# PREDICTING GENDER IN INTRODUCTORY CS

## PROJECT OVERVIEW

As I have often said, one of the biggest challenges of this century is successfully getting people into the technical workforce. We are now in the age of automation with autonomous cars driving down our streets and Alexa and Siri becoming an extension of our homes. With the increase of these systems, it is likely that we are also going to see an increase in technical jobs.

This shift in the workforce towards highly skilled, technical knowledge workers poses a challenge on the supply side; mostly because of a lack of presence of computer science in K-12 education; the underproduction of post-secondary degrees in computer science; the underrepresentation of women and the underrepresentation of ethnic minorities.

One of the solutions that have been proffered for this problems is redesigning introductory computer science to broadening participation.

As part of my doctoral study, I decided to study the socio-curricular factors that affect the decision to participate in introductory computer science through a data-driven lens. To do this, I designed a research study investigating the role of computer science self-identity centered around the experiences of undergraduates in two introductory computer science classes at UC Berkeley.

## PROBLEM STATEMENT

With this project, the problem I am interested in investigating is the gendered experience of these the two CS classes in my study. Using machine learning algorithms, I want to predict and identify the most salient variables that govern the experience of men versus women in introductory CS at an elite research university like Berkeley.

To predict gender in intro CS at Berkeley we suggest the following strategy:

(a) Explore the dataset to ensure its integrity and understand the context.

(b) Identify features that may be used. If possible, engineer features that might provide greater discrimination.

(c) With the understanding that this a "classification" task, explore a couple of classifiers that might be well suited for the problem at hand.

(d) Once a classifier has been selected, tune it for optimality.

### METRICS

Predicting gender in intro CS is a supervised learning problem. To determine the performance of the model, I have selected “accuracy" as the metric of most interest. In addition to that, we will take a look at the confusion matrix for the output of each model to give us more insight into how good our classifiers are at discriminating the data based on gender.

# ANALYSIS

This project uses a dataset which I created as part of my doctoral research. The dataset consists of survey responses. The survey instruments were developed to measure participants' self-reported efficacy along several dimensions.

(a) Self-reported attitudes about CS

(b) Gendered belief about CS ability

(c) Career driven beliefs about CS

(d) Self-reported attitudes about computational thinking

(e) Self-reported attitudes about CS class belonging

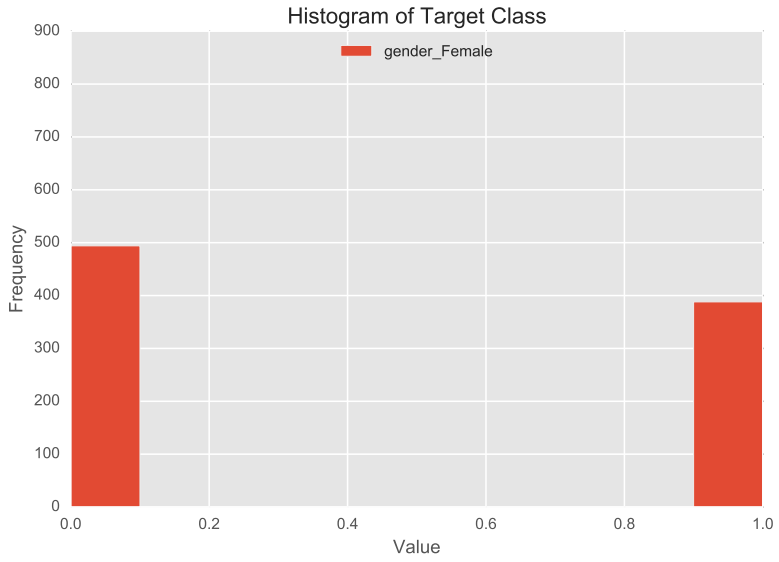
(f) Self-reported beliefs about collegiality

(g) Prior collegiate CS exposure

(h) CS mentors and role models

Majority of the questionnaire uses a 5-point Likert scale (where 1 = Not Really, 3 = Neutral and 5 = Absolutely). The dataset consists of 882 instances with no missing data.

An interesting aspect of this dataset is that the class labels for our classification are slightly unbalanced at a ratio of around 1:1.2 for male students as can be seen in figure 0.1a.




(a)

Figure 0.1: **Gender plot.** *The histogram shows a slightly unbalanced target dataset with 494 values of {0: male} and 388 values of {1: female}.*

### EXPLORATORY VISUALIZATION

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

- Have you visualized a relevant characteristic or feature about the dataset or input data?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

### DATA EXPLORATION

# METHODOLOGY

## DATA PREPROCESSING

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

- If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
- Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
- If no preprocessing is needed, has it been made clear why?

## IMPLEMENTATION

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?
- Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?
- Was there any part of the coding process (e.g., writing complicated functions) that should be documented?

## REFINEMENT

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as

any significant intermediate results as necessary. Questions to ask yourself when writing this section:

- Has an initial solution been found and clearly reported?
- Is the process of improvement clearly documented, such as what techniques were used?
- Are intermediate and final solutions clearly reported as the process is improved?

# RESULTS

(approximately 2 - 3 pages)

## MODEL EVALUATION AND VALIDATION

When I did feature selection using SelectPercentile on top 10% of the features, the model over-fitted, giving a 100% score on all the three classifiers used.

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model,Äôs solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
- Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
- Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
- Can results found from the model be trusted?

## JUSTIFICATION

In this section, your model,Äôs final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- Are the final results found stronger than the benchmark result reported earlier?
- Have you thoroughly analyzed and discussed the final solution?
- Is the final solution significant enough to have solved the problem?

# CONCLUSION

(approximately 1 - 2 pages)

### FREE-FORM VISUALIZATION

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- Have you visualized a relevant or important quality about the problem, dataset, input data, or results?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

### REFLECTION

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- Have you thoroughly summarized the entire process you used for this project?
- Were there any interesting aspects of the project?
- Were there any difficult aspects of the project?
- Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

### IMPROVEMENT

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to

make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- Are there further improvements that could be made on the algorithms or techniques you used in this project?
- Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
- If you used your final solution as the new benchmark, do you think an even better solution exists?