MathWorks Math Modeling Challenge
A program of siam

# PREVIEW PAPER: EXCELLENT*

The team has a good executive summary which includes results and provides insights into the methods the team used. The team provides strong responses to the first two questions. The response to question three is good, but the discussion lacks important details about their simulations. For example, it was not clear what rules were put in place, and it would be difficult to repeat their simulations based solely on the discussion. The team did a good job of annotating graphs. The team used citations but was not consistent throughout the paper. The different subsections, such as strengths and weaknesses and sensitivity, are clearly labeled, and they provide good insights into their analysis of their models.

In question one, the team's assumptions are well reasoned and appropriate. They have a broad view of the context. For example, the team notes the impact of Brexit. They include a good discussion of their regression methods. They discuss the residuals and note the patterns found in the residuals rather than simply rely on R-squared values. The team also discusses the differences in short-term versus long-term predictions, but they conclude an exponential model is better for the longer term which is problematic.

The team's approach to question two is somewhat unique in our sample. The factors considered are clearly stated. They make use of a random forest regressor and include a brief overview of the method, and they also include a good description of how they used the method. The methods used by the team's response to address question three, however, is more difficult to determine. The rules used in the simulation were not completely stated, and the results are not stated in a manner consistent with the stochastic nature of the simulations.

**from among the screened sample of papers examined during pre-triage work.*

# 1 Executive Summary

Environmental awareness has seen large increases globally in the past few decades. Due to measurable changes such as increased $CO_2$ emissions and the decline in certain wildlife populations, sustainable energy has become a much greater focus both for governments and society as a whole. Although electric cars often get most of the attention, e-bikes are quietly taking over and contributing to a new generation of sustainability. E-bikes offer an alternative form of transportation to those living in an urban setting, and are not only significantly healthier than gasoline cars for the environment, but also offer health benefits to those who ride daily.

As e-bikes are a continuously growing market, the sales of e-bikes will not remain constant throughout the next five years. To model this growth, we developed exponential models for both the US and the UK. Using these models, we found that the US will purchase 1,922,281 e-bikes in 2025 and 3,969,391 e-bikes in 2028; the UK will purchase 2,555,919 e-bikes in 2025 and 5,271,485 e-bikes in 2028. These predictions were realistic as they represented around 3% to 8% of the total population. While these percentages were on the high side, they were still deemed reasonable since annual car sales hover around 3% of the total population.

We then incorporated several factors to determine the growth of e-bike sales in the UK. We took into account environmental concerns, popularity, disposable income per capita, and gas prices to create a random forest regression model for total e-bike sales. This model allowed us to compute the feature importance of each factor to determine their significance on e-bike growth. The relative importance of environmental concerns, popularity, disposable income per capita, and gas prices were 0.033, 0.772, 0.160, and 0.036, respectively. From these values, we determined that popularity and disposable income were significant reasons for the e-bike growth witnessed, while gas prices and environmental concerns were less significant reasons for e-bike sales.

We then had to quantify the impact of e-bikes as alternatives to bikes and cars. We focused on carbon dioxide emissions, traffic congestion, and health effects due to exercise changes. To quantify these, we ran a Monte Carlo simulation to find the number of bikes and cars replaced annually. We assumed that these replacements were caused by the reason for purchasing the e-bike in the first place (people seeking popularity would replace their bikes, buyers driven by gas prices and environmental concerns would replace their cars, and buyers motivated by increases in disposable income would simply retain both). We used the Monte Carlo simulation to account for changes in these preferences throughout the next five years and found that e-bikes resulted in savings of 228,494 metric tons of carbon dioxide, 669 million miles of car travel, and 16,394 million more calories burnt for the entire population of the UK in 2028. For a five-year span from 2024-2028, this resulted in 734,001 metric tons of carbon dioxide saved, 52,590 million more calories burnt, and 2.148 billion fewer miles traveled by car. This was reasonable since the UK contributes approximately 478 million metric tons of carbon dioxide per year; thus, e-bikes contributed to a 0.05% decrease in carbon dioxide emissions, which is reasonable. However, this also shows that e-bikes alone cannot solve the problem of greenhouse gas emissions, and other measures need to be taken such as reducing factory emissions and implementing greener energy generation.

# Contents

# 2 Introduction

Due to the recent COVID-19 pandemic and the upward trend in environmentalism, general interest in electric vehicles has amped up at a shockingly fast rate. The end of the COVID-19 pandemic and other global events, such as the ongoing conflict between Russia and Ukraine, led to rapid inflation in the gasoline industry in the past 2-3 years. Further, environmental awareness has been a focus of many governments and activist organizations. These changes have prompted many people to look into electric vehicles. While celebrities like Elon Musk have popularized electric cars, electric bicycles have actually become more popular than electric cars [21]. Many prefer the use of electric bikes due to the health, environmental, and cost benefits that they offer.

# 3 Part I: The Road Ahead

## 3.1 Restatement of the Problem

In this problem, we are tasked to make a model that predicts the number of e-bike sales in 2025 and 2028. We chose to look at these numbers for the United States and the United Kingdom.

## 3.2 Assumptions

1. *The United Kingdom leaving the European Union does not have a significant impact on any economies.* All data that is used was collected before Brexit, so it will not be possible to account for any large economic impacts that resulted.

2. *UK's portion of total EU e-bike sales is equal to the UK's portion of the total EU GDP (15-16% every year [1]).* A country's GDP is a good indication of its purchasing power, so the UK should contribute a similar proportion to the EU GDP and e-bike sales.

3. *There will be no significant changes in legislation for e-bikes (tax credits, subsidies, etc.)* Government regulation of the e-bike market is highly random and very difficult to predict. This would make it impossible to make an accurate model as future government sentiment and legislation cannot be accounted for.

4. *The UK and the EU will experience economic growth at the same rate; US growth should also follow in the same pattern but not the same rate* These economies (US and UK) are highly interconnected so they should grow in similar fashions [2]. The UK was a part of the EU so their economic growth should be proportional to each other. This will be verified later in this section.

5. *The e-bike industry in the US and the UK is still in its growth phase, and will remain so for at least 5 years* The e-bike industry as a whole is still in its early stage, with new innovations and adoptions being released on a regular basis [3]. As such, the purchase of these bikes should not slow down or level off for another 5 years.

## 3.3 Model Development

### 3.3.1 Variables

| Variable | Definition | Units |
|---|---|---|
| S | Number of e-bike sales in a given year | Thousands of e-bikes |
| t | Years after 2005 (t=1 being 2006) | Years |

### 3.3.2 Regression Analysis

We begin the investigation of e-bike sales over time by considering Europe's e-bike sales in the past 14 years [4]. We plan to do a simple linear regression since this allows for the most convenient extrapolation of future data based on the past data that we have. However, beginning by plotting sales as a simple function of year, it is clear from the curvature of the residual plot that a simple linear regression will not be appropriate for modeling this data (Plots shown below).

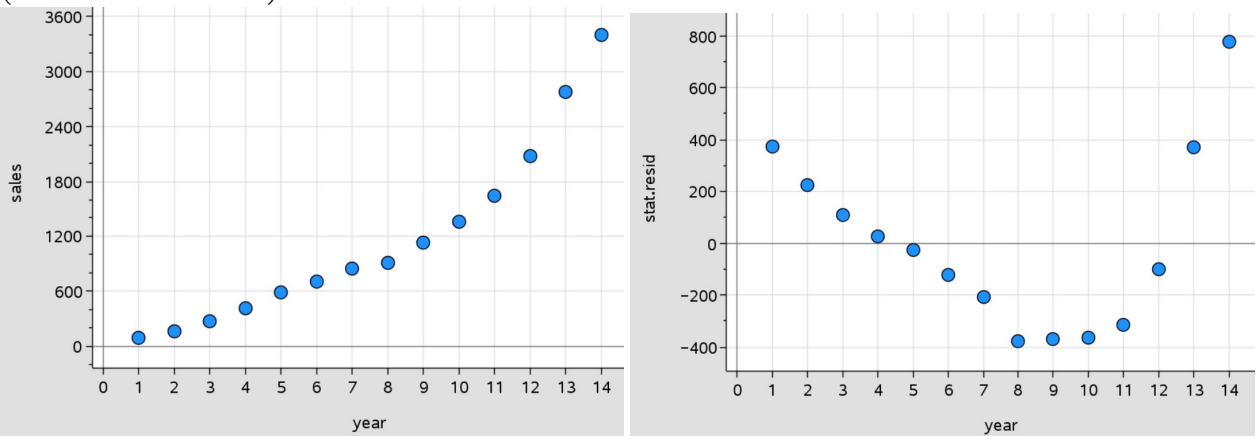



Figure 1: Plot of S versus t

After linearizing the data, the two most appropriate residual plots appear by graphing ln(S) as a function of t and ln(S) as a function of ln(t). The results from these two plots and their corresponding residual plots are shown below.

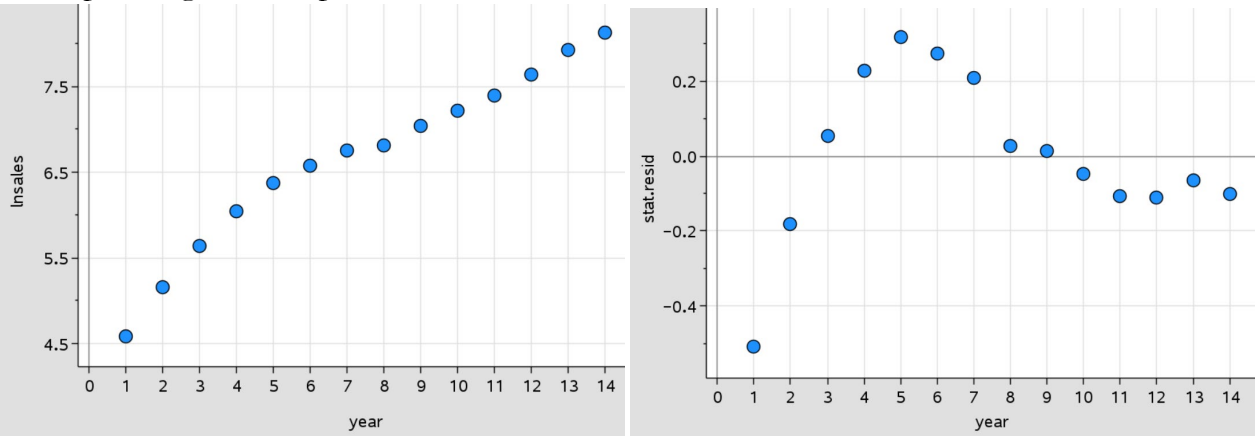



Figure 2: Plot of ln(S) versus t

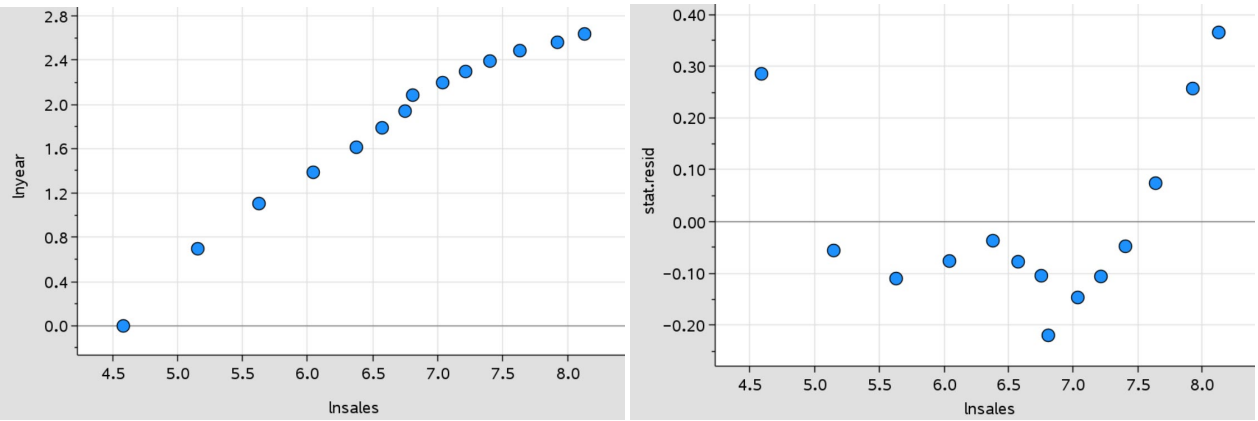



Figure 3: Plot of ln(S) versus ln(t)

Fitting a linear model to each of these, the former (Model1) has an $r^2$ of 0.9562, while the latter (Model2) has an $r^2$ of 0.9704. Since both of these values are comparably high, we will choose between the two models by looking at the residual trends at higher time values, since we are trying to predict e-bike sales in the future. Of the two residual plots, the one giving ln(S) against ln(t) has much more deviation at higher time values, whereas the plot giving ln(S) against t seems to decrease in deviation as time increases. Therefore, we chose Model2—ln(S) versus t—for our final model, with the following equation:

$$ln(S) = 4.8527 + 0.2413t$$

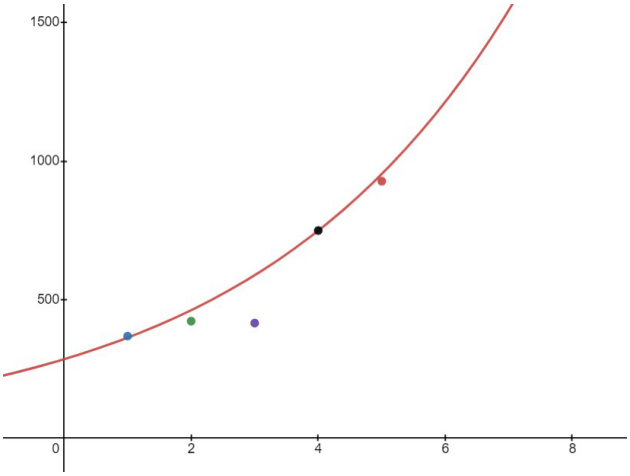
Solving for S, this yields

$$S_{Europe}(t) = 128.092e^{0.2413t}$$

Taking into account the British portion of the EU's total consumption of goods (specifically e-bikes) yields

$$S_{UK}(t) = 20.4947e^{0.2413t}$$

Doing the same regression analysis for the US proves a difficult challenge, since there are only 5 data points to work off of. This presents an issue with determining the type of growth of bike sales over time; however, if we know the form of the e-bike sale growth in the US, this problem would be minimized. While we did assume that the growth of the US e-bike sales would be the same fashion as Europe's (i.e. exponential), we want to verify this by comparing Europe's growth to the US's growth for the time period in which there is data for the United States. To do this, we shift Europe's function to intersect with the US's point in the year 2021, with 750 thousand e-bike sales (this point on the US curve is chosen because it is after the COVID-19 regulations began winding down, leading to more people going outside and using e-bikes). The following graph is obtained, showing that the US does, in fact, change in an exponential manner, just as Europe's function:




One notable exception to this trend is the year 2020 for the US. In this year, the scare of COVID-19 caused a large proportion of the country to stay indoors, dramatically decreasing the demand for e-bikes [5]. Because large-scale economic disruptions like this are very infrequent, and will likely not happen in the time frame of our concern, we considered removing this year from the data for our model, suspecting that this year is a significant outlier. Although a simple hypothesis test gives a p-value of 0.136, indicating that we should not delete 2020 from our data, we still delete it from our data, since we have very compelling real-world evidence that this data point is indeed an outlier. Doing a regression analysis on the remaining 4 data points, the following plot and equation are obtained:

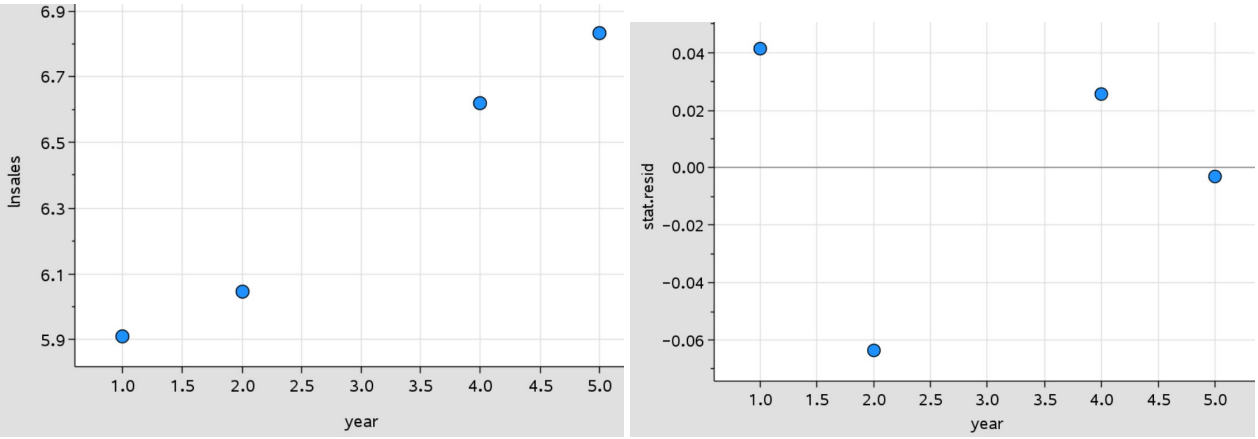



Figure 4: US plot of ln(S) versus t; $ln(S) = 5.6277 + 0.2417(t - 12)$

Note that, since the US data starts in 2018 instead of 2006 for the UK data, we manually added a 12-year shift in the input year.

This simplifies to an equation of $S_{US}(t) = 278.013e^{0.2417(t-12)}$ and an $r^2$ of 0.9891.

## 3.4 Results

According to the model, the UK will purchase 2,555,919 e-bikes in 2025 and 5,271,485 e-bikes in 2028. The US will buy 1,922,281 e-bikes in 2025 and 3,969,391 e-bikes in 2028.

As expected, the total number of e-bikes purchased increased in both countries over the 5-year time frame. In the final year of our model, approximately 5 million e-bikes were purchased in the UK with a population of around 67.3 million people [6]. This represented 7.8% of the population, which is reasonable considering the UK purchased 2.3 million passenger cars in 2019 [7], or 3.4% of the population, and cars are a much more expensive purchase with a longer shelf life and thus turnover between purchases. Comparing it to Model1, which yielded values of 602,757 and 724,323 respectively for 2025 and 2028. Comparatively, these values seem too low for the growth of the e-bike industry (hovering around 0.001% of the total population), further cementing Model2 as the more appropriate model for predicting the future of the e-bike market. The US totals for e-bikes are an even lower percentage of the population, which supports the notion that the US is in an earlier stage of e-bike growth. This is also likely because the US has a lower population density and thus it is less likely for individuals to be close enough to commute with an e-bike instead of a regular gas vehicle.

### 3.4.1 Model Validation

In order to validate the models, we looked at the $r^2$ values and residual plots, in addition to doing sensitivity analysis. The $r^2$ values for each UK model that we tested were high, where $r^2 = 0.9562$ for the model where ln(S) is a function of t and $r^2 = 0.9704$ for the model where ln(S) is a function of ln(t). These values mean that for the former model, time accounts for approximately 95.62% of the variability in ln(S). For the latter model, ln(time) accounts for approximately 97.04% of the variability in ln(S). As for the residual plots of the two models, both show nonlinear trends. Residual plots should theoretically have no visible trend, but the plots for our models clearly have nonlinear trends. The residual plot model we chose, $S_{UK}(t) = 20.4947e^{0.2413t}$, has smaller residuals as the number of years increases. This means that the model is better at predicting higher years, which hopefully means that the model can predict accurately in the future.

### 3.4.2 Sensitivity Analysis

The model only has one input (time), but our UK model is also heavily reliant on assumption 4. For example, if the UK percentage of EU GDP was changed by 1% to 15% (akin to its percentage in 2021 [8]), our values for 2025 and 2028 become 2,396,174 and 4,942,017. If the percentage was instead changed to 17%, our predicted values would be 2,715,664 and 5,600,952 respectively. Based on these numbers, this model is highly sensitive to changes in the proportion of Europe's GDP attributed to the UK.

## 3.5 Strengths and Weaknesses

The strength of our model comes from its ability to predict short-term growth, its simplicity, and its ability to be generalized to other countries. The e-bike industry in both the US and the UK appears to still be in the growth stage, but at different points for each. At this point, the UK's e-bike industry is rapidly growing, and our exponential model is able to predict how much the industry will grow in the short run. The model is also simple, which makes it a lot easier to use. It only requires one input, the year, which will minimize the amount of data collection needed to make a prediction.

Our model struggles to predict long-term growth in the e-bike market. In the long run, e-bike sales will eventually level off as a larger proportion of the population has purchased e-bikes. However, the current data shows no signs of slowing sales, so it is impossible to predict when that slowing in growth will occur. Thus, a logistic curve could not be fit to model the e-bike sales. There is also a lack of data on US e-bike sales. It is generally recommended to have at least 10 data points for each independent variable in a model. The UK has 14 years of data for one independent variable, which is above the recommended amount. However, the US only has 5 years of e-bike data, which is not enough to create a strong model. This lack of data will make it far more challenging to accurately predict future e-bike sales in the US. Lastly, both of the models for the UK had residual plots that resembled quadratic functions. Although one was upward facing and one was downward facing, these indicated that the exponential model might not be appropriate for the beginning and far future of the e-bike market.

# 4 Part II: Shifting Gears

## 4.1 Restatement of the Problem

In this problem, we were tasked with modeling one or more factors that may have been a significant reason for the recent e-bike growth. We then were to determine which factors were the most significant.

## 4.2 Assumptions

1. *Income, gas price, climate change concern, and popularity are the only factors that influence potential buyers of e-bikes.* While there may be other factors that influence buyers' decisions, they are either already included in the larger categories, or cannot be accounted for. Variables that are not accounted for will be discussed in the weaknesses.

2. *No technological advances occurred in the non-e-bike industries.* Other alternatives for e-bikes have generally reached a technological equilibrium phase where new changes are mostly cosmetic or superficial and will not have an effect on e-bike growth [9].

3. *The UK and the US share similar sentiments and opinions.* In the age of the Internet, opinions and propaganda are easily shared through social media. Both countries have

very similar demographics and as such should share similar opinions [10][11].

4. *The UK contribution of European e-bike consumption is equal to their GDP ratios* Since the growth of e-bikes was found in Part I, the same assumption will be used for Part II.

5. *No e-bikes are resold, thrown away, or exported out of the country following the initial purchase* E-bikes and other personal belongings are usually retained throughout their effective life span [12]. The purchase of an e-bike should represent an equal increase in e-bike usage.

6. *All e-bikes are perfect substitutes for each other.* Market share data is not available for e-bikes. Consumers will first decide to purchase an e-bike, and assuming they want an e-bike, the cost will not be a barrier since there is a wide range of different bikes making them accessible to all members of the public.

7. *The social popularity of e-bikes can be approximated by Google Search Trends.* Google is the largest search engine in the world accessible by a large population and thus is an accurate approximation for the "coolness" factor of e-bikes [13]. As e-bikes become more popular they will show up on more social media and thus also increase the search popularity.

8. *Features that can be used to predict e-bike sales are significant to the causes for the increase in e-bike usage.* Since e-bike usage is defined by e-bike sales, any factor that is significant to the boom in e-bike sales will also be significant to the increase in e-bike usage.

## 4.3 Model Development

### 4.3.1 Variables

| Variable | Definition | Units |
|---|---|---|
| S | Number of e-bike sales in a given year | Thousands of e-bikes |
| E | People who "fairly" or "greatly" care about environment | Percentage |
| D | Disposable income per capita | British pounds |
| G | Average price of premium motor spirit per liter | British pence |
| P | Relative popularity of "E-Bike" on Google Trends | % |

### 4.3.2 Factor Identification

The factors with the largest impact on e-bike sales were determined to be the average yearly gas prices, disposable income per capita, percentage of people who care about the environment (all from the datasets provided [14][15][16]), and the popularity of e-bikes (measured using Google Trends). The popularity of e-bikes approximates the "coolness" factor of e-bikes. People's attitude towards e-bikes, in general, is often reflected in their searches, and this is the best way to quantify the "coolness" associated with owning an e-bike.

We did not include the battery price as a significant factor because there was not enough data to give accurate predictions of how it would change in the future [17].

## 4.4 Model Development

The importance of each of the aforementioned factors was determined by their impact on a random forest regressor used to predict electric bike sales. If a factor was able to be effectively utilized for bike sale prediction, it was thought to be significant to the growth of the bike sale market.

To begin, Python's SciKit Learn library was used to create a random forest regressor with the following classes: Care for the environment, popularity, disposable income, and gas prices. The target was the sales of electronic bikes. Both the features and target data were taken yearly from 2006 through 2021. Random forest regression was chosen for a few primary reasons.

To preface, random forest regression makes predictions through ensemble learning, a process that involves the outcomes of several decision trees that are then used to make one final prediction. Each decision tree randomly selects features to use, and because each tree is independent, their individual errors are remedied by only outputting the most optimal result after balancing out the extreme errors. Our regressor included 200 trees, which we found was most consistent across seeds and minimized regular mean squared error. Other models, such as SVM and KNN were attempted, but they yielded higher errors and were less effective.

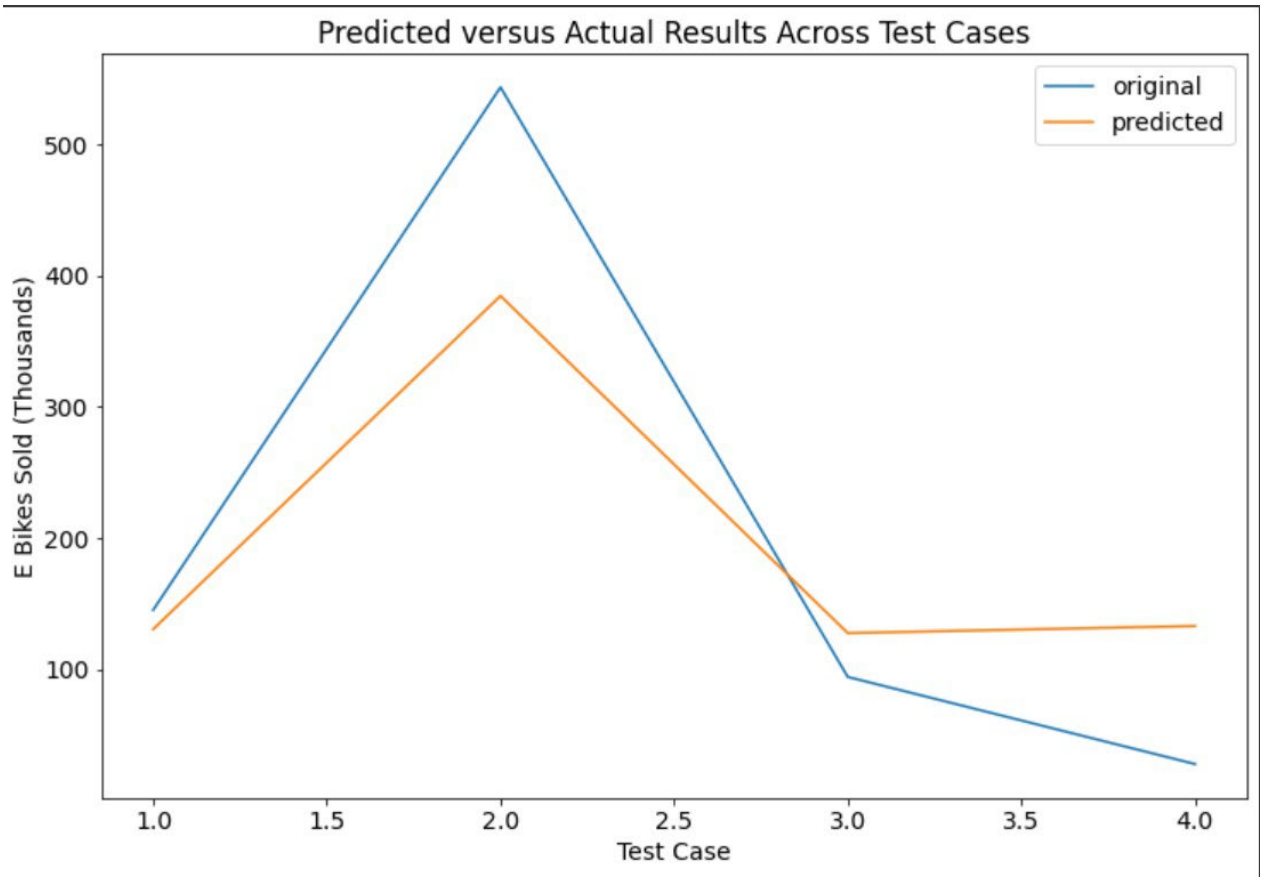
Then, using the regressor, we weighed the importance of each feature on the final prediction. Permutation feature importance was used to determine the significance of the features. The function that determines feature importance first finds the initial accuracy of the model using mean squared error. Next, it shuffles a column of selected data to corrupt the data of that feature. Then, it computes the error again and compares this with the original error. Columns with a higher change in error when corrupted are said to be more important than those with less change in error as they are more useful for making a prediction.

In addition to using this function, we measured the relative importance of each of these features by determining their effect on the root mean squared error. Each one of the features was dropped, and the model was then run on the remaining three. The features determined to be more important were those where the error increased the most when dropped. This is because, if the presence of that feature reduced the error significantly, it indicates that it was useful in making a prediction.

## 4.5 Results

The regressor was able to predict e-bike sales with a root mean squared error of about 73, which is acceptable for the range of inputs provided. This indicates that the given factors are relevant in the e-bike market. A graph of the predicted versus actual e-bike sales across test cases is shown below. The results of the permutation feature importance demonstrate that the most significant factors are ranked as follows: Popularity, disposable income, gas prices, and environmental concerns. The feature importance was, respectively: 0.772, 0.160, 0.036, and 0.033 (these do not sum to 1 due to rounding). Thus, we can confidently say that popularity and disposable income are the most important factors out of the four due to their

high percentages. It is important to consider that one aspect of the random forest regressor is variation across seeds. This comes about as a result of the random nature of the decision trees. These results showcase one seed of the regressor, although the number of trees selected somewhat accounts for seed variations. Furthermore, most seeds seem to follow the general trends shown in these results. They also comply with the model validation shown below.




The graph demonstrates the difference between the original and predicted data for e-bike sales (in thousands). These data points are set over several test cases (several runs of the random forest regressor) that happen in some time sequence only known to the program.

### 4.5.1 Model Validation

To validate and check the results of the model, we ran the random forest regression while removing each of the four variables. By removing the popularity factor, the error increased by 102.16; removing environmental concerns caused the error to increase by -15.5; removing disposable income increased the error by 25.54; removing gas prices increased the error by 23.76. By observing how errors changed with each change in the model, we were able to verify that popularity had the largest impact on sales, followed by disposable income, with gas prices and environmental concerns being the least important. The negative value for the environment suggests that it could be detrimental to include as the error is lower when it is excluded.

These changes supported our idea that disposable income and popularity are significant

reasons for e-bike growth. Removing these two had a much bigger effect on the error when fitting the regression model on e-bike sales. Comparatively, removing gas prices and environmental concerns had almost no effect on its error. As such, we concluded that the only two significant reasons for the growth in e-bike usage were the changes in UK disposable income per capita and the popularity of e-bikes on social media, which served as an approximation for the ”coolness” factor.

### 4.5.2 Sensitivity Analysis

The model is highly sensitive to parts of assumption 1. When removing either disposable income or popularity, the errors and the model changed significantly. However, when removing gas prices or environmental concerns, the model remained close to its original. The model is also reliant on the assumption that Google Trends is an accurate approximation of popularity, which may not always be true.

## 4.6 Strengths and Weaknesses

The strength of our model lies in its simplicity and applicability over time. The random forest regression means that one only needs the four variables in addition to total e-bike sales to determine the significance of each factor. Furthermore, the model will be applicable even as time goes on and more data is made available. We would not need to conduct more surveys to find out the motivation factors behind e-bike growth. Also, since feature importance was used, we can easily determine the relative significance of each factor, instead of merely whether a factor was significant to the growth or not. This way, we know how much more important popularity was compared to gas price changes, or how disposable income changes are much more significant than changes in environmental concerns when determining the growth of e-bike usage.

Our model fails to take into consideration factors other than the aforementioned 4. Many other factors should be taken into account, such as a person’s commute time as well as their health, among other things. Additionally, the model is based on regression instead of a survey of prospective buyers. However, privacy laws prevent such surveys from being publicly available, so it is impossible to determine the most significant factors prior to purchase. Also, we were limited to one seed in the random forest regression so that the results would not constantly change during each run.

There were also issues with the 4 variables that we decided to use. The popularity (”coolness” factor) was determined by looking at the relative number of Google searches of the term ”e-bike” over time. While this data should indicate the overall popularity of e-bikes, it does not directly quantify the ”coolness” factor that would prompt someone to buy an e-bike. This ”coolness” factor is very challenging to directly quantify without having information on distinct social media trends or the opinions of individual buyers. Also, Google Trends might be more reflective of the total purchases instead of ”coolness” since people often search for pictures and/or more information about goods prior to purchase. The proportion of ”people who care about the environment” is also very subjective, as it comes from surveys. Each

individual's definition of "fairly" or "greatly" caring about the environment will differ, so this variable is not perfectly objective.

Lastly, we were limited to using 100 trees in the random forest regressor. This was due to the fact that Google Colab was used in the processing of the model. Google Colab has resource limitations on RAM that slows down the runtime and processing of code.

# 5 Part III: Off the Chain

## 5.1 Restatement of the Problem

This problem tasked us to find and quantify the impact of increasing e-bike sales and e-bikes as alternatives to bikes and traditional vehicles. We decided to focus on pollution, traffic congestion, and health effects due to exercise routines.

## 5.2 Assumptions

1. *Each bike is only used by one person and one person will own no more than one bike.* E-bikes are not approved to be ridden by more than one person at a time. Each person will not need more than one e-bike as they all serve the same function.

2. *The reason people purchase e-bikes will determine whether it replaces a car, a bike, or neither. However, each bike cannot replace more than one bike or car since they are not shared among individuals.* People who purchase an e-bike due to popularity and "coolness" will likely continue to use their car and will instead replace their bike to seem cooler to society. People who purchase the e-bike due to environmental concerns or gas prices likely did so to replace their car, while people who purchased an e-bike due to changes in their disposable income will likely continue to use cars and bikes since they can afford to retain all of them.

3. *$CO_2$ is the only source of carbon emissions.* Carbon dioxide is the most prevalent greenhouse gas emitted by cars and is generally used as a measure of pollution by vehicles [18].

4. *Congestion changes are proportional to changes in total miles traveled by car throughout the car.* Traffic congestion is generally caused by an overabundance of cars on the road, so by reducing miles traveled by car, traffic congestion should also decrease in the same fashion.

5. *Health and wellness impacts can be approximated by the changes in kCal burnt.* The greatest concern to health in modern developed countries is obesity, which is combated by burning more calories throughout the day.

6. *The only buyers and users of e-bikes are people in urban areas of people with commutes short enough to replace a car with an e-bike.* Rural individuals tend not to have any incentive to purchase an e-bike since it will rarely be used. Even in the rare instances where they purchase one, it will likely be due to an abundance in disposable income, resulting in no net

change in bikes or cars and thus no net effect on emissions, congestion, or health.

7. *Carpooling is independent of trip length.* People carpool for short regular commutes and long road trips, so they should offset each other when compared to single-passenger trips.

8. *Individuals still travel the same length of average per year, per person.* The geography of a developed country, especially in urban settings, is relatively constant and so the trip length is unlikely to change within the next 5 years.

9. *COVID effects are outliers that merely stunted growth and are not going to be the start of a new pattern.* COVID effects have since recovered, and most people are back to similar commutes as before. The COVID-induced effects on travel have rebounded quickly enough to justify it simply being a temporary slowdown as opposed to the start of a new kind of travel [19][20].

## 5.3 Model Development

### 5.3.1 Variables

| Variable | Definition | Units |
|---|---|---|
| $p_E$ | Probability that enviro. concern will induce a random person to buy an e-bike | Percent |
| $p_G$ | Probability that gas prices will induce a random person to buy an e-bike | Percent |
| $p_P$ | Probability that popularity will induce a random person to buy an e-bike | Percent |
| $f_b$ | Forgone bike distance in a year due to replacing bikes with e-bikes | Miles |
| $f_c$ | Forgone car distance in a year due to replacing cars with e-bikes | Miles |

### 5.3.2 Monte Carlo Simulation

Per assumption 2, we predict that the number of bikes and cars swapped for e-bikes each year will depend only on the reason that someone bought their e-bike. Since we already found the relative importance of different factors in leading to a person buying an e-bike in the previous section, we will use the proportions generated by the AI as our value for p. However, since we do not expect these probabilities to be constant over time, we decided to induce randomness and use a Monte Carlo simulation to find the average number of bikes and cars replaced with e-bikes in a given year. To induce randomness, we re-ran the AI from the previous section 100 times, effectively simulating an independent random sample proportion of a population of these probabilities. We then calculated the means and standard deviations of the sample proportions, as shown in the table below:

| Factor | Sample Mean | Sample Standard Deviation |
|---|---|---|
| Popularity | 0.6824 | 0.1003 |
| Gas | 0.0681 | 0.0355 |
| Environment | 0.0367 | 0.0217 |
| Income | 0.2028 | 0.0722 |

From here, we assumed that the sampling distribution for each of these factors is normal. Although this is justified by the large counts condition for popularity and income, it isn't justified for gas and environment. However, since this was all just a means of introducing randomness in the probabilities, the normality of the probability distribution should have a negligible effect on the final model; so long as the probabilities randomly land on values with higher frequency density near the mean, the model should work.

With the assumption that these sample proportions are approximately normally distributed, we can randomly generate p for one run of the Monte Carlo simulation by generating a random decimal number between 0 and 1, then using an inverse normal function to work out the z-score that corresponds to this tail length. From here, a combination of the sample means and the sample standard deviations can be used to find the randomly generated values for p for each simulation.

Within a simulation, we calculated $f_c$ by adding the values $p_G$ and $p_E$. Then, we multiplied this number by the number of e-bikes sold from section 1. Finally, $f_c$ can be found by multiplying by the average distance traveled by a UK citizen per year in a car. Similarly, we calculated $f_b$ by multiplying $p_P$ by the number of e-bikes sold (popularity was the only factor that we identified would cause someone to replace their bike with an e-bike) with the total number of e-bikes sold (also from section 1). Using values found online that an average UK citizen drives a car for 7400 miles in a year [26] and a bike for 88 miles in a year [25], we converted these numbers to the total number of miles traveled by car and bike in a year in the UK. Thus, we found the number of forgone car or bike miles traveled in a year due to e-bikes ($f_c$ and $f_b$). By assumption 4, this value for cars was used as an indicator for traffic congestion in the UK. We used both the actual data from the past 10 years as well as the extrapolated data; the past data was used to make sure the Monte Carlo simulation itself was realistic and functional.

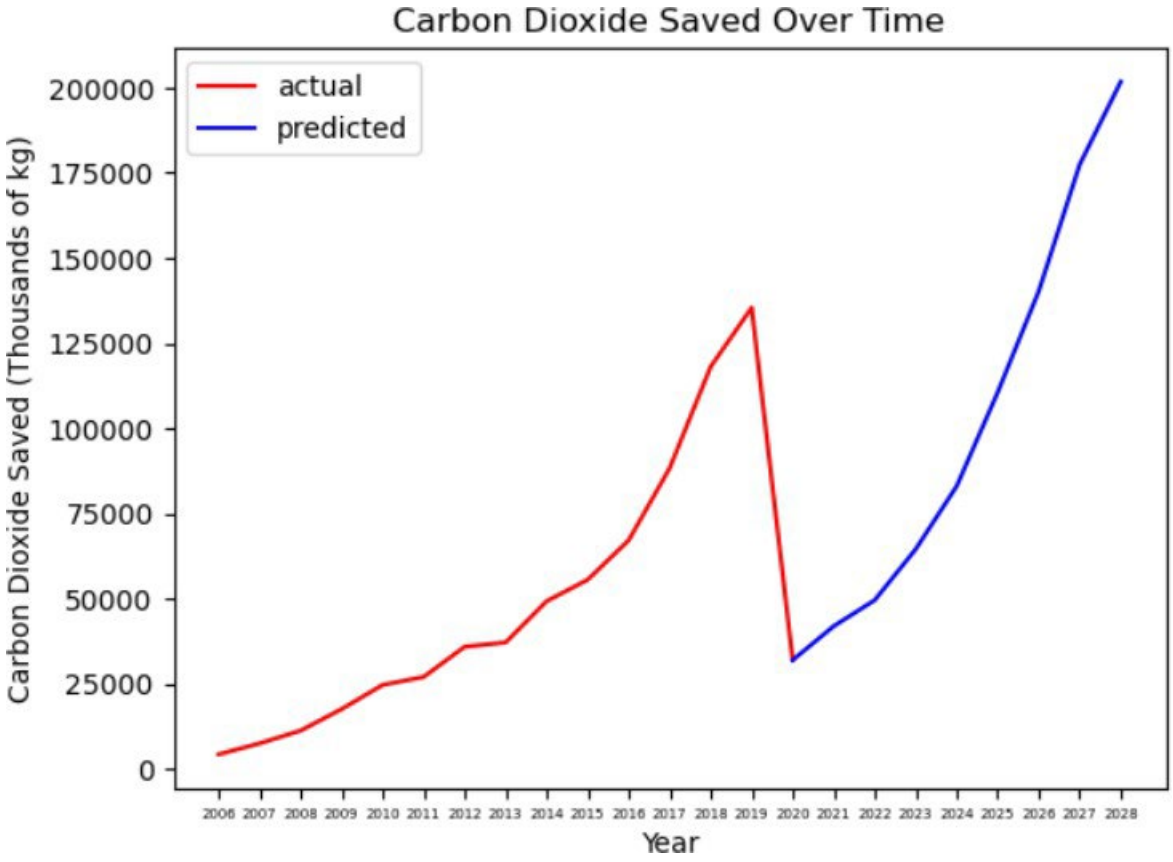
Knowing $f_c$ and $f_b$ in aggregate miles, we can then calculate the amount of $CO_2$ emissions saved and the extra number of kCal burned across the UK in a given year. To do this, we used online data to find that the average $CO_2$ emissions due to producing the electricity for an electric bike is 6.022 g of $CO_2$ per mile traveled [22][23], and 348 g of $CO_2$ per mile traveled for a car [24]. The saved carbon emissions are then given by $348f_c - 6.022f_c - 6.022f_b$, where $f_c$ and $f_b$ are in aggregate miles. We used a similar method to convert these two numbers into the aggregate number of kCal burnt, using numbers from the internet that biking burns 39.28 kCal per mile while e-biking burns 25.53 kCal per mile [27]. Since driving a car involves practically no pedaling, it was assumed that driving a car burns no calories.

After all of these values were calculated, we ran the Monte Carlo simulation 100 times for each year. The results from this simulation are summarized in the following section.

## 5.4 Results

The model predicted that the UK will replace 566,269 bikes and 90,402 cars in 2028. Using the other parts of the model, this resulted in 228,494 metric tons saved in carbon dioxide, 669 million miles less travel length, and thus traffic congestion, as well as 16,394 million

kCal for the entire population of the UK. For the five-year span 2024-2028, this would have resulted in total savings of 734,001 metric tons of carbon dioxide, 52,590 million kCal, and 2,148 million miles traveled saved for congestion for 290,413 cars replaced and 1,880,588 bikes replaced. These values were reasonable since the UK contributes approximately 478 million metric tons of carbon dioxide per year; thus, the e-bikes only contributed a 0.05% decrease in carbon dioxide emissions for the year 2028, which is reasonable.




The graph demonstrates how Carbon Dioxide (in thousands of kg) changes over time from the years 2006-2028.

### 5.4.1 Sensitivity Analysis

Since the means and standard deviations for p had the greatest impact on the model, we decided to explore our model's sensitivity to these values. We both doubled and halved the standard deviation of p to see how spread affects the three values. We also equalized all the means at 0.25 (Equal means) to see how the values would change if the population cared about all four factors equally, and ran the simulation where the mean for environmental concern was assigned a mean of 0.7 in order to simulate a world where everyone is extremely conscious about the environment (Enviro-Conscious). In the table below, only values for the year 2028 are shown, since this is sufficiently representative of the data as a whole. The values in the table demonstrate that our model is relatively insensitive to changes in the

standard deviation of p but highly sensitive to changes in the means. The extreme scenario where all the means are equal had significantly higher values than any of the changes to the standard deviation. Finally, the scenario where the world is more environmentally conscious has the highest values, suggesting that such a world would benefit the most from its reduction in carbon emissions and congestion as well as its increase in calories burned. All of these suggest that it is very important to increase environmental concerns in society in order to increase sustainability.

| Changed Factor | Carbon Emissions (metric tons) | Calories (millions of KCal) | Congestion (millions of miles) |
|---|---|---|---|
| $\sigma/2$ | 221183 | 15846 | 648 |
| $2\sigma$ | 239986 | 17235 | 703 |
| Equal means | 1065773 | 79308 | 3117 |
| Enviro-Conscious | 1047398 | 127351 | 4993 |
| Unchanged | 228494 | 16394 | 669 |

## 5.5 Strengths and Weaknesses

The strengths of our model include its applicability to a variety of communities as well as its accounting of variance in human preferences. Since the model uses likelihoods found from an AI-generated model, that model can also easily be trained on a new set of data for a new community. This would make finding new means and standard deviations quite easy and thus make the model highly adaptable. Furthermore, the Monte Carlo simulation accounts for random changes in human preferences, which makes it more realistic especially for long-term use.

However, the model is also heavily reliant on previous models and thus might struggle as the sales of e-bikes begin to level off. Also, due to the lack of time, the model only finds the number of bikes and cars replaced. From here, we had to use constants to calculate how much carbon dioxide emission, congestion, and health changes really occurred due to e-bike sales. We did not have time to model and extrapolate these values using real-life data, which would have made for a stronger model.

# 6 Conclusion

## 6.1 Further Studies

Our first model relies on the assumption that e-bike sales throughout the EU and the US are still in their early adoption growth phase. However, this is not always going to hold true, so further studies are required to find out when and where this growth will start to level off.

The second model uses a random forest regression with many layers of decision-making trees. Therefore, it is not easy to follow the algorithm that eventually arrived at our output, which

could be further studied to sharpen the predictor variables. Also, we could run surveys in the future to provide real-life data and verify the regression model's relative importance of the factors.

Although our Monte Carlo simulation takes into account random changes in purchase reasons, these reasons are not always going to stay constant. For example, geopolitical issues and new scientific discoveries will continue to shift the price of gas and environmental concerns, so we expect the relative importance to shift with time. Creating models to extrapolate these changes will be important in the future to accurately predict these changes and thus the changes in replaced vehicles/bikes. Also, as future technology is improved, commute times and emission constants will also vary, which will need to be studied and accounted for. The small percentage decrease in total carbon dioxide emissions in the UK in 2028 also shows that e-bikes alone cannot solve the problem of greenhouse gas emissions; other measures must be taken in the future to effectively decrease carbon dioxide emissions.

## 6.2 Summary

Using various models and simulations, we predicted the number of e-bikes that will be purchased in the next few years, the factors that will most significantly contribute to the purchasing of e-bikes, and how future e-bike purchases will impact other factors such as $CO_2$ emissions, overall health and fitness, and traffic congestion.

In order to predict the number of e-bike sales in 2025 and 2028, we used exponential regression based on data from the past 14 years to extrapolate data into the future. We found that the UK will purchase 2,555,919 e-bikes in 2025 and 5,271,485 in 2028, while the US will purchase 1,922,281 in 2025 and 3,969,391 in 2028. While these are relatively reasonable for these time frames, extrapolating any further is likely dangerous due to the eventual leveling-off of these values, which will likely begin around 2028, or possibly even sooner. If this data were to be used to predict any further than 2028, it would have to be refactored in the future when more information about the inflection point is known. We then used machine learning to analyze the correlation between yearly environmental concerns, disposable income, e-bike popularity, and gas prices as they all relate to the sales of e-bikes, both in the past and the projected future from the prior section. We found that popularity was the most important factor, then the disposable income, then gas prices, and then environmental concerns. Using the relative importance of these factors, we finished by assigning a factor's relative importance to the probability that a person's e-bike will either replace a car or a bike, and used these probabilities, combined with both the data from our first model and constants from the internet, found that, in the next 5 years, the UK will emit 734,001 less metric tons of carbon dioxide, will burn 52,590 more kCal across all its citizens, and will save 2,148 million miles of cars being on the road, thus reducing traffic congestion. It is significant to note that, while modifying the numbers for this final model, it was found that having a society that weighed environmental concern as the most important factor in buying an e-bike would lead to an increase in all of these factors, indicating that such a society would be highly beneficial to the well-being of the planet.

# 7 Appendix

## 7.1 References

[1] https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20170410-1

[2] https://www.bbc.com/news/business-44802666

[3] https://www.globenewswire.com/en/news-release/2022/09/28/2524061/0/en/Electric-Bike-Market-Size-to-Grow-Worth-USD-92-19-Billion-at-a-CAGR-of-12-6-by-2029-Fortune-Business-Insights.html

[4] https://www.unionsportcycle.com/fr/les-actualites/2019-04-09/observatoire-du-cycle-engouement-pour-le-velo-se-confirme

[5] https://www.cnbc.com/2020/12/23/covid-19-crippled-us-auto-sales-in-2020-but-it-could-have-been-worse.html

[6] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/ populationestimates/bulletins/annualmidyearpopulationestimates/mid2021

[7] https://carsalesbase.com/united-kingdom-car-sales-data/

[8] https://www.ons.gov.uk/economy/grossdomesticproductgdp

[9] https://ebike-mtb.com/en/innovations-for-2022-what-to-expect/

[10] https://www.kcl.ac.uk/news/british-and-us-public-more-likely-to-say-economic-costs-of-climate-change-will-be-greater-than-cost-of-measures-to-reduce-it

[11] https://blogs.shu.edu/thediplomaticenvoy/2021/02/11/how-social-media-influences-global-political-movement/

[12] https://easyebiking.com/how-fast-e-bikes-depreciate-how-fast-e-bikes-lose-value/

[13] https://www.frontiersin.org/articles/10.3389/frma.2021.670226/full

[14] https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/mwb7/ukea

[15] https://news.gallup.com/poll/391547/seven-year-stretch-elevated-environmental-concern.aspx

[16] https://www.gov.uk/government/statistical-data-sets/oil-and-petroleum-products-monthly-statistics

[17] https://www.researchgate.net/publication/284929881_The_energy-storage_frontier_Lithium-ion_batteries_and_beyond

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

[19] https://www.reutersevents.com/sustainability/rethinking-commute-work-post-covid-19

[20] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8651520/

[21] https://electrek.co/2022/01/26/electric-bicycles-are-now-outselling-electric-cars-and-plug-in-hybrids-combined-in-the-us/

[22] https://www.cyclevolta.com/understanding-e-bike-power-range-and-energy/

[23] https://www.eia.gov/tools/faqs/faq.php

[24] https://www.energy.gov/eere/vehicles/articles/fotw-1223-january-31-2022-average-carbon-dioxide-emissions-2021-model-year

[25] https://www.gov.uk/government/statistics/walking-and-cycling-statistics-england-2020/walking-and-cycling-statistics-england-2020

[26] https://www.nimblefins.co.uk/cheap-car-insurance/average-car-mileage-uk

[27] https://www.nytimes.com/2021/05/19/well/move/bikes-exercise-workouts.html

[28] https://www.ons.gov.uk/economy/environmentalaccounts/bulletins/ukenvironmentalaccounts/2022

[29] https://technation.io/report2020/

[30] https://www.politico.eu/article/france-britain-uk-tech-digital-emmanuel-macron-theresa-may-london-paris-week-google-facebook-europe/

## 7.2 Code for Problem 2

```
#import statements
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import sklearn
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.inspection import permutation_importance
from sklearn.metrics import mean_squared_error # for calculating the cost
function
import math
import matplotlib.pyplot as plt

#Stabilizing values
np.random.seed(3)

#Setting up data frame with names and features
names = ['E-Bikes Sold','Disposable Income','Gas Prices',
'Popularity','Environment']

df =
pd.read_excel('https://docs.google.com/spreadsheets/d/e/2PACX-1vRyQ1BoIXOd
Xgma-fkGOS0ehzC-fWV4L2y-wz3oVsqGbhJGf_xiOymIXD-59qUnkcwwFFHZXkLRaK8i/pub?o
utput=xlsx',names = names)
data = pd.DataFrame()
```

```
#X and y, predicting y of E-bike sales per year
y=df['E-Bikes Sold']
X = df.drop('E-Bikes Sold', 1)

#Split dataset into testing and training
X_train, X_test, y_train, y_test = train_test_split(X, y)

#Creating a random forest regressor to predict e bike sales
RandomForest = RandomForestRegressor(n_estimators = 100, oob_score=True)
RandomForest.fit(X_train, y_train)
predictions = RandomForest.predict(X_test)

#Prints Regular Mean Squared Error
```

```
print('RF RMSE:', math.sqrt(sklearn.metrics.mean_squared_error(y_test,
predictions)))

#Prints importance of each given feature as a percentage
print(RandomForest.feature_importances_)

#Plots graph of predicted and actual test values
plt.xlabel("Test Case")
plt.ylabel("E Bikes Sold (Thousands)")
plt.title("Predicted versus Actual Results Across Test Cases")
x_ax = np.array([1,2,3,4])
plt.plot(x_ax, y_test, label="original")
plt.plot(x_ax, predictions, label="predicted")
leg = plt.legend(loc='upper right')

#Drop certain values to measure their importance:
for i in range(4):
  X = X.drop(names[i+1],1)
  #Split dataset into testing and training
  X_train, X_test, y_train, y_test = train_test_split(X, y)

  #Creating a random forest regressor to predict e bike sales
  RandomForest = RandomForestRegressor(n_estimators = 100, oob_score=True)
  RandomForest.fit(X_train, y_train)
  predictions = RandomForest.predict(X_test)
```

```
  #Prints Regular Mean Squared Error
  print('RF RMSE Without', names[i+1],
math.sqrt(sklearn.metrics.mean_squared_error(y_test, predictions)))

  #Prints importance of each given feature as a percentage
  print('RMSW Without', names[i+1],RandomForest.feature_importances_)

  X = df.drop('E-Bikes Sold',1)
```

## 7.3 Code for Problem 3

```
from scipy.stats import norm

import numpy as np
import pandas as pd
import random
import math
import matplotlib.pyplot as plt

# Array for Graphing
co2 = []

# Defining constants

#grams of CO2 per mile
co2_per_mile_car = 348
co2_per_mile_ebike = 6.022

# kcal per mile
cal_per_mile_ebike = 25.53
cal_per_mile_bike = 39.28

# miles per year
total_miles_per_year_car = 7400
total_miles_per_year_bike = 88
# Simulate one year 100 times
def simulate(sold):
    sum_co2 = 0
    sum_cal = 0
    sum_con = 0
    sum_cars = 0
    sum_bikes = 0

    n = 100

    # Run one year 100 times
    for i in range(n):
        # Generating random probabilities for each factor considered in
Question 2
        randx = random.uniform(0, 1)
```

```
        randy = random.uniform(0, 1)
        randz = random.uniform(0, 1)
        randw = random.uniform(0, 1)

        z_x = norm.ppf(randx)
        z_y = norm.ppf(randy)
        z_z = norm.ppf(randz)
        z_w = norm.ppf(randw)

        actual_x = (z_x * 0.0722) + 0.2028
        actual_y = (z_y * 0.0355) + 0.0681
        actual_z = (z_z * 0.1003) + 0.6824
        actual_w = (z_w * 0.0217) + 0.0367

        # Calculating total e-bikes sold per year
        ebikes_sold = 0.16 * sold * 1000

        # Calculating the number of cars and bikes that are replaced by
e-bikes
        cars_saved = (actual_y + actual_w) * ebikes_sold
        bikes_saved = (actual_z) * ebikes_sold

        # Calculating the number of miles that would have been travelled
by a replaced car/bike
        total_miles_saved_car = 7400 * cars_saved
        total_miles_lost_bike = 88 * bikes_saved
        # Calculating the CO2, total KCal burned, and Traffic Congestion
that are forgone
        co2_saved = total_miles_saved_car * co2_per_mile_car -
total_miles_lost_bike * co2_per_mile_ebike - total_miles_saved_car *
co2_per_mile_ebike
        total_cal_burned = total_miles_lost_bike * cal_per_mile_ebike -
total_miles_lost_bike * cal_per_mile_bike + total_miles_saved_car *
cal_per_mile_ebike
        congestion_saved = total_miles_per_year_car * cars_saved

        # Averaging values
        sum_co2 += co2_saved / (1000000)
        sum_cal += total_cal_burned / (1000000)
```

```
        sum_con += congestion_saved / (1000000)
        sum_cars += cars_saved
        sum_bikes += bikes_saved


    print("Carbon Dioxide Saved in Millions: ", sum_co2/n)
    print("Total Calories Burned in Millions: ", sum_cal/n)
    print("Congestion Saved in Millions: ", sum_con/n)
    print("Cars Replaced: ", sum_cars/n)
    print("Bikes Replaced: ", sum_bikes/n)

    co2.append(sum_co2/n)


# Simulation runs for 23 years starting in 2006 to 2028
for i in range(23):
    print(i + 2006)
```

```
    # Simulate with e-bike sales values given in the data
    if(i == 0): simulate(98)
    if(i == 1): simulate(173)
    if(i == 2): simulate(279)
    if(i == 3): simulate(422)
    if(i == 4): simulate(588)
    if(i == 5): simulate(716)
    if(i == 6): simulate(854)
    if(i == 7): simulate(907)
    if(i == 8): simulate(1139)
    if(i == 9): simulate(1364)
    if(i == 10): simulate(1637)
    if(i == 11): simulate(2074)
    if(i == 12): simulate(2767)
    if(i == 13): simulate(3397)

    # Simulate with e-bike sales values from the equation calculated in
Part 1
    if (i > 13):
        x = i + 1
        simulate(0.16 * 128.092 * math.e**(0.2413*x))
```

```
# Years for the actual data
year_1 = ['2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013',
        '2014', '2015', '2016', '2017', '2018', '2019', '2020']

# Years for the predicted data
year_2 = ['2020', '2021','2022', '2023', '2024', '2025', '2026', '2027',
'2028']

# Graphing the CO2 data over time
plt.figure(0)
plt.plot(year_1, co2[0:15], "-r", label="actual")
plt.plot(year_2, co2[14:], "-b", label="predicted")

# Labelling the graph
plt.legend(loc="upper left")
plt.xlabel('Year')
plt.ylabel('Carbon Dioxide Saved (Thousands of kg)')
plt.title('Carbon Dioxide Saved Over Time')

plt.tick_params(axis='x', which='major', labelsize=5)

plt.show()
```