4-2022

Ethical Use of Machine Learning in Higher Education Admission

Siqi Fang '23

Ethical Use of Machine Learning in Higher Education Admission

Siqi Fang
Spring 2022   Math in Social Context
Hamilton College Mathematics Department

Introduction:

A machine learning model called GRADE was used for PhD admission at UT Austin from the year 2013-2020. The model was trained a small set of past admission decisions which are already bias and was used immediately without further tuning or human validation. The model will score all applicants and the decision is made without further human assessment for applicants with the highest and lowest score.
Only 362/588 full human reviews are conducted with a few people admitted and the majority of the rest being rejected by algorithm.

Model Overview:

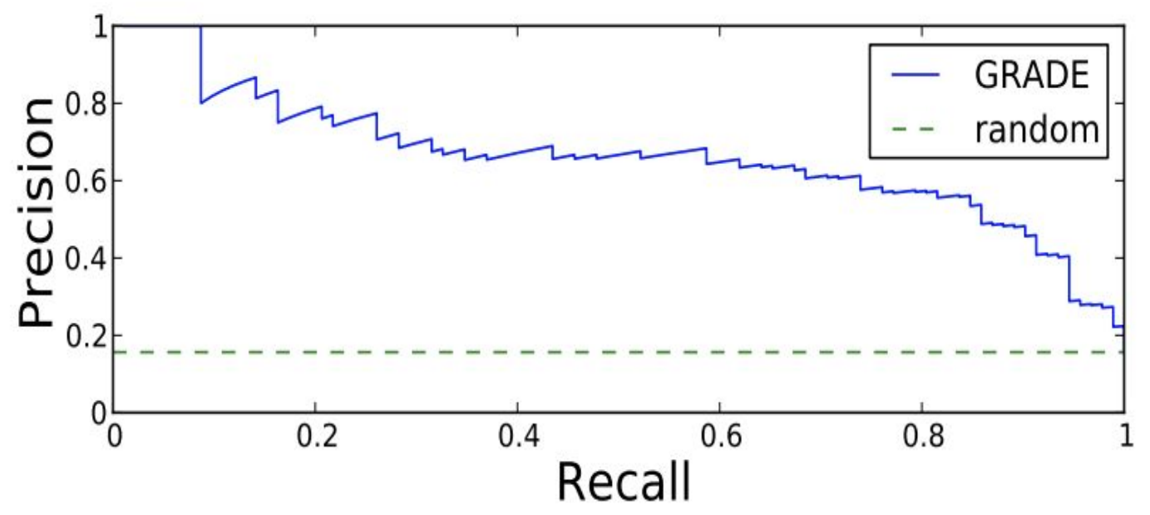
Model: Logistic Regression model with L1-regularization
Features:
total of 427 features (58 have non-zero learned weight)

- *numeric features* (GRE, GPA) are standardized
- *categorical features* (name of school, preferred advisors) *are* encoded either using *one-hot-encoding or log-odds(*probability of admission given the feature value based on past results)
- *text data* (personal statement, recommendation letters) personal statement is omitted and letters of recommendation are encoded with outdated NLP techniques (Bag of words).

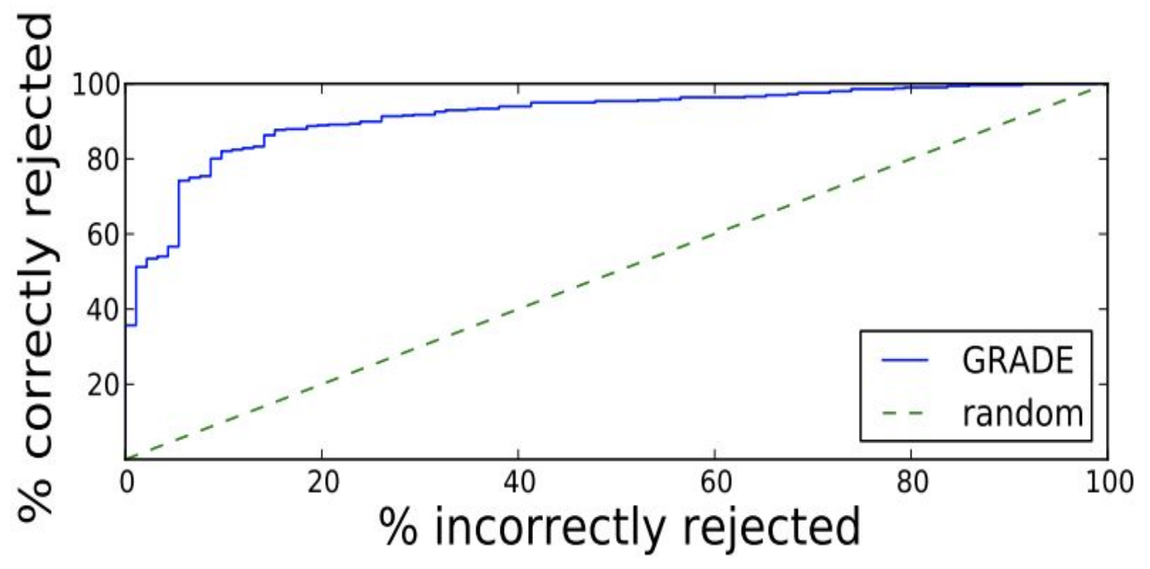
Training Data: Past admission data
Output: probability of admission that is mapped to a 0-5 score, 5 being a very competitive profile

Results

Classification performance of GRADE in 2013 shows that the model does a better job at identifying students that the admission committee would reject.(Figure a. has a low AUC)
Issue: In this case, when human review is available and the cost of rejecting an application is higher, the ability to identify students that the university **should admit** is more important than identifying rejects.

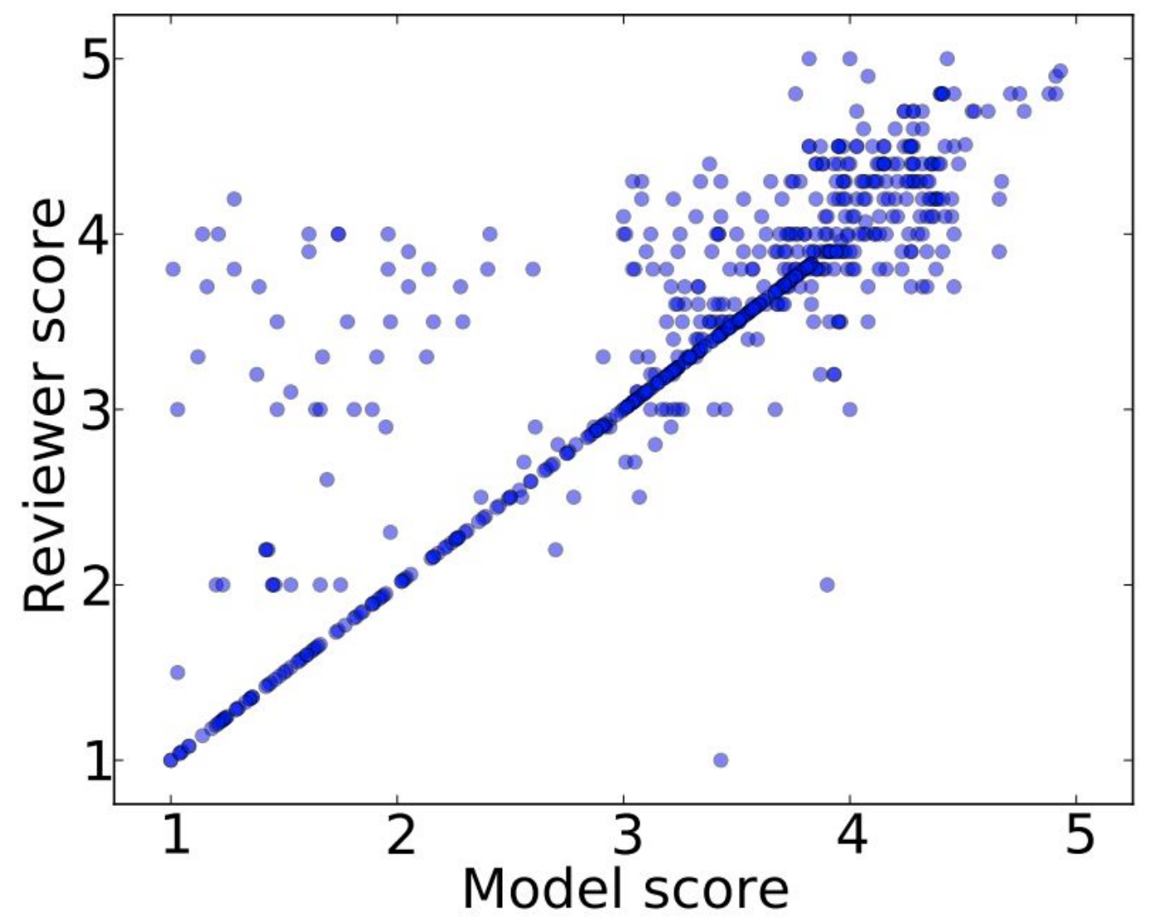


(a)

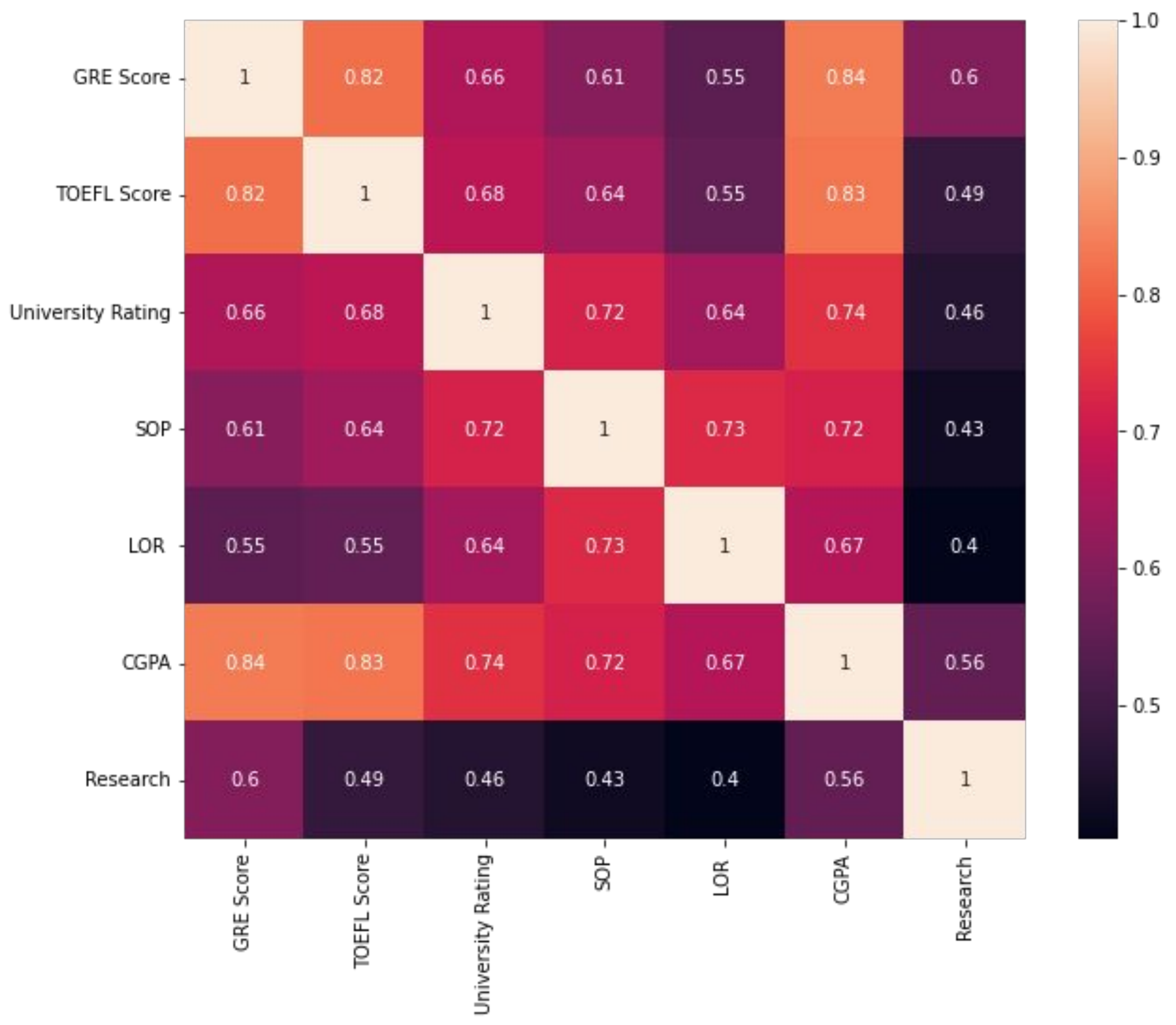


(b)

Score agreement of the model versus human reviewer. The model score is within 0.2 of the human score 40% of the time, and human reviewer scores agree 50% of the time. In most cases of disagreement, model produced a score **lower** than human reviewer.
Issue: The model has the tendency to underestimate, and applicants with low scores are less likely to have a second round of human review.



A Closer Look: Multicollinearity



Multicollinearity means that some features are correlated. Here is the heatmap for correlation of some common college admission materials. We can see that GPA is highly correlated with standardized test scores. This mean that we double count the contribution of standardized test scores. While this does not necessarily means lower performance, it corrupts the coefficient of our model, making it less interpretable. The source of data is from kaggle.

A Better GRADE Model

Remove correlated features: Some features have intrinsic correlation, for example, GPA has a correlation with most test scores. Meaning that students with a below average school performance is punished twice. Also, one-hot encoding introduces correlation, so features such as research area and preferred faculty advisors are have a higher actual weight (than reported feature importance). This does not mean biased since weighing materials are highly organic for human, but the issue is that the model weights are meaningless and misleading which makes the model not interpretable.

Remove biased information: Although demographic information all had zero learned weights, they are not removed in training. In addition, features such as the name of the institution had high learned importance, and can introduce implicit bias.

Better Model: With basic data engineering, a logistic regression does not have the complexity to replace an individualized assessment. It calculates a global weight for all applicants while human reviewers weigh each application materials differently depending on the applicant. Also, logistic regression does not handle correlated features and categorical features well(without proper feature engineering which this model lacked). An ensemble algorithm with tree based models could be a better model for this task.

Takeaway for all Machine Learning Tasks:

ML engineers should understand how and why each features are being used. For example, many admission processes involve prioritizing students from target schools for efficiency, however, for an algorithm, there is no efficiency problem. So this potentially biased feature should be removed. The computation power of machines should be used to improve the fairness of admission instead of continuing an existing biased human algorithm. Instead of letting regularization coefficient make the decision, model designer should specify their goals and come up with more reasonable feature selection and engineering that helps make the model fair.

The Big Picture

The Risks of Using AI in Higher Education Admission

Garbage in, garbage out: A model should be carefully reviewed if it uses biased data. There should be standards for feature selection. The importance of domain knowledge should be addressed so engineers are aware of possible biases when selecting features and training the model.

Challenges of Evaluating the model: A model can label a picture of cat and we can tell if it did right or wrong. However, there is no correct answer in assessing a person's profile, so the use of parametric models in these tasks is questionable. This is a common issue among the tasks that requires a holistic assessment. Better evaluation standards, for example, a certain amount of human review should be adopted.

Lack of standards: Having a relatively small dataset and number of features, the people who designed GRADE could easily run more tests with different models and data engineering techniques to train a more optimized model. The fact that such an evidently flawed and biased model is put into production and used for years without being challenged signals the lack of standards in ethical use of machine learning. A lot of the time, issues are only pointed out by people when the harm has been done. In addition, discovering issues is difficult because institutions keep their models in a black-box. While confidentiality is necessary for businesses and institutions, there should be standards for a model to be put into production.

References

Waters, Austin, and Risto Miikkulainen. "GRADE: Machine Learning Support for Graduate Admissions." The AI magazine 35.1 (2014): 64-75. CrossRef. Web.

Mahto, Krishna Kumar. "One-Hot-Encoding, Multicollinearity and the Dummy Variable Trap." Medium, Towards Data Science, 20 July 2019, https://towardsdatascience.com/one-hot-encoding-multicollinearity-and-the-dummy-variable-trap-b5840be3c41a.

Software

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

Kluyver, T. et al., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp. 87–90.

Dataset

Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019