Mary Grace Albright
Will Przedpelski
Charlie Reiney
CSC-272
April 23, 2021

Going Green: An Analysis of Student Printing Data

Introduction

What does it take for a college campus to become carbon neutral? The simple answer to that question is "balance carbon emissions with carbon removal so that the school can achieve net-zero carbon emissions." However, the reality of that answer is that it takes a great deal of change in countless areas to even come close to carbon neutrality. Reaching carbon neutral status has been an ambitious goal for Furman in recent years, with its sights set on 2026. Our team set out to find out how we could contribute to this goal, even in a seemingly insignificant way. We discovered that, even in our modern, computerized world, printing makes up a substantial portion of the university's carbon emissions. This project attempts to find a way to reduce the school's carbon emissions by learning more about our fellow students' printing habits.

Dataset Description

Our dataset was compiled from 10 years worth of print logs from Furman University. The original print logs contain an instance for each time a student used a printer anywhere on campus. Naturally, this was a massive amount of data to sort through and the process by which we accomplished this is detailed in the Data Preparation section. The dataset we compiled contained 11 attributes: AssignedID, Gender, Class, Semester, Major1, Major2, Division or type of major (i.e. NS for natural sciences), TotalPages printed, [number of print] Jobs, AveragePages per print job, and total CO_2 produced (which was directly calculated from TotalPages). In total the dataset contains 26,952 instances. Each instance represents a single student for a given semester. In other words, one row shows a student's information and their printing totals for one semester. Table 1 below includes a description of each attribute.

Attribute	**Description**	**Data Type**
AssignedID	Unique ID given to each student (separate from Furman ID) so that their identity was kept anonymous	Nominal
Gender	Two values: 1. M (male) 2. F (female)	Nominal
Division	Describes type of major. Six values:	Nominal

	1. FA (fine arts) 2. HU (humanities) 3. NS (natural science) 4. SS (social science) 5. UNDC (undecided) 6. ICP (individualized curriculum program)	
Class	Describes year of undergraduate study. Four values: 1. FR (freshman) 2. SO (sophomore) 3. JR (junior) 4. SR (senior)	Nominal
TotalPages	The total number of pages printed by a single student during one semester.	Numeric
Jobs	The total number of jobs sent to the printer by a single student during one semester.	Numeric
AvgPages	The average number of pages printed by a single student per print job. This was calculated by dividing TotalPages by Jobs.	Numeric
Major1	The first major of a given student. Forty-two values: 1. Sustainability 2. Physics 3. Computer science 4. Psychology 5. Chemistry 6. Business Administration 7. History 8. Undecided 9. Health & Exercise Science 10. Theatre Arts 11. Biology 12. English 13. Economics 14. French 15. Spanish 16. Asian Studies 17. Neuroscience 18. Politics and Intl Affairs 19. Education 20. Sociology	Nominal

	21. Religion 22. Art 23. Information Technology 24. Accounting 25. Communication Studies 26. Mathematics 27. Individualized Curriculum Pgrm 28. Earth & Environmental Sciences 29. Philosophy 30. Music 31. Pre-Engineering 32. Classics 33. Mathematics-Economics 34. German 35. Urban Studies 36. Japanese Studies 37. Latin 38. Anthropology 39. Applied Mathematics 40. Art History 41. Chinese Studies 42. Public Health	
Major2	The second major of a given student. If the value was missing, this meant that the student did not have a second major. Forty-three values: did not include the following from Major1: 1. Undecided 2. Pre-Engineering but contained the following values not included in Major1: 1. Health Journalism 2. Film & Median Studies 3. Digital Media	Nominal
Semester	The semester to which the data was attributed. Twelve values: 1. Fall 2011 2. Spring 2012 3. Fall 2012 4. Spring 2013 5. Fall 2014 6. Spring 2015 7. Fall 2015 8. Spring 2016	Nominal

	9. Fall 2016 10. Spring 2017 11. Fall 2017 12. Spring 2018	
CO_2	The number of pounds of Carbon Dioxide released into the air by a single page of paper. This was calculated by multiplying TotalPages times 0.0092, which is the average number of pounds released per page. *Note:* This number is approximation, as we were not able to calculate the specific carbon footprint of printing one page on Furman's campus based on the type of printer, printing software, and the type of paper that is used.	Numeric

Table 1: A list and description of all attributes in the dataset

Data Preparation

This dataset required a great deal of pre-processing to prepare it for analysis. The data that we received came from Dr. Dripps in the Earth, Environmental, and Sustainability Sciences Department at Furman. Some of it was already preprocessed, but there was much still left to be done. Each semester had to be processed individually then combined before we could overlay it with student demographic information. The initial print logs were large files that contained an instance for each time a student printed at Furman and how many pages were printed. To extract the information we needed, we used the programming software R to create a new dataset for each semester that only contained one instance per student. We created a function that counted how many times a student's ID appeared (which turned into the Jobs attribute) and how many total pages they printed (TotalPages). To determine the average number of pages printed per print job (AvgPages), we divided the total number of pages by the number of jobs. We then added the Semester attribute to each of our new datasets before merging them vertically in R. Before this was done, however, we had to change the attribute names of the previously made semester datasets, since they were not all spelled exactly the same.

Once this new dataset was created, we had to overlay this data with the demographic dataset that we were given for each of the students. Some of this was done in R again, while some was done in Excel. However, this process was complicated, as we only had demographic data for a few of the years, which meant that we had to backcast the data so it matched for students of the previous years. Because of this, we were not able to use the print logs for the Fall 2013 and Spring 2014 semesters, as we were not able to confidently match enough of the printing data with the student demographics from a future year. Each demographic dataset contained the students' assigned IDs along with their gender, class, first major, second major (if applicable), and division. The class was easy to backcast, but the majors were difficult, as we could not confidently say that a student retained the same major from freshman to senior year. Additionally, each of these datasets spelled everything differently. One spelled out the class (ex. freshman) while another

used an abbreviation (ex. FR). Similarly, most majors were spelled differently between the different datasets and occasionally the major names changed from one year to the next. Thus, a significant amount of time was spent creating a uniform dataset. These changes to spelling were made in Excel. We also chose to make changes such as combining all of the different types of music majors into one "Music" major, since they were similar and did not have very high counts of individuals for each of the specialized majors.

We also knew that it would potentially be an issue for a model to interpret Major1 and Major2 as two of the same attributes for a single student. For example, some CS-Math double majors listed Major1 as Computer Science and Major2 as Mathematics, and others were listed in reverse. So, in Excel, we created new columns for each individual major that assigned a binary value (1 or 0) to each student based on whether that was their major or not. For example, if a student's first or second major was Applied Mathematics, they would be assigned a 1 for that column and, if not, they were assigned a 0.

Real-world data is messy.

Data Analysis

A description of the analysis we performed, which includes a description of the algorithms that we used.

Association Learning

Association learning (sometimes called market basket analysis) is a kind of data analysis technique that tries to find relationships between different attributes within a dataset. For example, someone who buys a candle might also buy a lighter, which could be a useful piece of information. In the case of our research, it would be useful to know if people of certain majors and genders tend to print more, and association learning creates rules that could determine just that.

Linear Model

Linear regression models work by finding the *line of best fit* through a set of data points. For instance, if we have a number of data points arranged on an *xy* axis, we can try to find a linear relationship between the two attributes represented by *x* and *y*. We end up with a formula that can take into account many different variables at once. Using this formula, we can compare a predicted class value with a real value within a test set to determine the level of accuracy of the model. This is measured on a scale from 0 to 1 by a *correlation coefficient*.

Neural Networks

Neural Networks, or multilayer perceptrons, involve vectors and regression, but the weights applied to those vectors can change over time. They utilize multiple layers of calculations with many weight matrices that can be adjusted using a technique called backpropagation. They are

notoriously difficult to implement but are very powerful, with many high-performance artificial intelligence systems using multilayer neural nets.

Nearest Neighbor

Nearest Neighbor is a classification technique that takes every instance of a dataset and turns it into a point in space, and when the algorithm encounters a new data point, it tries to find the next-closest point to it. This can be extended to a *k-Nearest Neighbor* algorithm, in which our new data point is compared with more than one (*k*) next-closest data points and acquires the class value of the majority.

Results

Association Learning

In order to create association rules, we had to discretize the class attribute of CO_2. Our data was skewed, so we discretized our values using bins of equal frequency. We chose to use five bins, which were: (-inf - 0.9798), (0.9798 - 2.3322), (2.3322 - 4.3286), (4.3286 - 7.521), (7.521 - inf). When we used a support of 0.01 and confidence of 0.80, we were able to produce the following association rule:

Major1 = Psychology, CO2 = (7.521 - inf) 360 → Gender = F 311

In other words, Psychology majors who printed a lot (7.521 pounds of CO_2 is equivalent to about 818 pages) tended to be female. In fact, they were female 311/360 times, or around 86% of the time. This was the only rule in this set that we were interested in, since it involved the class attribute. We then lowered the support and confidence to 0.008 and 0.75, respectively. This gave us the following rules:

Major1 = Psychology, CO2= (4.3286 - 7.521) 287 → Gender = F 252
Major1 = Communication Studies, CO2 = (7.521 - inf) 307 → Gender = F 262
Major1 = Biology, CO2 = (7.521 - inf) 399 → Gender = F 332
Major1 = Business Administration, CO2 = (-inf - 0.9798) 556 → Gender = M 616

However, these rules might tell us more about students' gender than they do about how much certain majors print. For example, it might simply be the case that Business Administration students tend to be male, while Biology, Psychology, and Communication Studies students tend to be female. A little investigation shows that this holds true. Female Business Administration students only make up 532/1572 of the majors, or around 34%. Similarly, Communication Studies students are 73% (962/1319) female, Biology majors are 72% (754/1051) female, and Psychology students are 86% (795/920) female. It appears that the Apriori algorithm just happened to find subsets of the CO_2 attribute such that there was a slightly higher-than-average concentration of female students so that it would push the confidence over 75%. On the surface, these association rules seem like they may be helpful to understanding the printing patterns of

students, but really, it looks like they are just exploiting patterns in the data unrelated to the printing information.

Linear Regression Model in R

In addition to our analysis in Weka, we wanted to see if any of the attributes, or (in statistics terms) predictor variables, were significantly impacting the class attribute, or response variable. To do this, we created a linear model in R using the lm() function. We ran four different models with the following inputs:

```
lm.1 <- lm(CO2 ~ Major1, data = printing)
summary(lm.1)

lm.2 <- lm(CO2 ~ Major1 + Gender + Class, data = printing)
summary(lm.2)

lm.3 <- lm(CO2 ~ Gender + Class, data = printing)
summary(lm.3)

lm.4 <- lm(CO2 ~ Division + Gender + Class, data = printing)
summary(lm.4)

lm.5 <- lm(TotalPages ~ Major1 + Gender + Class, data = printing)
summary(lm.5)
```

Linear Model 1 (lm.1), only used Major1 to predict CO_2, whereas Model 2 (lm.2) used Major1, Gender, and Class, model 3 used Gender and Class, and model 4 used Division, Gender, and Class. Model 5 was almost the same as model 2, except we predicted for TotalPages instead of CO_2. In order to determine which model was best, we determined the AIC values for each one. Typically, the model with the lowest AIC value is the "best" model. Using the AIC() function, we were able to determine that model 2 had the lowest AIC and was therefore the model whose output we would look at.

```
> AIC(lm.1, lm.2, lm.3, lm.4, lm.5)
     df      AIC
lm.1 44 183890.4
lm.2 49 182816.4
lm.3  7 183169.5
lm.4 14 183132.6
lm.5 49 435548.1
```

The output for this model is listed below in Table 2:

	CO2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.13	-11.13 – 17.39	0.667
Major1 [Accounting]	-3.42	-5.72 – -1.12	**0.004**
Major1 [Anthropology]	-5.51	-8.54 – -2.48	**<0.001**
Major1 [Applied Mathematics]	-5.09	-7.67 – -2.52	**<0.001**
Major1 [Art]	-5.67	-8.02 – -3.32	**<0.001**
Major1 [Art History]	-5.31	-8.05 – -2.56	**<0.001**
Major1 [Asian Studies]	-3.11	-5.58 – -0.65	**0.013**
Major1 [Biology]	-2.81	-5.07 – -0.56	**0.015**
Major1 [Business Administration]	-3.95	-6.20 – -1.70	**0.001**
Major1 [Chemistry]	-1.50	-3.78 – 0.79	0.199
Major1 [Chinese Studies]	-4.71	-8.05 – -1.37	**0.006**
Major1 [Classics]	-4.35	-7.29 – -1.42	**0.004**
Major1 [Communication Studies]	-3.49	-5.75 – -1.24	**0.002**
Major1 [Computer Science]	-5.06	-7.39 – -2.73	**<0.001**
Major1 [Earth & Environmental Sciences]	-4.46	-6.81 – -2.12	**<0.001**
Major1 [Economics]	-2.72	-5.00 – -0.43	**0.020**
Major1 [Education]	-2.87	-5.15 – -0.58	**0.014**
Major1 [English]	-2.69	-4.98 – -0.39	**0.022**
Major1 [French]	-3.30	-5.79 – -0.81	**0.009**
Major1 [German]	-3.38	-6.08 – -0.69	**0.014**
Major1 [Health & Exercise Science]	-2.41	-4.66 – -0.17	**0.035**
Major1 [History]	-2.54	-4.82 – -0.26	**0.029**
Major1 [Individualized Curriculum Pgrm]	-3.44	-6.53 – -0.34	**0.029**

Major1 [Information Technology]	-4.81	-7.36 – -2.26	**<0.001**
Major1 [Japanese Studies]	-4.14	-6.94 – -1.34	**0.004**
Major1 [Latin]	-5.02	-10.47 – 0.44	0.071
Major1 [Mathematics]	-4.85	-7.25 – -2.45	**<0.001**
Major1 [Mathematics-Economics]	-4.13	-6.54 – -1.72	**0.001**
Major1 [Music]	-2.70	-4.97 – -0.43	**0.020**
Major1 [Neuroscience]	-1.91	-4.23 – 0.41	0.106
Major1 [Philosophy]	-2.74	-5.08 – -0.40	**0.022**
Major1 [Physics]	-5.35	-7.69 – -3.01	**<0.001**
Major1 [Politics and Intl Affairs]	-1.81	-4.06 – 0.44	0.115
Major1 [Pre-Engineering]	-5.49	-8.43 – -2.55	**<0.001**
Major1 [Psychology]	-2.51	-4.77 – -0.24	**0.030**
Major1 [Public Health]	-3.33	-5.78 – -0.88	**0.008**
Major1 [Religion]	-2.81	-5.15 – -0.48	**0.018**
Major1 [Sociology]	-2.01	-4.32 – 0.30	0.088
Major1 [Spanish]	-3.89	-6.22 – -1.57	**0.001**
Major1 [Sustainability]	-3.76	-6.08 – -1.45	**0.001**
Major1 [Theatre Arts]	-4.75	-7.23 – -2.27	**<0.001**
Major1 [Undecided]	-2.62	-4.87 – -0.38	**0.022**
Major1 [Urban Studies]	-3.19	-6.74 – 0.35	0.078
Gender [F]	3.44	-10.64 – 17.53	0.632
Gender [M]	1.81	-12.27 – 15.89	0.801
Class [JR]	2.36	2.09 – 2.63	**<0.001**
Class [SO]	1.49	1.23 – 1.74	**<0.001**
Class [SR]	3.86	3.60 – 4.13	**<0.001**
Observations	26952		
R^2 / R^2 adjusted	0.062 / 0.061		

Table 2: Results of linear model in R

In Table 2, bolded p-values indicates that the variable is significantly impacting the class attribute (which in this case is CO_2). Here, we can see that not all attributes are significant, indicating that there is likely some relationship between our predictor attributes and the class attribute. However, we then took our analysis to Weka to determine how predictive such a linear model is when presented with new data.

Holdout Method Implementation

Before performing classification learning on our dataset, we had to create separate training and testing files. To do so, we began by opening the full dataset in Weka. Then, we applied the *Randomize* filter (*weka.filters.Unsupervised.Instance.Randomize*) to make sure the data was not structured in any way that might skew the results of our classification. We did this four times just to be sure (even though we probably didn't need to). After that, we used the *RemoveRange* filter to separate into training and test sets (*weka.filters.Unsupervised.Instance.RemoveRange*). Since our original dataset had approximately 27,000 instances, we decided to take the first 9,000 instances and put them into a test set file (see attachments, testSet.csv [original holdout test set] or testSetSimplified.arff [with only a subset of attributes for testing]), and the remaining approximately 18,000 instances were saved as a training set (trainingSet.csv or trainingSetSimplified.arff). Since we cannot remove attributes or modify the test set after we have separated them, we ended up with multiple training and test sets. Also, interestingly, Weka would not recognize a test set that was in .csv format, and required us to save it as a file in .arff format.

Classification -- Linear Regression

In Weka, we first followed a similar model to the one that was used to create the output in Table 2. To do so, we loaded our cleaned dataset into Weka as a CSV file then removed all attributes but Gender, Class, Major1, and CO_2 (class). Under Classify, we chose the linear regression function, which was found in weka.classifiers.functions.LinearRegression. We accepted the default settings and applied the holdout method using the test set of 33%. The full model can be found in the Appendix section as Linear Regression Model 1. The results of this model were as follows:

=== Summary ===

Correlation coefficient	0.2282
Mean absolute error	3.3598
Root mean squared error	7.797
Relative absolute error	91.5132 %
Root relative squared error	97.3604 %
Total Number of Instances	17952

Unfortunately, this means that the model we created in Weka would not be predictive at all, since there was only a correlation coefficient of 0.2282. In an attempt to increase the predictive ability

of this model, we then ran the same model again, but included the *Semester* attribute in the linear regression. This left us with the following results:

=== Summary ===

Correlation coefficient	0.3068
Mean absolute error	3.182
Root mean squared error	7.622
Relative absolute error	86.6705 %
Root relative squared error	95.1758 %
Total Number of Instances	17952

By adding an attribute, we were able to increase our accuracy to 0.3068, which is still not very good (significantly worse than a coin flip). But we still wanted to increase this value, so we decided to take into account the fact that some students had two majors. So, we loaded the CSV file into Weka that contained the binary numbers for each major to create a model off of a different representation of the same data. We used the same linear regression function with all the same attributes, but instead of Major1, we used the binary representation (each major having its own column). This allowed us to include Major2, since now a student could have a value of 1 for more than one majors. The results we received for this new model were as follows:

=== Summary ===

Correlation coefficient	0.3093
Mean absolute error	3.1729
Root mean squared error	7.6157
Relative absolute error	86.4248 %
Root relative squared error	95.0966 %
Total Number of Instances	17952

It is important to note that the binary dataset contains a significant amount more data than the original dataset. Weka consistently crashed when trying to produce models based on this set. We also tried to experiment with some of the parameters, including the method for attribute selection with regard to the regression, Weka crashed when we tried to change to a greedy method. Ultimately we were able to increase our correlation coefficient to 0.3093, but this was significantly below what we hoped to get out of performing regression. The full model for this regression can also be found in the Appendix section as Linear Regression Model 2.

Neural Network

After no luck with the linear regression model, we decided to play around with a multilayer perceptron technique. We used the same attributes from the last linear model (Gender, Class,

Semester, Major [Binary] and CO_2), but quickly found that this would be worse than the regression tried previously (see below). We got a correlation coefficient of only 0.2256, which was worse than our worse regression model.

=== Summary ===

Correlation coefficient	0.2256
Mean absolute error	3.4091
Root mean squared error	9.5806
Relative absolute error	92.0083 %
Root relative squared error	97.9831 %
Total Number of Instances	6104

IBk (Nearest Neighbor(s))

As a last-ditch effort to try and save our correlation coefficient, we implemented the Nearest Neighbor algorithm in Weka (IBk). Initially, we started with all attributes, though we later realized that some attributes were simply duplicates of others. With the duplicate attributes, we had achieved an accuracy of around 97%, but upon removing those not-so-independent columns of the dataset, we ended up with accuracy ratings that actually significantly better than either the linear regression, association, or multilayer perceptron attempts, with a correlation coefficient of 0.5138. When we experimented with the value of *k*, the correlation coefficient decreased to 0.4592, so we decided to stick with *k=1*.

=== Summary ===

Correlation coefficient	0.5138
Mean absolute error	2.487
Root mean squared error	6.8704
Relative absolute error	67.7398 %
Root relative squared error	85.7903 %
Total Number of Instances	17952

While this is a significant improvement over the previous models, it still is not much better than a coin flip, and tells us very little about what attributes are most predictive (since we don't get any trees or rules that could help us in solving the real-world problem of carbon emissions). Regardless of the model we used, it seems that the attributes may not be predictive of printing frequency.

Discussion & Data Visualizations

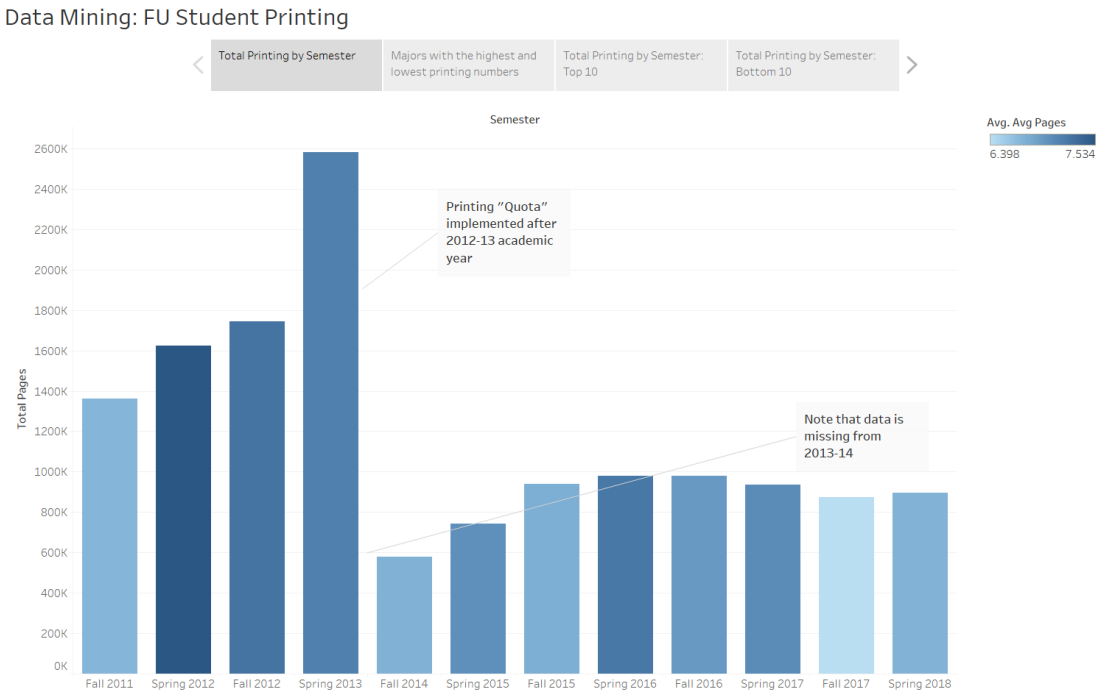

Figure 1: Total number of pages printed by Furman students by semester

Figure 1 shows the printing totals from each semester. Each bar represents one semester and the height of the bar represents the total number of pages printed by all students. The hue indicates how many pages an average student printed per print job, so a darker color suggests that students were printing a higher number of pages per print job. One of the most intriguing things that we learned from this chart is that the implementation of a printing "quota" (or fake-financial disincentive) was visibly effective as there is a steep drop-off after the quota was put in place. Additionally, we can see that the bar for Spring 2013 is the tallest by a significant margin. This could suggest that analysis of Spring 2013 printing data could have caused the print quota to come into effect, but further digging would be required to verify this.

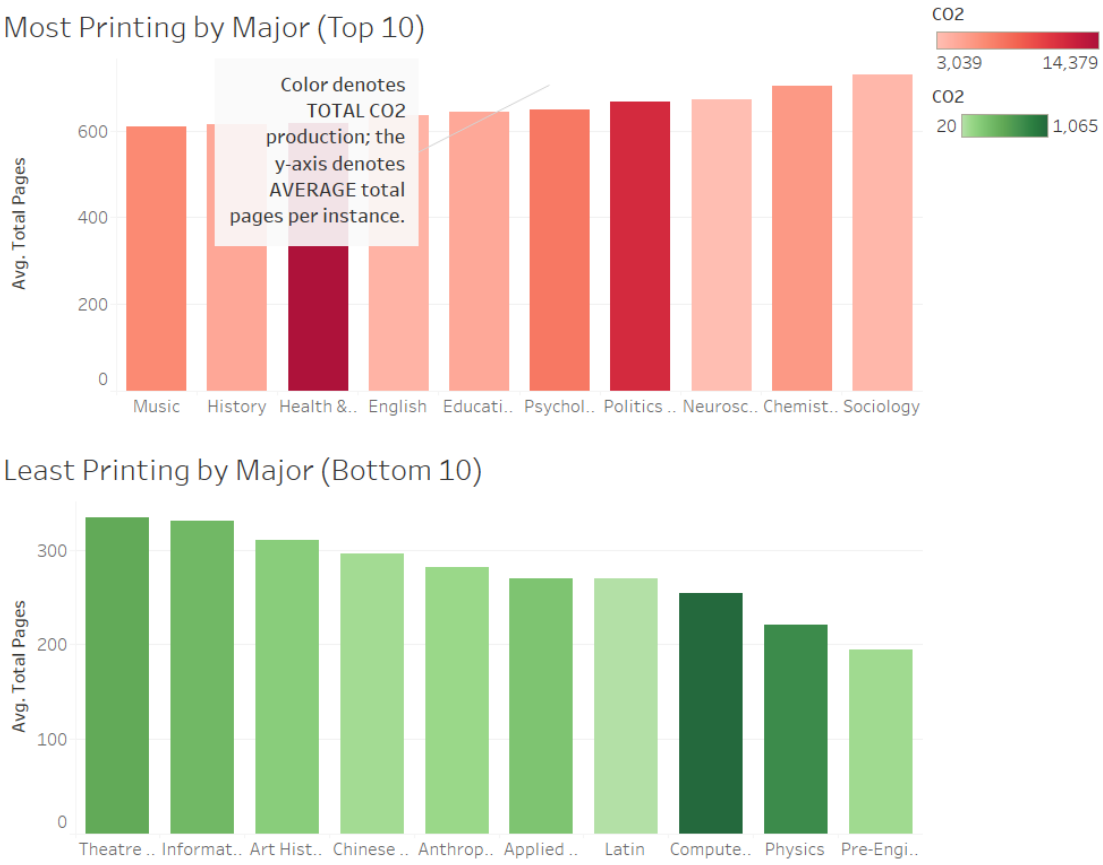

Figure 2: The average number of pages printed per instance for the top and bottom ten majors

Figure 2 shows the top and bottom 10 majors by average total printing numbers. We decided to sort by average total pages printed by each student to account for the fact that some majors have far more students than others. However, the color indicates the total sum of CO_2 produced, hence why the more popular majors tend to have darker bars. One of the interesting things we took away from these charts is that 5 of the bottom 10 majors are math/technology related (IT, Applied Math, CS, Physics, and Pre-Engineering). While we cannot definitively determine the reason for this phenomena, we hypothesized that it is due to a higher percentage of assignments being on a computer as opposed to on paper.

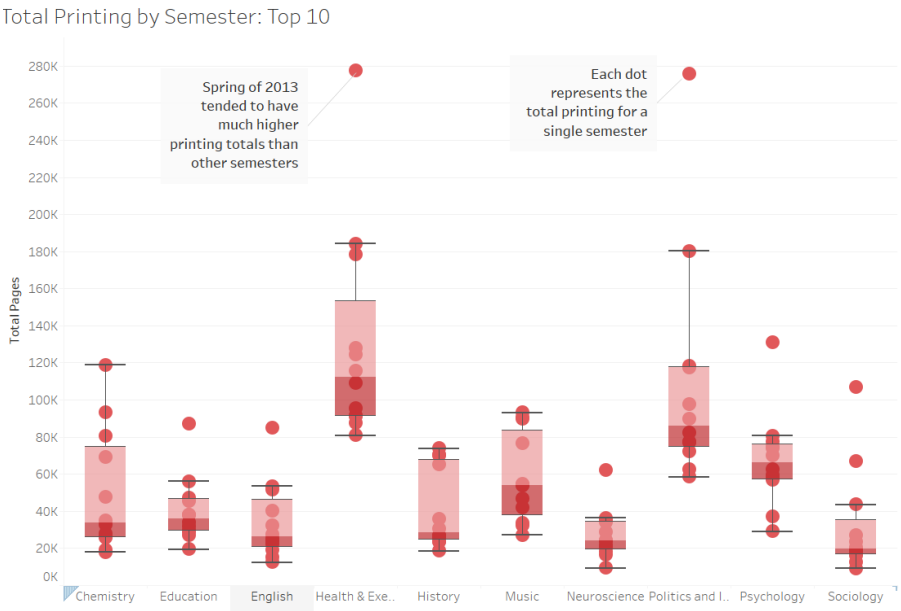

Figure 3: Total number of pages printed per semester for the 10 majors with the highest printing averages

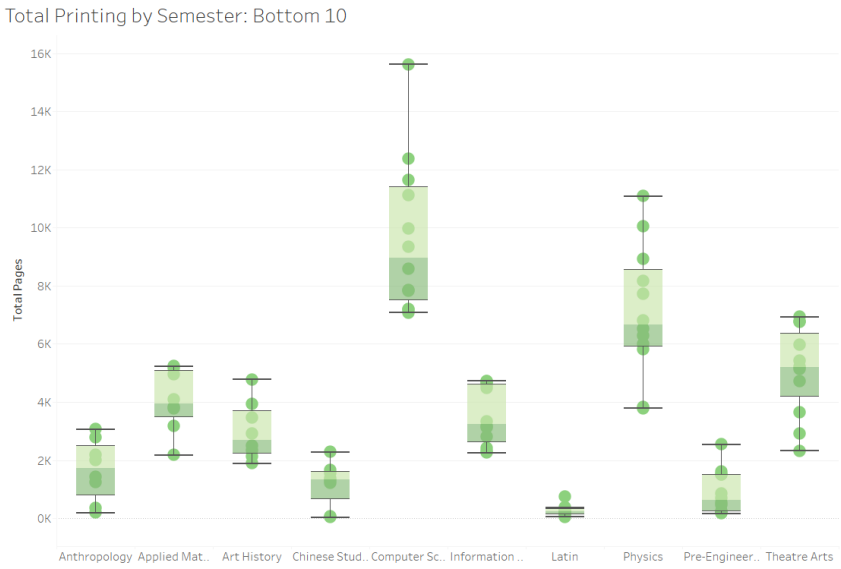

Figure 4: Total number of pages printed per semester for the 10 majors with the lowest printing averages

Figure 3 and Figure 4 combine the information in Figure 1 and Figure 2 in a box and whisker plot format. Each of the dots represents the total pages printed in a single semester. The boxes and tails represent the four quartiles of the data, showing how the printing totals were distributed over time. We chose to include these charts to visualize outliers and explore trends in distribution between majors. As previously predicted, we were able to clearly see that Spring 2013 was an outlier and contained much higher printing numbers than any other semester. Additionally, we noticed that the more popular majors (Health Sciences, Politics, CS) had a more spread-out distribution than other majors. We are not sure what contributed to this trend, but we hypothesized that the Spring 2013 instances could have something to do with it as the majors with spread distribution also tend to have particularly high totals for Spring 2013.

Based on the results from the machine learning algorithms, it seems unlikely that the attributes collected along with the printing data are particularly predictive. Although it would be nice to be able to estimate the number of pages a Classics student might print in a given semester, we were unable to improve the accuracy of any of our models to do so above 51% accuracy, with most of them falling below 35%, despite varying parameters and selecting different attributes. If this holds to be true, then it may be the case that there is no hidden knowledge to be gained from this data without expanding beyond those few attributes such as class, major, gender, and semester. However, we can gather some valuable information just from basic analysis of this data. For example, Politics & International Affairs and Health Sciences majors tend to print a whole lot more than Applied Math majors, and no machine learning is required to figure that out. Hence, we can apply this knowledge to have an effect on reducing the carbon footprint by specifically targeting the departments with students who print most on average.

Conclusions

In this project, we prepared our sizeable dataset for analysis by overlaying multiple immense datasets (with hundreds of thousands of instances) with demographic data, we experimented with four different kinds of machine learning models (Linear Regression in R and Weka, Multilayer Perceptron, Nearest Neighbor, and Association Learning), we implemented the holdout method to separate into training and test data, we analyzed the effectiveness of the models we used, and we created aesthetic visualizations that help us realize what we can learn from this difficult dataset. A major takeaway from this project is that sometimes the most essential information is on the surface, but you will never know for certain if hidden knowledge exists within your data unless you dig in and start experimenting.

Appendix

Linear Regression Model 1

CO_2 =
1.6062 * Gender=F +
1.4822 * Class=SO,JR,SR +
0.874 * Class=JR,SR +
1.5018 * Class=SR +
0.8623 * Major1=Japanese Studies, Mathematics-Economics, Earth & Environmental Sciences,Undecided,Business Administration, Classics, Sustainability, German, Spanish, Accounting, Public Health,Communication Studies,Economics,French,Philosophy,Religion,Individualized Curriculum Pgrm,Asian Studies,Urban Studies,Biology,Music,History,Health & Exercise Science,English,Education,Psychology,Politics and Intl Affairs,Neuroscience,Chemistry,Sociology +
1.6904 * Major1=Undecided, Business Administration, Classics, Sustainability, German, Spanish, Accounting, Public Health, Communication Studies, Economics, French, Philosophy, Religion, Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
-1.3421 * Major1=Business Administration, Classics, Sustainability, German, Spanish, Accounting, Public Health, Communication Studies, Economics, French, Philosophy, Religion, Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
0.4129 * Major1=Sustainability, German, Spanish, Accounting, Public Health, Communication Studies, Economics, French, Philosophy, Religion, Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
0.8316 * Major1=Economics, French, Philosophy, Religion, Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
-0.5669 * Major1=French, Philosophy, Religion, Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
0.5141 * Major1=Philosophy, Religion, Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
-0.3999 * Major1=Individualized Curriculum Pgrm, Asian Studies, Urban Studies, Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
0.575 * Major1=Biology, Music, History, Health & Exercise Science, English, Education, Psychology, Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
0.8221 * Major1=Politics and Intl Affairs, Neuroscience, Chemistry, Sociology +
-0.2155

Linear Regression Model 2 (much simpler)

CO2 =

1.5748 * Gender=F +
1.0181 * Class=SO,JR,SR +
0.7043 * Class=JR,SR +
0.9757 * Class=SR +
0.5634 * Semester=Spring 2015,Fall 2015,Spring 2016,Spring 2017,Fall 2011,Spring 2012,Fall 2012,Spring 2013 +
-0.2887 * Semester=Fall 2015,Spring 2016,Spring 2017,Fall 2011,Spring 2012,Fall 2012,Spring 2013 +
1.409 * Semester=Fall 2011,Spring 2012,Fall 2012,Spring 2013 +
0.6552 * Semester=Spring 2012,Fall 2012,Spring 2013 +
3.7965 * Semester=Spring 2013 +
0.7001 * Accounting +
-1.7562 * Art +
1.0785 * Biology +
2.0648 * Chemistry +
0.5759 * Communication Studies +
-0.9903 * Computer Science +
-1.2658 * Physics +
-2.1698 * Pre-Engineering +
-0.6087 * Earth & Environmental Sciences +
0.8347 * Economics +
1.0083 * Education +
0.8491 * English +
1.0776 * German +
1.5188 * Health & Exercise Science +
1.019 * History +
1.8536 * Individualized Curriculum Pgrm +
-0.848 * Information Technology +
-1.1546 * Mathematics +
1.2464 * Music +
1.9244 * Neuroscience +
0.7765 * Philosophy +
1.6967 * Politics and Intl Affairs +
1.5151 * Psychology +
0.9806 * Public Health +
0.703 * Religion +
1.5458 * Sociology +
-0.7058 * Theatre Arts +
-0.5057 * Undecided +
0.599