CSCI 183

Final Project Report

Jack Corley, Savannah Balistreri, Lauren Vu

**Analyzing Bias in Hiring Records using Machine Learning**

*Background*:

Our project was largely inspired by an AI Amazon created back in 2018 to “rank” resumes of job applicants. Data was submitted on different Amazon employees’ resumes over the course of 10 years. As most of this data submitted was from males, the resumes that contained the word “women” were penalized by the AI model. The model was eventually tossed due to its biases (BBC). The idea that biased data, like in the case of Amazon’s resume ranking dataset, can vastly influence an AI model performing tasks, made us curious about other areas in which we could recreate this problem.

We found a dataset used by scientists in 2001 and 2002, documenting generated resumes belonging to white and African American individuals submitted to different job applications within Boston and Chicago. The scientists in this study created virtually identical resumes, but assigned the resumes to people with stereotypically female, male, white, and African American names (Emily is used as a white, female name, while Darnell is used as an African American male name). Using this as well as a few other points of data, like zipcode, the scientists were able to conclude from this data that African Americans were being racially discriminated against when judged by the callbacks they received compared to their white counterparts. Specifically, people with stereotypically white names receive about 1 callback for every 10 resumes submitted, while African Americans receive 1 callback for every 15 (Bertrand, M., & Mullainathan, S.). It is our goal to create an algorithm trained on this data that will predict whether someone should get a callback, and to determine if the AI is quantifiably biased.

*Design*:

First, we discussed the different columns in the dataset. There were 30 features, and most were qualitative. We determined which features were necessary for our model, and which ones were not. Here is a table of those determinations:

| Feature Name | Definition of Feature | Qualitative vs Quantitative | Determination |
| --- | --- | --- | --- |
| Received\_callback | Whether the applicant received a callback | Qualitative. {0,1} | TARGET |
| Job\_ad\_id | Job ID (assigned to job) | Quantitative | Remove |
| Job\_city | Job’s City | Qualitative {Boston, Chicago} | Remove |
| Job\_industry | Job’s Industry | Qualitative {strings} | Include |
| Job\_type | Type of Job (i.e. secretary, manager) | Qualitative {strings} | Include |
| Job\_fed\_contractor | Indicator if job is a federal contractor | Qualitative {0,1} | Include |
| Job\_equal\_opp\_  employer | Indicator if the job is an equal opportunity employer | Qualitative {0,1} | Include |
| Job\_ownership | Type of the company (non-profit, for-profit) | Qualitative {strings} | Include |
| Job\_req\_any | Indicator if there are any requirements for the job | Qualitative {0,1} | Include |
| Job\_req\_  communication | Indicator of communication requirements for the job | Qualitative {0,1} | Include |
| Job\_req\_education | Indicator of the education required for the job | Qualitative {0,1} | Include |
| Job\_req\_min\_  experience | Indicator of the minimum experience required for the job | Qualitative {0-5, some} | Include |
| Job\_req\_computer | Indicator of computer skills required for the job | Qualitative {0,1} | Remove |
| Job\_req\_  organization | Indicator of organizational skills required for the job | Qualitative {0,1} | Include |
| Job\_req\_school | Level of education required for the job | Qualitative {string} | Remove |
| Firstname | Name assigned to applicant | Qualitative {string} | Remove |
| Race | Race assigned to applicant | Qualitative {string} | Include |
| Gender | Gender assigned to applicant | Qualitative {f,m} | Include |
| Years\_college | Years applicant was in college | Quantitative | Include |
| College\_degree | Indicator of educational degree obtained by applicant | Qualitative {0,1} | Include |
| Honors | Indicator of whether applicant has been awarded honors | Qualitative {0,1} | Remove |
| Worked\_during\_  school | Indicator of whether applicant worked while attending school | Qualitative {0,1} | Include |
| Years\_experience | Applicant’s prior experience in the field measured in years | Quantitative | Include |
| Computer\_skills | Indicator of applicant’s level of computer skills | Qualitative {0,1} | Remove |
| Special\_skills | Indicator of whether an applicant has special skills | Qualitative {0,1} | Remove |
| Volunteer | Indicator of whether an applicant had volunteered | Qualitative {0,1} | Remove |
| Military | Indicator of whether the applicant is or was in the military | Qualitative {0,1} | Include |
| Employment\_holes | Indicator of whether there were holes in the applicant’s employment history | Qualitative {0,1} | Include |
| Has\_email\_  address | Indicator of whether the applicant lists an email address | Qualitative {0,1} | Include |
| Resume\_quality | Indicator of whether the resume was classified as a higher or lower quality resume | Qualitative {string} | Include |

We decided to eliminate some features from the dataset: for example, we decided that job\_id and city were irrelevant to our target (receiving a callback or not), as ID would not affect our predictions at all and only two cities were included in our dataset. For a wide spread dataset, location could possibly have an impact (to answer questions such as, would people from a more affluent area be more likely to be hired than those from less affluent areas?) but that is both outside the scope of our project and is unavailable in our dataset. Skills such as computer skills, special skills, and whether they volunteered or not are also not closely related to the subject of our study so we chose to remove these as well. We also thought about dropping the job’s requirement on minimum years of experience and whether or not the job was from a federal contractor or not as these columns were mostly NaN values, reaching 56% and 36% NaN values respectively, but we were able to keep these two features by using the means and modes of the feature. We thought these two columns would be useful, but as a large amount of job postings didn’t include this, we were unable to utilize this data in our model. We also had to drop first names from the dataset, as those first names were created by the scientists to convey both race and gender, so that could unintentionally create bias in the data.

We then discussed which machine learning algorithms we thought would work the best on our data. Since our data is largely categorical, a lot of data would most likely overlap, and we needed an algorithm that would do well despite the lack of variety. We decided on SVM, and will utilize many kernel transformations in order to find which one can most accurately model our dataset. We chose to use SVM as one of our models due to SVM being suitable for more than two features, with the kernel allowing us to project data points into higher dimensions to accommodate this. SVM also chooses the most optimal hyperplane and is a more advanced classification model than logistic regression, matching our needs.

We also decided to try K-NN on our dataset, as K-NN can outperform more complicated algorithms like SVM. K-NN is also suitable due to its relatively easy implementation and the dataset size not being overly large. With a model created on a dataset that is proven to contain bias, if we accurately model the dataset, we will also be accurately modeling the bias as well.

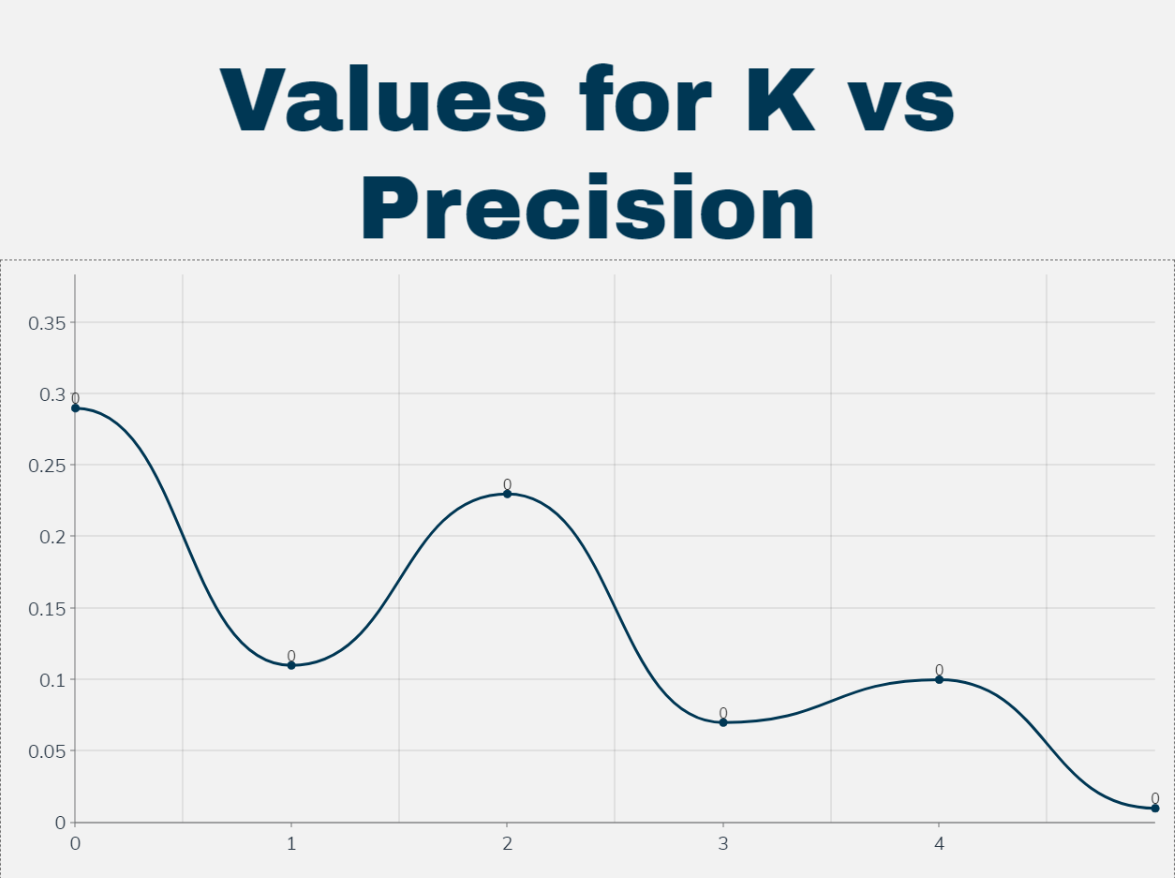
*Implementation:*

The first and one of the most upsetting issues regarding our implementation had to do with our dataset: it is incredibly unbalanced. In the dataset itself, about one in every ten white people received a callback, while only one in every fifteen African Americans received a callback. Our target variable didn’t have many instances of receiving a callback, about 9.3% across the board among 4869 people, which created a multitude of inaccurate and underfit models. As a result, we had to utilize different kernels and methods in order to overcome this unfortunate part of our dataset.

To handle the underfitting issue specifically in the SVM model, we had to adjust some hyperparameters. We first set the class weight to balanced, which gives more weight to the infrequent classes, in our case this was 'received callback'. Other hyperparameters used were gamma and regularization, which affects the influence of a single training example and how lenient the model is when it comes to accepting certain rows respectively. The regularization, or C parameter, allows us to choose how tolerant we are of outliers, decreasing the chance of overfitting. Lower values of C allows for a larger margin, and accordingly a less complicated decision boundary. Higher values of C are stricter with the classification of data points, however this may lead to overfitting in the long term. A program was created and ran which tested out several values and combinations of these to find the optimal balance. Although the first model built without such hyperparameters was accurate, it received very low precision and recall scores due to it classifying everyone as not receiving a callback. We sought to maximize precision, recall, and F1 score, but also understood that precision gave a very valuable indication on whether or not a certain testing group was receiving many unearned call backs.

For our SVM model, we used sklearn and pandas. We combed through the columns (features) and checked whether they were categorical or not and if they were categorical, we one hot encoded them. Then the NaN values were taken care of. If a column was numerical, the average value would replace the NaN, and if the column was categorical, the mode would replace the unknown category. After formatting the data appropriately, we split the dataset into training and test sets, with test size being 20%. The hyperparameter optimization function was ran separately, and after the optimal values were calculated, the SVM model is built with these hyperparameters. After training our model, we test it using the test data and record the number of True Positives, True Negatives, False Positives, and False Negatives, as well as calculating model evaluation metrics (accuracy, precision, F1 score).

We chose to set the SVC initialization parameter to kernel linear, as the SVM kernel can project features into higher dimensions in order to find the optimal plane for classifying the target using more than two features at once. This was also necessary as underfitting was a known issue at this point.

For our K-NN model, we utilized scikit-learn’s K-NN algorithm, and one-hot encoded all the categorical columns with strings depicting the feature. We divided the dataset into test and training, test size being 30%. We then tested the model with k through 1-10 by examining the four model evaluation scores. We focused mainly on precision, recall, and f1 as accuracy does not properly convey that a model is underfitting. After testing, we utilized the elbow method to choose k = 3, while looking at the graphs of values of k compared to the precision, the recall, and the f1 scores.

We also had 3 CSV files (derived from a separately written python script), which are all identical except for the race and gender columns. Our first CSV file is the default CSV file, which we used to train our models. We then altered this CSV file into two more, one in which everyone was turned into a white man, and a second in which everyone had been turned into an African American woman. We hypothesized that these two groups would receive the least and most racial and gender discrimination respectively, so we hoped using these two CSV files as testing data would give us more information about the potential bias we were creating within our model. We also created two additional CSV models, in which one only had the white men from the default CSV file, and one only had the African American women from the default CSV file. Then, we had one last CSV file, in which both the race and the gender of the applicants were removed.

*Results:*

In order to test for bias, we had to attempt to differentiate between bias in the model, and bias in the actual hiring process. To do this, we trained one model with race and gender columns included, and one model without these columns. To test the model built with race and gender included, we created additional copies of 'resume.csv' where everyone had either been turned into a white man, or a black woman. Although we created more files with only one category changed, we found the most significant results comparing white men to black women. In order to compare the bias of the model to the bias of the data, we then built a model with 'resume.csv' but with the race and gender column removed. To test this, we then created subsets of the original data, where only white men remained, and only black women remained, instead of artificially changing their race or gender. We then removed the race and gender columns from these subsets, so we knew we were testing with white men and black women, but the model had no indication of what race or gender they were. If the model trained on race and gender produced different results, it would be the model that is biased, but if the model built without race and gender produced different results on the subsets, we would know it is the data itself that is biased.

Below are the confusion matrices for our models. Each row’s model was built 4 separate times and tested, and those 4 separate models were then averaged to produce these results:

| Model Type | CSV File | Number of True Positives | Number of False Positives | Number of False Negatives | Number of True Negatives |
| --- | --- | --- | --- | --- | --- |
| SVM | All Races + Genders | 90  (~1.8%) | 413  (~8.5%) | 302  (~6.2%) | 4064  (~83.5%) |
|  | White Men | 99.5  (~2%) | 498.5  (~10.2%) | 296.5  (~6.1%) | 3979.5  (~81.6%) |
|  | African American Women | 62.5  (~1.3%) | 257.5  (~5.3%) | 329.5  (~6.8%) | 4220.5  (~86.6%) |
| SVM (no race or gender) | All Races + Genders without the Race or Gender Column | 78.25  (~1.6%) | 321.25  (~6.6%) | 313.75  (~6.4%) | 4156.75  (~85.4%) |
|  | Subset of White Men Without a Race or Gender Column | 9.5  (~1.7%) | 33.5  (~5.8%) | 41.5  (~7.2%) | 490.5  (~85.3%) |
|  | Subset of African American Women Without a Race or Gender Column | 27.5  (~1.5%) | 106  (~5.6%) | 97.5  (~5.1%) | 1665  (~87.8%) |
| K-NN | All Races + Genders | 139.75  (~2.9%) | 57.75  (~1.2%) | 277.25  (~5.7%) | 4420.25  (~90.3%) |
|  | White Men | 155.75  (~3.2%) | 68.75  (~1.4%) | 236.25  (~4.9%) | 4409.25  (~90.5%) |
|  | African American Women | 156  (~3.2%) | 64.25  (~1.3%) | 236  (~4.8%) | 4413.75  (~90.6%) |

This also allows us to calculate the average accuracy, precision, recall, and F1 score of each model, as shown in this table:

| Model Type | CSV File | Accuracy | Precision | Recall | F1 Score |
| --- | --- | --- | --- | --- | --- |
| SVM | All Races + Genders | ~0.853 | ~0.179 | ~0.230 | ~0.201 |
|  | White Men | ~0.837 | ~0.166 | ~0.251 | ~0.200 |
|  | African American Women | ~0.879 | ~0.195 | ~0.159 | ~0.176 |
| SVM  (no race or gender) | All Races + Genders without the Race or Gender Column | ~0.870 | ~0.196 | ~0.200 | ~0.198 |
|  | Subset of White Men Without a Race or Gender Column | ~0.870 | ~0.221 | ~0.186 | ~0.202 |
|  | Subset of African American Women Without a Race or Gender Column | ~0.893 | ~0.206 | 0.220 | ~0.213 |
| K-NN | All Races + Genders | ~0.932 | ~0.708 | ~0.335 | ~0.455 |
|  | White Men | ~0.937 | ~0.694 | ~0.397 | ~0.505 |
|  | African American Women | ~0.938 | ~0.708 | ~0.397 | ~0.510 |

*Interpretation:*

As is evident through these two tables, our SVM model is biased based on race and gender. African American women on average received 33 more false negatives than white men, and received 241 less false positives. Additionally, when we utilized the CSV file with only white men in it on a model blinded to race and gender, the false positive percentage went down from 10.2% to 5.8%.

Additionally, our K-NN model did not contain the bias that our SVM model did. African American women and white men received the same percentages of true positives, false positives, false negatives, and true negatives, within 0.1% of each other. It is also interesting to note that K-NN had a higher accuracy, precision, recall, and F1 score than our SVM model, even considering the dataset that the model was built on was shown to contain bias.

It is certainly evident that our models aren’t incredibly accurate, and part of that inaccuracy can be attributed to the unbalanced nature of our dataset, and the inability to use the data provided to predict a callback with certainty. While the accuracy scores of our models are high, the precision, recall, and F1 score are all low for our SVM model, and our recall and F1 score is low for the K-NN model as well. So while we were able to successfully build models off of a biased dataset, it is unfortunate that the callback rate of job applications is so low that it made this dataset and all other realistic ones on this subject difficult to accurately model.

*Our Final Conclusions*

In regards to SVM, the model itself is biased.

We took care to differentiate between bias in the model and bias in the data. Because there was significant bias present in the model trained with race and gender, and gave more deserved and undeserved callbacks to white men compared to identical black women. The SVM model trained without race or gender did not show bias when tested with real white men and black women.

KNN however, is not biased.

When testing the same data on the KNN which accounted for race and gender during training, bias did not appear. White men and black women were both classified similarly. Because of this, we felt we did not need to test for bias in the data by training a KNN model blind to race and gender. This is due to the model being biased, and us concluding the model is not biased as the previous SVM trained without race or gender was not biased.

Why the difference?

While we can not conclude the reason for the difference based on our data, it could be due to multiple factors. For one, SVM plots one hyperplane, which relies on all the data. Thus, overall trends can influence the categorization of new points, and thus correlation is found where causation does not exist. KNN however, handles classification on a case by case basis, and as long as K is not set too high to where such factors are considered, it could escape this uniformly present but irrelevant trend. This is, however, just a hypothesis.

*Future Research:*

In the future, to make sure that our KNN model is suitable for our dataset, we would try several k values and implement a dendrogram using the elbow method to verify that certain k values work. We would also see if more features are related to each other, whether the bias is related to education level, do certain features in the dataset tend to relate closely to each other than with others.

It also isn’t ideal to modify the original dataset we had in order to conduct further testing and analysis on our model, such as turning everyone into a black woman or white man, so ideally in any future research we would do on this project, we would gather more data. As the columns were many and unique to this dataset, it was not within the scope of the project to obtain data for every column. This way we could have an entire new dataset of just African American women or white men to run on our model, instead of modifying the current dataset to be all African American women or white men.

**References**

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. The American Economic Review, 94(4), 991–1013. <http://www.jstor.org/stable/3592802>

BBC. (2018, October 10). Amazon scrapped “sexist AI” tool. BBC News. <https://www.bbc.com/news/technology-45809919>

Parikh, N. (2023, October 5). Council post: Understanding bias in AI-enabled hiring. Forbes. <https://www.forbes.com/sites/forbeshumanresourcescouncil/2021/10/14/understanding-bias-in-ai-enabled-hiring/?sh=3eb32b907b96>