Ethical Analysis of Predicting Grades with Machine Learning

Kyra Thomas

Syracuse University

CIS 400: Ethics of Machine Learning

Dr. Sucheta Soundarajan

May 13, 2021

Abstract

The importance of tools that help in understanding the predictions and decisions a machine learning model makes is growing. There is a need to be able to understand the outcomes of a model and the ethical implications of it. In this paper, I analyze the quality and effectiveness of visual explanation tools provided by Yellowbrick and present a framework that can help analyze and address the ethical implications of a model. The results show that Yellowbrick did provide quality explanations that aided in understanding the models used. The explanations aligned with what social science says about the relationship between sex/gender and mathematics grades. Furthermore, I found that the ethical framework aided in analyzing ethics regarding the grade predicting model.

Ethical Analysis of Predicting Grades with Machine Learning

Gender and Mathematics

The existence of the stereotype that women have less mathematical ability then men can have a negative effect on women pursuing careers in STEM related fields. Studies have been done regarding the relationship between gender and mathematics and have found little to no correlation between the two (Hyde & Mertz, 2009; Lindberg et al., 2010; Melhuish et al., 2008). In this paper I will focus specifically on the relationship between gender and mathematics grades. Information used to analyze this relationship include a student's learning environment and their parent's education.

The Need for Explainable Machine Learning and Its Visual Representation

As machine learning models become more complex, it becomes harder to understand and interpret them. This can be a serious issue as models are being used in the criminal justice system or in health care. It is important to know why and how a model made a certain prediction. To help with this, there are tools and libraries that provide explanations for machine learning models (Choo, 2018). These tools can give an explanation in the form of text, numbers, or graphs and charts. Consider class balance where having an imbalance can result in poor prediction accuracy for the minority class. If a model is making poor predictions, this metric can help determine why (because the classes are unbalanced) and how to fix it (e.g. resampling the dataset). The Shannon entropy equation can be used to calculate the class balance of a dataset resulting in a numerical explanation. Yellowbrick's class balance visualizer can also be used to calculate the class balance of a dataset resulting in a visual explanation. In this paper I will focus on visual explanations.

Analysis of a Models Ethics

Let's consider potential real-world applications of predicting students' grades. Universities could use this model to predict students grades and determine what students will need additional academic support. Or the model can be used in deciding who to give scholarships to. It can also be used by companies wanting to predict senior year students grades before hiring them for a full-time job.

From the above examples, it is not obvious that the model can be used for bad or have a negative impact. However, it is possible. Especially if protected socioeconomic attributes such as gender or race are used in making the predictions. The long history of systematic racism, sexism, and

other forms of discrimination will negatively affect minorities. And each attribute in a person's intersectionality can multiply that affect.

It would be wrong to give an algorithm some data, create a model, and ship the model out to be used in the real world without proper evaluation of its performance and implications. Ensuring a model is accurate and ethical is important. To do this we have to interpret, understand, and question the models. This is the heart of the ethical framework I am proposing. A more formal definition will be discussed in the Framework for Ethical Analysis section.

Related Works

Research related to the relationship between gender and mathematics performance indicate that women do not have less mathematical ability then men. The analysis of 242 separate studies by Lindberg et al. found that girls are performing as well as boys are in mathematics. There are various factors that can contribute to the existence of a gender gap in complex problem solving, a skill highly used in mathematics. Parents and teachers tend to have higher expectations in the abilities of boys than girls and the results of these expectations can affect the children's estimates of their own abilities (Lindberg). Additionally, there are strong indicators that the gap is due to sociocultural and environmental factors and not due to biology or gender (Hyde & Mertz, 2009).

The features most important to the models created for my experiment were absences, study time, and number of failures (see figure 4b). Through work by Melhuish et al., mother's education, home learning environment, primary school effectiveness, and socioeconomic status had the highest effects on predicting student's mathematics achievement. While these factors are not the same, they do correlate with each other. This is from an assumption that study time and absences are linked to a student's home learning environment and failures linked to primary school effectiveness.

Experiment

Methods

The goal of this experiment is to compare 3 different models on how accurate they are in predicting students' grades, learn how sex as an attribute affects the results, and assess the quality of explanations provided by Yellowbrick. The predictions will be made using two combinations of attributes. Attribute set A includes "studytime", "failures", "schoolsup" or school support, "activities" or number of extracurricular activities, "internet" or internet access, and "absences". Attribute set B includes all the attributes in set A plus the student's sex. I choose the attributes based on how explicitly relevant I believed they are to grades. The models are predicting the students' final grade, labeled "G3" in the dataset.

Libraries and Models Used

For the predictions I used the nearest neighbors, SVM, and random forest models. The names of these models from scikit-learn are NearestCentroid, SVC, and RandomForestClassifier respectively. I computed the predictions and accuracy for each model on attribute set A and attribute set B.

I used Yellowbrick as my explainable machine learning technique. Yellowbrick's Rank2D visualizer was used to compute correlations between features. Yellowbrick's FeatureImportances visualizer was used to find the importance of features to a model. Unfortunately, scikit-learn's NearestCentroid and SVC do not have a feature importance parameter so the FeatureImportance was only calculated for the RandomForestClassifier.

Dataset

The dataset used contains attributes of students in secondary education of two Portuguese schools. The data was collected using school reports and questionnaires (Cortez). The subject the grades came from was math.

Table 1

Dataset label types

Label	Type
grades	numeric: 0-20
study time	numeric: 1-10 hours
failures	numeric: either 1-3 or 4
school support	binary: yes or no
extracurricular activities	binary: yes or no
internet access	binary: yes or no
absences	numeric: 0-93
sex	binary: 'F' - female or 'M' - male

To help forecast if the models would have poor predictive performance, I calculated the class balance. Class balance is a measure of if there are an equal number of examples for each class. I used Shannon entropy as a measure of this balance (2). The balance for attribute set A was 0.872802710377115 and the balance for attribute set B was 0.6677527043802982. With 0 being unbalanced and 1 being balanced, I would say both attribute sets were fairly balanced.

Equations 1 and 2 were used to compute the balance. n is the number of instances in the dataset, k is the number of classes with a size c_i.

$$H = -\sum_{i=1}^{k} \frac{c_i}{n} \, log \frac{c_i}{n} \tag{1}$$

$$Balance = \frac{H}{\log k} \tag{2}$$

Setting Up and Running the Experiment

I created the Python code in Visual Studio. I used an Anaconda environment to run the code. I used scikit-learn's train_test_split() function to split the data into training and testing data. Half of the data was withheld to do the accuracy evaluation. I did not do any cross validation.

The steps of the program are as follows:

1. Do computations for attribute set A
    a. Convert dataset from csv to array, only keeping the needed attributes
    b. Convert any string/text attribute values to a number (i.e. "yes"/"no" to 1/0)
        i. Attributes converted: schoolsup, activities, internet
    c. Split data into training and testing set
    d. For each model
        i. Create model
        ii. Fit the training data to the model
        iii. Make predictions
        iv. Compute accuracy score
2. Do computations for attribute set B
    a. Convert dataset from csv to array, only keeping the needed attributes
    b. Convert any string/text attribute values to a number (i.e. "yes"/"no" to 1/0 and "F"/"M" to 1/0)
        i. Attributes converted: schoolsup, activities, internet, sex
    c. Split data into training and testing set
    d. For each model
        i. Create model
        ii. Fit the training data to the model
        iii. Make predictions
        iv. Compute accuracy score
3. Plot results of the accuracy scores (seen in figure 1)
4. For each attribute set, compute explanations
    a. Compute correlation between features
    b. Plot the results (seen in figures 4a and 4b)
    c. Compute SVM feature importance
    d. Plot the results (seen in figures 3a and 3b)

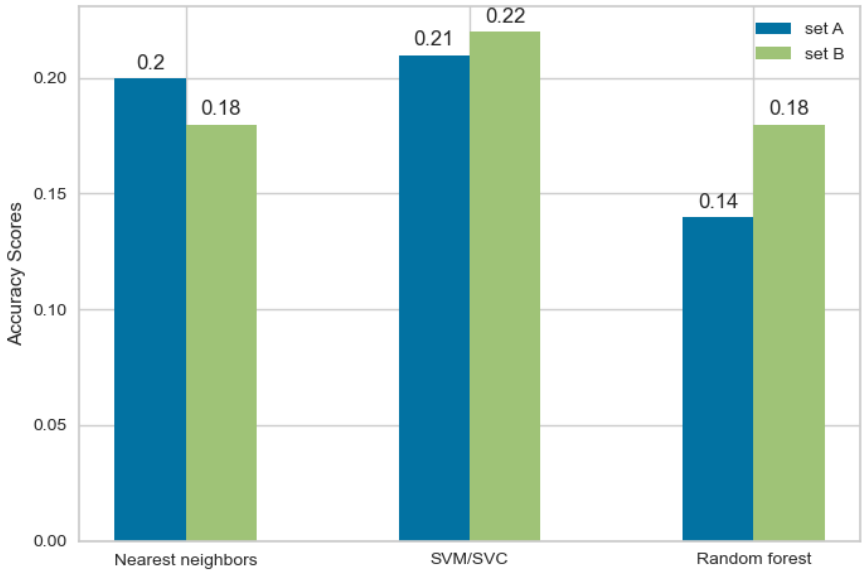
Experiment Results

Model Accuracies

Overall, the accuracies were very low which isn't a good thing. Attribute set B had a higher accuracy for 2 out of the 3 models. For both attribute sets, SVC had the highest accuracy. On the other hand, random forest had the lowest accuracies for both attribute sets (figure 1).

Figure 1

Accuracy Scores of Each Model for set A and B



Explainable Machine Learning

Sex did not appear to affect the relationship between study time and absences or failures and absences. There is a strong relationship between sex and study time. For both attribute sets, absences, study time and failures have a strong importance in the random forest classifier. For attribute set B, sex had the 4th highest importance and activities, internet, and school support had none.

Analysis

Analyzing the Experiment Results

The average accuracy (average of the accuracies from the 3 models used) of the attribute sets increased from .18 to .19 for set A and B respectively. Therefore, sex does increase the accuracy, by 5.55%, in predicting students' grades.

From comparing figure 3a and figure 3b, the importance of activities, internet, and school support dropped significantly when including sex as an attribute for the random forest classifier. On the other hand, the importance of absences, study time, and failures did not appear to be affected by having sex as an attribute.

Figure 3a

Feature Importance for Attribute set A

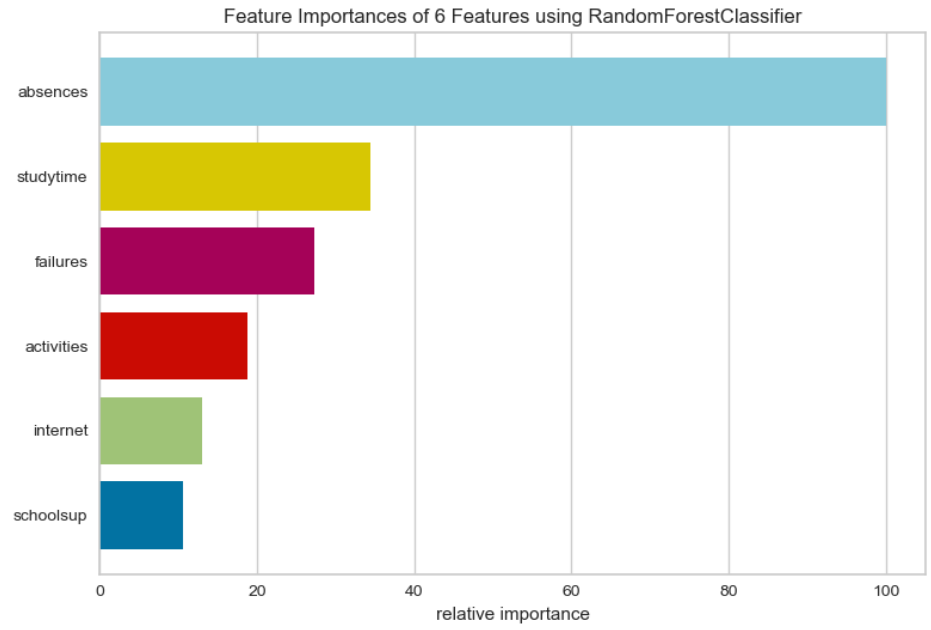
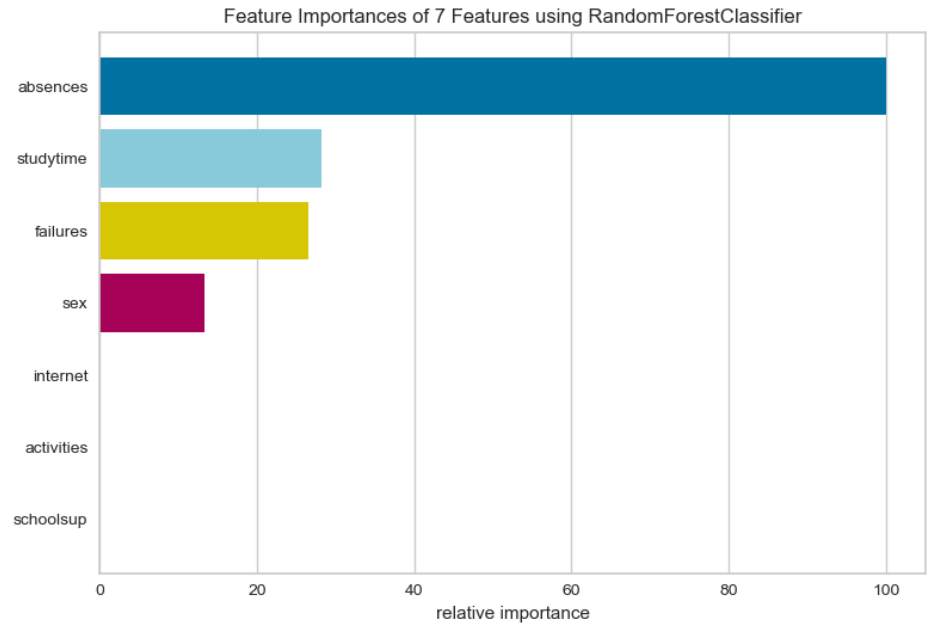


Figure 3b

Feature Importance for Attribute set B



I believe sex led to a higher accuracy because sex has a strong relationship with study time as seen in figure 4b.

Figure 4a

Colinear Relationships between Features without Sex

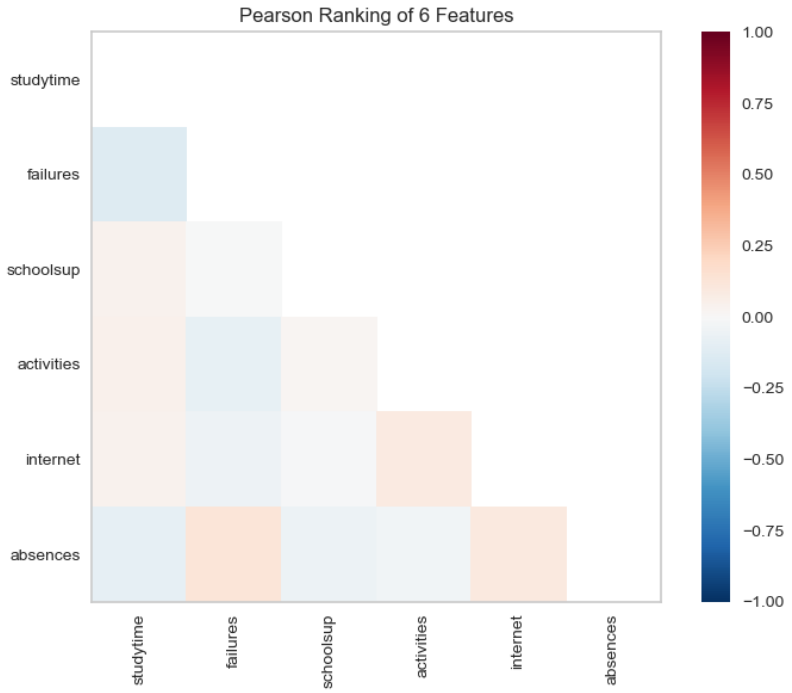
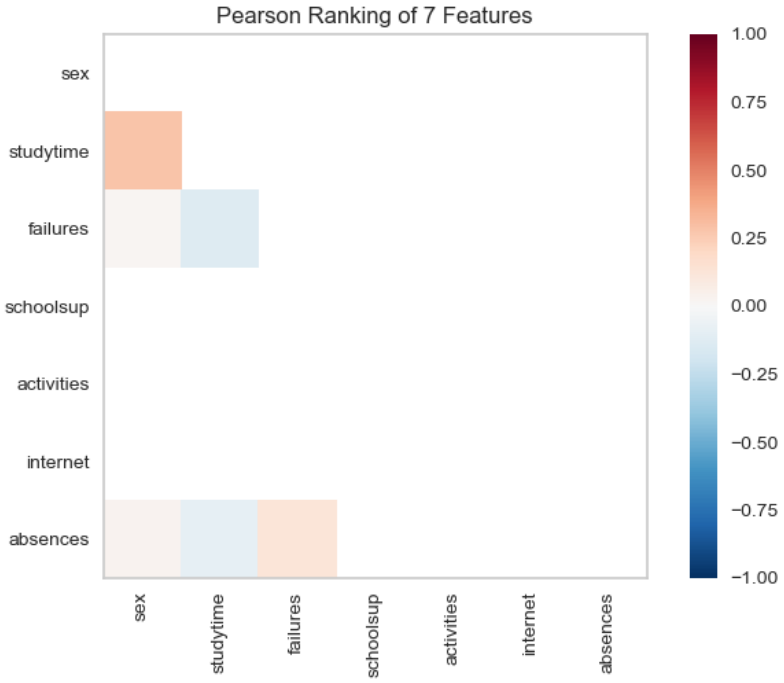


Figure 4b

Colinear Relationships between Features with Sex



From figures 3a and 3b, study time has a strong correlation with a students' grades and therefore has a high importance in both attribute sets.

Analyzing the Explanations

Yellowbrick's Visual Explanation Techniques

A general analysis of the visualizations used in Yellowbrick show that they convey the intended information. This was gathered from the types of the visual explanations (see table 2) and the purpose of the Yellowbrick visualization (see table 3).

Table 2

Different types of data visualizations and what are they good at conveying

Type	Information Conveyed
Bar/column	Shows a comparison among different sets of data
Line	Reveals trends, progress, or changes that occur over time, best with continuous data
Matrix chart	Shows relationships between two or more variables in a data set
Scatter	Shows a correlation between two variables, trade-offs, and outliers[a]

[a]Spence (2014)

Table 3

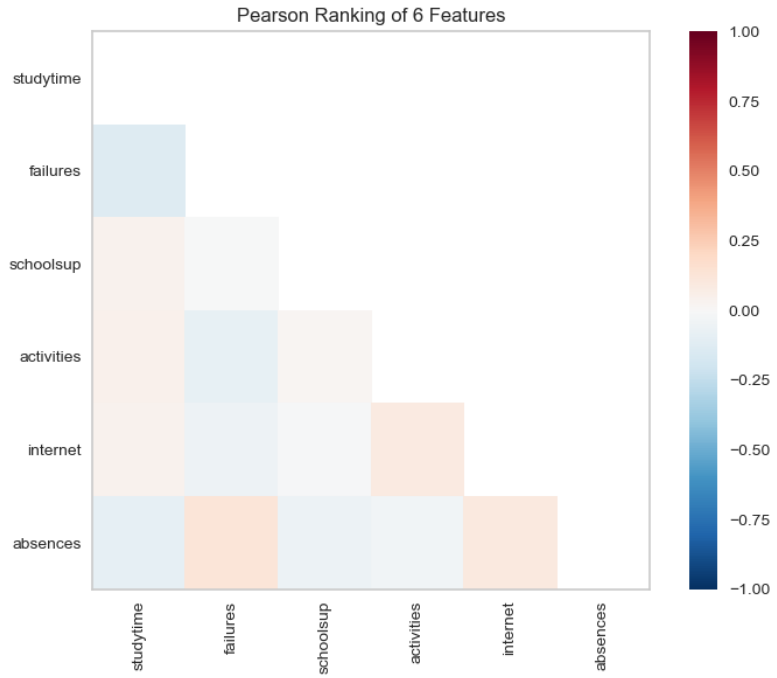
Yellowbrick visualizations grouped by the type of data visualization

Type	Yellowbrick Visualization
Bar	feature correlation, class balance, feature importance
Matrix	rank features, classification report, confusion matrix
Line	alpha selection, discrimination threshold
Star	RadViz (used to detect separability between classes)
Scatter	PCA projection, manifold
Stem	cook's distance
Silhouette	Silhouette visualizer
Lexical dispersion plot	Same name as the type
Parallel coordinate plot	parallel coordinates

For example, Yellowbrick uses a matrix graph to show the relationship between features (see figure 5a).

Figure 5a

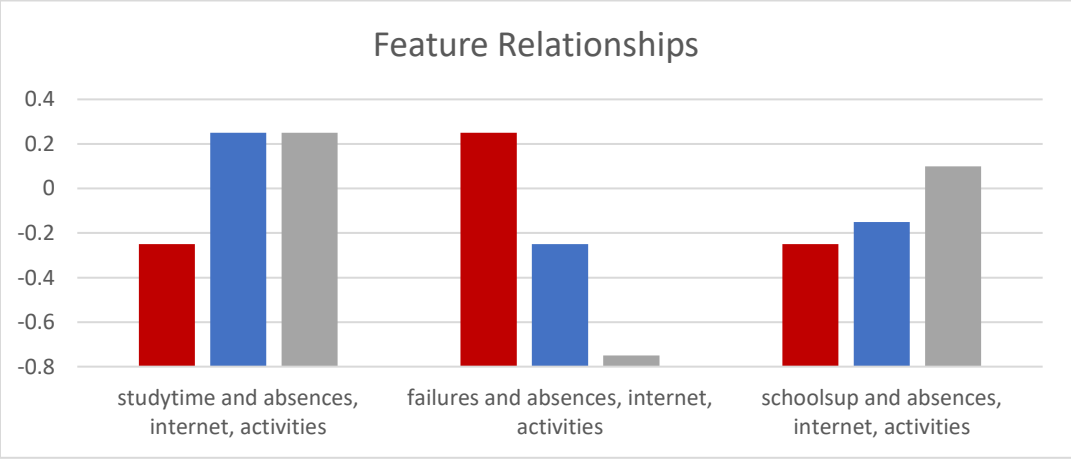
Colinear Relationships between Features without Sex



Alternatively, a bar chart could have been used where groups of bars represent a comparison of one feature to a few others and their relationship is determined using the y-axis (see figure 5b).

Figure 5b

Feature Relationships



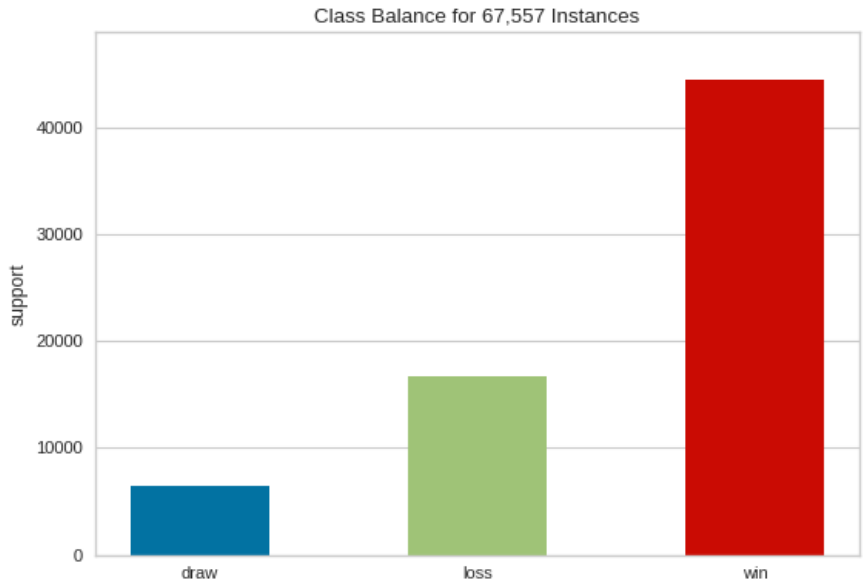
An example of how feature relationships can be conveyed via a bar chart

However, a matrix graph makes it easier to see and understand the relationships features have with a larger number of features.

Another example is the use of a bar chart to depict class balance (see figure 6a).

Figure 6a

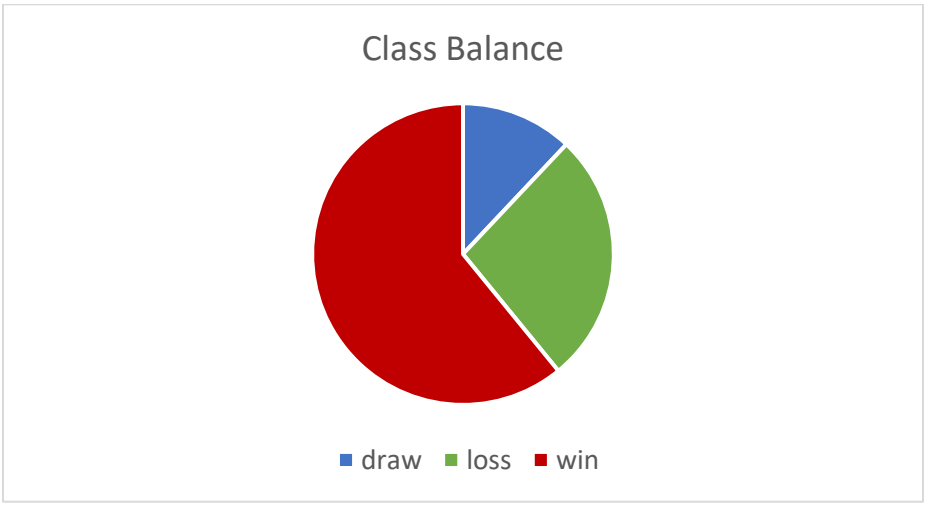
Class Balance example chart from Yellowbrick's documentation



A pie chart could be used to convey the same information (figure 6b), but the use of a bar chart clearly shows the frequency of the classes whereas the pie chart only conveys a ratio or percentage.

Figure 6b

A pie chart representation of the data from the chart in figure 6a



Even if numbers were added to a pie chart, a bar chart would display the information more clearly when showing large numbers of classes.

Explanations from the Experiment

Analyzing Yellowbrick's visualizations shows that the experiment's explanations, provided by Yellowbrick, have a good quality. The Rank2D visualizer allowed me to see how strong the relationships are between features. The FeatureImportances visualizer allowed me to see how important certain features were to the random forest classifier.

Social Science vs. Explainable Machine Learning Techniques

There are similarities between what social science says the relationship between sex and grades is and what the explanations say the relationship is. According to the explanations, sex increased the accuracy in predicting students' grades by 5.55%. Sex led to a higher accuracy because of its relationship with a students' study time. When ranking feature importance, sex had the 4th most importance in making the predictions.

According to social science, gender is not a strong predictor for mathematics performance (Lindberg, 2010). A study by Melhuish et al. (2008) compared how birth weight, gender, socioeconomic status, mother's education, father's education, family income, quality of the home learning environment, preschool effectiveness, and elementary school effectiveness effect predicting mathematics performance. In a similar study by Lindberg (2010), gender was the weakest of the predictors.

Both the explanations and social science agree that sex does not have a large impact on predicting students' grades. To keep my experiment simple, I excluded mother's and father's education levels from the training and testing data. I believe that if it were included in training and making predictions, the explanations would match what social science says even more.

Framework for Ethical Analysis

Machine learning models should be evaluated for their accuracy and ethical implications. Those implementing the models should consider the effects of using it in the real world. To do this they need to know why and how a model made a decision or reached a certain prediction. Defining this process can help guide machine learning developers in evaluating the accuracy and ethics of a model. The evaluation consists of three steps: interpret, understand, and question. After the questioning step, if any flaws in the model are found, changes should be made, and the process should repeat.

Interpret

Interpretation consists of gathering information on how and why a model made the decisions it made. This was done using Yellowbrick's Rank2D and FeatureImportances visualizers which showed the relationships between features and their importance to the model. Other useful information found was the accuracy scores in predicting students grades with and without sex as an attribute.

Understand

Understanding is the process of analyzing the information gathered when interpreting a model. This step might require further investigation into the data used or the model. Splitting the datasets used in the experiment by gender, the accuracy was higher in predicting female students' grades. Analyzing the datasets, the average scores of females was 10 and the average scores of males was 11. Alone, this information might not have much value. But questioning it can lead to important discoveries.

Question

Questioning means considering the ethical implications of a model and challenging the decisions it makes. It is important to verify that the explanations from interpreting a model make sense and can be backed by conclusions from previous studies on relevant subjects. Even if they are saying the same thing, the ethical implications of the model should be considered.

The information gathered in the understand step posed questions such as:

- What if the model had higher accuracy in predicting female students' grades because the grades were lower?
- Would the model still have the same accuracy if their grades were higher?
- If not, what does this mean?

A model that anticipates that female students will achieve lower mathematics grades despite data showing that females have achieved higher scores can be harmful. To consider what this means ethically, let's look at the use of a model that expects female students to achieve lower grades in real-world applications. From the applications mentioned in the Analysis of a Models Ethics section, this would have a very negative impact when used in the hiring process or for scholarships. A university using the model to decide what students need academic support would result in female students receiving more help than they might need. This alone does not imply any harm can be done but it can if the university places at risk students on probation.

Future Work

The proposed framework is intended to help model creators evaluate and make changes to a model and does not indicate when to stop this process. Future work can contribute to determining a way to measure when a model is accurate and ethical. Additional work can also evaluate more applications of the ethical framework presented.

Limitations

One of the explanations I selected for the models was feature importance. A stronger type of explanation for understanding the relationship between features would be a decision tree. Using a decision tree would have given me more insight on how features are connected and relate to each other and the model.

Conclusion

The explainability of a machine learning model is just as important as its accuracy. With machine learning applications being used in industries like helathcare, it is also important for the model to be ethical. Through application, a framework that consists of interpreting, understanding, and questioning models analyzed the ethical implications of a model that predicts students' grades.

In creating the grade predicting models, I found that introducing sex as attribute to the models resulted in a 5.55% increase in prediction accuracy. Visualizations provided by Yellowbrick showed that absences, study time, and failures were the most important features to the RandomForestClassifier. Evaluating Yellowbrick's visualizations showed that it gave suitable explanations for the information it intended to convey. In questioning the models, I found that the explanations and social science agree that sex does not have a large impact on predicting students' grades. To reach this conclusion I used Yellowbrick's visualizations along with research done by Hyde and Mertz (2009), Lindberg et al. (2010), and Melhuish et al. (2008).

There are risks associated with applications of a model that predicts students' grades. Making these types of predictions appear harmless at first and might have a positive impact in some cases. However, any model is at risk of harming the individuals it is trying to predict. Although the entirety of the proposed framework is important, the step of questioning is critical in making sure models are accurate and ethical because it can bring to light potential negative impacts and create an opportunity of improvement.

Annotated Bibliography

Choo, J., & Liu, S. (2018). Visual analytics For Explainable deep learning. *IEEE Computer Graphics and Applications, 38*(4), 84-92. doi:10.1109/mcg.2018.042731661

This paper discusses current techniques, potential challenges, and future research directions for deep learning visual explanations. This paper will provide me with information on current techniques. The more complex a model is, the more challenging it might be to represent it and its explanations visually. I am hoping this paper will show me how visual explanations are used in real world applications of machine learning since my project is simple in comparison.

Cortez, P. and Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. From https://www.kaggle.com/dipam7/student-grade-prediction

This is where the dataset was attained from.

Hansen, C. D., & Johnson, C. R. (2005). *The visualization handbook*. Amserdam Pays-Bas: Elsevier Butterworth Heinemann.

This book provides an overview of the field of visualization. It discusses current visualization software systems and research topics that are advancing the field. I plan on using this book to gain knowledge on creating visualizations. I believe this source might aid me in creating my own visual explanations.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences, 106*(22), 8801-8807. doi:10.1073/pnas.0901265106

This paper addresses if there are gender differences in mathematics performance. Gender performance in mathematics might vary across geographical locations. Since the dataset I used contained data from Portuguese students, I think using sources that don't focus only on U.S. data would help improve accuracy in my analysis.

Ippolito, P. (2020, October 13). Machine learning visualization. Retrieved April, 2021, from https://towardsdatascience.com/machine-learning-visualization-fcc39a1e376a

This article goes over a few visual techniques to represent different aspects of the machine learning pipeline. The article provides examples and code for various techniques. I plan on using this article to aid me in creating my own explanation techniques.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135. doi:10.1037/a0021276

This paper analyzes gender differences in mathematics performance. This is relevant since the grades my model predicted were mathematics grades. I plan on using the paper to understand what the relationship between sex and grades is.

Melhuish, E. C., Sylva, K., Sammons, P., Siraj-Blatchford, I., Taggart, B., Phan, M. B., & Malin, A. (2008). The early Years: Preschool Influences on Mathematics Achievement. *Science, 321*(5893), 1161-1162. doi:10.1126/science.1158808

This paper analyzes influences on mathematics performance. This is relevant since the grades my model predicted were mathematics grades. I plan on using the paper to understand what the relationship between sex and grades is.

Quade, M., Isele, T., & Abel, M. (2020). Machine learning control — explainable and analyzable methods. *Physica D: Nonlinear Phenomena, 412*, 132582. doi:10.1016/j.physd.2020.132582

This paper discusses in-depth mathematical analysis of an ML models decisions. The research extends previous work on symbolic regression methods to infer the optimal control of a dynamical system given one or several optimization criteria, or cost functions. I plan on using this research to get a better understanding of the mathematical analysis of explainable techniques.

Samek, W. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham, Switzerland: Springer.

This book discusses models, theory, and applications of interpretable and explainable AI and AI techniques. I am particularly interested in the chapter titled "Software and Application Patterns for Explanation Methods" which is about software and application patterns for explanation techniques. This source can give me more insight on libraries for and the implementation of current explainable AI techniques.

Spence, R. (2014). *Information Visualization an Introduction*. Cham: Springer International Publishing.

This book is about information visualization. I used the chapter titled “Representation” to gain insight on ways data can be transformed into images that are easily understood. The chapter provided me with information on a few representation techniques like scatter plots and star plots.

Visualizers and API. (n.d.). Retrieved April, 2021, from https://www.scikit-yb.org/en/latest/api/index.html

This is the API for Yellowbrick. Yellowbrick is the API I used to get explanations from my models. I used the documentation to see what types of visual representations were used for the different explanation tools Yellowbrick offers.

Zhou, J., & Chen, F. (2018). *Human and Machine Learning*. Cham: Springer International Publishing.

This book discusses the links between interpretability, trustworthiness, transparency, explanation, and visualization for machine learning. The chapter titled “Critical Challenges for the Visual Representation of Deep Neural Networks” is about the visual representation of neural networks and the representational challenges presented by the models. I plan on using this chapter to learn about past and current visual representations used.