Smoking and Diabetes

Causal Inference and Machine Learning Analysis

Eliyahu B. Addess

Johns Hopkins University

625.726 Theory of Statistics II

Dr. Tom Woolf and Dr. Burhan Sadiq

May 2023

Smoking and Diabetes

Causal Inference and Machine Learning Analysis

Eliyahu B. Addess

Abstract

Diabetes is a condition in which the body cannot adequately control blood sugar. Too much blood sugar can eventually lead to severe health problems such as heart disease, kidney disease, and vision loss (CDC, 2023). As such, it is of vital interest to understand the risk factors and causes of diabetes. It is known that obesity and lack of physical activity are risk factors for diabetes (CDC, 2022). In this project, we explore a potential connection between smoking and diabetes.

Using data, we demonstrate that there does exist an association between smoking and diabetes. This does not answer the more fundamental question of whether or not there exists a causal link from smoking to diabetes. To address the question of causality, we first develop the basic statistical theory of causal inference. We then use machine learning models to estimate the ATE (average treatment effect) of smoking on diabetes.

Linear regression is a classical model for estimating the ATE. For binary outcomes, logistic regression is the model of choice. The relatively new method of Bayesian Additive Regression Trees (BART) has been suggested as an effective model for causal inference (Hill, 2011). In this project, we use linear regression, logistic regression, and BART models for estimating the ATE of smoking on diabetes. Finally, the results from each model are contrasted. Please note that the analysis in this project was conducted at the level of the course and is not intended to be clinical research.

Contents

1 Introduction

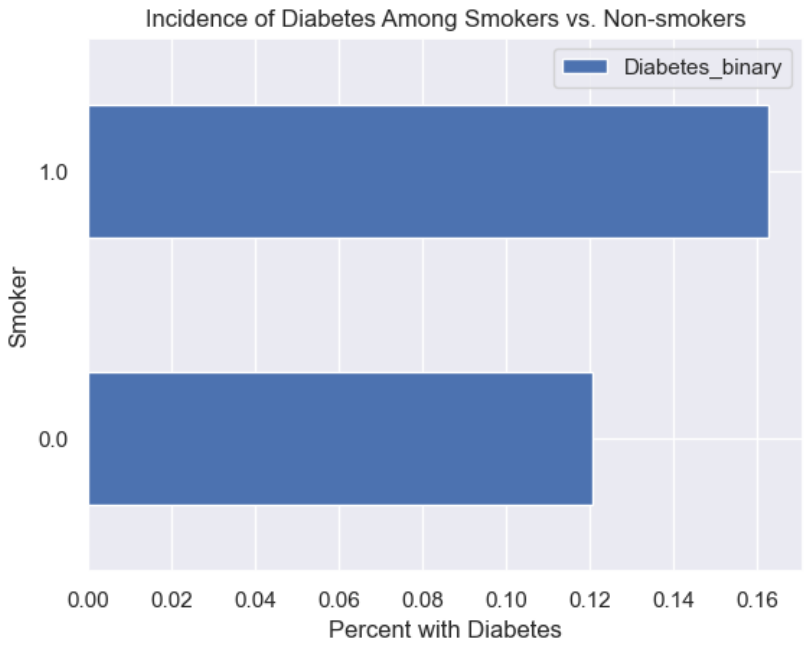
This project began with my assigned paper Hill, 2011. The paper discusses the theory of causal analysis and the challenge of accurately estimating causal effects. Hill proposed the use of Bayesian Additive Regression Trees (BART) as a practically effective method for estimating said effects. To demonstrate this, Hill performed a sophisticated Monte Carlo simulation, in which BART performed remarkably well.

Following Hill, 2011, the project goal is twofold. One, to explore the theory of causal inference and Bayesian Additive Regression Trees. Second, to apply these methods to real data. I am interested in biomedical data science applications and chose to work with a data set in that arena.

In this project, we analyze the Diabetes Health Indicators Dataset, available here on Kaggle.com. This data set is a cleaned and curated version of the Behavioral Risk Factor Surveillance System (BRFSS) 2015 data set, available from the CDC here on Kaggle. The Diabetes Health Indicators Dataset data set was prepared by Alex Teboul. Teboul provided several data sets for analysis. We are using the diabetes _ binary _ health _ indicators _ BRFSS2015 data. This data set has a binary target variable, with 0 indicating no diabetes and 1 indicating diabetes or pre-diabetes. There are 21 covariates pertaining to different aspects of the subjects medical history and nutrition. These include high blood pressure, physical activity, alcohol consumption, smoking, and vegetable intake. All of the variables are categorical, with most binary. The data set contains 253,680 observations.

2 Exploratory Analysis

We are specifically interested in using our data to investigate connections between smoking and diabetes. We first note that there seems to be a higher incidence of diabetes among smokers than among non-smokers. This is demonstrated in the following figure.



We see that about 16% of smokers have diabetes, while only 12% of non-smokers have diabetes. This observed imbalance gives rise to our primary question:

Does smoking have a causal effect on diabetes?

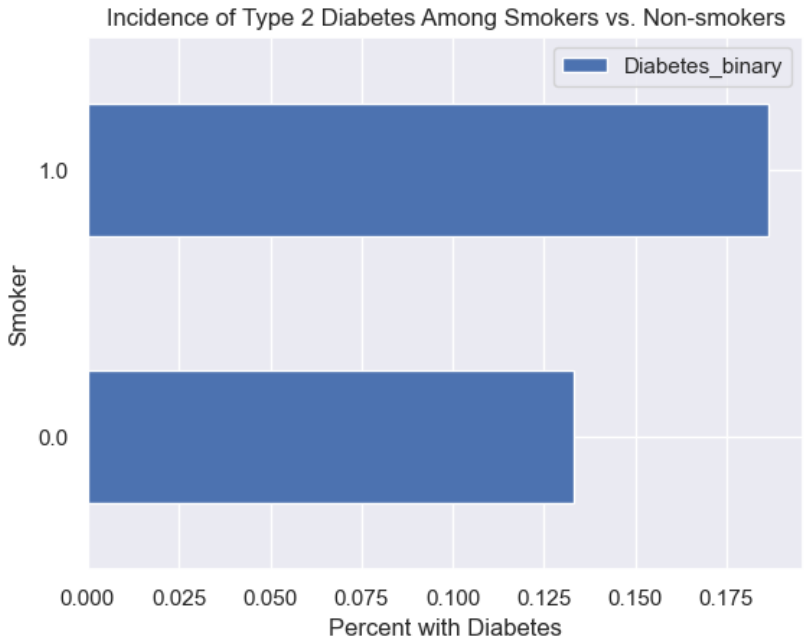
Smoking certainly does not necessarily cause diabetes. This is evident from the fact that only 16% of smokers have diabetes and not all smokers. However, we would like to know if smoking increases the risk of diabetes.

Our analysis can be further refined. There are several types of diabetes, namely type 1, type 2, and gestational diabetes. Type 1 diabetes generally develops early on in life and does not seem to be linked to lifestyle related health factors. Only about 5-10% of people diagnosed with diabetes have type 1. Gestational diabetes is pertinent specifically to pregnant women. Type 2 diabetes, however, is linked to various health and nutritional factors

(CDC, 2023, CDC, 2022). When researching risk factors and causation, we are primarily interested in analyzing type 2 diabetes [1].

Limiting our attention to type 2 diabetes presents a practical challenge because our data set does not differentiate between types of diabetes. When analyzing similar BRFSS data, Xie, 2019 built several models assuming that diabetic subjects who were over 30 years old and not pregnant had type 2 diabetes. We follow a similar convention. Our data set does not have a pregnancy variable, so to designate type 2 diabetes, we subset the data to include only men over 30. This new data set has 105,144 observations.

The following figure summarizes the observed incidence of type 2 diabetes among smokers and non-smokers.



We see that about 19% of smokers have type 2 diabetes, while only 13% of non-smokers have type 2 diabetes. Now that we have focused specifically on type 2 diabetes, we observe a slightly larger difference than before. This reinforces our original question:

Does smoking have a causal effect on diabetes?

[1] There is also a condition known as prediabetes (see CDC, 2023). Our data set does not differentiate between prediabetes and type 2 diabetes. Hence, for the purposes of this project, we will not differentiate between prediabetes and type 2 diabetes.

To attempt to answer this question, we need to first develop some basic causal inference theory, which we proceed to do in the upcoming section.

2.1 Bibliographical notes

The EDA was conducted in Python (Van Rossum & Drake, 2009) using the NumPy (Harris et al., 2020), pandas (pandas development team, 2022), Matplotlib (Hunter, 2007), and seaborn (Waskom, 2021) packages.

3 Causal Inference Theory

Consider a binary treatment T and a binary outcome Y. Define $Y(t)$ to be the *potential outcome* of doing treatment t. In other words, $Y(0)$ is the outcome that would occur if $T = 0$, and $Y(1)$ is the outcome that would occur if $T = 1$. For a particular unit i in the population, define the *individual treatment effect* (ITE) as

$$\tau_i \equiv Y_i(1) - Y_i(0).$$

The ITE is an intuitive measure of causality. Consider the situation where the outcome Y will equal 1 regardless of treatment. Then $Y_i(1) = Y_i(0) = 1$, which implies that $\tau_i = 0$. This means that the treatment has no causal effect on the outcome Y. Conversely, if Y will equal 1 if and only if $T = 1$, then $Y_i(1) = 1$ and $Y_i(0) = 0$, implying that $\tau_i = 1$. This means that T does have a causal effect on Y.

The ITE is a purely theoretical quantity because for any given subject, we cannot observe both $Y_i(1)$ and $Y_i(0)$. This inability to directly collect data on individual treatment effects is "the fundamental problem of causal inference" (Neal, 2020). To begin to address this problem, we turn our attention to the *average treatment effect* (ATE). The ATE is defined as

$$\tau \equiv \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

Estimating the ATE is also non-trivial. We must find conditions under which the ATE can be identified as a statistical quantity that can be estimated from data.

One such quantity is the association, defined as:

$$\alpha = \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$$

In general, $\tau \neq \alpha$, that is "Association is Not Causation" (Wasserman, 2004). This lack of equality is because the potential outcome, $Y(t)$, is not independent of treatment. Intuitively, $\mathbb{E}[Y(1)]$ is the average outcome if the entire population were to receive the treatment, while $\mathbb{E}[Y|T=1]$ is the average outcome among those who actually did the treatment. Data on the treatment group average does not amount to evidence about the population average. This is because there may be some factor that is causing the treatment group to both do the treatment and result in the particular outcome that took place. Perhaps if the non-treatment group would have done the treatment, they would have had an altogether different average outcome. This is known as *confounding*.

The assumption of no confounding is known as *ignorability*. Formally, ignorability means that $Y(t) \perp\!\!\!\perp T$. We can now show that this assumption implies that the ATE will equal the association.

$$\begin{aligned}
\tau &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&= \mathbb{E}[Y(1)|T=1] - \mathbb{E}[Y(0)|T=0] && \text{(by ignorability)} \\
&= \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] && \text{(def. of } Y(t)\text{)} \\
&= \alpha
\end{aligned}$$

In general, ignorability only holds in randomized trials where subjects are randomly assigned treatment. In observational studies however, confounding is a very real possibility and we certainly cannot assume ignorability. Our diabetes data set is an observational study. Hence we must search for further tools to ascertain causality in our data.

Another possible assumption is *conditional ignorability*. This assumption means that within particular values of covariates X, the potential outcomes are independent of treatment. Formally, conditional ignorability means that $Y(t) \perp\!\!\!\perp T|X$. This assumption also allows us to identify the ATE in terms of statistical quantities. Consider the *conditional average treatment affect* (CATE),

$$\tau_c \equiv \mathbb{E}[Y(1) - Y(0)|X].$$

Using conditional ignorability, we can write:

$$\begin{aligned}
\tau_c &= \mathbb{E}[Y(1) - Y(0)|X] \\
&= \mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X] \\
&= \mathbb{E}[Y(1)|T=1, X] - \mathbb{E}[Y(0)|T=0, X] \\
&= \mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X]
\end{aligned}$$

This last quantity can be estimated from data. Moreover, we can recover the unconditional ATE from the CATE using the law of iterated expectation (See Wasserman, 2004 Theorem 3.24).

$$\begin{aligned}\mathbb{E}[\tau_c] &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X]] \\ &= \mathbb{E}[Y(1) - Y(0)] \\ &= \tau\end{aligned}$$

Thus, under the assumption of condtional ignorability, we can write:

$$\begin{aligned}\tau &= \mathbb{E}[\tau_c] \\ &= \mathbb{E}[\mathbb{E}[Y|T = 1, X] - \mathbb{E}[Y|T = 0, X]]\end{aligned}$$

Neal, 2020 calls this result the *"adjustment formula"*.

The adjustment formula gives us a practical way to estimate the ATE from data. Given data on treatment T, outcome Y, and covariates X, assume that $Y(t) \perp\!\!\!\perp T|X$. We can then use a machine learning model to estimate $\mathbb{E}[Y|T, X]$. As we learned in the course, estimating $\mathbb{E}[Y|X]$ is the same as estimating the regression function.[2] We then set T to 1 and for each data point, use our model to predict values for $\mathbb{E}[Y|T = 1$, X]. We do the same for $T = 0$. Subtracting the predicted values of $\mathbb{E}[Y|T = 0, X]$ from the predicted values of $\mathbb{E}[Y|T = 1, X]$ gives us an estimate of the CATE for each level of the covariates X. We can then simply take the average to obtain an estimate of the ATE.

Conditional ignorability is the main assumption for understanding how to transition from causal quantities to estimable statistical quantities. There are several other assumptions that are necessary for the above equations to be rigorous. These assumptions will be discussed in the later sections.

It is important to note that while conditional ignorability is more plausible for observational data, there is no way of knowing if our data contains sufficient covariates to guarantee that this assumption holds.

As mentioned previously, we use machine learning models to estimate the ATE. As always, there is a question of which model to use. In the assigned paper, Hill, 2011 demonstrated that Bayesian Additive Regression Trees (BART) are an excellent model for causal inference. In the upcoming section, we give a brief introduction to BART.

[2] see Wasserman, 2004 chapter 13 and section 20.4

3.1 Bibliographical notes

The theory and notation in this section were primarily based on Neal, 2020 and Wasserman, 2004. I used some terminology from Hill, 2011.

4 BART Theory

As their name suggests, Bayesian Additive Regression Trees are based on tree models.[3]. Single tree models are not as accurate as other machine learning models. To address this problem, several improved tree-based methods have been developed, which use an ensemble of trees to construct a far more accurate model. These methods include bagging, boosting, and random forests.[4] The BART model was motivated by these methods, particularly by boosting (H. A. Chipman et al., 2010). We introduce BART in the context of regression, although it can be used for binary outcome classification as well.[5]

When fitting a BART model, we first select the number of trees to grow and the number of iterations the algorithm should run. Let K be the number of trees and B be the number of iterations. Let $\hat{f}_k^b(x)$ be the predicted value at tree k and iteration b. These predicted values are summed at the conclusion of each iteration to give a total iteration predicted value of $\hat{f}^b(x)$.

In the first iteration, all the trees are set to a single node with value $\hat{f}_1^b(x) = \frac{1}{nK}\sum_{i=1}^n y_i$. This gives a total predicted value of

$$\hat{f}^1(x) = \sum_{k=1}^{K}\left(\frac{1}{nK}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

which is simply the sample mean. In the following iterations, the algorithm updates each tree by first computing a "*partial residual*" (James et al., 2021). For the kth tree, each partial residual is obtained by taking the predicted values from the remaining $K-1$ trees and subtracting them from the the observed response y_i. That is, for observation i,

$$r_i = y_i - \sum_{k'<k}\hat{f}_{k'}^b(x_i) - \sum_{k'>k}\hat{f}_{k'}^{b-1}(x_i).$$

[3]For a basic introduction to tree models see Wasserman, 2004 and James et al., 2021.
[4]See Hastie et al., 2009.
[5]The extension to binary outcomes is developed in H. A. Chipman et al., 2010.

Instead of fitting a new tree to this partial residual, the kth tree from the previous iteration undergoes a random perturbation. These perturbations may include altering the size of the tree and/or altering the predictions $\hat{f}_k^b(x)$. The perturbation process is formally based on drawing from a Bayesian posterior distribution, which is computed using Markov Chain Monte Carlo (MCMC). The details of the regularization prior, and the MCMC algorithm are discussed in H. A. Chipman et al., 2010. Hill, 2011 notes that treating the model parameters as formal statistical parameters is in contrast to "much of the data-mining literature", where parameters are primarily dealt with algorithmically.

Due to the iterative nature of BART, the output is a set of models,

$$\hat{f}^b(x) = \sum_{k=1}^{K} \hat{f}_k^b(x), \ b = 1, 2, \ldots, B.$$

Since the models from the earlier iterations are typically not very accurate, we do not use them. These first L iterations are called "the burn-in period"(James et al., 2021). The results from the remaining iterations can then be averaged to give a one-number prediction. The percentiles of the output models can also be used to establish confidence bands for the predicted values.

Having introduced causal inference and BART, we can return to our analysis of smoking and diabetes in the upcoming section.

4.1 Bibliographical Notes

BART is due to H. Chipman et al., 2006 and H. A. Chipman et al., 2010. The theory and notation in this section were primarily based on James et al., 2021.

5 Methods

We left off our data analysis with the observation that while 19% of smokers had Type 2 Diabetes, only 13% of non-smokers had Type 2 Diabetes. This led us to the question: **Does smoking have a causal effect on diabetes?** With causal inference theory in hand, we can conduct a formal analysis.

We begin by computing the association α. Wasserman, 2004 writes that α can be estimated by $\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$, where $\bar{Y}_1$ is the average response among

the subjects who did the treatment, and $\bar{Y_0}$ symmetrically defined for the control group. In our case,

$$\alpha = 0.19 - 0.13 = 0.05.$$

This 5% association seems small, but not entirely negligible. As discussed before, the association is not equivalent to a causal quantity in an observational study such as ours. To establish clear methods for causality, we summarize the steps in an algorithm.

Causal Inference Algorithm

1. Make causal inference assumptions (including conditional ignorability).
2. Fit a machine learning model from response data Y to treatment and covariate data (T, X).
3. Set T to 1 and generate predicted $\mathbb{E}[Y|T = 1, \mathrm{X}]$. Denote this vector by pr_1.
4. Set T to 0 and generate predicted $\mathbb{E}[Y|T = 0, \mathrm{X}]$. Denote this vector by pr_0.
5. Subtract pr_0 from pr_1.
6. Estimate the ATE by taking the average of the differences.

The first step is to check assumptions. Our data set has 21 covariates. Given the number of covariates, conditional ignorability may be a valid assumption, however we cannot know this with certainty. It also turns out that there are serious concerns that our data does not satisfy the remaining causal inference assumptions. These issues will be discussed in the conclusions section. Had this been clinical research, we would conclude our analysis here. If the assumptions are not met, then any inference is dubious. However, as Dr. Woolf has stated many times, the primary purpose of this project is to generate learning. I believe that there is much to be gained by continuing the analysis and implementing the algorithm developed above. Hence we will proceed as if the necessary assumptions were satisfied.

5.1 Linear and Logistic Regression Models

I implemented the Causal Inference Algorithm with three distinct machine learning models, namely linear regression, logistic regression, and BART. Logistic regression is the classic model for binary data, but linear regression may seem like a strange choice for modelling our binary response variable. While it is true that the linear model $Y = r(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ certainly does not apply to binary data (Wasserman, 2004), I believe that it is still a useful model for our purposes, as we will explain.

Binary data follow a Bernoulli distribution where $\mathbb{E}[Y]$ is equal to the probability of success. Hence the linear regression line $\mathbb{E}[Y] = r(x)$ can be interpreted as the probability of $Y = 1$. One issue with modelling probabilities with $r(x)$ is that for many x values, $r(x)$ will be less than 0 or greater than 1. This would be a major problem if we were trying to predict probability values outside the support of the data. However, within the range of the data, $r(x)$ should intuitively be bounded between 0 and 1. Fortunately, when determining causal effects we are not concerned with out-of-sample predictions. Although, we do need to generate predictions for $\mathbb{E}[Y|T = 1]$ and $\mathbb{E}[Y|T = 0, X]$, these are really in-sample prediction due to the *positivity* assumption. In short, the positivity or overlap assumption asserts that estimating the ATE is not mathematically viable unless $0 < P(T = 1|X = x) < 1$. In other words, for each level of the covariates X there must be non-zero probability of both treatment and control. Thus for or around every level of X, we should have some data for both $T = 1$ and $T = 0$. See Gomila, 2021 for further justification of using linear regression for causal inference regarding a binary response.

The linear and logistic legression ATE estimations were implemented in Python (v3.9.13; Van Rossum and Drake, 2009). The ATE was estimated for both the original full diabetes data set and the modified type 2 diabetes data set. Results are presented below.

5.2 BART Model

I found BART difficult to implement in Python, and instead implemented the algorithm in R (v4.2.2; R Core Team, 2022) using the BART package (Sparapani et al., 2021). BART seems to be a computationally expensive algorithm, and running it for all 253,680 observations was impractical. Instead, a random subset of 5000 observations was used for this analysis. The

BART algorithm was run with 50 trees per iteration. I used the pbart (Probit BART) option, which is designed for binary outcomes. Here too, the analysis was conducted both on 5000 observations from the original diabetes data set and on 5000 observations from the modified type 2 diabetes data set. Results are presented below.

5.3 Bibliographical notes

The linear regression discussion was inspired by Gomila, 2021 and Wasserman, 2004. However, most of the theoretical musings are my own. The definition and details of the positivity assumption are from Neal, 2020.

In my Python implementation, I used the NumPy (Harris et al., 2020), pandas (pandas development team, 2022), and scikit-learn (Pedregosa et al., 2011) packages. For the R implementation, I used the tidyverse (Wickham et al., 2019) and BART (Sparapani et al., 2021) packages.

6 Results

The estimated ATE (rounded to four decimal places) for both data sets is displayed in the following table.

ATE Estimates	Linear Regression	Logistic Regression	BART
Diabetes Data	-0.0059	0.0000	0.0148
Type 2 Diabetes Data	-0.003	0.0000	0.0148

As we can see, the results for both data sets were similar. Linear and logistic Regression provided ATE estimates of approximately 0. This would indicate that smoking has no causal effect on diabetes. The BART estimate was slightly higher, about 0.01. This is still significantly smaller than the estimated association which was 0.05. For a binary outcome, the ATE can range between -1 and 1. An ATE estimate of 0.01 indicates that there may be a minute causal effect of smoking on diabetes. This effect may be negligible.

7 Conclusions

Based on our analysis, we would conclude that smoking does not increase the risk of type 2 diabetes. However, the National Center for Chronic Disease

Prevention and Health Promotion (US) Office on Smoking and Health, 2014 provides strong evidence that indeed smoking does increase the risk of type 2 diabetes. In light of this, we review some limitations of our analysis.

Neal, 2020 states that there are four assumptions necessary for causal inference. They are "Unconfoundedness", "Positivity", "No Interference", and "Consistency". Unconfoundedness was discussed above as conditional ignorability. Positivity was also introduced in the section on linear regression modelling. The no interference assumption asserts that the outcome of subject A is not influenced by the treatment of subject B. Without going into the mathematical details, the consistency assumption implies that there is only one version of the treatment T.

The final assumption presents a major issue with our analysis. We used the binary Smoker variable as the treatment T. However, in the CDC BRFSS 2015 codebook, this variable is defined as "Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]". This means that anyone who smoked at least 5 packs of cigarettes was classified as a smoker in our data set. Within this group there are obviously many levels. These may range in quantity of cigarettes smoked and also in the time period over which they were smoked. The consistency assumption is certainly not satisfied.

It is possible that the no interference assumption is violated as well. For example, consider the case where subject A did not smoke but his friend, subject B, did smoke. In this case, subject A did not do the treatment, but his outcome may be influenced by subject B's treatment due to their proximity and the effects of second hand smoke. For information on verifying the positivity assumption, see Petersen et al., 2012.

In conclusion, our analysis does not provide clinical evidence regarding the effects of smoking on type 2 diabetes. However, it has been a eye-opening academic experience in the areas of causal inference and machine learning.

8 Appendices

8.1 Data

The data used in this project can be accessed here: https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook

8.2 Code

The project code can be accessed here: https://github.com/eb-addess/causal-project

9 Bibliography

References

CDC. (2022). Diabetes Risk Factors. Retrieved April 18, 2023, from https://www.cdc.gov/diabetes/basics/risk-factors.html

CDC. (2023). What is Diabetes? Retrieved April 18, 2023, from https://www.cdc.gov/diabetes/basics/diabetes.html

Chipman, H., George, E., & Mcculloch, R. (2006). Bayesian Ensemble Learning. *Advances in Neural Information Processing Systems*, *19*. Retrieved April 25, 2023, from https://papers.nips.cc/paper_files/paper/2006/hash/1706f191d760c78dfcec5012e43b6714-Abstract.html

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian Additive Regression Trees [Publisher: Institute of Mathematical Statistics]. *The Annals of Applied Statistics*, *4*(1), 266–298. Retrieved April 25, 2023, from https://www.jstor.org/stable/27801587

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*, 700–709. https://doi.org/10.1037/xge0000920
Place: US Publisher: American Psychological Association

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference [Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/jcgs.2010.08162].

Journal of Computational and Graphical Statistics, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (Second edition). Springer. https://doi.org/10.1007/978-1-0716-1418-1

National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. (2014). *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Centers for Disease Control; Prevention (US). Retrieved April 19, 2023, from http://www.ncbi.nlm.nih.gov/books/NBK179276/

Neal, B. (2020). *Introduction to causal inference from a machine learning perspective*.

pandas development team, T. (2022). *Pandas-dev/pandas: Pandas 1.4.4* (Version v1.4.4). Zenodo. https://doi.org/10.5281/zenodo.7037953

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & Laan, M. J. v. d. (2012). Diagnosing and responding to violations in the positivity assumption [_eprint: https://doi.org/10.1177/0962280210386207]. *Statistical Methods in Medical Research*, *21*(1), 31–54. https://doi.org/10.1177/0962280210386207

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software*, *97*, 1–66. https://doi.org/10.18637/jss.v097.i01

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Xie, Z. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109