Honors Thesis

Honors Program

5-4-2022

Generating a Dataset for Comparing Linear vs. Non-Linear Prediction Methods in Education Research

Jack Mauro
Loyola Marymount University, jmauro1@lion.lmu.edu

Elena Martinez
Loyola Marymount University, emarti78@lion.lmu.edu

Anna Bargagliotti
Loyola Marymount University, Anna.Bargagliotti@lmu.edu

Recommended Citation

Mauro, Jack; Martinez, Elena; and Bargagliotti, Anna, "Generating a Dataset for Comparing Linear vs. Non-Linear Prediction Methods in Education Research" (2022). *Honors Thesis*. 446.
https://digitalcommons.lmu.edu/honors-thesis/446

Loyola Marymount University
University Honors Program

Generating a Dataset for Comparing Linear vs. Non-Linear Prediction Methods in Education Research

A thesis submitted in partial satisfaction

of the requirements of the University Honors Program

of Loyola Marymount University

by

Jack Mauro

May 4, 2022

Generating a Dataset for Comparing Linear vs. Non-Linear Prediction Methods in Education Research

Elena Martinez[*], Jack Mauro[†]

Faculty Advisor: Anna Bargagliotti[‡]

Abstract - Machine learning is often used to build predictive models by extracting patterns from large data sets. Such techniques are increasingly being utilized to predict outcomes in the social sciences. One such application is predicting student success. Machine learning can be applied to predicting student acceptance and success in academia. Using these tools for education-related data analysis, may enable the evaluation of programs, resources and curriculum. Currently, research is needed to examine application, admissions, and retention data in order to address equity in college computer science programs. However, most student-level data sets contain sensitive data that cannot be made public. To help facilitate research and the application of machine learning models to this field, we generate an artificial student-level data set of 50,000 students to simulate college admissions data. We generate this data set for public access and without privacy concerns. Once the data is generated, we then analyze it using logistic regression, K-Nearest Neighbor, random forest, neural networks, and XGBoost techniques to demonstrate and compare the type of analyses that can be conducted on data sets of this type. Finally we provide an analysis on whether the predictive gains of machine learning models outweigh the potential loss of interpretability in comparison to classical statistical methods.

Keywords : machine learning; data generation; education; linear prediction; non-linear prediction

Mathematics Subject Classification (2021) **:** statistics; data analysis

1 Introduction

While data analysis offers important insights for decision making, society has become increasingly aware of the consequences of sharing private information with the public. Although providing access to information such as patient histories or student academic performance supports research in academic and professional settings, consumer protection

[*] Loyola Marymount University Applied Mathematics and Computer Science double major

[†] Loyola Marymount University Applied Mathematics and Philosophy double major

[‡] Loyola Marymount University Professor of Mathematics

acts prohibit sharing of patient and customer data. One example of such protection acts is the Family Educational Rights and Privacy Act (FERPA). FERPA is a federal law that protects students' educational records and profiles. This regulation limits the ability to publish education data sets with student information [3] and thus limits the testing of models that might best capture student behavior. Artificial data sets, however, can be simulated and can provide a playground for researchers to implement predictive models, train algorithms, and test hypotheses. Such artificial data sets do not violate privacy concerns and can be legally published for public use. In this paper, we discuss the construction of such a data set and then compare the use of machine learning algorithms and statistical models to analyze it. More specifically, the objectives for this paper are threefold

1. Generate an artificial dataset of over 50,000 college applicants. Make dataset publicly available through our Github link:
   `https://github.com/jmauro1/generated-education-dataset`

2. Analyze the artificial data set using classical statistical techniques, i.e logistic regression, and machine learning techniques: K-Nearest Neighbor, random forest, neural networks, and XGBoost techniques

3. Explore the relationship between the predictive gains of machine learning models and the value of interpretability statistics of logisitic regressions

The second goal of this paper focuses on comparing traditional statistical methodologies to those put forth in machine learning. While data analysis in the social science has a tradition of being focused on linear and logistic regression models, according to the American Academy of Political Science and Social Science, "machine learning methods now provide us with better alternatives." [2]. Machine learning models may introduce a plethora of new ways to predict the outcome of a student based on known features. Unfortunately, the equivalent of datasets like Imagenet [1] that pave the way for deep learning in other fields does not exist in the social sciences. Due to privacy concerns and the nature of social science data, it is rare to find a dataset that contains meaningful information that is accessible to the public. The first goal of the paper addresses this problem by providing a simulated realistic data set to work with.

2 Generating an Artificial Dataset

The NSF Grant entitled *Equity of Access to Computer Science: Factors Impacting the Characteristics and Success of Undergraduate CS Majors* (Grant no. 2031907) uses a large student-level data set consisting of admissions and retention from four universities in the Western United States. These data provide student application records, admission records, and for those who were admitted and attended, student course records for all years at a university. The data used in the grant is restricted and not available for public use. However, we generate a public use artificial data set to use to study different methodologies for predicting student success based on student characteristics.

The artificial data set we construct features 50,000 student entries where each entry contains five student-level features. Each entry in the artificial dataset represents an applicant to a university, and each feature represents a variable that captures information that would typically be included in a student's application to a university. We create the artificial dataset using the Pandas and NumPy packages in Python. The following sections explain in detail how each feature of the dataset was generated.

2.1 Dataset Variables

A total of eight variables are generated in the artificial data set (see Table 1). For each student applicant, their race, gender, GPA, SAT score, and a socio-economic status (SES) variable were generated. In addition, three acceptance output models were generated: random acceptance, trained acceptance, and modeled acceptance; all three measured on a binary scale of accepted or rejected.

Variable Name	Variable Codomain
Race	Race 1, Race 2, Race 3, Race 4, Race 5
GPA	[0.0, 4.0]
SAT	[400, 1600]
SES	$[0, \infty)$
Gender	Male, Female
Random Acceptance	0, 1
Trained Acceptance	0, 1
Modeled Acceptance	0, 1

Table 1: Features of Each Artificial Applicant

2.1.1 Race and Gender

To generate a realistic and authentic data set, we base our parameters off of the NSF grant's data set. The NSF data provides five racial categories across the institutions included: Race 1, Race 2, Race 3, Race 4, and Race 5. Table 2 illustrates the proportion of students in the data set of each different race by gender. Using the joint distribution, we can generate the race and gender features for the simulated data.

	Male	Female
Race 1	7.20%	6.39%
Race 2	20.68%	18.59%
Race 3	15.54%	23.64%
Race 4	2.72%	4.13%
Race 5	0.53%	0.58%

Table 2: Dataset Breakdown By Race and Gender

2.1.2 SES, GPA, and SAT Score

Next, we generate the SES, GPA, and SAT features. We jointly model these through a multivariate Gaussian distribution. Table 3 shows the means and standard deviations used for each of the features given a specific race and gender combination.

Race, Gender	SES Mean	SES STD	GPA Mean	GPA STD	SAT Mean	SAT STD
Race 1 Male	121428	149881	3.60	0.476	1515	308
Race 2 Male	68397	155398	3.59	0.476	1532	312
Race 3 Male	51514	66423	3.46	0.459	1263	277
Race 4 Male	64207	89952	3.33.	0.531	1307	293
Race 5 Male	835756	131112	3.52	0.47	161	318
Race 1 Female	104251	134816	3.71	0.439	1443	307
Race 2 Female	114165	143540	3.71	0.43	1365	313
Race 3 Female	44077	54385	3.54	0.449	1192	256
Race 4 Female	53708	75904	3.45	0.493	1252	282
Race 5 Female	74666	157799	3.58	0.484	1251	333

Table 3: SES, GPA, and SAT Means and Standard Deviations by Race

In addition, the correlations among these variables are extracted from the NSF data and shown in Table 4.

	GPA	SES	SAT
GPA	1	0.0770	0.2446
SES	0.0770	1	0.2073
SAT	0.2446	0.2073	1

Table 4: GPA, SES, and SAT Correlations

We can visualize the pairwise projection of the multivariate Gaussian distribution in Figures 1, 2, and 3. This shows the joint distributions of each of pair of features in the SES, SAT, and GPA.

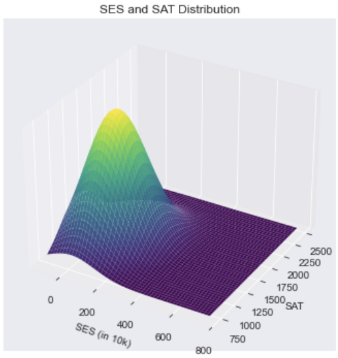


Figure 1: Projection of distribution of SES and SAT Scores

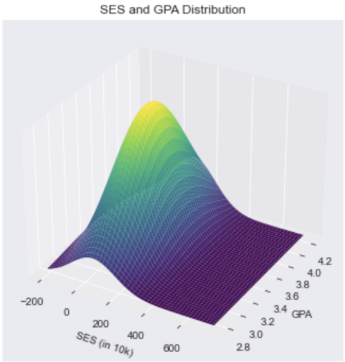


Figure 2: Projection of distribution of SES and GPA

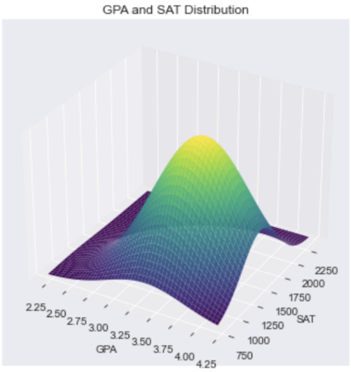


Figure 3: Projection of distribution of GPA and SAT Scores

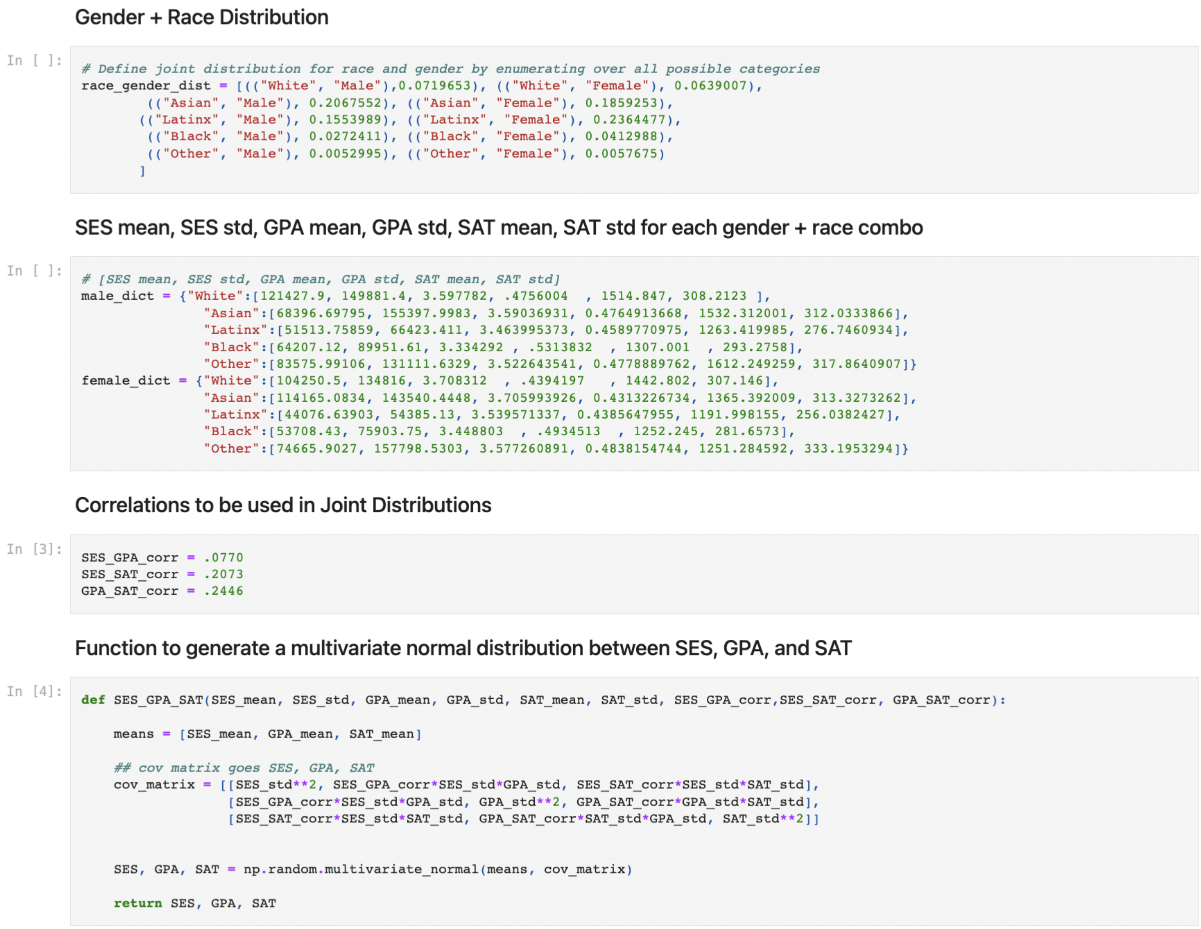
2.2 Data Generation Process

2.2.1 Independent Variables

In order to generate each artificial student, first we simultaneously select a gender and race for the student using the race and gender joint distribution described above. Based on this selection, we then use the SES, GPA, and SAT means and standard deviations that pertain to the selected race gender combination of the student. Using these values and the correlation values listed in Table 4 we generate the following covariance matrix.

$$\begin{pmatrix} (SES\ \sigma)^2 & (SES\ GPA\ corr) \times (SES\ \sigma) \times (GPA\ \sigma) & (SES\ SAT\ corr) \times (SES\ \sigma) \times (SAT\ \sigma) \\ (SES\ GPA\ corr) \times (SES\ \sigma) \times (GPA\ \sigma) & (GPA\ \sigma)^2 & (GPA\ SAT\ corr) \times (GPA\ \sigma) \times (SAT\ \sigma) \\ (SES\ SAT\ corr) \times (SES\ \sigma) \times (SAT\ \sigma) & (GAP\ SAT\ corr) \times (SAT\ \sigma) \times (GPA\ \sigma) & (SAT\ \sigma)^2 \end{pmatrix}$$

We then run the multivariate normal command in Python to extract an SES, GPA, SAT for this student from the corresponding covariance matrix. The Python code to perform the steps described above can be found in Figure 4 below.



Gender + Race Distribution

```
# Define joint distribution for race and gender by enumerating over all possible categories
race_gender_dist = [(("White", "Male"),0.0719653), (("White", "Female"), 0.0639007),
        (("Asian", "Male"), 0.2067552), (("Asian", "Female"), 0.1859253),
       (("Latinx", "Male"), 0.1553989), (("Latinx", "Female"), 0.2364477),
        (("Black", "Male"), 0.0272411), (("Black", "Female"), 0.0412988),
        (("Other", "Male"), 0.0052995), (("Other", "Female"), 0.0057675)
       ]
```

SES mean, SES std, GPA mean, GPA std, SAT mean, SAT std for each gender + race combo

```
# [SES mean, SES std, GPA mean, GPA std, SAT mean, SAT std]
male_dict = {"White":[121427.9, 149881.4, 3.597782, .4756004  , 1514.847, 308.2123 ],
               "Asian":[68396.69795, 155397.9983, 3.59036931, 0.4764913668, 1532.312001, 312.0333866],
               "Latinx":[51513.75859, 66423.411, 3.463995373, 0.4589770975, 1263.419985, 276.7460934],
               "Black":[64207.12, 89951.61, 3.334292 , .5313832  , 1307.001  , 293.2758],
               "Other":[83575.99106, 131111.6329, 3.522643541, 0.4778889762, 1612.249259, 317.8640907]}
female_dict = {"White":[104250.5, 134816, 3.708312  , .4394197   , 1442.802, 307.146],
               "Asian":[114165.0834, 143540.4448, 3.705993926, 0.4313226734, 1365.392009, 313.3273262],
               "Latinx":[44076.63903, 54385.13, 3.539571337, 0.4385647955, 1191.998155, 256.0382427],
               "Black":[53708.43, 75903.75, 3.448803  , .4934513  , 1252.245, 281.6573],
               "Other":[74665.9027, 157798.5303, 3.577260891, 0.4838154744, 1251.284592, 333.1953294]}
```

Correlations to be used in Joint Distributions

```
SES_GPA_corr = .0770
SES_SAT_corr = .2073
GPA_SAT_corr = .2446
```

Function to generate a multivariate normal distribution between SES, GPA, and SAT

```
def SES_GPA_SAT(SES_mean, SES_std, GPA_mean, GPA_std, SAT_mean, SAT_std, SES_GPA_corr,SES_SAT_corr, GPA_SAT_corr):

    means = [SES_mean, GPA_mean, SAT_mean]

    ## cov matrix goes SES, GPA, SAT
    cov_matrix = [[SES_std**2, SES_GPA_corr*SES_std*GPA_std, SES_SAT_corr*SES_std*SAT_std],
                  [SES_GPA_corr*SES_std*GPA_std, GPA_std**2, GPA_SAT_corr*GPA_std*SAT_std],
                  [SES_SAT_corr*SES_std*SAT_std, GPA_SAT_corr*SAT_std*GPA_std, SAT_std**2]]


    SES, GPA, SAT = np.random.multivariate_normal(means, cov_matrix)

    return SES, GPA, SAT
```

Figure 4: Python Data Generation Code

2.2.2 Dependent Variables

Since this is an artificial dataset, there are no actual acceptance values for our ”students.” Therefore, we generated three different approaches to determining acceptance.

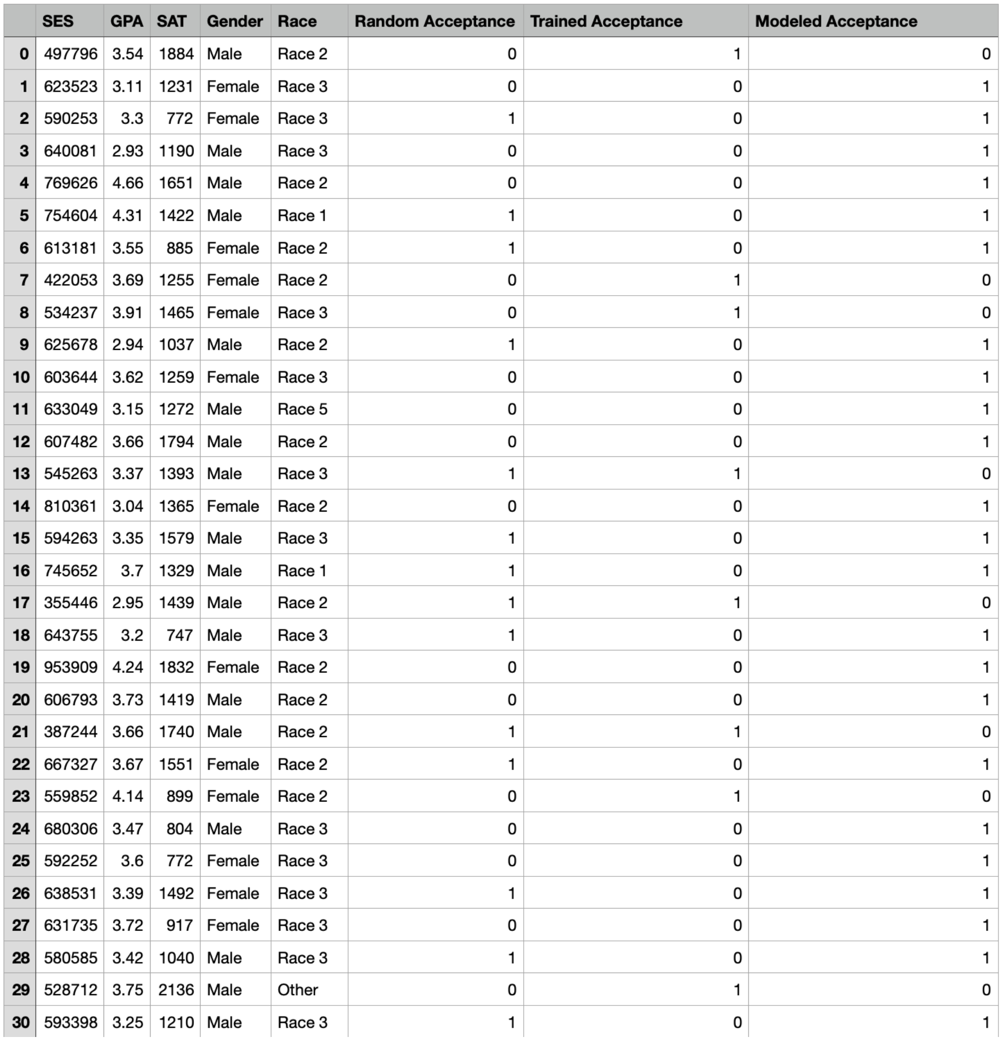
1. **Random Acceptance:** The random acceptance model randomly determines if an artificial student is accepted to the university. We specify that each artificial student has a 40 % chance of being accepted and 60% chance of being rejected from the university. The random acceptance model acts as a baseline to assure that all of the prediction methods in the following section of the paper are functioning properly.

2. **Trained Acceptance:** In this acceptance model, a logistic regression was run on 10% of the NSF grant data with the input features for the logistic regression being the exact same as the input features for the artificial students in the generated dataset (race, gender, SES, SAT, GPA). Once estimated, we use the coefficients found from the logistic regression and use them to determine whether an artificial student will be accepted or rejected. A student's acceptance is determined using the outputted logit probability as the probability of acceptance.

3. **Modeled Acceptance:** This acceptance model follows the exact same process as the Trained Acceptance model, but the logistic regression is trained on the entire NSF grant data.

2.2.3 Final Dataset

The artificial data set that was generated has 50,000 students each with five features. Each of the features was generated using parameters found in real data. In addition, three different acceptance variables were generated per student. Figure 4 shows several entries of the final generated dataset.



	SES	GPA	SAT	Gender	Race	Random Acceptance	Trained Acceptance	Modeled Acceptance
0	497796	3.54	1884	Male	Race 2	0	1	0
1	623523	3.11	1231	Female	Race 3	0	0	1
2	590253	3.3	772	Female	Race 3	1	0	1
3	640081	2.93	1190	Male	Race 3	0	0	1
4	769626	4.66	1651	Male	Race 2	0	0	1
5	754604	4.31	1422	Male	Race 1	1	0	1
6	613181	3.55	885	Female	Race 2	1	0	1
7	422053	3.69	1255	Female	Race 2	0	1	0
8	534237	3.91	1465	Female	Race 3	0	1	0
9	625678	2.94	1037	Male	Race 2	1	0	1
10	603644	3.62	1259	Female	Race 3	0	0	1
11	633049	3.15	1272	Male	Race 5	0	0	1
12	607482	3.66	1794	Male	Race 2	0	0	1
13	545263	3.37	1393	Male	Race 3	1	1	0
14	810361	3.04	1365	Female	Race 2	0	0	1
15	594263	3.35	1579	Male	Race 3	1	0	1
16	745652	3.7	1329	Male	Race 1	1	0	1
17	355446	2.95	1439	Male	Race 2	1	1	0
18	643755	3.2	747	Male	Race 3	1	0	1
19	953909	4.24	1832	Female	Race 2	0	0	1
20	606793	3.73	1419	Male	Race 2	0	0	1
21	387244	3.66	1740	Male	Race 2	1	1	0
22	667327	3.67	1551	Female	Race 2	1	0	1
23	559852	4.14	899	Female	Race 2	0	1	0
24	680306	3.47	804	Male	Race 3	0	0	1
25	592252	3.6	772	Female	Race 3	0	0	1
26	638531	3.39	1492	Female	Race 3	1	0	1
27	631735	3.72	917	Female	Race 3	0	0	1
28	580585	3.42	1040	Male	Race 3	1	0	1
29	528712	3.75	2136	Male	Other	0	1	0
30	593398	3.25	1210	Male	Race 3	1	0	1

Figure 5: Final Generated Dataset

With the artificial data set completed, then researchers can use it to investigate and compare methodologies for predicting acceptance.

3 Predicting Acceptance

Machine learning is often used to build predictive models by extracting patterns from large data sets. The introduction of machine learning into the social sciences can be applied to predicting student acceptances and success [4]. The artificial data set generated above provides the needed data to explore the use of machine learning models on public data. Using the generated data above, we implement both a classical statistical logistic regression and five different machine learning techniques to compare their accuracy and interpretability.

To do so, we first split our data into a training set consisting of 80% of the artificial data set and a test set consisting of 20% of the artificial data set. Next, we run the following models: (1) logistic regression, (2) support vector machine (SVM), (3) neural network (NN), (4) K-nearest neighbor algorithm, (5) random forest, and (6) XG Boost.

Each model predicts whether or not a student is accepted to the university the student applied to. The accuracy of a model is calculated by $100 * \frac{\text{number of correct predictions}}{\text{size of test set}}$ where a correct prediction signifies that the model correctly predicted whether or not the artificial student was accepted to the university the student applied to.

The neural network used was a three layer neural network with linear layers, a stochastic gradient descent optimizer, and ReLu nonlinearity between each layer. The logistic regression and the support vector machine were the linear models used to predict acceptance, and the neural network, K-nearest neighbor algorithm, random forest, and XG Boost were the nonlinear models used.

4 Logisitic Regression

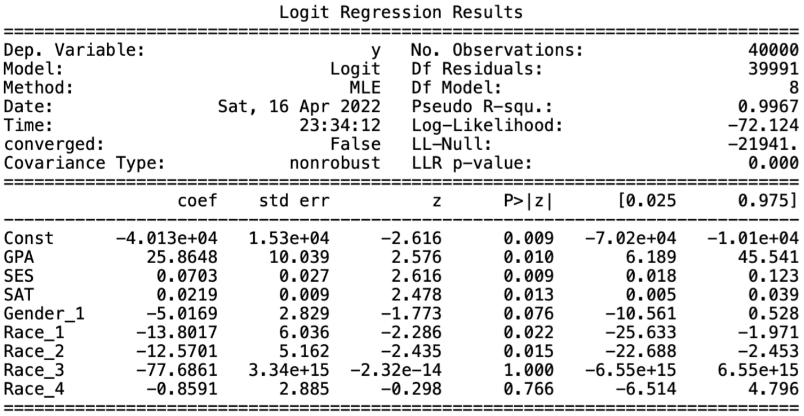
A benefit of using linear predictive methods over non-linear predictive methods is the advantage of interpretability. Linear predictive methods offer insight to which input features influences the decision of the classifier whereas nonlinear methods do not.
In a logistic regression model, a 1 unit increase in a variable X_i results in a predicted β_i, where β_i is the coefficient in the logit equation corresponding to X_i increase in the log odds ratio. In other words, a 1 unit increase in X_i results in a predicted $e^{\beta_i} - 1$ increase in the odds of getting accepted. The logistic regression output is given in Figure 6. From this, we see that a 1 unit increase in a student's SAT score results in a $e^{.0219} - 1 = 2.2$ % increase in the odds of getting accepted, and a 1 unit increase in a student's SES results in a $e^{.0703} - 1 = 7.3$ % increase in the odds of getting accepted (using the trained acceptance variable).

This interpretability is important in the social sciences as often the scope of research is to find which influential factors are associated with specific outputs.

5 Model Comparison

The test accuracy percentage of each model is given in the table below. Since we have three different possible dependent variables, there are 3 columns of accuracies: Random Acceptance, Trained Acceptance, and Modeled Acceptance.

Table 5 below reveals that the Random Acceptance accuracy is similar across all models - both linear and non-linear. In all cases, the accuracy reflects the 60 % probability that was specified to define the random acceptance variable. When acceptance is defined by Trained Acceptance and Modeled Acceptance, then there the predictability accuracy



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:                      y   No. Observations:                40000
Model:                          Logit   Df Residuals:                    39991
Method:                           MLE   Df Model:                            8
Date:                Sat, 16 Apr 2022   Pseudo R-squ.:                  0.9967
Time:                        23:34:12   Log-Likelihood:                -72.124
converged:                      False   LL-Null:                       -21941.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Const      -4.013e+04   1.53e+04     -2.616      0.009   -7.02e+04   -1.01e+04
GPA           25.8648     10.039      2.576      0.010       6.189      45.541
SES            0.0703      0.027      2.616      0.009       0.018       0.123
SAT            0.0219      0.009      2.478      0.013       0.005       0.039
Gender_1      -5.0169      2.829     -1.773      0.076     -10.561       0.528
Race_1       -13.8017      6.036     -2.286      0.022     -25.633      -1.971
Race_2       -12.5701      5.162     -2.435      0.015     -22.688      -2.453
Race_3       -77.6861   3.34e+15  -2.32e-14      1.000   -6.55e+15    6.55e+15
Race_4        -0.8591      2.885     -0.298      0.766      -6.514       4.796
==============================================================================
```

Figure 6: Output of the Logistic Regression with the Trained Acceptance as the dependent variable

varies. In both cases, the linear models accuracies are lower than the non-linear models. The Random Forest and the XG Boost both perform the best for both Trained and Modeled Acceptance. K-Nearest Neighbor also performs well for the Modeled Acceptance.

Model	Random Acceptance	Trained Acceptance	Modeled Acceptance
Logistic Regression	59.7 %	81.8%	81.3%
Support Vector Machine	59.7%	80.4%	80.4%
Neural Network	59.7%	96.3%	97.0%
K-Nearest Neighbor	53.7%	97.6%	99.9%
Random Forest	54.9%	98.3%	99.9%
XG Boost	59.7%	98.3%	99.9%

Table 5: Performance of different prediction models using the three acceptance models

6 Conclusion

6.1 Accuracy vs Interpretability

By comparing the performance of the linear and non-linear models, we see that the non-linear predictive methods outperform the linear predictive methods when predicting the non-random acceptance variables. This difference in accuracy, however, bears a trade off because the nonlinear predictive methods do not offer the same interpretability demonstrated usinf the linear predictive methods. For example, by using XG Boost, K-Nearest Neighbor, or Random Forest techniques we obtain a prediction accuracy of nearly 100%. We are not, however, able to understand the effect that changes to the independent variables have on our dependent variable. While we lose about 18.6% in accuracy using a logisitic regression, we are able to determine the approximate effect that a change to either SAT, GPA, or SES will have on whether or not a student is accepted using any of the three acceptance models. Future work to understand why the different accuracies occur across the different models is worthwhile and likely will provide insight into what model is best to use in what scenario. The artificial data set generated provides a playground for researchers to investigate these differences in predictability.

6.2 Application of Work

While many are well-aware of the lack of women and under-represented minorities in the tech industry, what is most alarming is that representation of these demographic groups is not increasing. According to a 2021 study on women in technology, "The percentage of computing roles women hold has largely declined in the United States over the past 25 years" and "unless we take action, the trajectory is unlikely to change." The lack of representation of women and minorities in computing is reflected in the share of black, Latina, and Native American women receiving computing degrees. This share has declined by one-third over the past decade, dropping from 6% to a 4% [6]. Research is needed to unpack what factors impact the success of undergraduate computer science majors thus allowing for policy to shift and help work towards addressing the lack of representation in computer science education.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. doi: 10.1109/CVPR.2009.5206848.

[2] Mason, W., Vaughan, J.W., Wallach, H, Computational social science and social computing. *Mach Learn* **95** (2014), 257–260. https://doi.org/10.1007/s10994-013-5426-8

[3] Bhumiratana, Bhume and Bishop, Matt. Privacy aware data sharing: Balancing the usability and privacy of datasets *Proc 2nd Intl Conf Perv Tech Rel Assist Envir* **01 2009** (2009).

[4] Hindman, Matthew. Building Better Models *The ANNALS of the American Academy of Political and Social Science* **659** (April 2015) 48-62.

[5] Verhagen, Mark D. A Pragmatist's Guide to Using Prediction in the Social Sciences SocArXiv (April 2021).

[6] Pivotal Ventures McKinsey and Company. Using CSR and Philanthropy to Close the Gender Gap in Tech (2021).

Elena Martinex
Loyola Marymount University
1 LMU Drive
Los Angeles, CA
E-mail: emarti78@lion.lmu.edu

Jack Mauro
Loyola Marymount University
1 LMU Drive
Los Angeles, CA
E-mail: jmauro1@lion.lmu.edu