

# Explainable Machine Learning for Public Policy

Sam Meyer	Shrestha Mohanty	Sung Joo Son	Monicah Wambugu
meyer_samuel@berkeley.edu	shrestha_mohanty@berkeley.edu	sungjoo_son@berkeley.edu	monicah_wambugu@berkeley.edu

## Abstract

Machine learning may provide opportunities for governments to make better decisions. However, with its inscrutability, it raises new challenges for accountability, which is essential for public policy. We built and iteratively improved a web application for analyzing logistic regression models. We ran focus groups to understand how the application would be used in public policy discussions. We found many ways to improve machine learning visualizations, but also found that an application that both lets users understand how a model works must also explain the concepts of the model. Our application was useful for public policy discussions among users experienced in machine learning, and with added educational tools, the application could be used by a wider audience.

# Introduction

---

Governments are beginning to use machine learning, and we created a user interface that allows the public to comment on logistic regression models used by the government. This is necessary because much of the public currently sees all machine learning as equally inscrutable, but our application shows how a model can be displayed, edited, and opened for public comment. While machine learning researchers focus on neural nets and other highly accurate models, governments have only begun to use machine learning, so logistic regression is in use by tools such as the COMPAS Recidivism Risk Scale.<sup>1</sup> While others have created visualizations specific to single datasets or for educational purposes, our goal was to build a tool that could be used by governments with a minimum of customization needed. Accordingly, it can train new models based on new datasets uploaded in a standard csv format.

## Background

---

### Motivation

There are many projects that apply machine learning to public policy. For example, public schools in Tulsa<sup>2</sup>, Mesa<sup>3</sup>, and Montgomery<sup>4</sup> use machine learning for early intervention to

---

<sup>1</sup> "COMPAS Scales and Risk Models Validity and Reliability." Northpointe Research and Development Department.  
<https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASSummaryResults.pdf>

<sup>2</sup> Huang, Charlotte. "Tulsa Public Schools: Preparing Kids for the Third Grade Turning Point." Data Science for Social Good. July 19, 2016. Accessed December 13, 2017.  
<https://dssg.uchicago.edu/2016/07/19/tulsa-public-schools/>.

prevent students from dropping out of school or graduating late. Local police departments including New York<sup>5</sup> and Miami<sup>6</sup> have adopted machine learning for crime forecasting. Northshore University Health System predicts cardiac arrest using electronic medical records.<sup>7</sup> The use of machine learning for public policy will very likely increase as the use of machine learning proliferates in various domains. When important elements of governance use software implementations, public policies should have accountability mechanisms, including explainability and contestability, to better earn the trust of the public. However, for the implementations that use machine learning, explainability and contestability are more challenging to implement.

## Accountability in Public Policy

A public policy may make decisions about individuals according to rules and conditions agreed upon by the public. Sarah Lister of the UN Development Program defined accountability as “the obligation of power-holders to take responsibility for their actions” and described the accountable relationship between citizens and their government as an exchange where citizens give power to their government, and the government returns the explanation and justification of its use of power to citizens for taking corrective measures if necessary.<sup>8</sup>

---

<sup>3</sup> Su, Edward. "Mesa Public Schools: Undermining Undermatching." Data Science for Social Good. December 09, 2015. Accessed December 13, 2017.

<https://dssg.uchicago.edu/2014/01/16/mesa-public-schools-undermining-undermatching/>.

<sup>4</sup> Bhanpuri, Nasir. "Early Warning Systems for Struggling Students." Data Science for Social Good. December 09, 2015. Accessed December 13, 2017.

<https://dssg.uchicago.edu/2014/11/20/early-warning-systems-for-struggling-students/>.

<sup>5</sup> Nahmias, Laura, and Miranda Neubauer. "NYPD testing crime-forecast software." Politico. July 08, 2015. Accessed December 13, 2017.

<http://www.politico.com/states/new-york/city-hall/story/2015/07/nypd-testing-crime-forecast-software-090820>.

<sup>6</sup> Smiley, David. "Not science fiction: Miami wants to predict when and where crime will occur." Miami Herald. April 23, 2015. Accessed December 13, 2017.

<http://www.miamiherald.com/news/local/community/miami-dade/article19256145.html>.

<sup>7</sup> Somanchi, Sriram, Samrachana Adhikari, Allen Lin, Elena Eneva, and Rayid Ghani. "Early prediction of cardiac arrest (code blue) using electronic medical records." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2119-2126. ACM, 2015.

<sup>8</sup> Lister, Sarah. "Fostering social accountability: from principle to practice - a guidance note." August 2010. Accessed December 13, 2017.

Accountability is what empowers citizens to correct their government's behavior. To have accountability, citizens must be able to know what the government is doing and argue with it.

Accountability can be represented in the form of a feedback mechanism shown in Fig. 1 (based on Steets 2010).<sup>9</sup> A government or public institute behaves according to its obligations and expectations. Information about the behavior can be provided voluntarily or on demand by citizens. The citizens then evaluate the information and return positive or negative sanctions to the government. Here “sanctions” include voting, writing letters to representatives, making political donations, or other political activity. The government should modify its behavior to avoid negative sanctions. The quality of citizen evaluation and sanctions will depend on the quality of information. If information is of low quality, it is hard for citizens to respond. This happens intentionally when the NSA hides activity that would be objectionable to citizens, but it can also happen unintentionally when algorithms are used to implement policy.

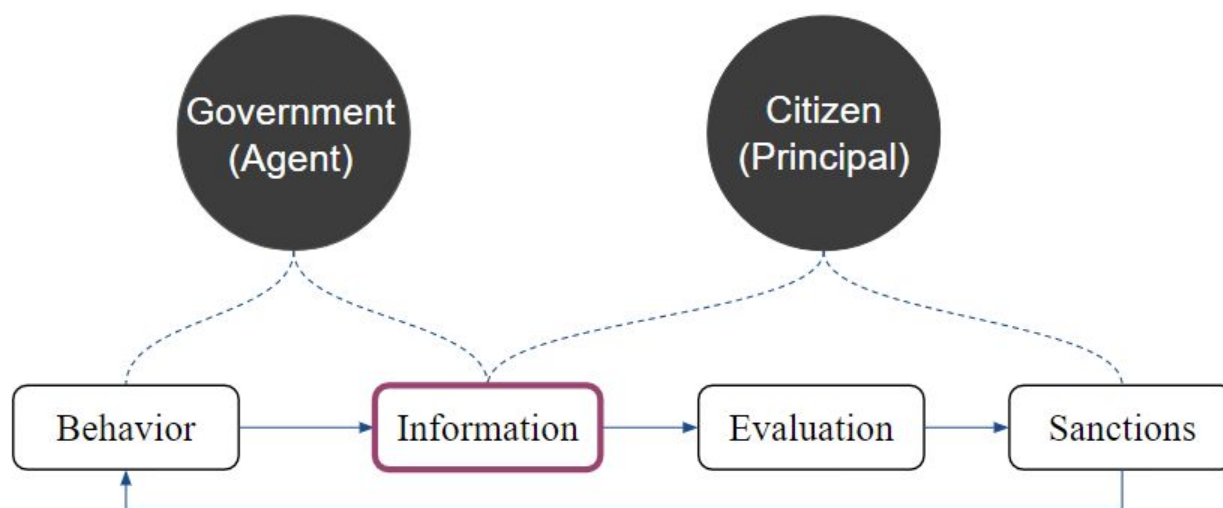


Figure 1. Accountability Mechanism (Steets, 2010)

<http://www.undp.org/content/dam/undp/library/Democratic%20Governance/OGC/dg-ogc-Fostering%20Social%20Accountability-Guidance%20Note.pdf>.

<sup>9</sup> Steets, Julia. Accountability in public policy partnerships. Palgrave Macmillan, 2010: pp. 14-26.

# Algorithmic Accountability

The World Wide Web Foundation defines accountability as the “obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.”<sup>10</sup> The definition of algorithmic accountability uses the same mechanisms as accountability in conventional public policy. When an algorithm is implemented into a information system for public policy, the system becomes the agent of behavior to whom a government delegates its authority.

However, public policies using algorithms have more obstacles to explainability than conventional ones without algorithms. Introna (2016) argues that algorithms are powerful and dangerous because of their inscrutability and executability.<sup>11</sup> Inscrutability means human cannot directly inspect object or machine-executable codes and relevant expertise is required to read source code. Moreover, some programs are too large to understand. For example, the Linux kernel version 4.14.5 consists of more than 20 million lines of code.<sup>12</sup> Executability means code can operate automatically without human intervention and be hidden in the background of society. Therefore, algorithms can be more dangerous actors because it is hard for citizens to notice when they have problems. This gives rise to the need for explainability of the algorithm.

---

<sup>10</sup> “Algorithmic Accountability: Applying the concept to different country contexts.” World Wide Web Foundation. July 2017. Accessed December 13, 2017. [http://webfoundation.org/docs/2017/07/Algorithms\\_Report\\_WF.pdf](http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf): p.11.

<sup>11</sup> Introna, Lucas D. "Algorithms, governance, and governmentality on governing academic writing." *Science, Technology & Human Values*, January 2016, Vol. 41

<sup>12</sup> Löhner, Christin. "Funny Statistics for the Linux Kernel - Lines of Code, Bad words, Good words - The Linux Counter Project - Statistics about Linux, its Users and more." Linux Counter. Accessed December 13, 2017. <https://www.linuxcounter.net/statistics/kernel>.

## Explainability

DARPA describes explainability as the ability to explain machine's decisions and actions to human users. The agency states that "new machine-learning systems will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future."<sup>13</sup> While DARPA viewed explainability from the perspective of utility and predictability, the EU's GDPR mentions explainability and accompanying contestability from the perspective of human rights. The regulation states that the subjects of automated individual decision making should have the right not to be subordinate to the decisions, to express their own perspectives, to obtain explanations about the decisions, and to challenge the decisions.<sup>14</sup>

Algorithmic explainability is particularly relevant with respect to machine learning because it can make detailed decisions automatically. At the same time, machine learning algorithms have more obstacles to explainability than other types of algorithms. The common solutions to help various stakeholders to understand an algorithm and to anticipate its behaviors are providing open source code and requesting expert review. However, each of these solutions has its own limit when it comes to machine learning.

## Limits of Open Source

Open source code was billed as the great solution to algorithmic opacity by many developers. As Gnu.org states, open code gives "The freedom to improve the program, and

---

<sup>13</sup> Gunning, David. "Explainable Artificial Intelligence (XAI)." Defense Advanced Research Projects Agency. Accessed December 13, 2017. <https://www.darpa.mil/program/explainable-artificial-intelligence>.  
<sup>23</sup> "General Data Protection Regulation - European Commission." European Commission. April 27, 2016. Accessed December 13, 2017.

<sup>14</sup> "General Data Protection Regulation - European Commission." European Commission. April 27, 2016. Accessed December 13, 2017.  
[http://ec.europa.eu/justice/data-protection/reform/files/regulation\\_oj\\_en.pdf](http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf).

release your improvements to the public, so that the whole community benefits.” Machine learning destroys this promise. Even when given the learned parameters of a model, it can be hard for skilled computer scientists to explain how a program made its decisions. However, open source code still has its place when using machine learning. Given that the input features for machine learning models are often computed from other values, open source code (or at least clear definitions of how features are calculated) may be required for machine learning algorithms to be truly explainable. Many implementations of machine learning use input data that has itself been computed from other factors. For example, an important factor for a model could be local prescription drug usage, normalized by population density and resident age. If an explainable machine learning algorithm uses this kind of computed inputs, the code for computing those factors must be open source for explainable machine learning to enlighten outsiders.

## Limits of Expert Review

As Burrell (2016)<sup>15</sup> points out, significant technical ability is required to understand most computer code, including the algorithms implemented by it. This is consistent with other complex systems that expect expert review for accountability. When using machine learning, other expert software developers may be able to replicate machine learning models given the same code and data, but they cannot always clarify the logic in a way that reveals the values of the model. Government experts are more used to creating clear public policy decisions that can be defended later, so they desire clearer rules than most machine learning provides (Veale 2017).<sup>16</sup> These public policy experts are often willing to sacrifice accuracy in machine learning to

---

<sup>15</sup> Burrell, Jenna. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society* 3, no. 1 (2016): 2053951715622512.

<sup>16</sup> Veale, Michael. "Logics and practices of transparency and opacity in real-world applications of public sector machine learning." *arXiv preprint arXiv:1706.09249* (2017).

create a simpler set of rules. They believe this set of rules is easier to measure for fairness and policy implications. When working in governments, explainable machine learning can enable expert review to elucidate the effects of machine learning models.

## Explainable Machine Learning

What machine learning models are explainable? Here we limit ourselves to models where the inputs are a list of factors, avoiding the cases where inputs are image or sound, such as in Hendricks (2016).<sup>17</sup> Freitas (2015)<sup>18</sup> considered the comprehensibility of decision trees, classification rules, decision tables, nearest neighbors, and bayesian network classifiers, and found decision trees to be most comprehensible. Martens (2011)<sup>19</sup> considered linear models (including logistic regression) to be just as explainable as decision trees. Lipton (2016)<sup>20</sup> questions whether we have any clear definitions of transparency and explainability, so any claim to explainability must clarify the goals. Because we want models that will increase trust in fair outcomes, it is important that explanations provide an overall view of the model (which Lipton calls transparency). Here, we believe logistic regression meets the requirements.

## Explainability and Contestability

Public policy has been focused on giving people the right to an explanation when decisions are made about them as individuals. However, an individual explanation alone does not allow for civil society to organize public responses to the important details of a model. This is

---

<sup>17</sup> Hendricks, Lisa Anne, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. "Generating visual explanations." In *European Conference on Computer Vision*, pp. 3-19. Springer International Publishing, 2016.

<sup>18</sup> Freitas, Alex A. "Comprehensible classification models: a position paper." *ACM SIGKDD explorations newsletter* 15, no. 1 (2014): 1-10.

<sup>19</sup> Martens, D., Vanthienen, J., Verbeke, W., and Baesens, B. Performance of classification models from a user perspective. *Decision Support Systems* 51(4): 782-793. 2011.

<sup>20</sup> Lipton, Zachary C. "The mythos of model interpretability." *arXiv preprint arXiv:1606.03490* (2016).



where we bring in the idea of contestability: can people not only understand what the algorithm is doing, but can they improve it? Contestability exists to a small extent when you tell Netflix that the movie it recommended is terrible, but systems that implement public policy should allow many parties to contest the overall decisions.

As Hirsch (2017)<sup>21</sup> points out, in high-stakes decision-making by machine learning systems, users need to not only report when the model makes poor predictions, but must “make arguments to powerful actors whose decisions are informed by those systems.” Simply reporting that a decision is wrong is not enough information to improve it. We apply this to arguments over about the aggregate effect of the model.

	<b>Explainability</b>	<b>Contestability</b>
<b>Question</b>	“How does it decide?”	“How do I fix it?”
<b>User</b>	Designer, engineer, auditor	Civil society, end user
<b>Purpose</b>	To understand	To improve

Table 1. Explainability and Contestability

## Public Trust in Contestable Machine Learning

Based on Lee and Baykal (2017),<sup>22</sup> people are more comfortable with decisions made through discussion even when algorithms can make decisions that are technically more fair. Explainability will allow people to begin with the decisions of a machine learning algorithm and let people argue over the decisions. This may allow people to better trust the results of the machine learning.

<sup>21</sup> Hirsch, Tad, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. "Designing Contestability: Interaction Design, Machine Learning, and Mental Health." In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pp. 95-99. ACM, 2017.

<sup>22</sup> Lee, Min Kyung, and Su Baykal. "Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division." In *CSCW*, pp. 1035-1048. 2017.

We want explainability when social decisions can make *better* choices even when those choices lead to lower accuracy. (“Better” in the sense of making value judgements.)

Trust will be built slowly by the public, but new tools could help to make contestable machine learning available to the public. If the public can interact with machine learning models that are in use, it will likely increase their trust in those models as government agents.

## Fair Machine Learning

In 2016, the White House released a report recommending safeguards that can be used both in the private as well as the public sector to prevent biased systems from being deployed.<sup>23</sup> In this paper, researchers called for “equal opportunity by design” which sought to prevent “unequal access to opportunity.” Understandably, it is in the government’s interest to ensure that “people of equal talents and ambition can achieve equal outcomes over the course of their lives.”<sup>24</sup> In academia, researchers are tackling the same problems by proactively exploring ways of “bias mitigation.”

One simple but naive approach is ignoring protected classes altogether. Although it is simple to implement, this approach ignores the fact that these protected classes are sometimes strongly correlated with other variables that are not protected. Historically, the correlation between ZIP codes and race for example led to redlining of neighborhoods in the US.<sup>25</sup> Banks and lending institutions did not explicitly consider race but still denied one particular group of people access to mortgages.

---

<sup>23</sup> Executive Office of the President, et al. *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President, 2016.

<sup>24</sup> Solon Barocas and Moritz Hardt. “Fairness in Machine Learning: NIPS 2017 Tutorial.” (2017). <http://mrtz.org/nips17/#/16>

<sup>25</sup> Zenou, Yves, and Nicolas Boccoard. “Racial discrimination and redlining in cities.” *Journal of Urban economics* 48.2 (2000): 260-285.

An alternative approach is demographic or statistical parity. In this case, even though protected classes may or may not be included (depending on the relevant policies), there is an added constraint of independence. Given two groups of people, the conditional probability of a certain prediction should be the same in both groups. In a hypothetical scenario of college admission, the probability of getting admitted  $P(C = 1)$  should be the same for both men and women.

$$P(C = 1 \mid A = 'male') = P(C = 1 \mid A = 'female')$$

Although demographic parity is a great improvement in attainment of fairness, it conflicts more directly with the business objectives of accuracy. Real world data samples are often skewed. In the hypothetical scenario of college admissions it might be that more data exists about men than women for example. A machine learning model with embedded demographic parity would do better to predict outcomes in the case of men as compared to women. Inevitably, the model would perform less optimally resulting in a fairness versus accuracy tradeoff. Secondly, if college admission is correlated with gender, enforcing the demographic parity constraint would also result in a significant decrease in accuracy.

Hardt et. al introduced a new constraint of separation.<sup>26</sup> This algorithmic constraint ensures that the predictor is independent of a sensitive attribute (e.g. gender) conditional on the target variable. This constraint allows for sensitive attribute to be correlated with the target variable, but does not address any pre-existing biases in the data. In this project, we implement this algorithmic constraint by requiring that the true positive rate and false positive rates across groups be the same. Our “*balance*” feature implements fairness by adjusting the thresholds

---

<sup>26</sup> Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems*. 2016.

between the two groups. The optimal thresholds are selected as the point of intersection of the ROC curves for the two groups defined by the sensitive attribute (Fig. 2).

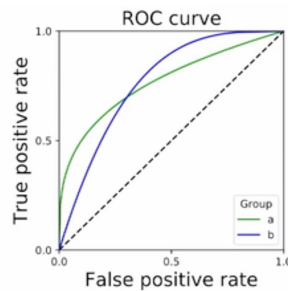


Figure 2. Finding the intersection of ROC curves to implement equality of opportunity. From Moritz, Solon NIPS 2017<sup>27</sup>

## System Design

---

We built a web application that allows users to visualize logistic regression models. They can then view, modify, and test changes to the logistic regression machine learning model by visualizing decision factors, accuracy, and confusion matrices of the model. Users may add comments to factors to share opinions or suggestions about them.

In an effort to make our system generalizable, new models can be added by uploading a new csv with the data. The new data must be in a standardized format, but users can use the existing datasets as examples.

Our code is open source (MIT license) and can be downloaded from our Github page.<sup>28</sup>

## System Architecture Overview

The system has an architecture of common 3-tier web applications (Fig. 3).

---

<sup>27</sup> <https://vimeo.com/248490141>

<sup>28</sup> [https://github.com/shresh02/explainable\\_ML\\_public\\_policy](https://github.com/shresh02/explainable_ML_public_policy)

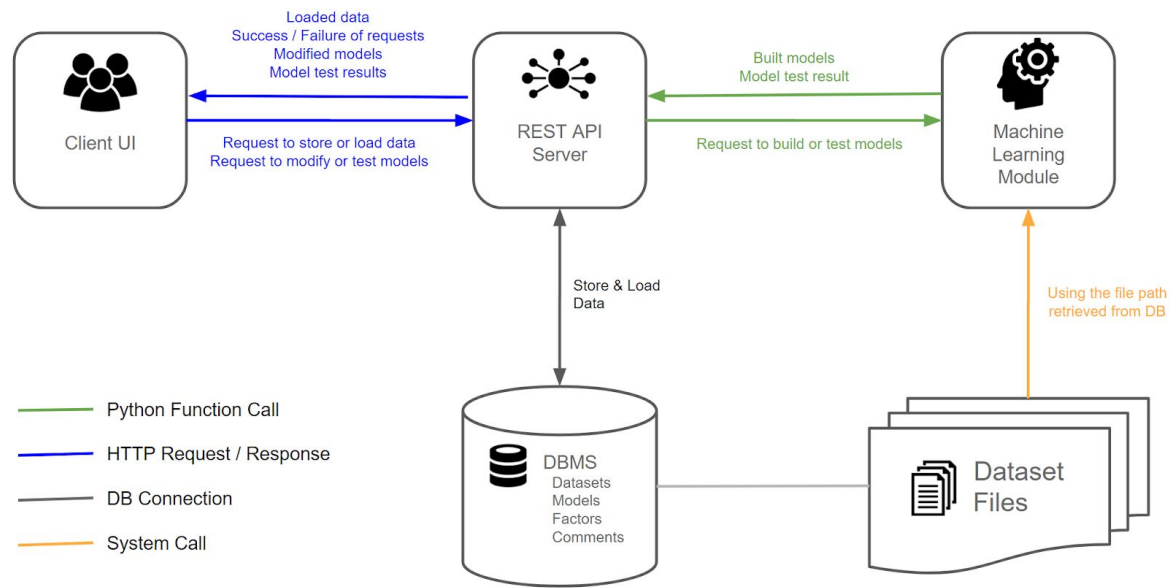


Figure 3. System Architecture Overview

Component	Technology	Description
Client UI	<ul style="list-style-type: none"> <li>HTML5/CSS3</li> <li>React</li> </ul>	<ul style="list-style-type: none"> <li>To support dynamic user interactions on the client side</li> </ul>
REST API Server	<ul style="list-style-type: none"> <li>Django REST framework</li> </ul>	<ul style="list-style-type: none"> <li>For faster implementation of database interactions</li> <li>For easier integration with machine learning module using Python</li> </ul>
DBMS	<ul style="list-style-type: none"> <li>SQLite</li> </ul>	<ul style="list-style-type: none"> <li>For faster and efficient implementation</li> <li>The website does not require heavy DBMS specifications for experiment</li> <li>DBMS can be easily changed because the website uses Django to interact with DBMS</li> </ul>
Machine Learning Modules	<ul style="list-style-type: none"> <li>Pandas</li> <li>Patsy</li> <li>Sklearn</li> <li>Scipy</li> </ul>	<ul style="list-style-type: none"> <li>Python is a popular language for machine learning</li> <li>Python machine learning and statistical analysis libraries</li> </ul>
Dataset Files	<ul style="list-style-type: none"> <li>CSV (Comma Separated Values)</li> </ul>	<ul style="list-style-type: none"> <li>Easy to find open datasets in CSV format</li> <li>High compatibility</li> <li>Python libraries support CSV</li> </ul>

# Components

## **Client UI**

The client UI is composed of a single-page web application. The interface is designed for a personal computer environment using a large screen, keyboard, and mouse. Users can create, review, modify, and give comments on logistic regression machine learning models. It uses React to implement real-time user interaction and to minimize server load by handling the most of the data manipulations on the client side. For example, users can adjust weights of decision factors easily and quickly without waiting for a response from the server. The UI calls server APIs only to store, load, train, and test models.

## **REST API Server**

REST API is a common way to implement web based application. It is easy to integrate server application modules with web-based client modules because the client and server communicate through HTTP requests. The API server connects the UI, DB, and machine learning module to each other. By using the Django REST framework, we were able to reduce the amount of code to build APIs. We did not need to write or execute any SQL query to create or manipulate the database because all database transactions were automated by the framework. As the server modules for Django were written in Python, Python function calls interact with machine learning modules which are also written in Python.

## **DBMS**

SQLite is a lightweight DBMS which stores the whole database in a single file. It is the default database for the Django framework. As we are only using the database to store data and do not

need to support and manage heavy workloads, SQLite works well. Moreover, the DBMS can be easily changed if the application needs to be scaled to support more users. The Django framework makes it easy to replace SQLite with MySQL, PostgreSQL, or other standard databases better suited for a larger deployment of the application.

## **Machine Learning Modules**

The machine learning module is used to train and test logistic regression models with given user settings. Our models use logistic regression with L2 regularization when training models, based on the sklearn implementation. We randomly used 75% of the uploaded csv for the training set and the remaining 25% as the testing set. This prevents users from seeing changes due to new random training sets being selected rather than changes to the model. In addition, to implement fair machine learning, we calculated ROC curves for both negative and positive classes of the variable of interest, then found the intersection point. With small datasets, it is possible that there will be multiple intersection points, so we chose the points closest to 0.5, the default threshold. As you can see in Fig. 4, multiple overlaps may occur due to the discrete nature of the ROC curve, especially near the minimum and maximum threshold values.

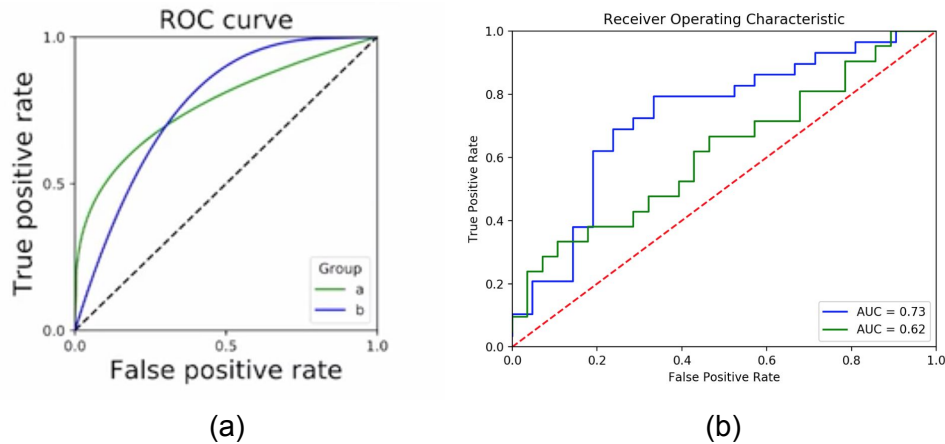


Figure 4. (a) Ideal ROC curves for dataset with an infinite number of testing points compared to (b) the ROC curve for our real-world student grade performance dataset, with 300 rows in the training set. The ROC curve for male students is in blue, and the curve for female students is in green.

## Dataset Files

We used a student performance dataset<sup>29</sup> for our experiment. The dataset was in CSV format. CSV is a simple and common format for data. Most of datasets in public domain supports CSV format. Datasets could be uploaded to the DBMS through the client UI. As the file itself would be stored in the server file system while the DB had store metadatas like file path, Machine Learning Modules directly accessed the file using the metadata from DBMS.

One of the core goals of the project was to make a system that would generalize to data sets other than our original test set. We have tested this by using it on a second set of data: DHS data from US AID. The model predicts which households are poor on the basis of a variety of factors.

<sup>29</sup> P. Cortez and A. Silva. 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. <http://www3.dsi.uminho.pt/pcortez/student.pdf>



## Database Design

The database of the system consists of 5 tables. The Dataset table stores information about the dataset. CSV files for the datasets were stored in the local file system of the server. The MIModel table stores information about logistic regression machine learning models. Factors are stored in a separate table because a model might have more than one factor. Each factor might have more than one comment from user. The comments are stored in a separate table. MIModelDetail was used for additional attributes such as test results. A full description can be found in Appendix C.

## User Interface Design

We used the following steps to develop the UI design (Fig. 5).



Figure 5. UI development process

### User Story Map

The following image shows the the user story map which lists functional requirements and corresponding tasks (Fig. 6)

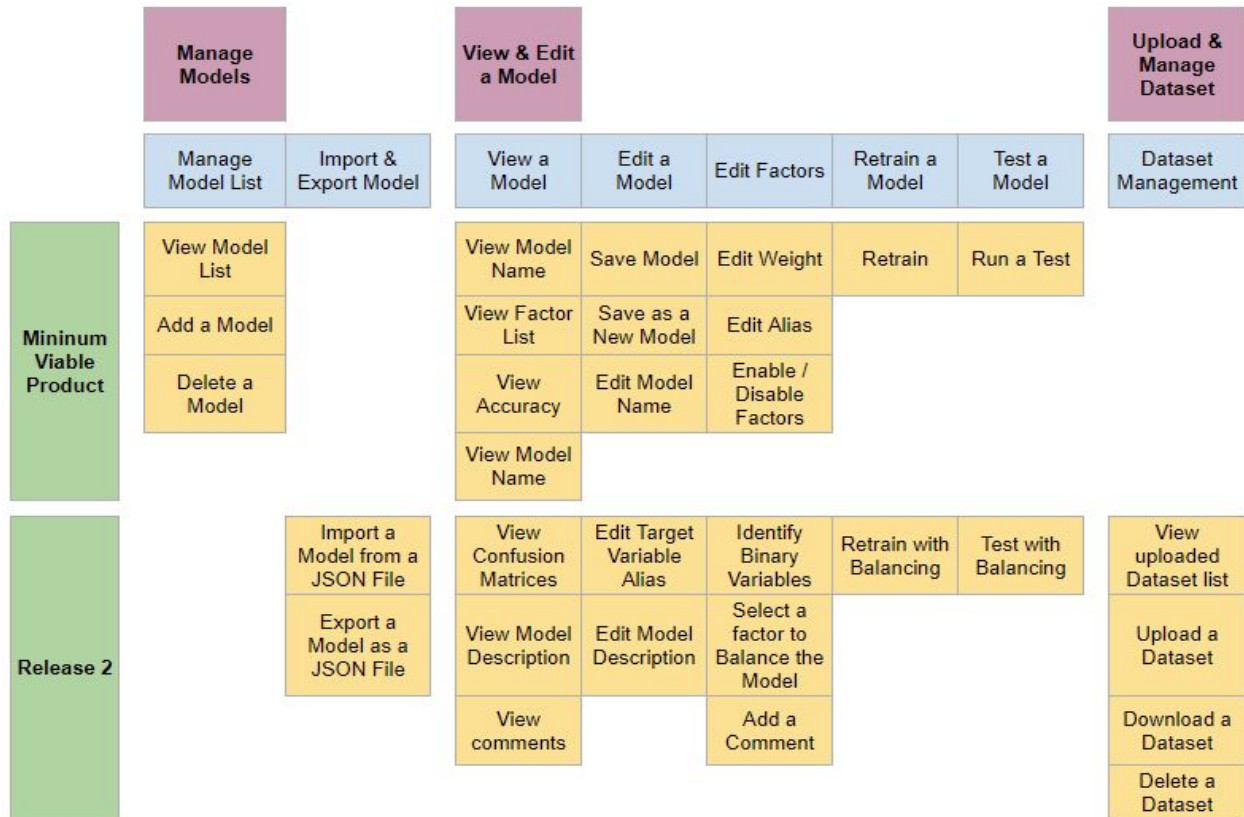


Figure 6. User Story Map

## Paper Sketches

Paper sketches were used while ideating for the first draft of key user interfaces (Fig. 7).

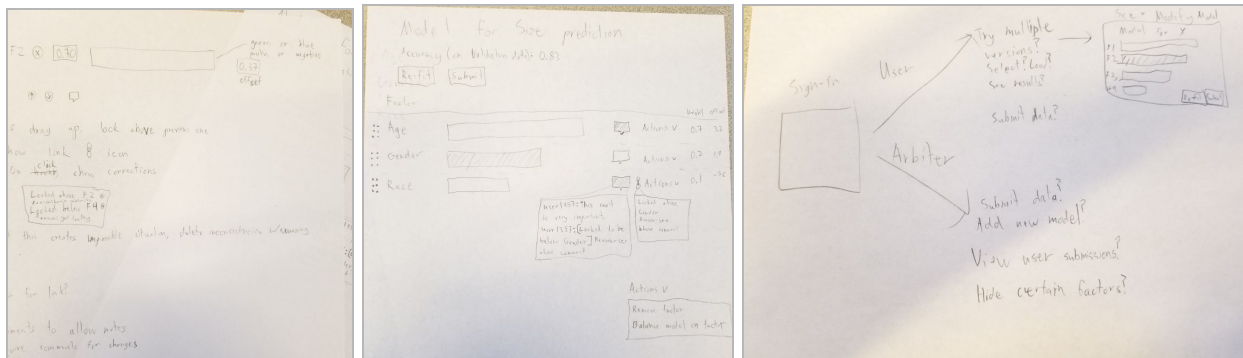


Figure 7. Paper Sketches

## UI Wireframes

After a few iterations on the UI design, we converted paper sketches into UI wireframes. It helped us discuss user interaction and the aesthetics of visual elements (Fig. 8).

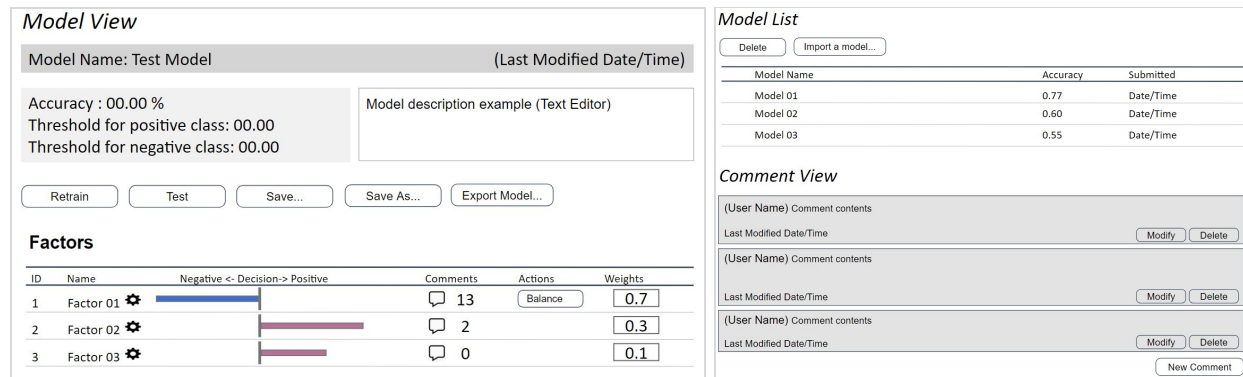


Figure 8. UI Wireframes

## Implementation and user testing

The user interface was implemented in HTML5/CSS3 and React based on the wireframes. We started user testing after the initial prototype was created, and the application has continuously evolved based on user feedback.

## User Interface Overview

### Model View

The “Model View” is the interface to view and edit a model (Fig. 9). It is the key interface of the website. All information about a model is presented in a single webpage. The upper part of the page displays the model type, the name of the predicted variable, a description about the model, accuracy, and confusion matrices. The lower part of the page displays a factor list where users interact with the system by exploring and adjusting weights on factors.

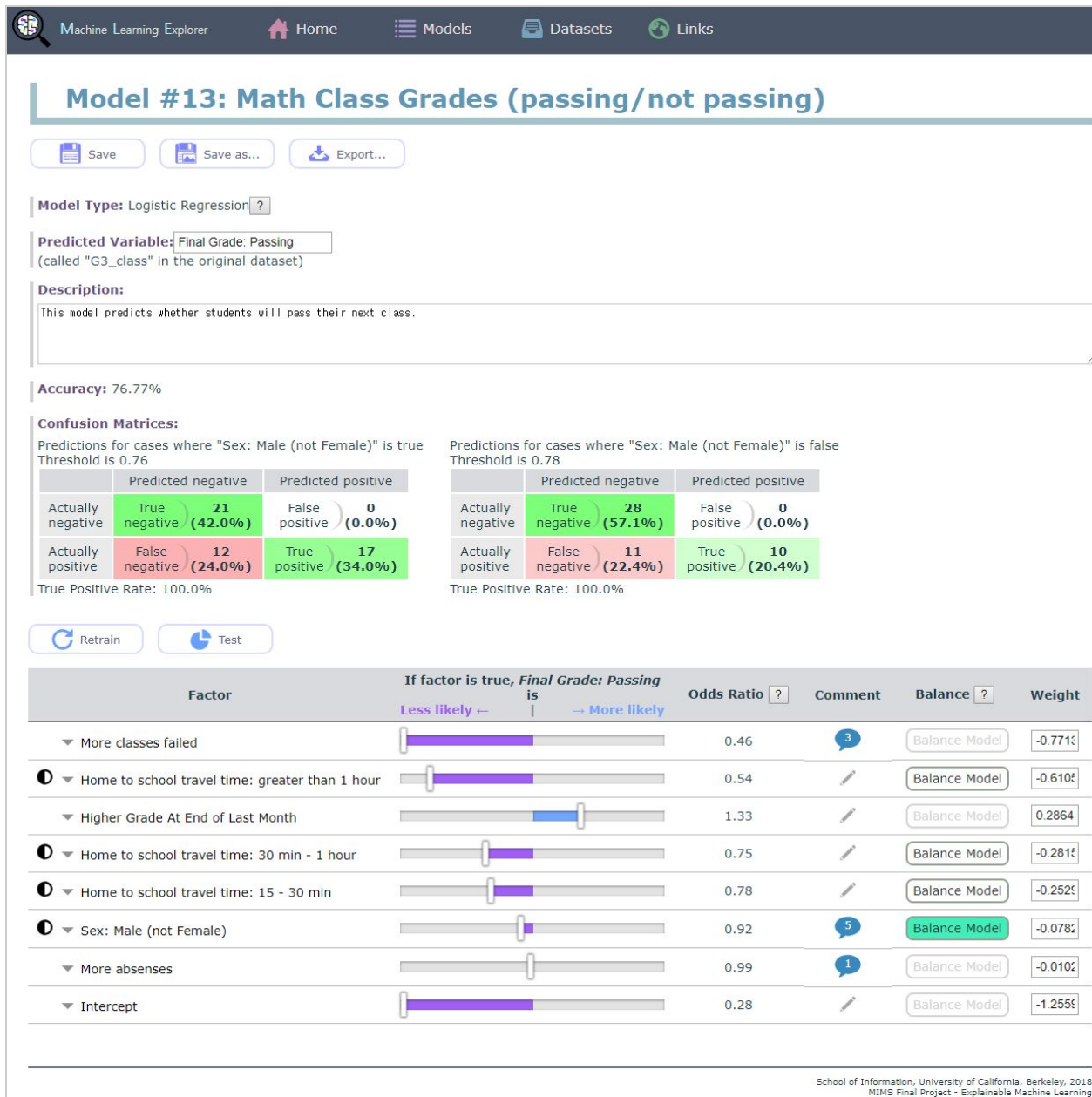


Figure 9. Screenshot of Model View

Users can retrain model weights, respecting any disabled factors, by clicking the “Retrain” button. To enable or disable a factor, a user needs to click the triangle beside the factor name and use the checkbox in the factor property popup to modify its state (Fig. 10). The circle with contrasting black and white labels a row as a binary factor.

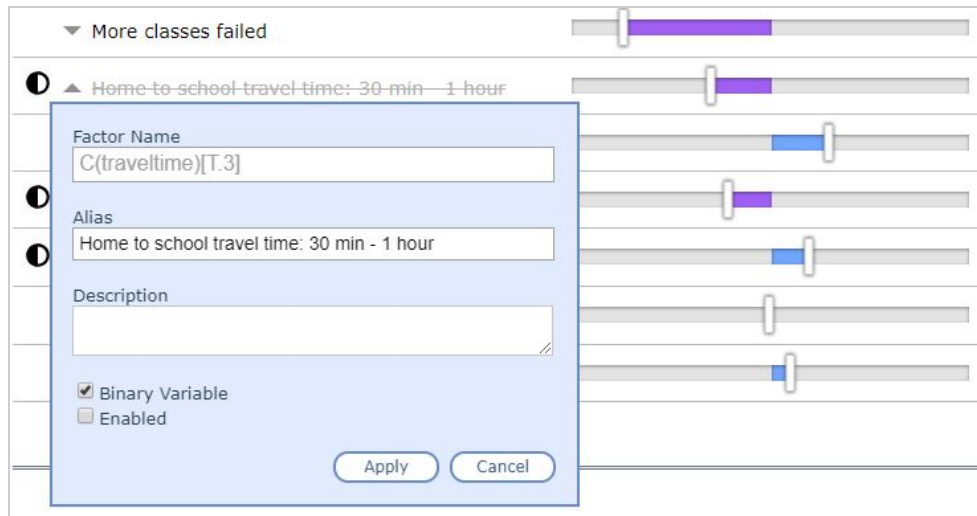


Figure 10. Factor Property Popup

After retraining a model, a user can test the model’s accuracy by clicking the “Test” button right beside the “Retrain” button. The test result will be displayed in a popup and in the upper part of model view as well (Fig. 11). Both show the same result but are represented differently. The popup supports a more intuitive visualization of the confusion matrix and was designed for users with lesser experience in machine learning. If users change the weight of a factor, they can run a “Test” to see what effect that change has. To change the weight, users drag the sliders to the desired location or edit weight values in the rightmost column.

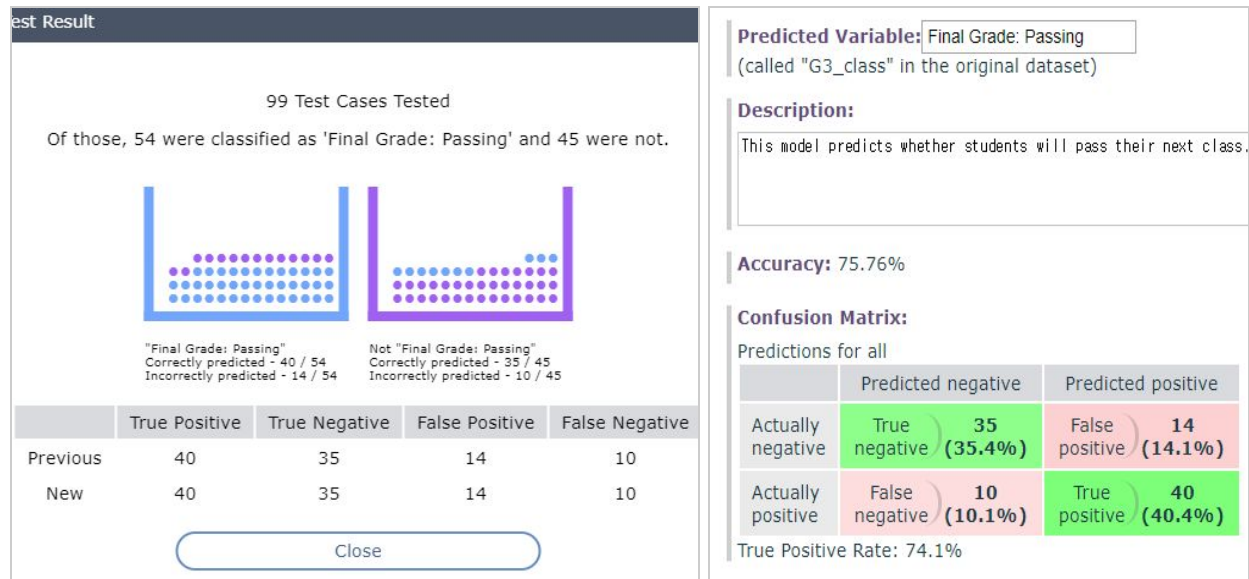


Figure 11. Test Result in a Popup and in the Model View Page

The “Balance” buttons in each row act like radio buttons. The buttons are enabled only for binary variables. If a user selects a factor to be balanced and retrain the model, the test result will be split in two parts according to the value of the factor, one for true and the other for false. For example, if a user selects “sex” for balancing, the results for men and women will be shown separately (Fig. 12).

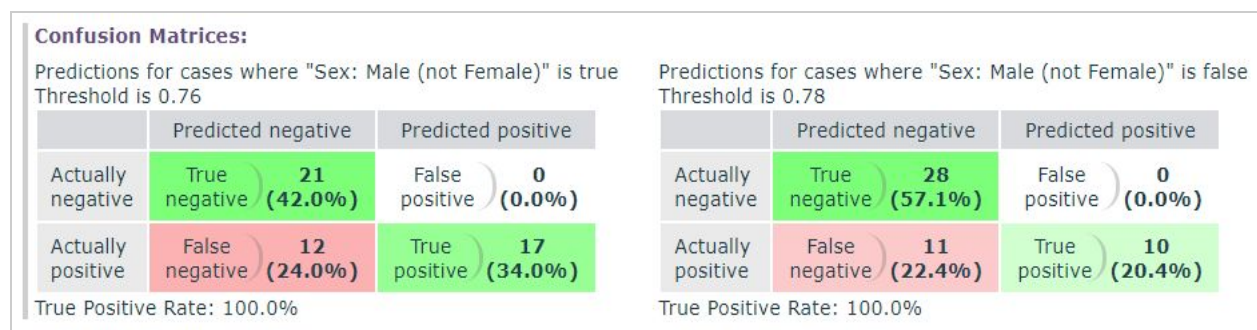
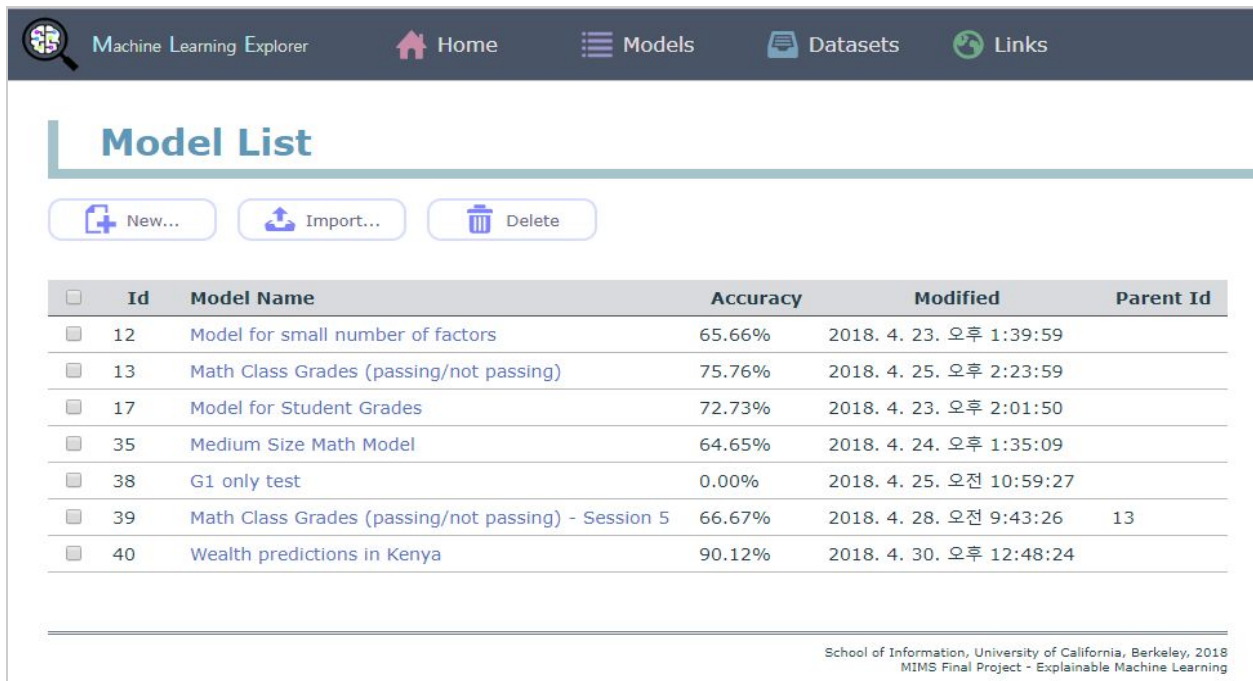


Figure 12. The Result Balanced by “Sex” Factor

Users can export a model to their local machine in JSON format text file. Help buttons for logistic regressions, odd ratio, and balance are provided to give more explanation to users.

## Model List

The “Model List” page shows the list of models stored in the database (Fig. 13). On clicking the model name the page is redirected to the model view page. A new model can be created by clicking the “New” button. The “Parent Id” column means that the model is derived from another model. Users may import a JSON format text file from their local files to add a new model to the database.



<input type="checkbox"/>	Id	Model Name	Accuracy	Modified	Parent Id
<input type="checkbox"/>	12	<a href="#">Model for small number of factors</a>	65.66%	2018. 4. 23. 오후 1:39:59	
<input type="checkbox"/>	13	<a href="#">Math Class Grades (passing/not passing)</a>	75.76%	2018. 4. 25. 오후 2:23:59	
<input type="checkbox"/>	17	<a href="#">Model for Student Grades</a>	72.73%	2018. 4. 23. 오후 2:01:50	
<input type="checkbox"/>	35	<a href="#">Medium Size Math Model</a>	64.65%	2018. 4. 24. 오후 1:35:09	
<input type="checkbox"/>	38	<a href="#">G1 only test</a>	0.00%	2018. 4. 25. 오전 10:59:27	
<input type="checkbox"/>	39	<a href="#">Math Class Grades (passing/not passing) - Session 5</a>	66.67%	2018. 4. 28. 오전 9:43:26	13
<input type="checkbox"/>	40	<a href="#">Wealth predictions in Kenya</a>	90.12%	2018. 4. 30. 오후 12:48:24	

School of Information, University of California, Berkeley, 2018  
MIMS Final Project - Explainable Machine Learning

Figure 13. Screenshot of Model List

## Dataset List

The “Model List” page manages dataset files in CSV format. Users can upload, delete, and download dataset files (Fig. 14).



## Dataset List

 Upload...

 Delete

<input type="checkbox"/>	Id	Dataset Name	Uploaded	File URL
<input type="checkbox"/>	15	Math Dataset with binary columns	2018. 4. 7. 오전 8:38:10	<a href="#">df_math_binary_columns.csv</a>
<input type="checkbox"/>	16	Small Math Dataset	2018. 4. 7. 오전 8:38:42	<a href="#">df_math_cleaned_smaller.csv</a>
<input type="checkbox"/>	17	Math Dataset with Grade 1 included	2018. 4. 7. 오전 8:42:38	<a href="#">df_math_cleaned_G1.csv</a>
<input type="checkbox"/>	18	Medium Math Dataset	2018. 4. 24. 오후 1:17:16	<a href="#">dataset_math_medium.csv</a>
<input type="checkbox"/>	19	G1 only	2018. 4. 25. 오전 10:58:55	<a href="#">df_math_cleaned_G1_only.csv</a>
<input type="checkbox"/>	20	Kenya wealth	2018. 4. 30. 오전 11:29:20	<a href="#">kenya_HS2.csv</a>
<input type="checkbox"/>	21	Wealth	2018. 4. 30. 오후 12:01:10	<a href="#">kenya_HS2_cleaned.csv</a>

School of Information, University of California, Berkeley, 2018  
MIMS Final Project - Explainable Machine Learning

Figure 14. Screenshot of Dataset List

## Research Methodology

Our research included both exploratory research on public perceptions of machine learning and formative studies on how people use the user interface we built.

### Exploratory Research

We conducted interviews with six participants and surveyed 44 students on their feelings related to machine learning. While we found many participants were excited about the possibilities offered by machine learning, the main concerns expressed were about the complexity of machine learning and the danger of bias. This exploratory research informed our



tool, which focuses on helping people understand the details inside the complexity of machine learning.

While some interviewees brought up concerns about bias, our survey found that the median respondent thought using machine learning rather than a human decision-maker would result in slightly less bias. This showed the opposite trend from that found by Lee and Baykal (2017),<sup>30</sup> but the difference could be related to the different subject matter, or it could be related to having a question asked about government decisions rather than personal decision-making. While governments have been working on improving human-based decision-making for hundreds of years, they currently have low public trust.

## Experiment Design and Protocol

Our formative experimental design was intended to both improve the design and consider how it affected participant's views of machine learning. We ran focus groups as an extension of Lee and Baykal (2017),<sup>31</sup> which examined decision-making with Spliddit. Here, that algorithmic decision making was extended to machine learning algorithms. We ran three focus groups of two people each and four individual user testing sessions.

We used a dataset of education-based data to have users predict whether students will fail in the next semester (Cortez 2008).<sup>32</sup> It combines many usual features for education (previous grades and absences) with survey responses (future life plans, travel time to school). Due to the large number of factors and early testing that showed the number of factors was overwhelming for users without domain expertise, we reduced the dataset to only include these factors: "More classes failed", "Home to school travel time", "Grade from first semester", "Sex",

---

<sup>30</sup> Lee and Baykal, *ibid.*

<sup>31</sup> Lee and Baykal, *ibid.*

<sup>32</sup> P. Cortez and A. Silva, *ibid.*

and “Absences”. All factors had descriptions entered for users to dig deeper into the data. Because the “Grade from first semester” was so strongly predictive of the future grades, making other factors redundant, we added noise to the grades so other factors would also be needed for predictions.

Participants were given a hypothetical situation where they were asked to consider a machine learning model that would be used to decide which students were likely to fail a class. Participants were told that extra tutoring would be given to students predicted to fail. After receiving initial instructions on how to use the software, they were asked to make any changes or comments they would recommend to school officials.

After completion of the exercise, all participants were independently surveyed about the model. They were asked questions that analyzed feelings about the model shown, values that it would support (based on Friedman (2013)<sup>33</sup>), programming and machine learning experience, demographics, and measures of group leadership (Lee 2017<sup>34</sup>). See Appendix A for questions.

As the 3 focus groups were run, we updated our procedure to focus our learnings, following a procedure similar to Rapid Iterative Testing and Evaluation (RITE).<sup>35</sup> The first focus group had both participants working together for the entire hour, which resulted in significant time discussing the tool rather than the task. In the second and third focus groups, participants worked alone for the majority of the time, allowing usability testing to be run, then discussed their results after coming up with individual conclusions. Also, after the first focus group, we realized that some buttons were unclear and should be better documented. To simulate this documentation, we used a wizard-of-oz method where participants were read instructions that

---

<sup>33</sup> Friedman, Batya, et al. "Value sensitive design and information systems." in *Early engagement and new technologies: Opening up the laboratory*. Springer Netherlands, 2013. 55-95.

<sup>34</sup> Lee and Baykal, *ibid*.

<sup>35</sup> Medlock, Michael C., Dennis Wixon, Mark Terrano, Ramon Romero, and Bill Fulton. "Using the RITE method to improve products: A definition and a case study." *Usability Professionals Association* 51 (2002).

could, in the future, be added to the user interface. These instructions are included in the experimental script in Appendix B.

Participants were highly educated, with all participants either enrolled in or graduated from a graduate degree program. The participants skewed toward Asian (seven out of ten) and female (seven out of ten).

## Results

---

Our experimental results can be broken into usability testing, which found usability issues in the interface, and focus groups, which revealed how participants used the information from the interface in their discussions.

### Usability Testing

The usability testing sessions helped identify bugs and areas of improvement in the user interface, which we iteratively improved during the testing period. This allowed us to evaluate whether the changes made improved the user experience.

Some issues were found that required clarifying the internal state of the system. A common issue faced among participants was not knowing when to use the test or retrain button after changing the weights or disabling factors. This was improved by creating a pop-up dialog box that let participants know they should retrain the model when they disabled a factor and clicked on test. Similarly, all icons needed to have explanatory text for new users.

Some findings became clear as participants gained confidence with the tool. At first, users felt that having too many factors was overwhelming, but as they became more comfortable with the tool, having more factors was seen as valuable because participants

expected those extra factors to make the model more accurate. Also, participants often wanted to compare the previous results of running a test to the current version. We attempted to implement this change, but our implementation failed to improve the usability. More user research would be needed to optimize the design. Finally, participants who had explored the model completely tried to change the intercept, which biases the model toward the positive or negative class. However, it is often on a far different scale from the factor weights, so our interface had forced it to fit. As a result, participants would think they had moved it a small amount when it had actually changed significantly. We concluded that the intercept should be moved to a separate area of the screen and put on its own scale.

Participants who had prior knowledge or experience with statistics were able to comprehend the confusion matrix and logistic regression weights more easily. For less experienced participants, the UI attempted to explain topics such as logistic regression, odds ratio, and balancing models by clicking on the question mark icon next to the topics. Generally, the more description that was provided in the user interface, the more it helped participants understand the machine learning and statistics concepts. Participants also mentioned the need for better data and factor collection such as socioeconomic, home environment variables (time spent on extracurriculars, guardian) and nature of the course (mathematical or reading) which would be helpful in making the model predictions more reliable. Some participants had suggestions for how they wanted the factors sorted, possibly logically grouped or sorted from most positive to most negative, but there was no consistent request.

Finally, some issues were a result of our testing environment setup rather than the tool itself. These issues are relevant for anyone explaining machine learning models, even if they are not using our tool. Factor names and descriptions need to be clear. It is easy for data scientists to use machine-readable names when presenting data, but clear factor names and

longer descriptions are needed for any model presented to the public. Also, factor names and values should be selected to avoid any factor names with “not” in them, as negative weights result in confusing double negatives. The model itself must also be named well: we used coded names matched to our session number, and we had to explain our naming scheme to participants each time. Even the number of test cases matters: we showed the raw numbers of test cases rather than percentages so participants would think about the number of test cases, but because we had 99 test cases, users assumed the numbers were percentages. In the future, it would be better to have a test set with larger numbers to avoid this confusion.

During the experiment, we also found two bugs that were fixed before the next session. Issues have been created on our GitHub page<sup>36</sup> for all problems listed above.

## Focus Group

The tool was successful in keeping users engaged and served as a platform to voice opinions and contest the output of the machine learning model while providing transparency in decision making. Most participants used the features for commenting on factors and changing weights. Some modified the model when the default did not align with their prior experience or belief of how the system should work. For example, some participants adjusted the weight of the “Sex: Male” factor to be zero or disabled the factor entirely since they did not want this factor to affect the likelihood of a student to pass or fail. Similarly, some participants reversed the weight of the consumption of alcohol factor if it showed a correlation between consuming more alcohol and the probability of passing the class. However, other participants decided that their goal was to maximize accuracy, and chose to change none of the parameters. In discussions, different

---

<sup>36</sup> [https://github.com/shresh02/explainable\\_ML\\_public\\_policy/issues](https://github.com/shresh02/explainable_ML_public_policy/issues)

participants chose to argue for a higher accuracy model or a more curated model that took into account their domain knowledge.

The accuracy of the model was crucial and a concern for several participants. Some believed it was not accurate enough to take actions in the real world especially since it could be accompanied by a factor of stigmatization towards students who are predicted to fail. Others felt that despite the low accuracy, the model could still be used provided students or parents are given the choice to opt out of the intervention. Some indicated they would prefer using the results of the model as an additional data point rather than the primary decision making tool. There were also questions on what accuracy meant; these participants felt they would benefit from a longer explanation of how it was calculated.

No participants advocated for the use of equal opportunity provided by the “Balance” button. We expect that this is mainly due to a lack of full understanding of what it does, even after some participants read the long explanation. Common questions included whether participants could balance the model on multiple factors or on continuous (not binary) factors. A few participants eventually understand what it meant and modified the model to give equal opportunity across a single factor, which in most cases was the sex of the student. However, they tended to revert this change later after trying it out.

## Conclusion

---

Machine learning may provide opportunities for governments to make decisions well. However, with its inscrutability, it raises new challenges for accountability, which is essential for public policy. Our early survey suggested that people were both excited and worried about the rise of machine learning. In our experiment, we attempted to create a tool to give the public

control over machine learning used by the government by adding contestability. Iterative development while testing helped us to make our tool become more useful.

Unfortunately, in a tool designed to help the public understand the internals of a model, there are still many statistical concepts that users must understand. Even for the simplest class of machine learning model, logistic regression, users need to understand the meaning of terms like false positives. However, even though our application had not been optimized for usability, it helped users with a machine learning background to gain a better understanding of the algorithm and how it is being used in making decisions, allowing them to consider more than just a target accuracy. Although less experienced users must learn more about machine learning to use the application, it still increases transparency for any model shown in the tool. The application also succeeded in keep users engaged in inspecting and contesting a policy making process using machine learning. Some users continued to explore the user interface after their experiment session had concluded and most users expected to try the next version of the tool.

The experiment highlighted that while equality of opportunity or fairness is a valuable goal, implementing it will require significant training on exactly what fairness algorithms do. As our experiment showed, explaining machine learning is very hard. Nonetheless, even if the explanation is insufficient and discursive, it still helps to improve the transparency and fairness of machine learning. This tool could be successfully used by a group of public policy decision makers as they could be provided with basic training in statistics and regression analysis which would be tough to provide to the general public.

In conclusion, governments will need many tools to keep machine learning accountable, including educational tools, legal structures, and other frameworks that make this kind of

visualization possible. Tools like the one we developed are part of that toolkit needed for governments to use machine learning in an accountable way.

## Future Work

---

Our development and research have suggested future work within this space. Some future work relates directly to development of the tool, while other parts are focused on new uses. The tool itself currently does not implement user accounts and permissions, so there is no differentiation between administrators and the public. Before the tool could be deployed for explaining any model, user accounts would need to be secured. Similarly, while we put significant efforts into the model definition page, no user testing has been done on the process of creating a new model, and we would need to work to make that process more user friendly.

Seeing how participants used the interface to explore how a logistic regression model works, and how challenging it was to find a math-light explanation of logistic regression online, our tool could be extended to let participants explore changing the weight of a one-factor model with an intercept, then moving to a more complicated model. Letting them explore by trying to optimize by hand would make it feel like a game, and then it could be revealed how logistic regression optimizes for minimal loss.







# Appendix B: Experiment Protocol Script

Thank you for volunteering for our focus group today!

We (Shrestha, Sung, Sam and Monicah) are students in the School of Information. This focus group is part of our capstone project. We are being advised by Prof. David Bamman.

## **What is the project about?**

Machine learning solutions are increasingly being used to inform decisions in a variety of fields. In many cases, subject matter experts are not the same people as machine learning experts. When the government uses machine learning, citizens are the non-experts who need some control over the models. We have provided an interface for you to view and provide commentary on the underlying workings of machine learning models, which we intend to help those who want to know more about a model without having machine learning expertise.

## **How will the data be used?**

We will analyze notes that we take to understand how people interact with our tool and how they think about machine learning. All data will be anonymized. However, do remember that is a focus group, so you should not share other people's comments outside of this group. That said, we cannot ensure that other people here will not share things outside of this group, so please understand that confidentiality is not guaranteed.

You can leave this session at any time if you feel uncomfortable or for any other reason.

## **What is the experiment?**

### **1. Overview and Dataset**

Today, some high schools attempt to predict which students are likely to fail classes so that they can attempt an intervention such as sending a student to extra tutoring. Often, this is done by considering their current grade. In our exercise today, we will consider a hypothetical situation where much more data has been collected about past students, including survey data, and a machine learning model has been trained on that data to predict which current students will need help. The data are based on an anonymized real-world dataset from a single school, so you should not draw any general conclusions from what is shown in this data.

We want you to act as citizens in the neighborhood commenting on the model. On the web page we will show you in a moment, you can try out changes to the model, such as removing factors, changing weights, or adding comments that you would want a school

official to see. Please discuss as a group, and we will be listening and taking notes. We can answer questions if you are confused. When you believe that you have done everything you wish to with the web application, let us know. Then we will give you a follow-up questionnaire that you will fill out individually.

## **2. User Interface**

Here, we have a web application showing detail about this model. The user interface provided shows how much different factors affect student performance.

At the top, we note that this is a logistic regression model. This means that it will find correlations between each of factors the variable we are trying to predict. In this case, we are predicting whether or not students will pass their next class.

Now let's look at the list of factors. As an example, "more classes failed" shows that there is a correlation between students who have failed more classes in the past and students who are less likely to pass this class.

Try clicking the Test button. Take a look at what is shown and tell me what it means to you (Note: we can fill in here after we hear some responses).

Now try the Retrain button. This will ignore any by-hand changes you make to the model, but it will keep track of things like disabling factors or balancing the model on a single factor.

# Appendix C: Database details

## Entity-Relationship Diagram

ERD was used to understand the relationships among data objects and their attributes.

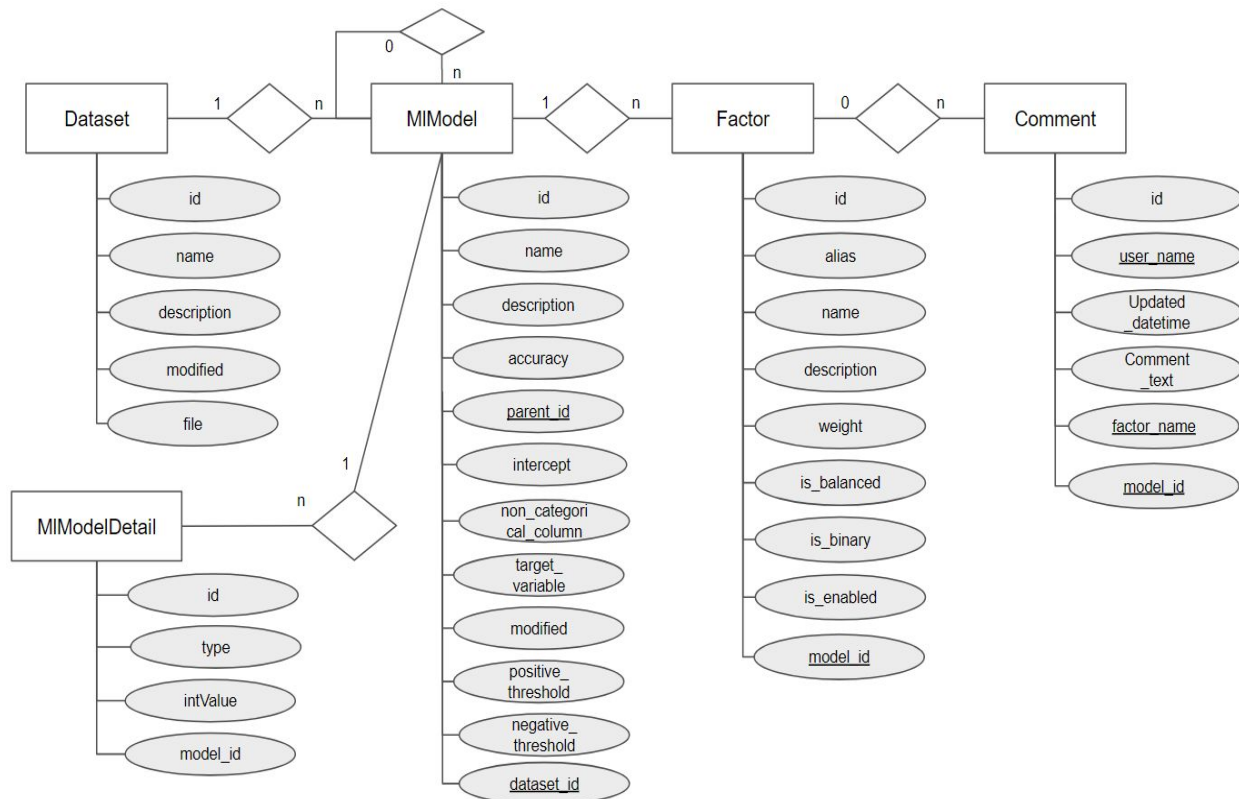


Figure ?. Entity-Relationship Diagram

## Data Dictionary

The data dictionary described how to implement actual database tables and their fields by defining types and constraints for each field.

### Dataset Table

Column Name	Type	Constraint	Description	Value Example
-------------	------	------------	-------------	---------------

id	INT	Primary key	Auto Increment	1
name	VARCHAR(255)	Not Null		Student data
description	TEXT			This is from a school.
modified	DATETIME	Not Null		2010-01-30 23:59:59
file	VARCHAR(100)	Not Null	Dataset File path	\\dataset\\sample.csv

## MIModel Table

Column Name	Type	Constraint	Description	Value Example
id	INT	Primary key	Auto Increment	1
name	VARCHAR(255)	Not Null		Decision model
description	TEXT			To decide passing or not.
accuracy	REAL			0.72
parent_id	INT	Foreign key	Model ID derived from	1
intercept	REAL			
non_categorical_columns	TEXT		List of factor names	col1, col2, col3
target_variable	VARCHAR(255)		Original CSV column name	col4
target_variable_alias	VARCHAR(255)		Human readable name	2010-01-30 23:59:59
modified	DATETIME		Time of the last modification	
positive_threshold	REAL			0.50
negative_threshold	REAL			0.28
dataset_id	INT	Foreign key	ID of a Dataset table entry	1

## MIModelDetail Table

Column Name	Type	Constraint	Description	Value Example
id	INT	Primary key	Auto Increment	1
type	VARCHAR(255)		Value type	all#true_negative_count
intvalue	INT		Value	65
model_id	INT	Foreign key	ID of a MIModel Table entry	1

## Factor Table

Column Name	Type	Constraint	Description	Value Example
id	INT	Primary key	Auto Increment	1
alias	VARCHAR(255)		Human readable name	Distance from school
name	VARCHAR(255)	Not Null	Original CSV column name	col1
description	TEXT			This factor is important
weight	REAL	Not Null		0.55464231
is_balanced	BOOLEAN			false
is_binary	BOOLEAN			false
is_enabled	BOOLEAN			false
model_id	INT	Foreign key	ID of a MIModel Table entry	1

## Comment Table

Column Name	Type	Constraint	Description	Value Example
id	INT	Primary key	Auto Increment	1
user_name	VARCHAR(255)			user01
updated_datetime	DATETIME			2010-01-30 23:59:59
comment_text	TEXT			This is not appropriate.
factor_names	VARCHAR(255)		Name of a Factor table entry	col1
model_id	INT	Foreign key	ID of a MIModel Table entry	1