MathWorks Math Modeling Challenge 2023

William Mason High School

Team #16968, Mason, Ohio

Coach: Colleen Everett

Students: Oliver Gao, Matthew Qiu, Jason Wang, Katie Wilson, David Zhang

M3 Challenge TECHNICAL COMPUTING THIRD PLACE $1,000 Team Prize

JUDGE COMMENTS

Specifically for Team # 16968 —Submitted at the Close of Triage Judging:

COMMENT 1: You demonstrated a thorough understanding of the subjects.

COMMENT 2: I like your work and results for Q1. I think you may have done a lot of extra work in Q2 for little result.

COMMENT 3: Well done; You have lots of fresh and creative ideas; Good job with assumption statements; Highly organized.

COMMENT 4: Nice summary!

Good job in Q1. Your strengths and weaknesses demonstrate your deep understanding of modeling.

Using random forest analysis in Q2 is good idea. Work on explaining your steps, what logic did you use when writing your code?

Specifically for Team # 16968 —Submitted at the Close of Technical Computing Judging:

The judges were very impressed with this team's solution to Part 1. They employed a Bass diffusion model to simulate the adoption of electric bikes in the population. To integrate the model with historical e-bikes sales data, the team fine-tuned model parameters to minimize its sum-of-squared prediction error. To do so, they wrote an objective function, which was minimized using a gradient descent method that was well implemented in Python. Overall, a very creative and effective use of computing! Judges found the team's solution to Part 2 less compelling -- it was difficult to follow the approach at times, and the use of a heavy duty machine learning method like Random Forest regression was not well justified given the amount of data available.

MathWorks Math Modeling Challenge
A program of siam

***Note: This cover sheet has been added by SIAM to identify the winning team after judging was completed. Any identifying information other than team # on a MathWorks Math Modeling Challenge submission is a rules violation. Further, this paper is posted exactly as submitted to M3 Challenge. Typos, odd formatting, or other mistakes may be attributed to the 14-hour time constraint.

Bikers? I Hardly Know Her!

Executive Summary

Although a bountiful amount of attention and discussion has been attributed to the growth of electric car use in the United States, the lesser-known but arguably more significant member of the movement towards electric vehicles is that of electric bikes. In the vast majority of urban areas in the United States, bike lanes and roads are becoming increasingly dominated by electric bikes, or e-bikes, becoming a core aspect of American transportation systems [1]. Due to this significant growth in recent years in the United States and worldwide [6], activists and policymakers have begun investigating the phenomenon, both the given motivations to cause recent trends, as well as the nationwide impacts of the fad. Ultimately, both of these results should be of great importance to the US Department of Transportation, helping to both promote additional e-bike sales, as well as alter policy based on the trends of its growth.

First, we developed a model to predict not only the trend of future e-bike sales but also the concrete values of such in the years 2025 and 2028. Viewing e-bikes as a new innovation entering the market, we opted to solve and alter the pre-existing Bass-Diffusion model to predict growth patterns. After solving, we utilized two variable regressions to calibrate the model to existing historical data. In order to quantify such values, examine long-term behavior, and minimize error, we then modeled this equation using a computational model in Python, utilizing both our mathematical expression, as well as gradient descent. Such resulted in an estimated 1.556 million e-bikes purchased nationwide in 2025, and 2.808 million e-bikes in 2028.

Next, we utilized random forest regression to determine which factors had the most significant influence on consumers purchasing e-bikes. Initially, before modeling, a quantified representation of environmental concern [6] as compared to e-bike sales quickly deemed that factor insignificant and was thus not considered in our computational model. Thus, we quantified cost and coolness (meaning an inclination to purchase due to popularity), with cost using a series of linear regressions to measure the amount of money saved when switching from other transportation methods to e-bikes, and coolness as a function of existing e-bikes. Our model yielded 4 cost importance values, one per each previously-utilized method of transport (19.98%, 19.70%, 20.42%, and 20.05%), as well as an importance of the previously mentioned ‘coolness’ factor (20.05%). Thus, our model predicted environmental concern to be negligible, but concerns of cost-efficiency and the coolness factor to be equally factored and both significant factors

in e-bike sales.

Finally, we widened the scope of the problem to the surrounding effects of increased e-bikes on the community and environment. Deciding carbon emissions and health to be the most significant outcomes of e-bike popularity, we created two separate models to articulate the overall impact on both of these areas. Using a system of multiple variables and weighted summation, each model output values to indicate either an improvement or worsening of both categories, with the magnitude indicating the extent of each. Our model predicted that the growth of electric bikes would increase calories burnt during travel by 17.45% in 2023, thus having an overall positive impact on health and wellness. As for emissions, in the year 2023, our model predicts that 3.81 fewer metric tons of carbon emissions will be produced during transportation.

Introduction

The adoption of electric bicycles, motivated by a wide variety of factors and arguably accelerated by the pandemic, has not only seen a rise in recent sales but is projected to have long-term benefits on national carbon emissions

In the first section of the problem, we were tasked with predicting e-bike sales in 2025 and 2028. Then, in the second section, we modeled the significance of various factors on the growth of e-bikes, notably money saved and the so-called "coolness factor". Finally, the environmental impacts were quantified, presenting the Department of Transportation with tangible impacts seen from the effects of parts 1 and 2.

As for global assumption, an electric bicycle is continually defined to be a two-wheeled vehicle with fully operable pedals, and a motor under 750 watts [18]. They primarily drive on roads or bike lanes of roads and are mainly used in urban areas in the United States [10].

1. Question 1: The Road Ahead

1.1. Defining the Problem:

In this problem, we were tasked with creating a model to predict the number of e-bikes sold two and five years from now, in 2025 and 2028, in the United States.

1.2. Assumptions:

1. People only own one e-bike: An individual has no reason to own more than a single e-bike. Bike rides should always be a round trip, so no benefit is gained by owning multiple e-bikes.

2. Assumes the US population is at a constant 320,000,000: According to the US Census Bureau, the population only increased by 0.1% in 2021 [2], thus the number of consumers stays relatively constant. Fluctuations in population also do not create significant changes in our model.

3. The number of e-bikes in the US before 2012 is 300,000: Given the limited data of e-bike sales prior to 2012, we found that an estimated 200,000 e-bikes existed in 2009 [5]. We assumed that 100,000 more e-bikes were sold in 2010 and 2011 based on sales in 2012 [3], leaving us with 300,000 e-bikes at the start of 2012.

4. *The initial proportion of people in the US with an e-bike is* $F_0 = \frac{370,000}{320,000,000}$: The number of e-bikes sold in the US before 2012 is 300,000 by assumption 3, and an additional 70,000 e-bikes were sold in 2012.

5. *Bass diffusion applies for at least 5 years:* Although in the long term, the Bass Diffusion model struggles with accuracy due to new innovation, in the time frame in which our model applies, we assume the model to hold [7]

1.3. Variables

Symbol	Definition	Unit	Value
$F(t)$	Proportion of Americans who own an e-bike at the end of a given year	#	
F_0	Proportion of Americans who own an e-bike in 2012 (constant)	n/a	$\frac{370,000}{320,000,000}$
$f(t)$	$\frac{dF(t)}{dt}$, or the change in the proportion of Americans who own an e-bike	n/a	
p	Coefficient of innovation (constant)	n/a	1.32×10^{-4}
q	Coefficient of imitation (constant)	n/a	1.98×10^{-1}
t	Number of years after 2012	years	
P	Population of the US	people	320,000,000

Table 1.3.1: Variable symbols, definitions, and units used in the model

1.4. The Model

The adoption of e-bikes in the United States represents a new product being adopted into a population and thus can be reflected through the Bass-Diffusion model. The Bass-Diffusion model is helpful in this situation, taking into account repeat or replacement purchases that contribute towards total purchases.

We employ the Bass-Diffusion model defined as:

$$\frac{f(t)}{1-F(t)} = p + qF(t)$$

After substituting in $f(t) = \frac{dF(t)}{dt}$ and setting the initial condition of $(t, F(t)) = (0, F_0)$, solving the differential equation for $F(t)$ resulted in:

$$F(t) = \frac{1-\frac{p(1-F_0)}{qF_0+p}e^{-t(p+q)}}{1+\frac{q(1-F_0)}{qF_0+p}e^{-t(p+q)}}.$$

Employing a multivariate regression in Python with data from 2012 to 2022 [3,4,6], we utilize pre-existing data to calibrate the model and calculate the values of p and q. The values of p and q are given in table 1.3.1.

Now, with all variables determined in the model of the proportion of citizens with e-bikes, we scale this to the population. The number of units sold each year is given by:

$$P \cdot (F(t) - F(t - 1)).$$

Recall that t is defined as the number of years after 2012, and thus in 2025, $t = 13$, and in 2028, $t = 16$. Results of such can be found in Table 1.5.1.

1.4.1 Technical Computing

Python was used to optimize parameters p and q. Due to the complex equation and multiple variables to optimize for, an exact analytical solution is likely not possible. Instead, an iterative descent gradient algorithm can be used to find a numerical solution.

The algorithm aims to minimize cost, C, which is defined as the sum of squares between expected and predicted values. Cost is minimized when the partial derivatives of cost with respect to p and q are equal to 0.

In addition to performing multivariate regressions and data analysis through computer programming tools (both of which are not otherwise possible), our model additionally utilizes gradient descent to minimize error within the model. By following incremental vectors of greatest decline along the function, gradient descent allows us to strengthen our model by best-minimizing error, as well as find the best fit between predicted and exact outputs. Each successive iteration is calculated as:

$$q, p(n + 1) = q, p(n) - LR * C' * C.$$

LR is the learning rate that controls how far the algorithm skips along the gradient. This value is kept small to ensure it doesn't jump over the minimum. To ensure as many local minimums are examined to find a global minimum, the algorithm is run many different times with different initial guesses of p and q. The p and q values found for the lowest cost are in table 1.3.1.

Gradient descent, an optimization algorithm, is rather inefficient and hefty when not performed by computational processes, and thus was best performed through our code. Our code was validated to be accurate by graphing the Bass Diffusion model with given data. Fig 1.5.2 shows the similarity between the data and the regression.

1.5. Results

	2025	2028
# of e-bikes sold in the US (thousands)	1,556	2,808

Table 1.5.1: Number of predicted e-bikes sold in the United States after 2 and 5 years, i.e. in the year 2025 and in the year 2028.

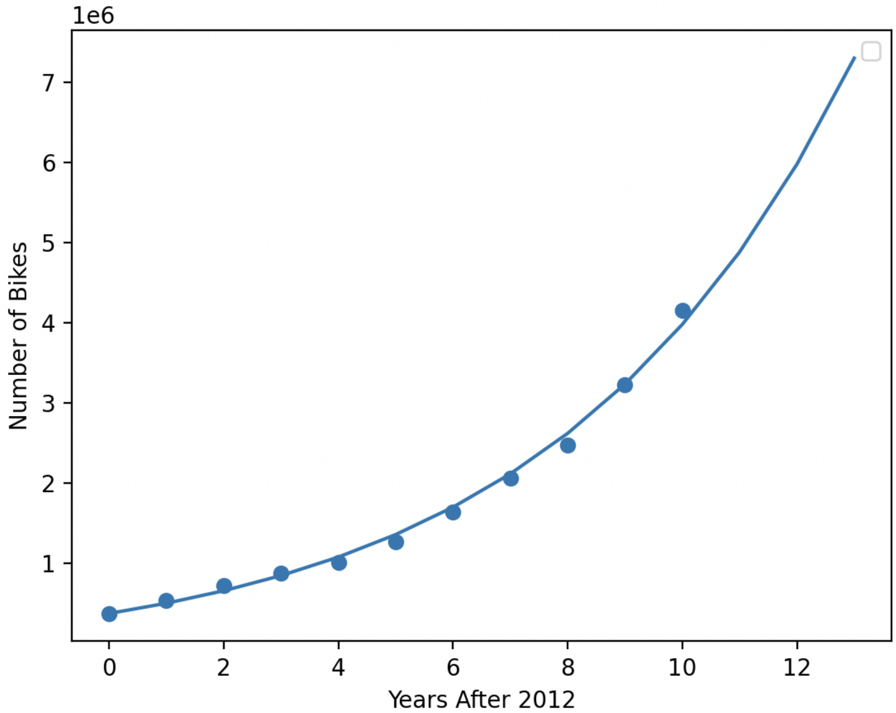

Fig 1.5.2: Model 1.4 until 2032, showing known data used to regress p and q specifically. Note that this data is the cumulative number of bikes sold, not the number of bikes sold per year. Not shown is additional growth past known points, showing points from table 1.5.1, as well as the eventual capping off of growth due to market saturation.

1.6 **Discussion**

Strengths

- As new data is acquired, it is easy to adjust the constants in our equation to refit our model. This allows our model to be updated in the future.
- The values of p and q are consistent with their typical values. p, the coefficient of innovation, has an average value of 0.03 but is often less than 0.01 [7]. Our value of q, the coefficient of imitation, is also similar to the average value of 0.38 [7].

Weakness

- The only factors truly accounted for by the Bass-Diffusion model are innovation, imitation, and initial condition. The model doesn't consider the effects of factors in the population, for example: average income, population density, age, etc. The values of p and q are an abstraction of many variables in real life, and thus makes the model difficult to adjust for changes in real life.
- Our model assumes that e-bikes as an idea catch on as a new product and are readily adopted into the population. However, it is possible for e-bikes to turn into a fad in the future, but as it is impossible to accurately assess society's opinion of e-bikes, our model still holds.

Sensitivity Analysis

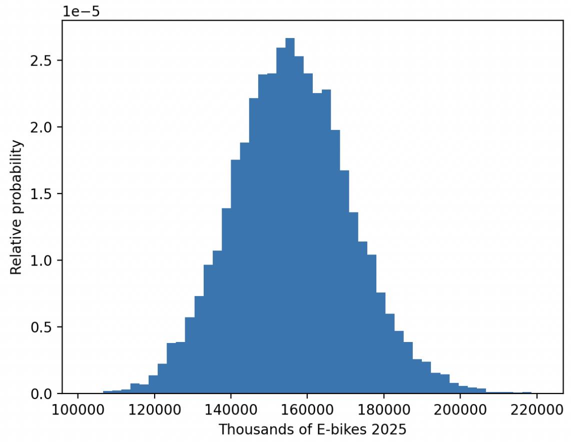

Fig 1.6.2: Probabilities of model's number of e-bike estimations in 2025 given random fluctuations in p and q.

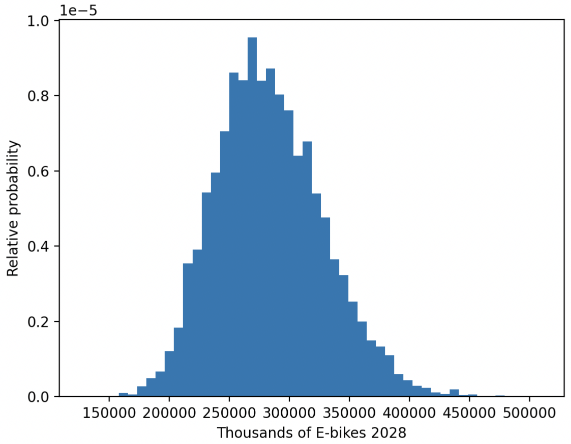

Fig 1.6.2: Probabilities of model's number of e-bike estimations in 2028 given random fluctuations in p and q.

We used a randomized Monte Carlo model to analyze the overall sensitivity of our results. We varied p and q each by a relative 10% using a normal distribution and simulated resulting bike sales. We repeated this random process 10,000 times. For 2025 and 2028, a 10% change in p and q yields a 10.09% and 15.87% standard deviation, respectively. With 10% changes being rather large in the context of the preciseness of the model, we deem these values to mark our model as being somewhat sensitive and resistant to small changes rather than larger changes.

Validation

- Assumption: The growth of electric bikes in France, USA, and Europe are similar to each other: France, USA, and Europe are most similar to each other in terms of economy, transportation culture, and development when compared with Asian countries.

Using the growth in other countries allows us to more accurately predict the future of the US. A meta-data regression produces a single regression connecting the 3 data sets together. To correct for differences in locations, all 3 sets of data were normalized to the population, to find electric bikes per capita. To account for the different stages of growth, all 3 sets of data were lined up to when their respective thousand bike sales per million persons were equal to 1.5. 1.5 was chosen because it was a common point. Linear interpolation was used to find a decimal year.

Finally, the meta-data regression was done on Desmos with an exponential function. This yielded the following regression with an R^2 of 0.87:

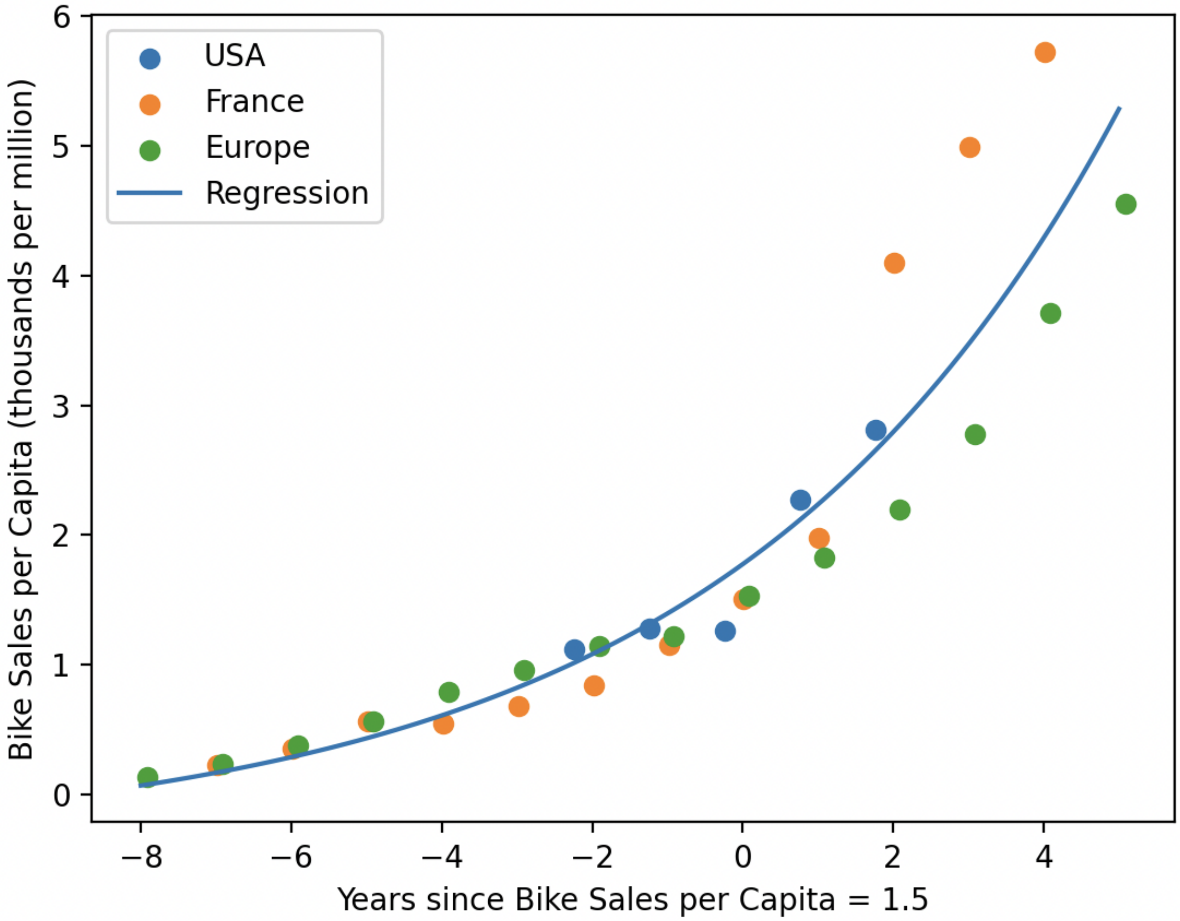

Fig 1.6.3: values yielded from alternative meta-data regression, echoing those of our model

This resulted in the following values for years 2025 and 2028:

	2025	2028
# of e-bikes sold in the US (thousands)	1,661	3,059

These values closely match the values produced by the Bass Diffusion Model, strongly validating our model's output results.

2. Question 2: Shifting Gears

2.1. Defining the Problem: Due to the debate surrounding what factors impact sales trends, we were tasked with analyzing the significance of various factors on increased e-bike sales.

2.2. Assumptions

1. *Americans drive 9 miles a day [9] at 25 miles a gallon [14], generalizing these average values to all Americans*. Thus, in one year, an average American will drive 3285 miles. The average cost of gas per mile [6] times miles per year equals the cost of gas per year (G_y).
2. *E-bike owners replace all car trips with e-bikes:* When focusing on urban areas, the most practical option will become e-bikes, allowing e-bikes to replace all car trips.
3. *The cost of charging an e-bike can be calculated each year:* E-bikes average 1.02 kWh per 100 miles and 1440 miles of travel per year [8], meaning an e-bike uses 14.688 kWh annually. Also, e-bike chargers consume as much energy as they supply, so the total electricity used is double [8]. Multiplying the annual cost of one kWh using data from the US Bureau of Labor [17] by 14.688, then doubling it, results in the annual cost of electricity for an e-bike (E_y).
4. *Environmental concern is not a significant factor in motivating consumers to buy e-bikes*: after quantifying survey data regarding environmental concern per year [6], data was relatively constant over time, as seen in Figure 2.2.1 Because of this, our simulation disregards environmental concerns, opting to analyze other factors.
5. *Only individuals without children will switch to e-bikes:* due to the lack of practicality with transporting children, when looking at ratios of the population as a whole, only those without children will consider purchasing e-bikes.
6. *All e-bikes have a lifespan of ten years and are well maintained at a negligible cost:* E-bikes can have a useful lifespan of up to ten years [15], so we assume all of the vehicles are properly taken care of at a very low cost compared to that of cars, allowing every e-bike to reach an age of ten.
7. *The price of purchasing an e-bike is assumed to be $182.50 per year:* An e-bike costs $1825 to buy on average [12], thus by distributing the cost across 10 years, $182.50 is paid per year for the price of the e-bike.
8. *Coolness is defined as the number of e-bikes sold already:* It makes sense that as the number of e-bikes sold increases, coolness increases since one sees more e-bikes around. People see how others are enjoying the product and are more likely to adopt the product themselves.

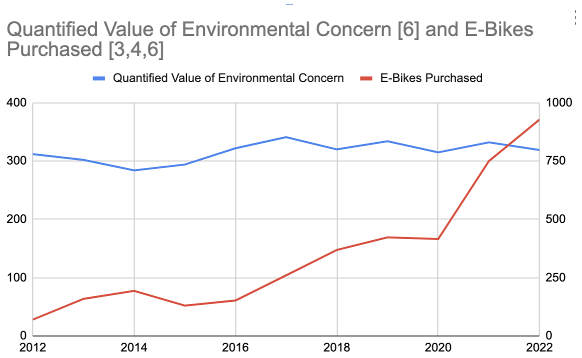

Fig 2.2.1: Lack of relationship between environmental concern in Americans and purchase of e-bikes.

2.3. Variables

Symbol	Definition	Unit	Value
G_y	Price of gas per year	$	
E_y	Price of charging e-bike per year	$	
P_y	Price paid for e-bike per year (constant)	$	182.50
T_y	Cost saved through saving time	$	
F_y	Cost saved through public transports fees (only applicable for those currently using public transportation constant)	$	1398 [24]
D_y	Disposable income per year	$	
f_{cost}	Cost input for switching from a certain form of transportation to using an e-bike	n/a	
$f_{coolness}$	“Coolness” input	n/a	
N	Number of cumulative e-bikes sold	e-bikes	

Table 2.3.1: Variable symbols, definitions, and units used in the model

2.4. The Model

2.4.1 Factor Quantification

Ultimately, we investigated the effects of cost and “coolness” (increased popularity increases demand [6]) on the predicted number of people that will switch to electric bikes in the future. We fed data starting from 2018 into a random forest model that identified the most important factor.

Cost:

First, we defined cost. We split the population who could possibly transition into e-bikes into four groups: those currently with a car, those currently with a regular bike, those currently taking public transportation, and those currently walking.

The cost input for each of the four situations is defined as the ratio of cost saved from switching to an electric car per year over the total disposable income per year. This is because it is the fraction of the disposable income each year that ultimately dictates the effect of cost on an individual. Ultimately, this value gets plugged into the random forest model.

$$f_{cost} = \frac{T_y + G_y + F_y - E_y - P_y}{D_y}$$

The projected values of E_y (assumption 2.2.3), P_y (assumption 2.2.7), and D_y [6] are given in table 2.4.1, performed via linear regressions. These values are the same regardless of the form of transportation used.

Year	E_y	P_y	D_y
2023	4.94	182.50	46,903
2024	4.59	182.50	47,493
2025	4.67	182.50	48,083
2026	4.74	182.50	48,673
2027	4.81	182.50	49,263
2028	4.89	182.50	49,853

Table 2.4.1: Projected values of E_y, P_y, and D_y to 2028

Now, we apply this model to each of the 4 situations described.

Situation 1: Currently using car

Year	T_y	G_y	F_y	f_{cost}
2023	0	457.92	0	0.00577
2024	0	468.51	0	0.00593

2025	0	479.10	0	0.00607
2026	0	489.69	0	0.00621
2027	0	500.28	0	0.00635
2028	0	510.87	0	0.00649

Table 2.4.2: Projected values of T_y, G_y, F_y and f_{cost} for those currently using cars from 2023 to 2028

In this case, T_y is assumed to be zero since traveling in a car is presumed to take the same amount of time as traveling via e-bike and F_y is zero since there is no public transportation fee for using a car. G_y was calculated via a linear regression (assumption 2.2.1).

<u>Situation 2: Currently with regular bike</u>

Year	T_y	G_y	F_y	f_{cost}
2023	5,167	0	0	0.106
2024	5,412	0	0	0.110
2025	5,573	0	0	0.112
2026	5,733	0	0	0.114
2027	5,893	0	0	0.116
2028	6,054	0	0	0.118

Table 2.4.3: Projected values of T_y, G_y, F_y, and f_{cost} for those currently using regular bikes from 2023 to 2028

In this situation, G_y is zero since a regular bike does not use gas. F_y is zero since a regular bike does not require a public transportation fine. T_y was calculated by multiplying average distance traveled per year [8] with regression-based US hourly wage [20] and dividing by the average difference in speed of an e-bike and a normal bike [19] and subtracting the two.

<u>Situation 3: Currently taking public transportation</u>

Year	T_y	G_y	F_y	f_{cost}

2023	3,217	0	1,398	0.0944
2024	3,370	0	1,398	0.0964
2025	3,469	0	1,398	0.0973
2026	3,569	0	1,398	0.0982
2027	3,669	0	1,398	0.0991
2028	3,769	0	1,398	0.0999

Table 2.4.4: Projected values of T_y, G_y, F_y and f_{cost} for those currently taking public transportation from 2023 to 2028

In this situation, G_y is assumed to be zero because the customer does not pay for the gas of public transportation. F_y was assumed to be a constant $1,398 dollars per year [24]. T_y was calculated via multiplying the average distance traveled per year [8] with regression-based US hourly wage [20] and dividing by the average difference in speed of the e-bike and normal bike [19][21] and subtracting the two.

Situation 4: Currently walking

Year	T_y	G_y	F_y	f_{cost}
2023	32,293	0	0	0.685
2024	33,828	0	0	0.708
2025	34,830	0	0	0.720
2026	35,832	0	0	0.732
2027	36,834	0	0	0.744
2028	37,836	0	0	0.755

Table 2.4.5: Projected values of T_y, G_y, F_y, and f_{cost} for those currently walking from 2023 to 2028.

In this situation G_y is zero since walking does not require gas and F_y is zero since walking does not require a public transportation fee. T_y was calculated by multiplying average distance traveled per year [8] with regression-based US hourly wage [20] and dividing by the average difference in speed of an e-bike and a normal bike [19][22] and subtracting the two.

"Coolness"

For simplicity, we defined coolness as the number of e-bikes that have already been sold (assumption 2.2.8).

$$f_c = N$$

2.4.2 Technical Computing

After determining contributing factors and quantifying them (as seen above), we utilized a random forest model to compare these outputs, and determine which had the most significant impact on projected e-bike sales. In short, via bootstrapping we taught each of the 'trees' a random section of our predicted data from 2.4.1 and projected growth from model #1. Input data for each of the cost methods were determined from a series of regressions as shown above, while coolness was determined from the number of cumulative e-bikes sold. Following input, 10,000 decision trees ensembled outputs from the 5 inputs and their significance towards contributing to e-bike sales. Clearly, the breadth of the number of computations performed and the sheer number of computations required the use of a computer.

The scikit-learn package was used for the random forest regression. The package hugely simplifies the process, saving time for coding and debugging. It takes the 5 inputs and 1 output across 10 years and generates the model. 80% of models were used for training. The model was verified for accuracy in table 2.6.1, yielding consistent predictions. Additionally, Scikit has a built-in function to examine feature importance. This essentially looks at each node in each decision tree and how well each factor is able to separate the data. The output of the feature importance function is as follows.

2.5. Results

Contributing element	**Feature Importance**
Saving cost from walking	19.98%
Saving cost from biking	19.70%
Saving cost from public transport	20.42%
Saving cost from car	20.05%
"Coolness" factor	19.85%

Table 2.5.1: Relative importance of each element contributing to e-bike sales.

Of the environment, cost, and coolness factors, we deemed that only environmental concerns had a negligible impact on e-bike sales, as determined in assumption 2.2.4. As a result, our model determined cost efficiency and coolness factor to both have an approximately equal impact on e-bike sales.

2.6. Discussion

Strengths

- Of those who work non-remotely and must commute regularly, our model accounts for the vast majority of them through our four scenarios. And in addition to this, the breadth of angles viewed and specific factors considered makes the model more encompassing of real-life externalities and impacting variables.
- The nature of the model itself, that being of randomness attributable to the computational model utilized, the model is rather robust and resistant to significant sway.
- Finally, also given the ensembling nature of the random forest model, our model prevents overfitting. As a result, not only does our regression fit the fewer parameters provided, but also serves as a more sound projective model.

Weaknesses

- Rather than determining some sort of statistical significance in relation to e-bike sales as a whole, our model instead compares modes to one another, finding a comparative significance. This falls a bit short of determining some sort of causal relationship, not providing a definitive correlation.
- Although we do agree that all of our tested nodes do, in fact, have significant impacts on e-bike sales, we acknowledge that the differences between such factors are not as significant as we would have desired, nor as what is most likely in a realistic setting.
- Our model only accounts for reasons why an individual would want to switch to an e-bike, rather than also considering reasons why an individual would be inclined to remain with their current method of travel, such as through some sort of attachment factor (that is, a quantified representation of any mode of travel's given likelihood of remaining with such).

Validation

In addition to providing values for feature importance, our model additionally utilized a regression that predicted e-bike sales for a given year, which can be compared to results from our diffusion model in 1.4.

Year	Bass diffusion result (1.4) (thousands of e-bikes)	Model 2 value (thousands of e-bikes)

2019	423	416
2021	1,090	928
2023	1,600	1,320

Table 2.6.1: Model #2's output values, as compared to outputs from model #1

Although the values vary slightly, given the same general trend and end behavior between the two models, as well as the relatively close values given the scale of the model as a whole, we find the similarity between the two to verify and validate our results in 2.5.

2.7 Model Refinement

Focusing on the weaknesses of the model, given additional time, drastic improvements can be noted, and potentially applied. One possible way to improve the model would be the addition of the attachment variable.

It is known that many individuals are attached to their vehicles and have established emotional connections to them. This could be a relevant factor in determining the proportion of people who decide to switch to e-bikes. In this case, we would have chosen to use a probability model instead of a random forest model, and calculated the probability each year that an individual would switch by weighing the benefits against the cons. In fact, this is the first model that we tried, but it was difficult to avoid overfitting the data in this case.

3. Question 3: Off The Chain

3.1. Defining the Problem: Scaling back out to the population as a whole, we were tasked with modeling the impact of e-bike growth on carbon emissions, as well as health and wellness.

3.2. Assumptions

1. *Traffic congestion will not be significantly altered by e-bike usage:* in the vast majority of US States, e-bike usage operations must occur on roadways [10], not detracting from traffic congestion.

2. *An individual riding an e-bike, riding a traditional bike, and taking public transport will have no impact on carbon emissions.* Neither of the biking options directly burns a fuel that contributes to carbon emissions. For public transportation, the public transportation will generally run regardless of if an individual stops riding, so when an individual switches to an e-bike, they don't decrease the emissions coming from public transport.

3. *An individual will only use a single mode of transportation:* We are modeling the

primary mode of transportation for individuals, so it is safe to assume that for the purposes of transportation, only one mode is used.

4. *An individual should burn 500 calories a day through exercise.* This is the recommended value given by [26].

5. *The calories burnt per hour for walking is 100, for biking is 500, and for e-biking is 300.* These are the values given by [25, 27, 28].

6. *Driving burns 0 calories.* The act of driving is generally not a form of exercise and contributes 0 calories to the 500-calorie goal.

7. *A person is equally likely to switch to riding e-bikes.* Based on the results of our previous model, many factors are equally important in determining if someone will switch to e-bikes, meaning that on average, any person is equally likely to switch to e-bikes.

3.3. Variables

Symbol	Definition	Unit	Value
M	Miles traveled per day	mi/day	9
H_e	Hours per mile by e-bike (constant)	hr/mi	$\frac{1}{28}$
C_e	Calories burned per hour by e-bike (constant)	cal/hr	300
H_0	Hours per mile by initial transportation	hr/mi	
C_0	Calories burned per hour by initial transportation	cal/hr	
c	Change in calories burned per day	cal	
P	% contribution to daily calorie goal	%	
w	Weight of change in calories	%	
$\%_n$	Net percent change in calories burned	%	
B_s	E-bikes sold in 2023 (constant)	#	1.09×10^6
p_c	Proportion of e-bikes sold converted from car (constant)	%	75.6%

E_c	CO_2 emissions per car per year (constant)	metric tons/yr	4.6
E_r	CO_2 removed per year	metric tons/yr	

Table 3.3.1: Variable symbols, definitions, and units used in the model

3.4. The Model

The factors we analyzed in this model are exercise and carbon emissions. One model calculates the average percent change in exercise as a result of the adoption of e-bikes, while another model determines the change in carbon emissions as a result of adopting e-bikes.

To calculate the effect of switching to e-bikes on exercise, the difference per day in the number of calories yielded by changing mode of transportation is calculated by:

$$M(H_e C_e - H_0 C_0) = c\ .$$

This amount is then converted into a percentage towards the 500 calories daily goal:

$$\frac{c}{500} * 100\% = P\ .$$

Finally, the percentages are multiplied with the proportion of e-bike converters that come from driving, biking, public transit, and walking, in order to weigh the factors out and get a total % change on daily caloric goal:

$$\sum_{n=1}^{4} wP = \%_n$$

The method for finding the change in carbon emissions was straightforward: since cars are the only source of carbon emissions considered, the total change in carbon emissions from switching to e-bikes is as follows:

$$B_s \cdot p_c \cdot E_c = E_r\ .$$

3.5. Results

	Change in calories per day (kcal)	% on daily goal	Weighted % on daily goal	CO_2 removed in 2023(metric tons)
Car	+96.43	+19.29%	+18.07%	$3.81 \cdot 10^6$
Biking	-278.57	-55.71%	-0.28%	0

Walking	-203.57	-40.71%	-0.94%	0
Public Transport	+96.43	+19.29%	+0.60%	0
Total			+17.45%	$3.81 \cdot 10^6$

Table 3.5.1: The results of the model

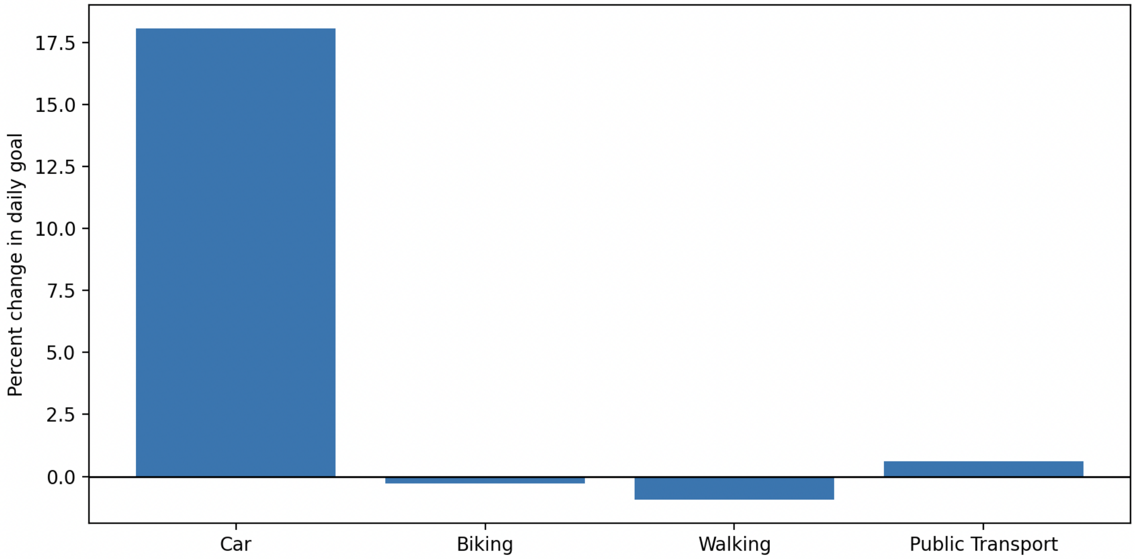

Table 3.5.2: A graphical representation of the contribution of each mode of transportation to the change in percent daily caloric goal

3.6. Discussion:

Strengths

- Our model is strong in simplicity, with only multiplication and addition. This makes it easy to understand and able to be adapted to many other situations. The premise of our model is rooted in the fundamental relationships between variables, which can be adjusted and applied to other different situations.

Weakness

- The major weakness of our model is that the model as a whole is strongly based on our assumptions, especially assumption 3.2.2. It is improbable for an individual to only own an e-bike and not a car, or to commute 9 miles a day by walking. However, there are many possible exceptions to our general assumptions which could have significant impacts on results.

Conclusion

The e-bike trend has been increasing recently, and it seems as if the trend will continue in the future. The state of transportation has never been more important than right now. The cost of infrastructure is at the forefront of public media, and while the global warming effects of automobiles become more noticeable, it becomes more important than ever for us as a society to develop alternatives and solutions to our problems. Bicycles and e-bikes represent the first step into the future of transportation and infrastructure.

Based on our models, we conclude that cost efficiency and the coolness factors have a similar importance in determining the future trends of e-bike sales. This reflects the complex dynamic we see in the sales industry and in much of society, as no trends are ever dependent or reliant on a single, identifiable factor. Then, when looking outwards, we also see that although the shift towards e-bikes is beneficial for health and wellness as well as for the environment, interestingly enough, this is not a major factor in determining trends in e-bike sales.

Looking analytically, our models serve to verify and validate one another, additionally supported by outside data. Thus, from an accuracy perspective, models 1-3 seem to accurately echo that of real life, which can be confirmed in each section's validation section above. However, most of the models fall short of reaching the full scope of contributing variables, instead prioritizing including the specific factors of considered variables. This is one continuous weakness across models, which, given additional time, could be addressed with the addition of parameters and terms.

Nevertheless, given the strategic avoidance of over-fitting and consistent usage of given data to verify, our models serve to successfully predict future trends, as well as analyze contributing factors and overarching societal benefits. All in all, e-bikes stand at the forefront of the development of American transportation, and seem as though they will continue to be such.

References

[1]https://www.calbike.org/wp-content/uploads/2019/02/A-North-American-Survey-of-Electric-Bicycle-Owners.pdf

[2]https://www.census.gov/library/stories/2021/12/us-population-grew-in-2021-slowest-rate-since-founding-of-the-nation.html

[3]https://www.statista.com/statistics/326124/us-sales-of-electric-bicycles/

[4]https://leva-eu.com/usa-electric-bike-market-reaches-26000-in-2017/

[5]https://en.wikipedia.org/wiki/Electric_bicycle

[6]https://m3challenge.siam.org/node/596

[7]https://en.wikipedia.org/wiki/Bass_diffusion_model

[8]https://www.efficiencyvermont.com/Media/Default/docs/white-papers/efficiency-vermont-electric-bike-white-paper.pdf

[9]https://www.nyc.gov/html/dot/downloads/pdf/nyc_greendividend_april2010.pdf

[10]https://www.bikeberry.com/blogs/learning-center/electric-bike-laws-a-state-by-state-breakdown

[11]https://www.bts.gov/content/average-fuel-efficiency-us-light-duty-vehicles

[12]https://theroundup.org/ebike-statistics/

[13]https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle

[14]https://www.businessinsider.com/most-fuel-efficient-cars-vehicles-best-gas-mileage-2019-11

[15]https://ride1up.com/how-long-do-electric-bikes-last/

[16]https://www.thezebra.com/resources/driving/average-miles-driven-per-year

[17]https://www.bls.gov/regions/midwest/data/averageenergyprices_selectedareas_table.htm

[18]https://www.federalregister.gov/documents/2020/11/02/2020-22129/general-provisions-electric-bicycles

[19]https://discerningcyclist.com/how-fast-are-electric-bikes/

[20]https://ycharts.com/indicators/us_average_hourly_earnings

[21]https://www.cato.org/policy-analysis/charting-public-transits-decline#transit-is-slow

[22]https://www.healthline.com/health/exercise-fitness/average-walking-speed#

[23]https://www.statista.com/statistics/242022/number-of-single-person-households-in-the-us/

[24]https://www.timeout.com/newyork/blog/the-cost-of-public-transportation-in-new-york-is-75-percent-higher-than-the-national-average-041816

[25]https://www.flyer-bikes.com/en/this-is-how-many-calories-you-burn-when-e-biking

[26]https://www.cnet.com/health/fitness/burn-this-number-of-calories-in-a-day-to-lose-weight-according-to-experts/.

[27]https://www.nerdfitness.com/blog/walking/
[28]https://www.bizcalcs.com/calories-burned-biking/

Code Appendices

Model 1: The Road Ahead

Gradient Descent Algorithm

```
#requires the exponential function from math module
from math import exp

#Learn rate of the gradient descent machine learning
learnRate = 10**-34

#Data we are given about bikes, as well as the calculated F0
bikeData =
[370000,529000,722000,872000,1004000,1267000,1636000,2059000,2475000,3225000,4153000]
F0 = 370000 / 320000000

#This is the bass diffusion equation.Given values of p, q and the year, this function
returns the expected number of bike sales
def func(q,p,t):
   likeTerm = ( 1 - F0 ) * exp( -t * ( p + q ) ) / (q * F0 + p)
   return( 320000000 * ( 1 - p * likeTerm) / (1 + q * likeTerm) )

#This is the cost function. Given the list of years, q, and p parameter, it calls the func
function to calculate the expected number of bike sales. It then compares that to the given
number of bike sales to calculate the square of the differences. The function finally
returns the sum of all the squared differences.
def cost(input,outputExpected,q,p):
   cost = 0
   for year in range(len(outputExpected)):
       cost += ( func(q,p,input[year]) - outputExpected[year] ) ** 2
   return(cost)

#Given an initial value of p, q, and the expected bike data, the function returns the
converged values of p and q.
def optimize(bikeData,p0,q0):
   #Initializes variables. The increment is used during numeric differentiation.
   increment = 10**-12

   #Initializes value of changes in cost, to be above the while loop threshold.
   changeCostp = 99999
   changeCostq = 99999

   #Intializes the values of p and q that will be iterated through.
   p = p0
   q = q0

   #Sets a counter for iterations done to avoid a large runtime
   iterations = 0

   #The while loop continues when the change in costs is above 10^-9, to a maximum of 1000
iterations. If the change in costs with respect to p and q is below 10^-9, the program has
```

```
converged and returns p, q, and cost of the model.
    while (abs(changeCostp) > 10**-9 and abs(changeCostq) > 10**-9) and (iterations < 1000):

        #Numeric partial differentiation with respect to p and q.
        changeCostp = ( cost(range(len(bikeData)),bikeData,q,p + increment) -
cost(range(len(bikeData)),bikeData,q,p) )  / increment
        changeCostq = ( cost(range(len(bikeData)),bikeData,q + increment,p) -
cost(range(len(bikeData)),bikeData,q,p) ) / increment

        #Calculates the current cost of the function based off of current p and q
        currentCost = cost(range(len(bikeData)),bikeData,q,p)

        #Iterates p and q as part of gradient descent model
        p = p - learnRate * changeCostp * currentCost
        q = q - learnRate * changeCostq * currentCost

        iterations += 1
    return(p,q,currentCost)


#defines initial starting point for p0 and q0
p0 = 1*10**-5
q0 = 1*10**-1

#initializes running cost counter, p tracker, and q tracker
runningCostCounter = 99999999999999
runningpqCounter = (0,0)
testq = q0
testp = p0

#Iterates through many starting values of p0 and q0 to assure many local minimums have been
checked for, to find a better guess for the absolute minimum of the cost function. It
iterates from p0 to 2p0, and q0 to 2q0, with 50 equal intervals.
for i in range(50):
    testq = q0
    for j in range(50):
        try:
            testResults = optimize(bikeData,testp,testq)

            #Updates p,q,and best cost, if the cost from the optimization function is better
than the current best cost.
            if testResults[2] < runningCostCounter:
                runningCostCounter = testResults[2]
                runningpqCounter = (testResults[0],testResults[1])
            testq += q0/50
        except:
            pass
    testp += p0/50
#prints results
print(runningpqCounter)
```

Monte Carlo for Sensitivity Analysis

```
#Imports necessary packages. Pyplot for generating graphs, numpy for normal distributions,
and math for the exp function.
from matplotlib import pyplot as plt
import numpy as np
from math import exp

#p0 and q0 determined from gradient descent model
p0 = 0.00013258321395629718
q0 = 0.19800034599222005

#data/constants
bikeData =
[370000,529000,722000,872000,1004000,1267000,1636000,2059000,2475000,3225000,4153000]
F0 = 370000 / 320000000

#Same function as gradient descent model for outputting predicted number of bikes
def func(q,p,t):
  likeTerm = ( 1 - F0 ) * exp( -t * ( p + q ) ) / (q * F0 + p)
  return( 320000000 * ( 1 - p * likeTerm) / (1 + q * likeTerm) )

#Creates a scatterplot for bike data, and graphs the predicted bass diffiusion curve
fig = plt.scatter(range(len(bikeData)), bikeData)
x = range(len(bikeData)+3)
y = []
for value in x:
  print (value)
  y.append(func(q0,p0,value))
fig = plt.plot(x,y)
plt.xlabel("Years After 2012")
plt.ylabel("Number of Bikes")
plt.legend()
plt.show()

#Monte Carlo sensitivity analysis. Iterates 10000 trials, where q and p are picked from a
normal distribution with a mean of q or p, and a standard deviation of 10%.
yearTwo = []
yearFive = []
for i in range(10000):
  q = np.random.normal(q0,q0/10)
  p = np.random.normal(p0,p0/10)
  yearTwo.append(func(q,p,2)-func(q,p,1))
  yearFive.append(func(q,p,5)-func(q,p,4))

#Plots histogram to show spread of bike sales after varying p and q.
plt.hist(yearTwo,density=True,bins = 50)
plt.xlabel("Thousands of E-bikes 2025")
plt.ylabel("Relative probability")
plt.show()

plt.hist(yearFive,density=True, bins = 50)
plt.xlabel("Thousands of E-bikes 2028")
plt.ylabel("Relative probability")
plt.show()

#Outputs standard deviations of outputs after varying p and q.
stdv2years = np.std(yearTwo) / np.mean(yearTwo)
stdv5years = np.std(yearFive) / np.mean(yearFive)
print(stdv2years,stdv5years)
```

Model 2: Shifting Gears

Random Forest Algorithm

```
#Imports necessary packages for the random forest model. Data will be stored and manipulated
in a Pandas dataframe. The scikit ensemble package will be used to make the random forest
model.
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier


#Constants defined in the mode
MMG = 25
MILESPERDAY = 9
BIKECOST = 182.5
PUBLICTRANSPORTCOST = 1398

#Data from regressions
YEARS = [2018,2019,2020,2021,2022,2023,2024,2025,2026,2027,2028]
DOLLARSKWHR = [4.00,4.00,3.97,4.14,4.67,4.94,4.59,4.67,4.74,4.81,4.89]
DOLLARSGAL =
[2.72,2.6,2.17,3.01,3.95,3.4849401,3.5655288,3.6461175,3.7267062,3.8072949,3.8878836]
DISPINC =
[43886,44644,47241,48219,46312.53,46902.645,47492.76,48082.875,48672.99,49263.105,49853.22]
EBIKES =
[1636000,2059000,2475000,3225000,4153000,4878754.786687736,5973197.687153384,7297219.6978950
55,8896341.865764502,10823883.981983375,13141730.250541717]
EBIKESALES =
[369000,423000,416000,750000,928000,1094442.9004656477,1324022.0107416706,1599122.1678694477
,1927542.1162188724,2317846.268558342,2779158.499445403]
WAGE = [26.71,27.59,28.43,29.92,31.63,33.03,34.6,35.625,36.65,37.675,38.7]

#Initializes dataframe to be used. Inputs all relevant lists into dataframe.
df = pd.DataFrame()
df['Years'] = YEARS
df['Dollars per kwhr'] = DOLLARSKWHR
df['Dollars per gal'] = DOLLARSGAL
df['Disposable Income'] = DISPINC
df['Ebikes'] = EBIKES
df['Sales'] = EBIKESALES
df['Wage'] = WAGE
```

```
#Converts Ebikes and Sales from regression into an integer. Random forest models require
integers for output training.
df['Ebikes'] = df['Ebikes'].astype(int)
df['Sales'] = df['Sales'].astype(int)

#Calculates the net cost saved from switching from each method
df['gas cost'] = df['Dollars per gal'] * MILESPERDAY * 365 / MMG
df['Bike Cost'] =  df['Dollars per kwhr'] + BIKECOST

df['Net Save per Income Car'] = (df['gas cost'] - df['Bike Cost'])/ df['Disposable Income']
df['Walk Save'] = ( (- MILESPERDAY * 365 / 28 + MILESPERDAY * 365 / 3) * df['Wage'] -
df['Bike Cost'] ) / df['Disposable Income']
df['Bike Save'] = ( (- MILESPERDAY * 365 / 28 + MILESPERDAY * 365 / 12) * df['Wage'] -
df['Bike Cost'] ) / df['Disposable Income']
df['Public Save'] = ( (- MILESPERDAY * 365 / 28 + MILESPERDAY * 365 / 15.3) * df['Wage'] -
df['Bike Cost'] + PUBLICTRANSPORTCOST ) / df['Disposable Income']

#Sets the 5 variables into the inputs
inputs = df.filter(['Walk Save','Bike Save','Public Save','Net Save per Income
Car','Ebikes'])

#Sets sales as output
output = df['Sales']

#Starts the random forest model with 10,000 trees.
randomForest = RandomForestClassifier(n_estimators = 10000)

#Trains the data on 80% of the given inputs.
Xtrain,Xtest,Ytrain,Ytest = train_test_split(inputs,output,test_size=.2)
randomForest.fit(Xtrain,Ytrain)

#Uses the remaining 20% of inputs to test validity, feature importances, and accuracy
Ypred = randomForest.predict(Xtest)
print(Ypred,Ytest)
print(randomForest.feature_importances_)
```