting TPR vs. FPR. FPR values lie on the x-axis and TPR values lie on the y-axis for different predicted output thresholds between 0.0 and 1.0. A ROC curve that bows up to the upper left side of the plot indicates a high possibility of prediction accuracy. A straight, linear ROC curve close to showing y = x indicates a random guess with a probability 0.5 for either class in a two-class classification problem. A model with perfect classification skills would show a ROC curve starting from the bottom left corner of the plot to the top left corner and then across the top to the top right corner as it represents a 100% probability of predicting correct positive and negative outputs (24).

The Area Under the ROC curve score (AUROC) represents the measure of separability/ classification. It is equal to the probability that a classifier model will rank a randomly selected positive instance higher than a randomly selected negative instance.

AUROC is scale-invariant and threshold invariant.

AUROC measures the area underneath the entire ROC curve from (0,0) to (1,1) on a graph and the range of values of AUROC is between 0 and 1. AUROC = 1 indicates that the prediction model is perfect. The higher the AUROC, the better the model's ability to predict True Positives and True Negatives. When AUROC is equal to or approximately 0.5, one can interpret the score as an indication that the model has no discrimination ability to distinguish between the positive and negative classes. If AUROC is less than 0.5, it indicates that the model's predictions are worse than random guessing and we can invert predictions for negative and positive outputs to improve performance (24). One needs the AUROC score even when ROC curve is presented because ROC curve is a probability curve, and AUROC gives a quantitative measure of how well the model can distinguish between the positive and negative classes. Another reason to use AUROC in addition to ROC curve is that when multiple ROC curves are present and overlapping, it might be hard to distinguish which model has better performance. Using AUROC in this case helps us identify which model performs better overall. The confusion matrix, ROC Curve, and AUROC score in this study were calculated and plotted in Python via PyCharm (Community Edition 2022.1.3).

3. Results

3.1 Dependent variable and independent variables

Following the above-described methodology, the independent variables (factors) selected from the correlation analysis are shown in Table 3. The dependent variable in this research is the Affordability of Medical and Healthcare Expenses, as measured by Question F3 in the NSCH2021 survey.

Table 3. Dependent Variable and Selected Independent Variables

Dependent variables				
	Name	**Definition**	**Question in 2021NSCH Survey**	**Value**
1	K3Q25	Affordability of Medical and Health Care Expenses	Did your family have problems paying for the medical or health care expenses of the child?	1 = Yes, 0 = No
Independent variables				
	Name	**Definition**	**Question in 2021 NSCH Survey**	**Value**
1	CUTHOURS	Cut work Hours because of Health Conditions	During the past 12 months, have you or other family members: Cut down on the hours you work because of this child's health or health conditions?	1 = Yes, 0 = No
2	AVOIDCHG	Avoided Changing	During the past 12 months, have	1 = Yes, 0 = No

		Jobs to Maintain Health Insurance	you or other family members: Avoided changing jobs because of concerns about maintaining health insurance for this child?	
3	FPL_I1	Family Poverty Ratio (total family income to the family poverty threshold)	Family Poverty Ratio	50 = FPL ratio 50% or less 51 = FPL ratio 51% … 399 = FPL ratio 399% 400 = FPL ratio 400% or more
4	K2Q01	General Health Condition of Child	In general, how would you describe this child's health?	1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent
5	HOWMUCH	Cost of Medical Health Care over the past 12 Months	Including co-pays and amounts reimbursed from Health Savings Accounts (HSA) and Flexible Spending Accounts (FSA), how much money did you pay for this child's medical, health, dental, and vision care during the past 12 months?	1= $0 (No medical expense) 2= $1-$249 3= $250-$499 4= $500-$999 5= $1000-$5000 6= more than $5000
6	ACE1	Basic Needs Affordability	Since this child was born, how often has it been very hard to cover the basics, like food or housing, on your family's income?	1 = never, 2 = rarely, 3 = somewhat often, 4 = very often
7	A1_PHYS HEALTH	Adult's Physical Health Condition	In general, how is the Caregiver 1's physical health?	1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent
8	A1_MENT HEALTH	Adult's Mental Health Condition	In general, how is the Caregiver 1's mental health?	1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent
9	K3Q20	Health Insurance - Benefits Cover Service Needs	How often does this child’s health insurance offer benefits or cover services that meet this child’s needs?	1 = never, 2 = sometimes, 3 = usually, 4 = always
10	K3Q22	Health Insurance - Allow to See Provider	How often does the health insurance allow them to see the health care providers they need?	1 = never, 2 = sometimes, 3 = usually, 4 = always

11	K3Q21B	How Often Costs Reasonable	How Often child's medical Costs Reasonable?	1 = never, 2 = sometimes, 3 = usually, 4 = always
12	ATHOMEHC	Hours Spent Providing Home Health Care	In an average week, how many hours do you or other family members spend providing health care at home for this child?	1 = No need for health care at home 2 = < 1 hour per week, 3 = 1 to 4 hours per week, 4= 5 to10 hours per week, 5 = ≥11 hours per week
13	TOTCSHCN	Count of Children with Special Health Care Needs in Household	Number of Children with Special Health Care in the household?	0 = 0, 1 = 1, 2 = 2, 3 = 3, 4 = 4 or more

3.2 Logistic Regression Analysis and result interpretation

In our analysis dataset, comprising 35,806 observations, we split the data into two datasets in an 80:20 ratio for training and testing purpose. Specifically, 28,645 (80%) observations were used for model training and development, and the remaining 7,161 (20%) observations were used for model testing and validation.

Figure 3 shows the importance of factors affecting a family's ability to afford medical expenses, based on coefficients of the logistic regression model built from the training dataset.

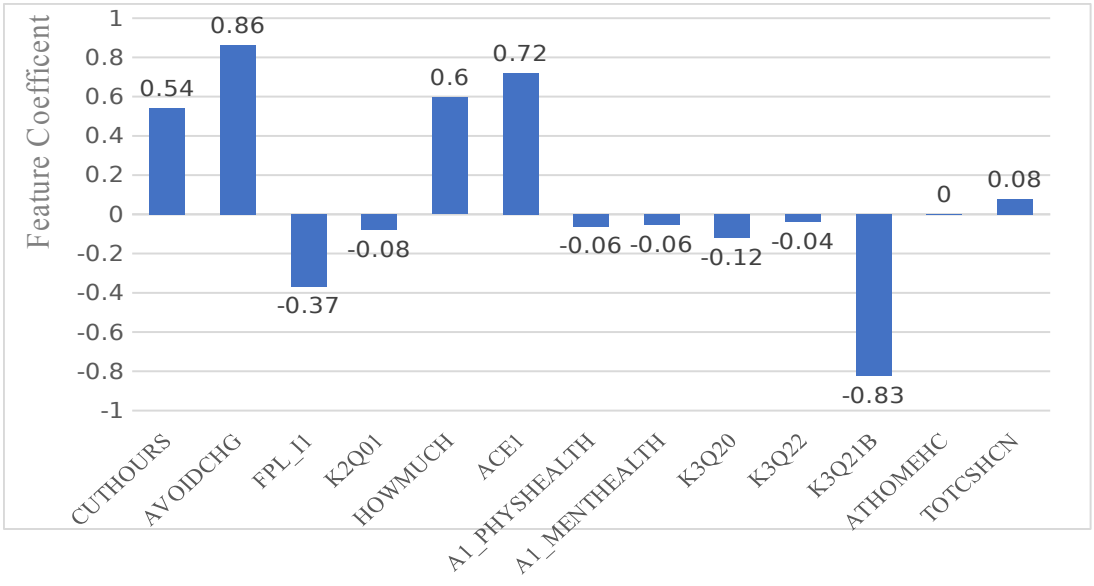

Figure 3. Logistic Regression Analysis Result _Coefficients of Independent Variables

Out of the 35,806 observations in the final analysis dataset, 3,897 (10.88%) families did have difficulty in paying their medical expenses (self-reported), and 31,909 (89.12%) families did not (self-reported). According to the result of logistic regression analysis shown in Figure 3, at a significance level of 0.05, the predictive model of family medical and healthcare costs affordability is:

Predicted logit of family medical costs affordability =

-1.2241+0.5410× Cutting work Hours because of Health Conditions

+0.8617× Avoided Changing Jobs to Maintain Health Insurance

-0.3674× Family Poverty Ratio

-0.0804× General Health Condition of Children

+0.5960× Cost of Medical Health Care over the past 12 Months

+0.7219× Basic Needs Affordability

-0.0616× Adult's Physical Health Condition

-0.0560× Adult's Mental Health Condition

-0.1183× Health Insurance Benefits Coverage of Medical Service Needs

-0.0394× Health Insurance Allowing to See Providers Needed

-0.8253× How Often Costs Reasonable

+0.0026× Hours Spent Providing Home Health Care

+0.0784× Number of Children with Special Needs in household

The following is an interpretation of the coefficients of the factors. At the 0.05 significance level,

1] On average, controlling other variables, family with caregivers cutting work hours in the past 12 months because of child's health condition has 71.6% ($e^{0.5410}$-1=0.7160) more likely of having problem paying for medical and health care expenses than families not cutting working hours in the past 12 months due to child's health condition. **2]** On average, controlling other variables, family with caregivers that have avoided changing job in the past 12 months to maintain health insurance is 137% ($e^{0.8617}$-1=1.367) more likely to have problem paying for medical and health care expenses than family with caregivers not to avoid changing jobs in the past months to maintain health insurance. **3]** On average, controlling other variables, an increase in 1 standard deviation in Family Poverty Ratio is associated with 30.7% ($e^{-0.3674}$-1 = - 0.3075) decrease in the odds of having problem paying for medical and healthcare expenses. **4]** On average, controlling other variables, an increase of 1 level in the health condition of child is associated with 7.73% ($e^{-0.0804}$-1 = - 0.0773) decrease in the odds of having problem paying for medical and healthcare expenses. **5]** On average, controlling other variables, an increase of 1 level in the cost of medical and healthcare expenses is associated with 81.5% ($e^{0.5960}$-1 = 0.8148) increase in the odds of having problem paying for medical and healthcare expenses. **6]** On average, controlling other variables, going up from 1 level of affordability for basic needs to the next is associated with an increase of 107.7% ($e^{0.7219}$-1 = 1.073) in the odds of affordability for medical and healthcare expenses. **7]** On average, controlling other variables, an increase of 1 level in the physical health condition of adult is associated with 5.97% ($e^{-0.0616}$-1 = -0.0597) decrease in the odds of having problem paying for medical and healthcare expenses. **8]** On average, controlling other variables, an increase of 1 level in the mental health condition of adult is associated with 5.45% ($e^{-0.0560}$-1 = -0.0545) decrease in

the odds of having problem paying for medical and healthcare expenses. **9]** On average, controlling other variables, an increase of 1 level in the frequency of health insurance benefits covering health service needs of child is associated with 11.6% ($e^{-0.1183}$-1 = -0.1157) decrease in the odds of having problem paying for medical and healthcare expenses. **10]** On average, controlling other variables, an increase of 1 level in the frequency of health insurance allowing the child to see health care provider needed is associated with 3.86% ($e^{-0.0394}$-1 = -0.0386) decreases in the odds of having problem paying for medical and healthcare expenses. **11]** On average, controlling other variables, an increase of 1 reasonable level of the medical costs is associated with 56.2% ($e^{-0.8253}$-1 = -0.5619) decrease in the odds of having problem paying for medical and healthcare expenses. **12]** On average, controlling other variables, an increase of 1 level in the average time that the caregiver spends for health care at home for child is associated with 0.26% ($e^{0.0026}$-1 = 0.0026) increase in the odds of having problem paying for medical and healthcare expenses. **13]** On average, controlling other variables, an increase of 1 unit of children with special needs is associated with 8.16% ($e^{0.0784}$-1 = 0.0816) increase in the odds of having problem paying for medical and healthcare expenses.

At a 95% confidence level, the model indicated that Cutting work Hours because of Health Conditions, Avoided Changing Jobs to Maintain Health Insurance, Family Poverty Ratio, Cost of Medical and Health Care Bills, Basic Needs Affordability, Adult's Physical Health Condition, Adult's Mental Health Condition, Reasonable Medical Costs, Health Insurance Benefits Coverage of Medical Service Needs, Number of Children with Special Needs, Health Insurance Allowing to See Providers Needed, and Health Condition of Children are significant predictors in predicting a family's affordability to pay medical bills. Especially Avoided Changing Jobs to Maintain Health Insurance, Cost of Medical and Health Care Bills, Basic Needs Affordability and Reasonable Medical Costs are significant predictors with high absolute value of coefficients in determining the affordability of medical and healthcare expenses.

For factors with negative coefficients, the greater their absolute value, the lower the probability of having a problem paying medical bills. For example, a high family poverty ratio, good health condition of child, the more the times that health insurance covers the required medical services, and reasonable medical costs may all reduce the probability of having problem paying for medical bills. Conversely, for factors with positive coefficients, the higher the value, the greater the probability of having a problem paying medical bills. For example, cutting down work hours because of health condition, avoiding changing jobs to maintain medical insurance, having more children with special needs, incurring higher medical and healthcare expenses, and experiencing more instances when the family cannot afford basic needs may all increase the probability of having problem paying medical bills.

4. Discussion

4.1 Logistic Regression Analysis of medical and healthcare expenses affordability

To test whether each independent variable is a significant predictor of the dependent variable, we ran the Wald test on the logistic regression

model. The Wald test works by testing the null hypothesis that the coefficient of the independent variable is zero. Rejection of the null hypothesis means that the target independent variable is a significant predictor. At a 95% confidence level, a high p-value (> 0.05) indicates a failure to reject the null hypothesis, while a low p-value (< 0.05) leads to the rejection of the null hypothesis.

Table 4. Wald Test Summary of the Logistic Regression Model

Name	Coefficient	Standard Error	Z	P>|Z|
CONSTANT	-1.2241	0.016	-75.656	0.000
CUTHOURS	0.5410	0.049	11.142	0.000
AVOIDCHG	0.8617	0.036	23.639	0.000
FPL_I1	-0.3674	0.012	-29.402	0.000
K2Q01	-0.0804	0.012	-6.607	0.000
HOWMUCH	0.5960	0.014	43.860	0.000
ACE1	0.7219	0.012	57.808	0.000
A1_PHYSHEALTH	-0.0616	0.015	-4.137	0.000
A1_MENTHEALTH	-0.0560	0.015	-3.830	0.000
K3Q20	-0.1183	0.014	-8.220	0.000
K3Q22	-0.0394	0.014	-2.895	0.004
K3Q21B	-0.8253	0.014	-57.441	0.000
ATHOMEHC	0.0026	0.012	0.223	0.824
TOTCSHCN	0.0784	0.012	6.509	0.000

According to Table 4 above, Time Spending for Health Care at Home (ATHOMEHC: Coefficient= -0.0025, P=0.824) is a factor with a p-value greater than 0.05, indicating that this factor does not statistically significantly influence the dependent variable. As such, we cannot conclude a relationship between this factor and the affordability of medical and healthcare expenses. Therefore, this factor should be excluded from our predictive model. Some potential reasons for this insignificance are limited variation and indirect effect. There might not be enough variation in the amount of time spent on home healthcare among the population studied. As most families spend roughly the same amount of time on home healthcare, this factor tends not to have a significant effect on the dependent variable that we are predicting. Another potential reason is that "time spending for health care at home" may indirectly affect medical and health care affordability by affecting factors like employment or the overall health condition of the family.

4.2 Evaluating the performance of the Logistic Regression Model

The performance of the logistic regression model in predicting the affordability of medical and healthcare expenses was evaluated by the Confusion Matrix (based on the test dataset), as shown in Figure 4, and the ROC curve and AUROC scores of the training and test datasets, as shown in Figure 5.

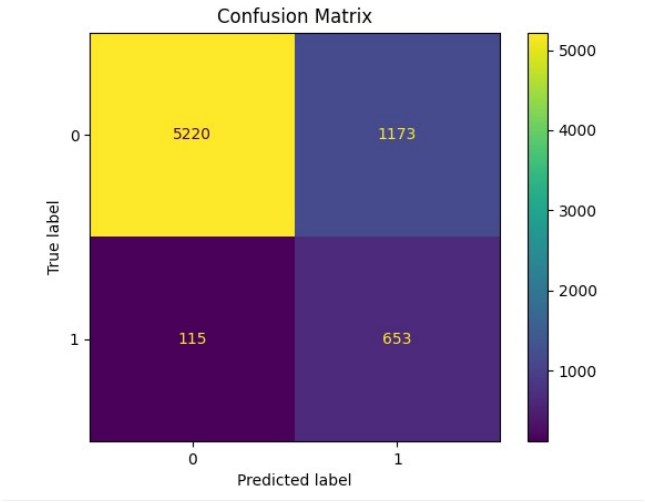

Figure 4. Confusion Matrix of the Logistic Regression Model

As we mentioned earlier in 3.2, 7,161 (20% of 35,806) observations were used for model testing. Based on the Confusion Matrix, the specificity of the predictive model on the test set is 0.82 (5220/ (5220+1173)), and the sensitivity of the predictive model on the test set is 0.85 (653/ (653+115)). The high sensitivity indicates that the model is good at detecting positive cases and thus can effectively identify families that need support most of the time. The high specificity indicates the model is good at correctly identifying negative cases and thus is less likely to falsely classify those families being able to afford medical and healthcare expenses as unaffordable.

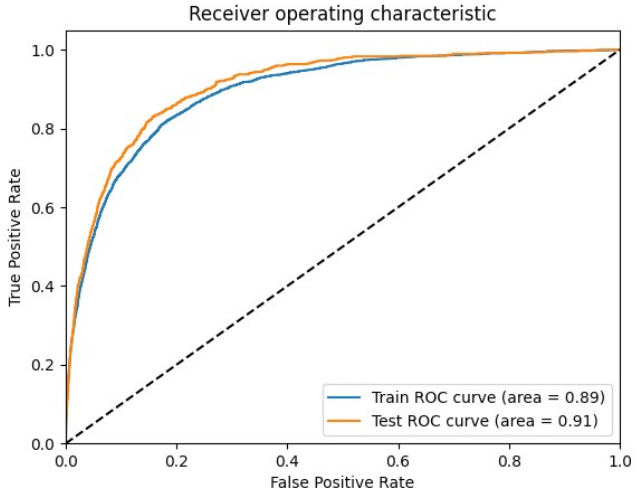

Figure 5. ROC Curves and AUROC scores of the Logistic Regression Model

The AUROC scores for the performance of the predictive model were 0.89 on the training set and 0.91 on the test set. The Receiver Operator Characteristic (ROC) curves for both the training and test sets show a bow up to the upper left of the plot, which indicateed a stable and high prediction accuracy, with a small prediction deviation compared to actual results.

5. Conclusion

In this study, we built a machine learning model to predict whether a family could afford medical and healthcare expenses by calculating the effects of different factors on the affordability of medical and healthcare expenses. The dataset of 2021 National Survey of Children's Health was used to identify potential predictors for the affordability of medical and healthcare expenses and to develop and validate the predictive model. The predictive model based on logistic regression was validated by various performance metrics, conclusively affirming its robust proficiency in predicting the affordability of medical and healthcare expenses. Overall, it achieved an AUROC score of 0.89 on the training set and 0.91 on the test set.

Other models, such as k-nearest neighbors (k-NN) and random forest, were not explored because they did not offer advantages over logistic regression for our research objectives. For example, k-NN, being a non-parametric model, it was typically slower than logistical regression. This is because k-NN requires storage of the entire dataset and must repeatedly compute distance to make predictions. Moreover, logistic regression has the added advantage of greater interpretability, as the coefficients of predictors can be used to indicate the significance and impact of the variables on the outcome (25). On the other hand, random forest is an ensemble method that relies on multiple decision trees to make predictions. It can be slower to train and is not as interpretable as logistic regression. Furthermore, with smaller datasets like ours, the complexity of random forest makes it more susceptible to overfitting (26). Therefore, given these considerations, we opted for the logistic regression model in this study.

The result of this study could have implications in different aspects. One potential application of this predictive model is its use in public or private insurance companies. By leveraging the results of this predictive model, private or public insurance companies may be able to correctly predict which families can afford medical and healthcare expenses. This can provide insights to insurance companies on how to determine appropriate insurance premiums. Moreover, by understanding the relationship between how often health insurance provides benefits to cover needed medical services, how often insurance allows to see health care providers needed, and the affordability of medical and healthcare expenses, health insurers can offer more comprehensive policies and expand their network of in-network healthcare providers, which can contribute to make healthcare more affordable and make their insurance policy more appealing to potential and current customers compared to other insurers.

Another possible application of this study is for governmental or non-governmental welfare institutions. The results of this predictive model can help these welfare institutions in identifying target families for support according to their respective economic

conditions and specific circumstances, and determining the extent of financial assistance that can be provided to different families. For instance, based on this predictive model, welfare institutions can anticipate which families may require financial assistance based on factors such as affordability of basic needs, cutting work hours due to health conditions, the number of children with special needs, and the costs of medical and healthcare expenses. They can further establish appropriate levels of support for different families. Although currently, some information such as cutting down working hours or avoiding changing jobs to pay for the healthcare costs may not be accessible to the insurance provider, which could reduce the utility of our predictive model, insurance companies can consider including this information in their requests to the insured parties.

For future studies, factors such as state median household income and family socioeconomic status may need to be incorporated into independent variable selection to improve model development and performance testing. Additionally, as primary survey data, there may be common method bias given the nature of the data. Future studies can include common latent variables or social desirability to control for common method bias. Finally, this study used a cross-sectional dataset to test the causality between the dependent variable and independent variables. Therefore, longitudinal data could also be collected and utilized.

6. References

1. Creamer, John, et al. "Poverty in the United States: 2021." *Census.gov*, 13 Sept. 2022, https://www.census.gov/library/publications/2022/demo/p60-277.html.

2. Vaidya, Anuja. "Survey: 1 in 5 Americans can't afford necessary care." *MedCity News*, 1 Apr. 2021, https://medcitynews.com/2021/04/survey-1-in-5-americans-cant-afford-necessary-care/.

3. Kreider, Amanda R, et al. "Quality of Health Insurance Coverage and Access to Care for Children in Low-Income Families." *JAMA Pediatrics,* U.S. National Library of Medicine, Jan. 2016, 170(1):43–51. doi:10.1001/jamapediatrics.2015.3028.

4. Emanuel, Ezekiel J., et al. "Measuring the Burden of Health Care Costs on US Families." *JAMA Network*, 21 Nov. 2017, https://jamanetwork.com/journals/jama/article-abstract/2661699.

5. Pollitz, Karen, et al. "Medical Debt Among People With Health Insurance." *KFF*, 07 Jan. 2014, www.kff.org/private-insurance/report/medical-debt-among-people-with-health-insurance/.

6. Hamel, Liz, et al. "The Burden of Medical Debt: Results from the Kaiser Family Foundation/New York Times Medical Bills Survey." *KFF*, 05 Jan. 2016, https://www.kff.org/health-costs/report/the-burden-of-medical-debt-results-from-the-kaiser-family-foundationnew-york-times-medical-bills-survey/view/print/.

7. Claxton, Gary, et al. “How Affordability of Employer Coverage Varies by Family Income.” *Peterson-KFF Health System Tracker,* 10 Mar. 2022, www.healthsystemtracker.org/brief/how-affordability-of-health-care-varies-by-income-among-people-with-employer-coverage/#Average%20share%20of%20family%20income%20going%20toward%20employer-based%20health%20insurance%20premium%20contributions%20and%20out-of-pocket%20costs,%202020.

8. Schwartz, Hope, et al. “Emergency Department Visits Exceed Affordability Threshold for Many Consumers with Private Insurance.” *Peterson-KFF Health System Tracker*, 16 Dec. 2022, www.healthsystemtracker.org/brief/emergency-department-visits-exceed-affordability-thresholds-for-many-consumers-with-private-Insurance/#Average%20cost%20of%20emergency%20department%20visits,%20by%20MSA,%202019.

9. Nyman, John A, and Helen M Trenz. “Affordability of the Health Expenditures of Insured Americans before the Affordable Care Act.” American Journal of Public Health, *U.S. National Library of Medicine*, Feb. 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4815608/.

10. Cree, Robyn A., et al. “Health Care, Family, and Community Factors Associated with Mental, Behavioral, and Developmental Disorders and Poverty among Children Aged 2–8 Years - United States, 2016.” *Centers for Disease Control and Prevention*, 20 Dec. 2018, www.cdc.gov/mmwr/volumes/67/wr/mm6750a1.htm.

11. Karabiber, Fatih. “Binary Variable.” *Learn Data Science - Tutorials, Books, Courses, and More*, https://www.learndatasci.com/glossary/binary-variable/.

12. *Dummy Variables in Regression*, https://stattrek.com/multiple-regression/dummy-variables.

13. “Correlation in Statistics: Correlation Analysis Explained.” *Statistics How To*, 3 Mar. 2023, www.statisticshowto.com/probability-and-statistics/correlation-analysis/.

14. Health-Care Utilization as a Proxy in Disability Determination, www.ncbi.nlm.nih.gov/books/NBK500097/.

15. Zach. “How to Test for Multicollinearity in SPSS.” *Statology*, 5 June 2020, https://www.statology.org/multicollinearity-spss/.

16. Luo, Shuyu. “Optimization: Loss Function under the Hood (Part II).” *Medium*, 7 June 2019, towardsdatascience.com/optimization-loss-function-under-the-hood-part-ii-d20a239cde11.

17. Pykes, Kurtis. “Oversampling and Undersampling.” *Medium*, Towards Data Science, 10 Sept. 2020, https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf.

18. Paul, Sayak. “Towards Preventing Overfitting.” *DataCamp*, 29 Aug. 2018, https://www.datacamp.com/tutorial/towards-preventing-overfitting-regularization.

19. Godwin, James Andrew. “Ridge, Lasso, and ElasticNet Regression.” *Medium*, 16 Apr. 2021, towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3.

20. “Logistic Regression (with Elastic Net Regularization).” *SAP Help Portal,* https://help.sap.com/docs/SAP_HANA_PLATFORM/2cfbc5cf2bc14f028cfbe2a2bba60a50/46effe560ea74ecbbf43bab011f2acf8.html.

21. Brownlee, Jason. “How to Develop Elastic Net Regression Models in Python.” *MachineLearningMastery.Com*, 11 June 2020, https://machinelearningmastery.com/elastic-net-regression-in-python/#:~:text=The%20benefit%20is%20that%20elastic,penalties%20to%20the%20loss%20function.

22. Sharma, Abhishek. “Cross Validation in Machine Learning.” *GeeksforGeeks*, 15 Feb. 2023, www.geeksforgeeks.org/cross-validation-machine-learning/

23. Indeed Editorial Team. “What Is a Confusion Matrix? (and How to Calculate One).” *Indeed Career Guide*, 18 Aug. 2021, https://www.indeed.com/career-advice/career-development/confusion-matrix.

24. “Classification: Roc Curve and AUC.” *Google*, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

25. Varghese, Danny. “Comparative Study on Classic Machine Learning Algorithms.” *Medium,* Towards Data Science, 10 May 2019, https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=Logistic%20Regression%20vs%20KNN%20%3A,can%20only%20output%20the%20labels.

26. “Logistic Regression vs Random Forest Classifier.” *GeeksforGeeks*, 8 June 2023, www.geeksforgeeks.org/logistic-regression-vs-random-forest-classifier/.