Profiling a Tech Worker

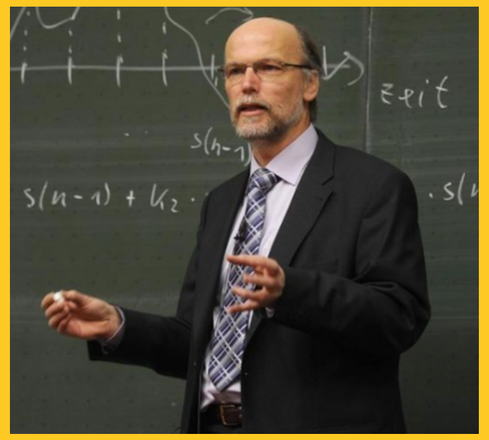
by Jiji Craynock

America the Melting Pot

The US has long been hailed as the melting pot of the world. Yet, stereotypes are as persistent among Americans as ever.

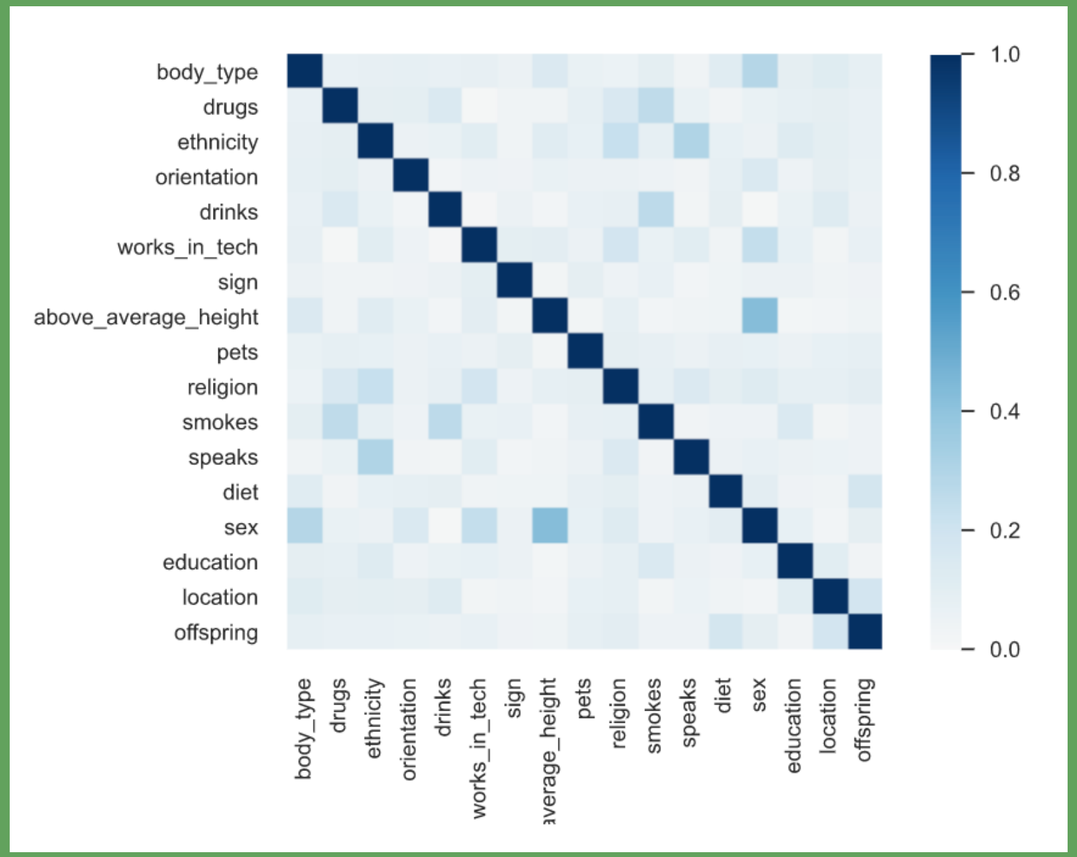
Google As A Reflection of Persistent Biases

When doing an image search for terms like teacher vs. professor, surgeon vs. nurse, and restaurant owner vs. restaurant worker, we are still faced with the same kind of white male leanings that have been the norm for decades



Case Focus

A model that predicts whether or not someone works in the tech industry based on features not usually found on a resume.



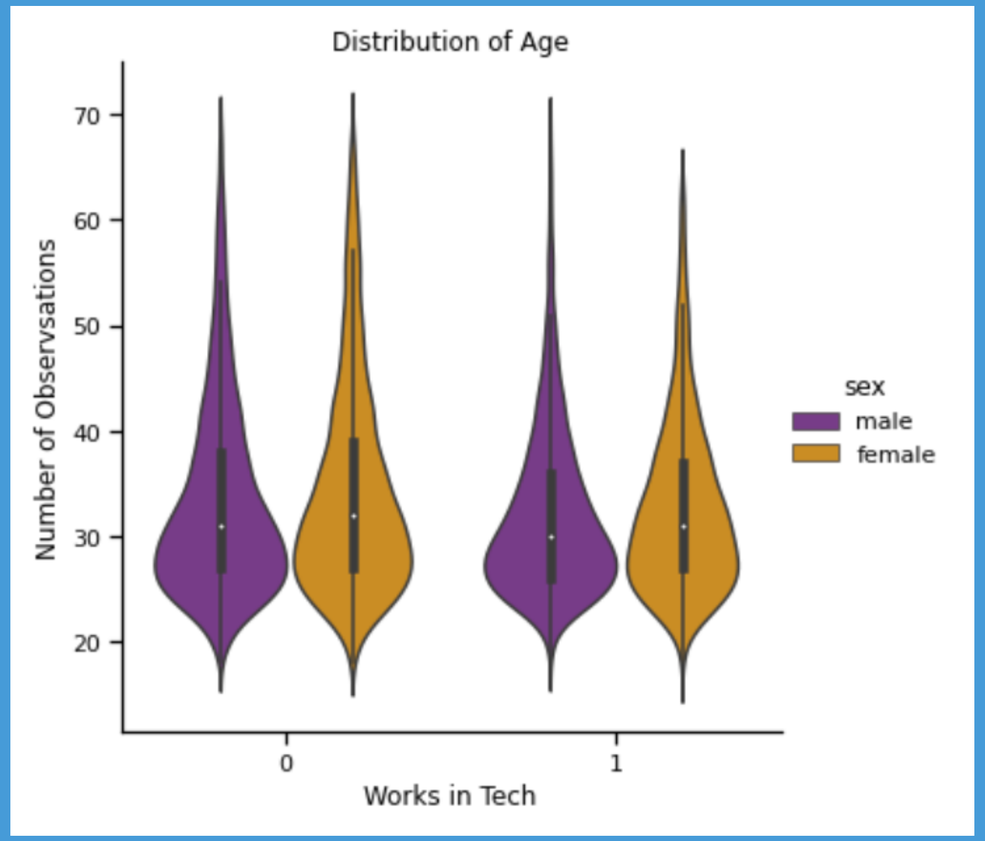
Pre Processing: Nearly as Complex as People Themselves

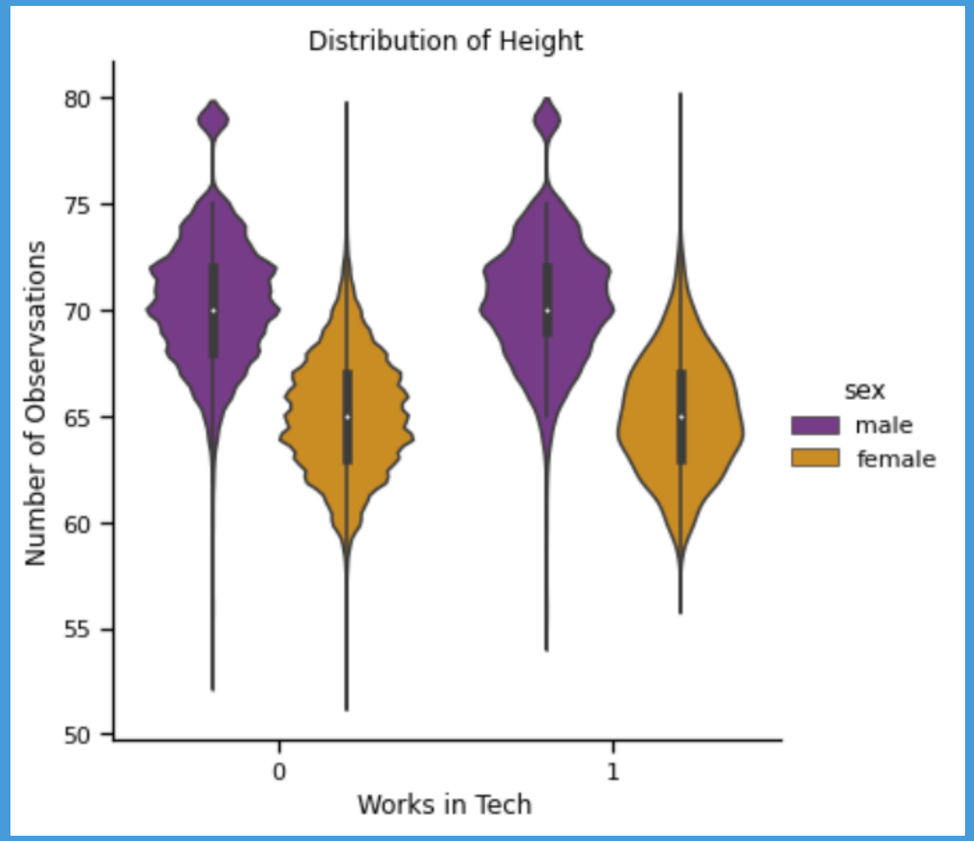
In order to run a predictive test, I need to use data that contained job title information as well as other demographic data. OkCupid, a popular dating website collects exactly this type of information.

However, users are not required to fill out the vast majority of questions and have a plethora of ways to answer when they do choose to. This meant that in order to prepare the data each feature needed to be re-engineered for modeling. The result was a whole host of categorical features, none of which at a glance said much about anything.

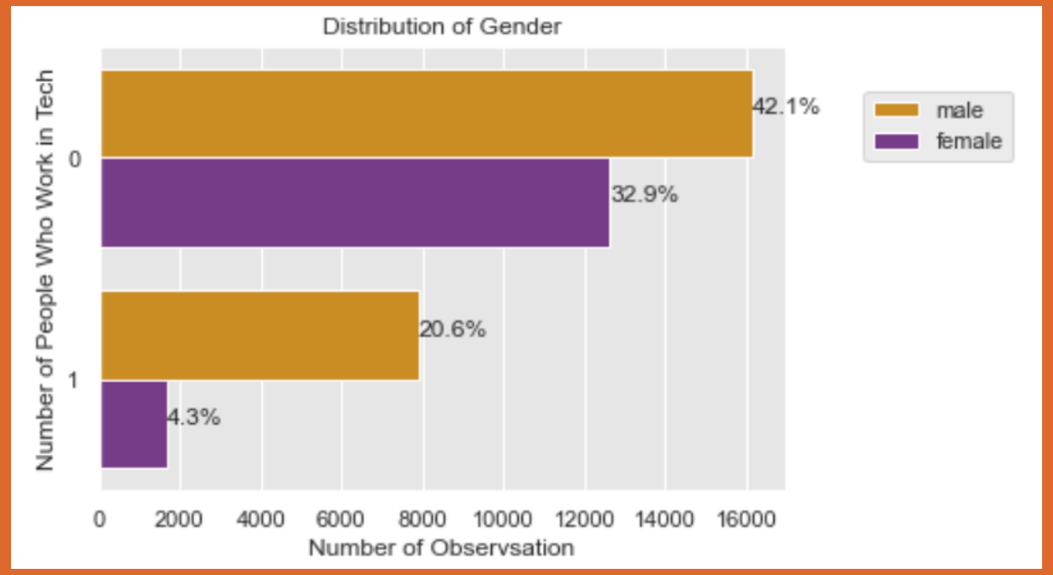
Exploratory Data Analysis

Once I began to visualize the data, patterns started to emerge, between those who work in tech and the general population. Here we can see that those who were in tech tend to a bit younger with a shorter range of ages when it comes to women. We see a similar pattern when it comes to height, with those that work in tech seeming to tend to be a bit taller. Notably, this is likely affected by the overall class sizes.

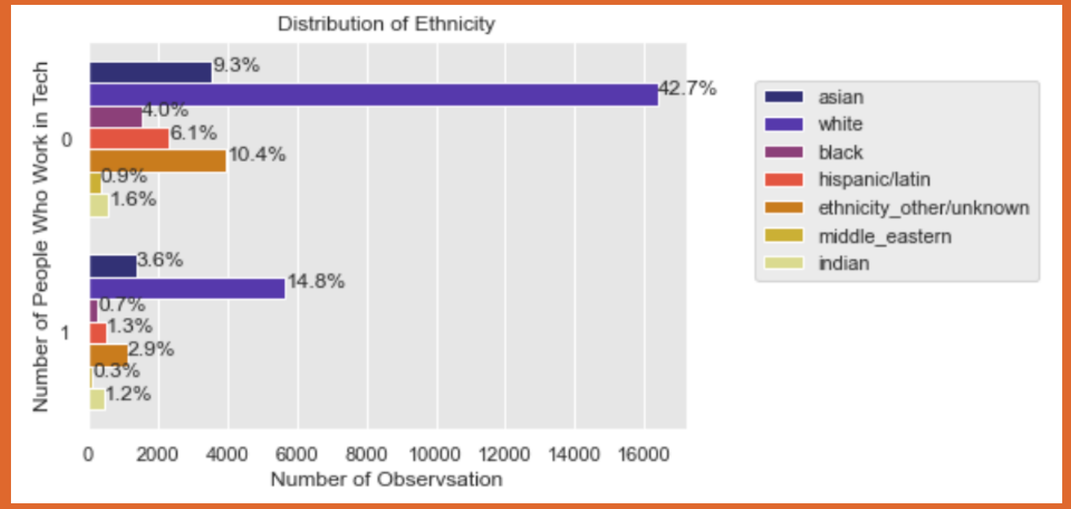




Distribution of Gender

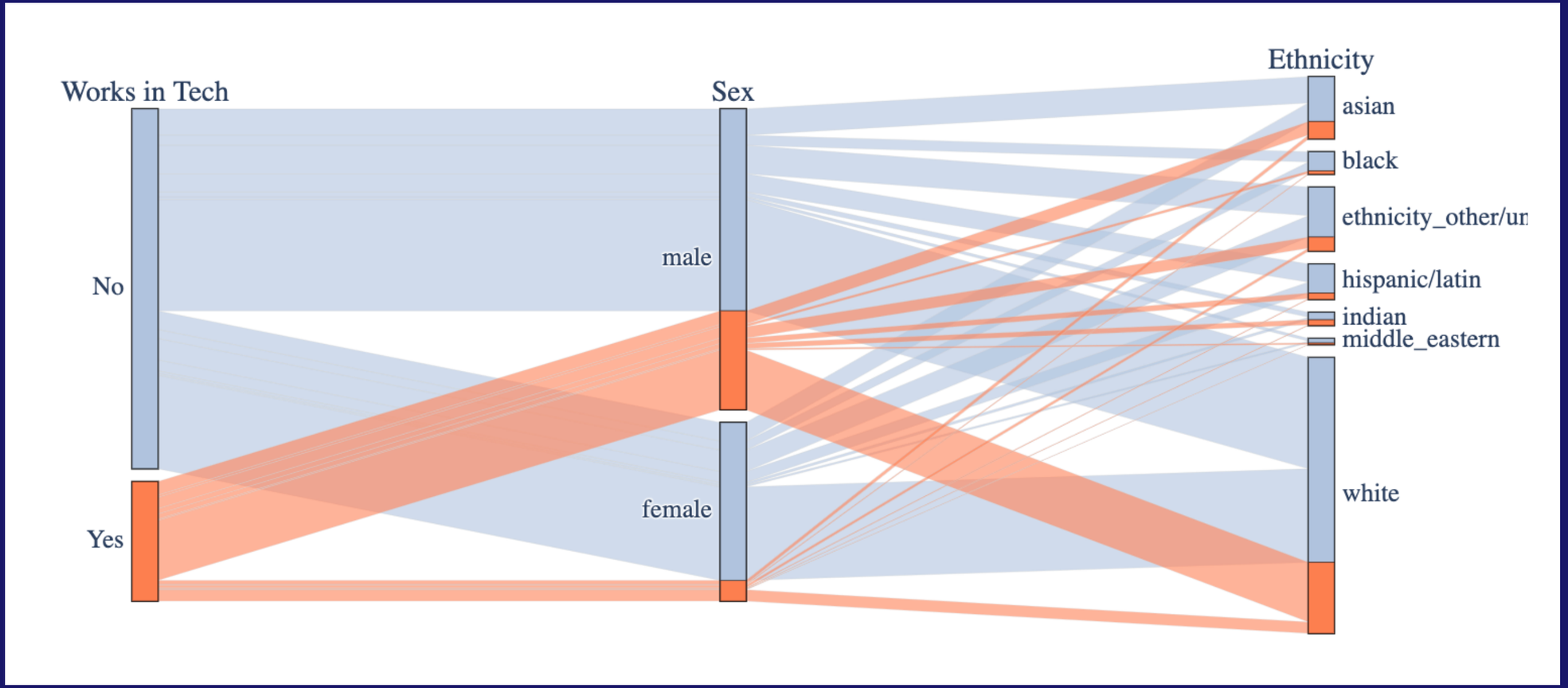


Distribution of Ethnicity



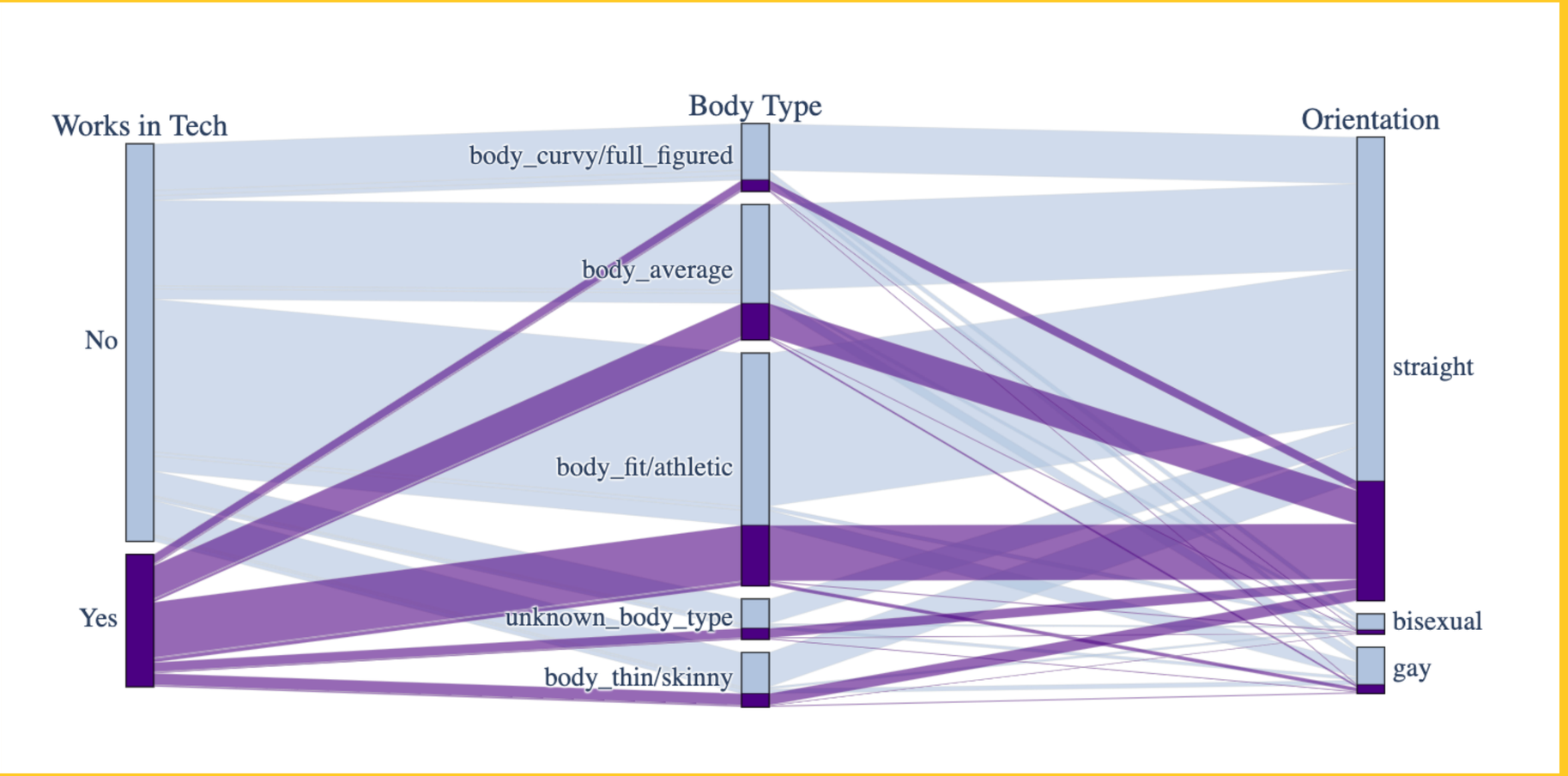
Sex and Ethnicity Distribution

As previously mentioned, there is a class imbalance in my data but even so, there is an emerging pattern when looking at the charts proportionally. Here we can clearly see that although the non-tech class is bigger the male-to-female ratio is very different. as is the ethnic distribution.



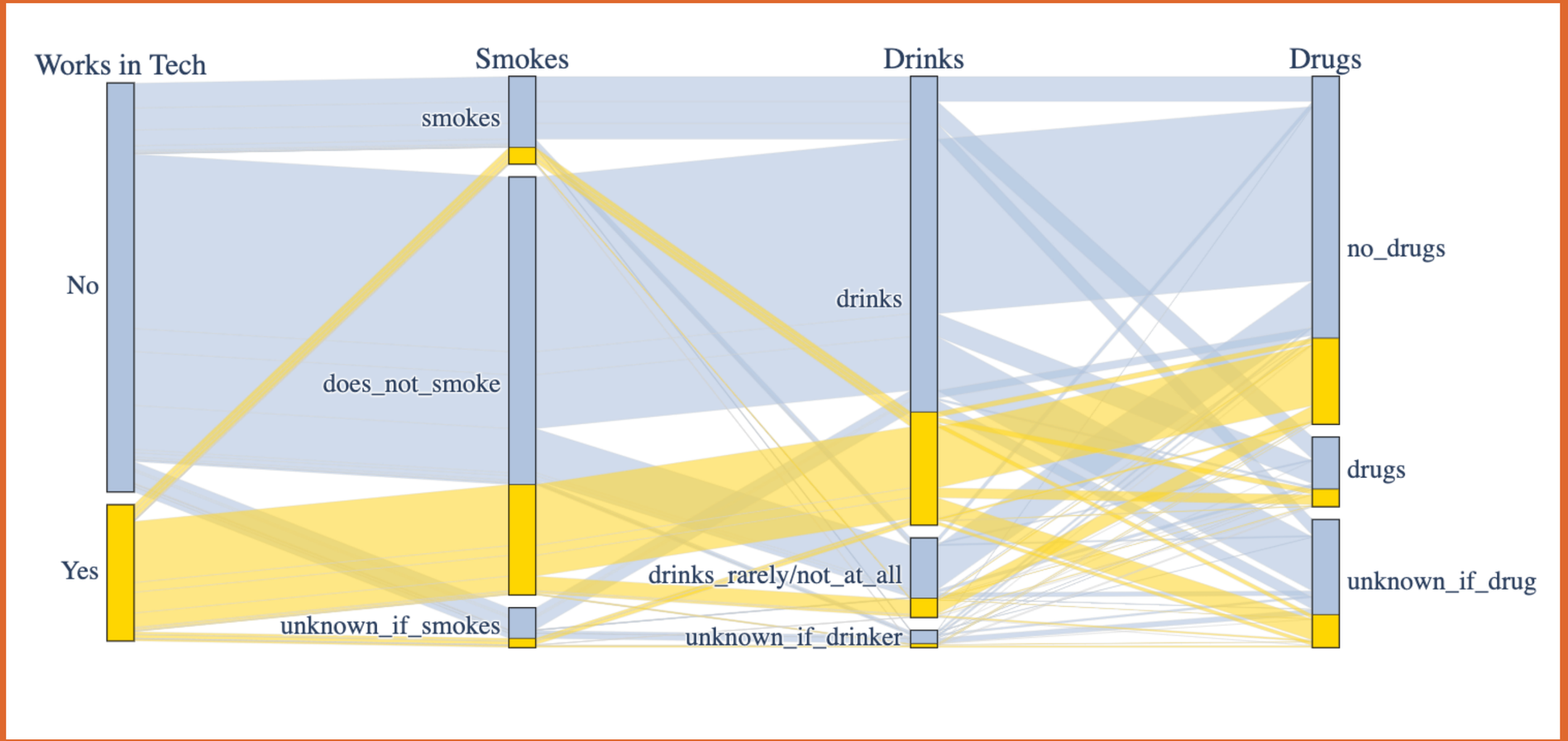
Body Type and Sexual Orientation Distribution

Between body type and orientation, the pattern is decidedly less pronounced but still marked between the two classes. Though both of those features also had a class imbalance which may be skewing the graph a bit.



Drugs, Drinking, & Smoking Distribution

In this chart, we can see that there isn't much difference between the two classes proportionally and both groups have a good amount of variation when it comes to drugs.



Model Test

Models Tested:

- Logistic Regression
- KNN
- Decision Tree

I tested a total of seven classifiers on this model. It quickly became apparent that my data was even noisier than I had initially thought as a result every model was overfit.

```
Model Type:  Logistic Regression
Training Precision:  0.6746399602714782
Testing Precision:  0.3907078266292485

Training Recall:  0.7088134266707248
Testing Recall:  0.6584340514976353

Training Accuracy:  0.6834862385321101
Testing Accuracy:  0.6602296450939458

Training F1-Score:  0.6913046243877615
Testing F1-Score:  0.4904109589041097
```

```
Model Type:  KNN
Training Precision:  0.5023812733072741
Testing Precision:  0.24864937871420853

Training Recall:  0.9723466237662507
Testing Recall:  0.9674198633736206

Training Accuracy:  0.5046088960389582
Testing Accuracy:  0.2660490605427975

Training F1-Score:  0.6624798187015833
Testing F1-Score:  0.3956162028580638
```

```
Model Type:  Decision Tree
Training Precision:  1.0
Testing Precision:  0.33168805528134254

Training Recall:  0.9997825992434454
Testing Recall:  0.3531266421439832

Training Accuracy:  0.9998912996217226
Testing Accuracy:  0.6627087682672234

Training F1-Score:  0.999891287804666
Testing F1-Score:  0.3420717739882922
```

Ensemble Model Test

The ensemble methods in some cases did perform better than the solo methods however they were all incredibly unstable. For this reason I chose not to use any of them for my final model.

```
Model Type:  Random Forest
Training Precision:  0.6752847255282528
Testing Precision:  0.3749644381223329

Training Recall:  0.8017739901734858
Testing Recall:  0.6925906463478718

Training Accuracy:  0.7081177442497499
Testing Accuracy:  0.6370041753653445

Training F1-Score:  0.7331133463205183
Testing F1-Score:  0.48652639350313776
```

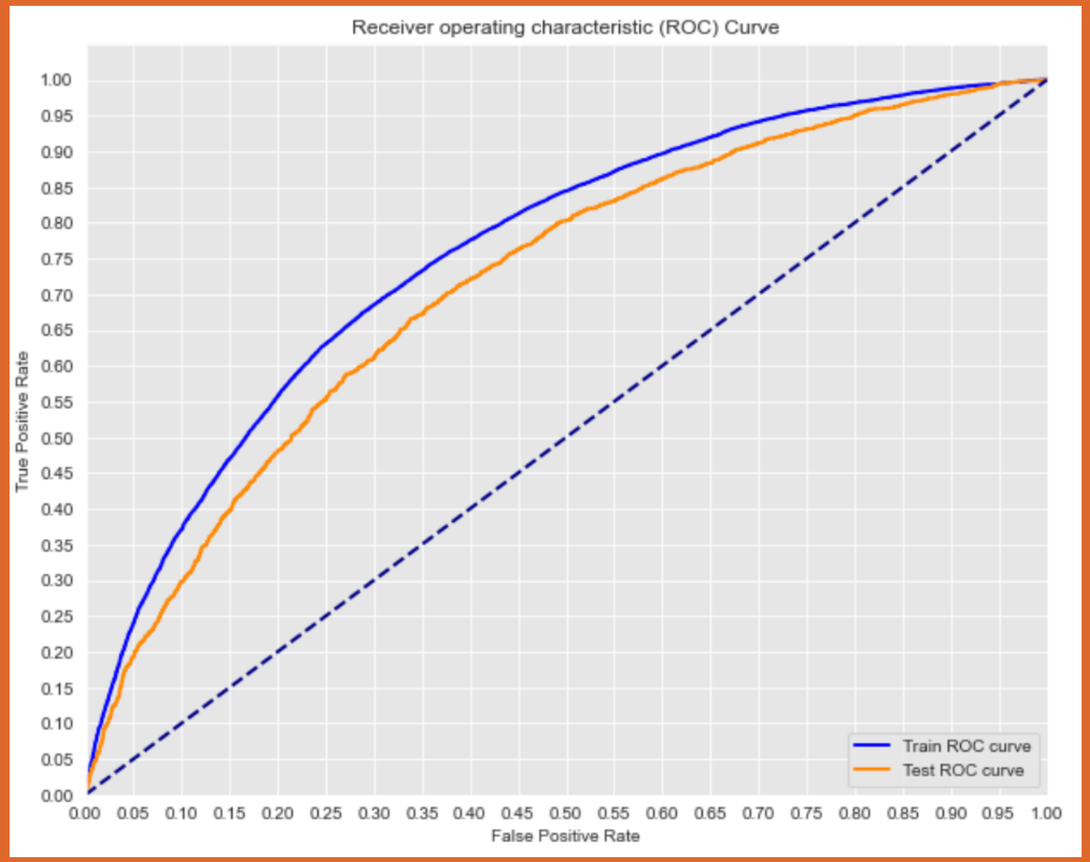
```
Model Type:  Adaboost
Training Precision:  0.7527481678880746
Testing Precision:  0.4090909090909091

Training Recall:  0.7860341753989304
Testing Recall:  0.5438780872306884

Training Accuracy:  0.7639245184573242
Testing Accuracy:  0.6916753653444676

Training F1-Score:  0.7690311602679996
Testing F1-Score:  0.46695240243627345
```

```
Model Type:  Gradient Boosting
Training Precision:  0.8356511328053049
Testing Precision:  0.4560989982321744

Training Recall:  0.7890343058393843
Testing Recall:  0.40672622175512346

Training Accuracy:  0.8169268229053437
Testing Accuracy:  0.7322546972860126

Training F1-Score:  0.8116739349211674
Testing F1-Score:  0.43
```

```
Model Type:  XGBoost
Training Precision:  0.8307439974486309
Testing Precision:  0.4532332563510393

Training Recall:  0.7928170790034349
Testing Recall:  0.41250656857593276

Training Accuracy:  0.8156441584416714
Testing Accuracy:  0.7305584551148225

Training F1-Score:  0.8113375456082583
Testing F1-Score:  0.43191196698762035
```

Models Tested:

- Logistic Regression
- KNN
- Decision Tree



The Logical Choice was Logistic Regression .... Sort Of

Logistic Regression was ultimately the model that performed the best on my data set, however, even after extensive feature selections and tuning it did not improve much from the baseline. This indicates to me that either this class can't be reasonably predicted from these features or different work needs to be done on the underlying data.

R|R

Conclusion & Next Steps

This project took many twists and terms and ultimately my model did not perform as well as I may have liked but I do think there are reasons for that beyond the model. The fact that the model had trouble predicting is a good thing. It means my target is varied enough to not be easily distinguished which leads me to think that the population of people who work in tech, is in fact, quite diverse. Though further testing would be needed to prove this theory.

Next steps: in order to take this project to the next level, I believe there are a number of things that could be done. Among the first steps would be to re-integrate many of the attitudes and more specific categories that were simplified early on in this project. Next, I think using data from a non-dating site would perhaps also give a clearer picture since it is hard to say if this sample is truly representative of the whole give where the data was sourced from.

Thank You.

Questions?

Email: dataonatangent@gmail.com
Twitter: @dataonatangent
Medium: dataontangent.medium.com