Stanford | ENGINEERING
Computer Science

Evaluating Predictive Models for Scores on the Exame Nacional de Desempenho de Estudantes

Breno Dal Bianco, Georgia Sampaio, Peter McEvoy

Motivation

Exame Nacional de Desempenho de Estudantes (ENADE)

- Exam used by the Brazilian government to assess the quality of higher education
- Self-reported sociodemographic data is collected alongside the administering of the exam
- The full annual results of the exam are publicly available
- How does a student's sociodemographic background influence her performance in university as measured by her permformance on the ENADE?

Problem Definition

Given:

$$\text{Training set } \mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid i \in [1, n]\} \text{ s.t.}$$
$$x^{(i)} \in \{0,1\}^d,\ y^{(i)} \in \mathbb{R}$$

Learn:

$$\text{Model } h : \{0,1\}^d \to \mathbb{R}$$

To predict:

$$h(x^{(i)}) = \hat{y}^{(i)}$$

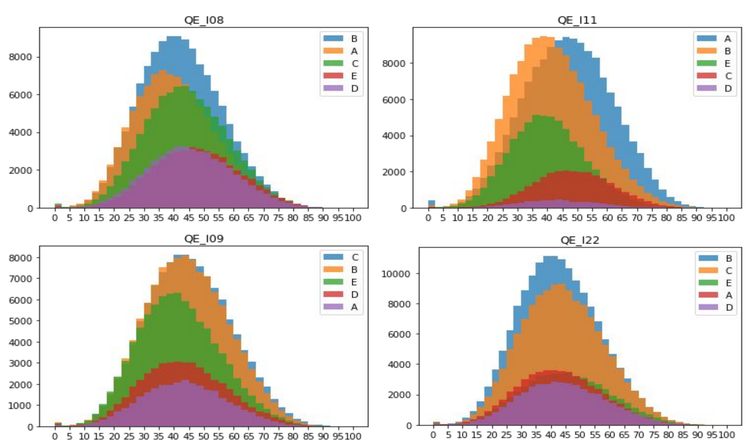
Given a training set consisting of pairs of one-hot encoded feature vectors representing a student's categorical answers to various sociodemographic questions and that student's score on the exam, learn a model that accurately predicts an exam score given a yet-unseen one-hot encoded feature vector.

Data

Sociodemographic questions:

- QE_I01: What is your relationship status?
- QE_I02: What is your race?
- QE_I04: What is the level of education of your father?
- QE_I05: What is the level of education of your mother?
- QE_I06: Where and with whom do you live right now?
- QE_I07: How many people live with you? (Include parents, siblings, spouse, children, and other relatives.)
- QE_I08: What is the total income of your household?
- QE_I09: How would you describe your financial situation?
- QE_I10: Do you currently work?
- QE_I11: Have you received scholarships or financial aid?
- QE_I12: Have you ever received housing aid as a student?
- QE_I15: Were you admitted through affirmative action?
- QE_I17: What type of school did you attend during high school?
- QE_I22: How many books did you read this year?
- QE_I23: How many hours per week do you study outside of classroom?
- QE_I25: Why did you choose this major?
- QE_I26: Why did you choose this university?

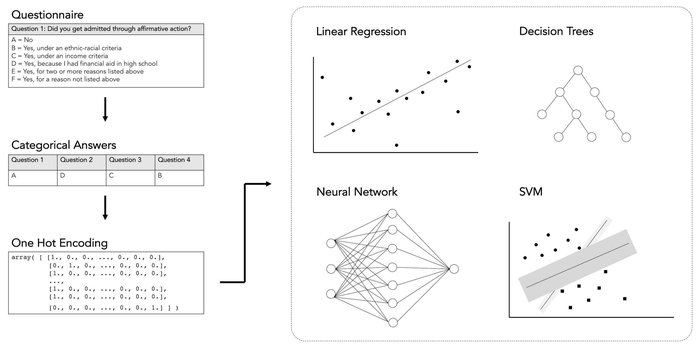
Figure 1: Histograms of the answer distributions for QE_I08, QE_I09, QE_I11, and QE_I22



Approaches

1. **Linear Regression**
   - Inputs: one-hot encoded answers to sociodemographic questions
   - Output: predicted exam score
   - Loss function: mean squared error
   - Model characteristics: bias term included
2. **Neural Network**
   - Inputs: one-hot encoded answers to sociodemographic questions
   - Output: one-hot encoded vector representing the predicted decile of the exam scores
   - Loss function: cross-entropy
   - Network architecture: one 111-node input layer, one 300-node hidden layer, one 10-node output layer
   - Model characteristics: mini-batch gradient descent, trained with and without L2 regularization
3. **Decision Trees**
   - Inputs: one-hot encoded answers to sociodemographic questions
   - Output: predicted exam score
   - Model characteristics: five models learned with maximum depths of 5, 10, 15, 40, and 100, respectively
4. **Support Vector Machine**
   - Inputs: one-hot encoded answers to sociodemographic questions
   - Output: one-hot encoded vector representing the predicted decile of the exam scores
   - Model characteristics: three models learned with RBF, linear, and 8-degree polynomial kernels, respectively

Figure 2: System overview.



Results

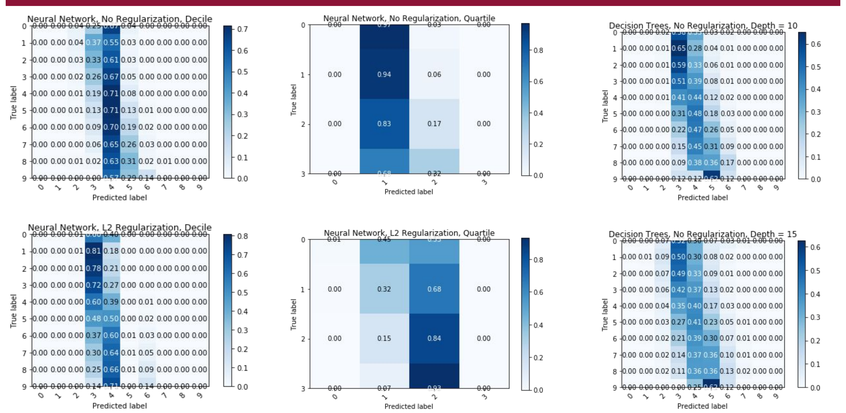


Figure 4: decile neural network confusion matrices

Figure 5: quartile neural network confusion matrices

Figure 6: decision tree classifier (depths = 10, 15) confusion matrices

Results

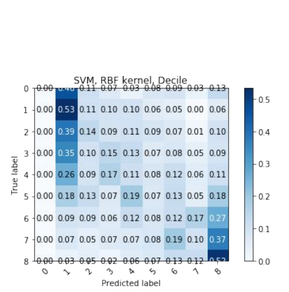


Figure 7: SVM (with radial basis function as kernel) decision matrix

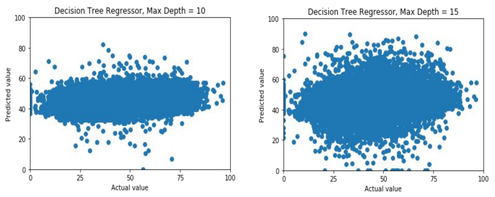


Figure 8: actual test scores vs. predicted test scores from decision tree regressor (depths = 10, 15), test set

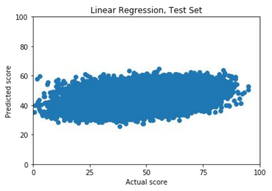


Figure 9: actual test scores vs. predicted test scores from linear regression on the test set

Model	Train Set Accuracy	Test Set Accuracy	Mean Squared Error (Test Set)	Comments
Linear regression	15.6%	16.0%	163.9	--
Neural network (classification)	60.0%	59.9%	--	Quartile
Neural network w/ L2 regularization (classification)	45.7%	45.7%	--	Quartile
Neural network (classification)	28.2%	28.0%	--	Decile
Neural network w/ L2 regularization (classification)	27.3%	27.9%	--	Decile
Decision tree (regression)	23.1%	2.6%	155.1	Depth = 15
Decision tree (classification)	37.7%	26.8%	--	Depth = 15
Support vector machine (classification)	25.9%	20.2%	--	Kernel = radial basis function

Table 1: Performance of the various models used.

Analysis

This project started as an analysis of bias on artificial intelligence algorithms. We had a large data set of scores along with socioeconomic data from undergraduate students in Brazil. Our expectation was to run a few models and probe them for their biases in areas such as race, field of study, household income.

Ironically, we were blind to our own biases, and expected the scores to have high correlation with the features given Brazil's problems with social inequality. For example, we expected a very high correlation between household income and test score. While the correlation does exist, it is not very strong, and the variance on the data shows that there are other factors which influence the score.

Because of our expectation of high correlation, we were surprised when our models performed so poorly. We reran and tweaked them extensively before realizing that we should look more closely into our data. Upon scrutiny, the score variance for each feature value was revealed to be very large, and the features largely uncorrelated with the score.

We then set out to extract the most we could out of this challenging dataset, looking at how different models perform in it. As expected, more complex models like Neural Networks performed better than simpler models like Linear Regression. However, even our highest R^2 value of ~60% turned out to be meaningless, as the model just learned to predict close to the mean regardless of the features.