Machine Learning Capstone

—

By Nicolai Banke

Exploring the data

Orientation and Religion

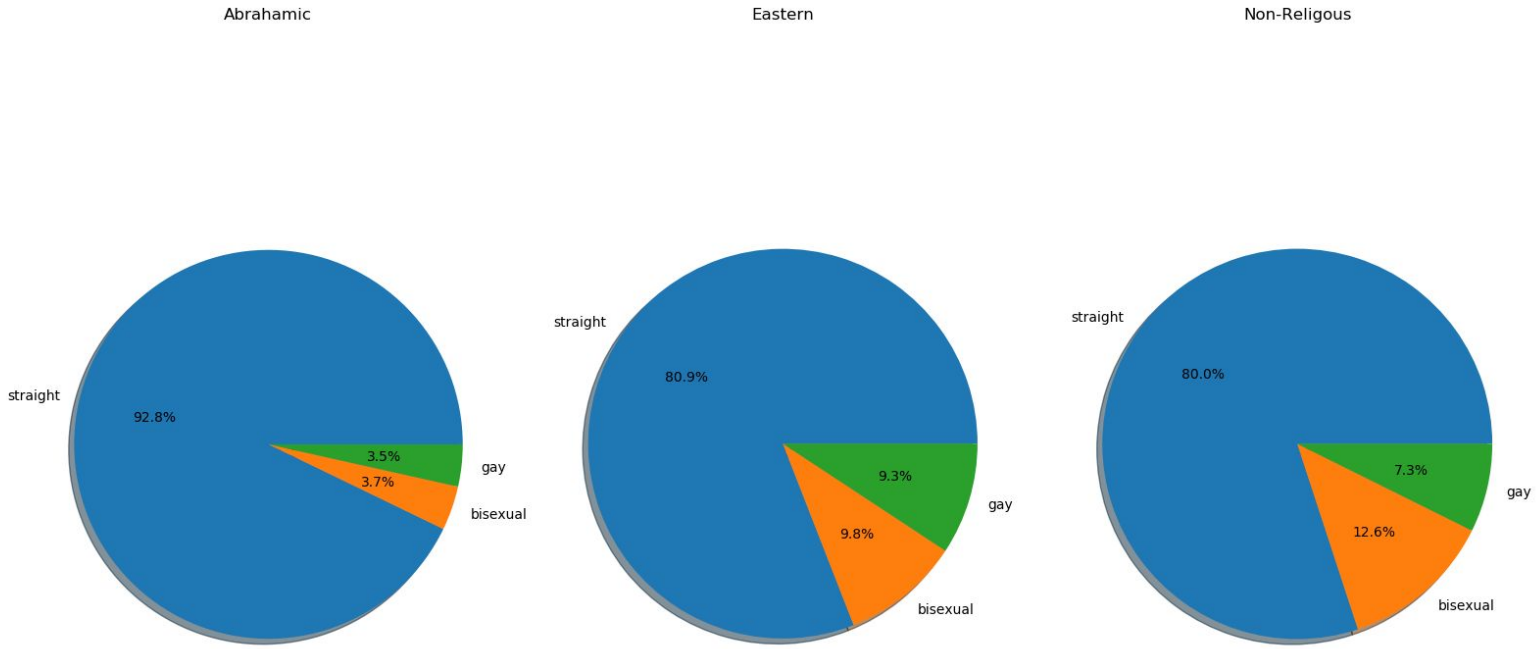
First we will look at the sexual orientations of the participants. Since it turned out that hardly any men identified as bisexual, for the remainder of this study only the female responses will be analyzed.

In the next slide are pie charts showing the distribution of orientations for three groups of religions which will be explained later. Interestingly, one group seems to have significantly more women identifying as straight, while another has more identifying as bisexual than as gay.

The new columns were created by

```
df = pd.read_csv("profiles.csv")
female = df[["drugs", "drinks", "religion", "smokes", "orientation"]][df.sex == "f"]
drugs_mappings = {"never": 0, "sometimes": 1, "often": 2}
drinks_mappings = {"not at all": 0, "rarely": 1, "socially": 2, "often": 3, "very often": 4, "desperately": 5}
smokes_mappings = {"no": 0, "trying to quit": 1, "when drinking": 2, "sometimes": 3, "yes": 4}
eastern_mappings = {
                "buddhism but not too serious about it": "little" ,
                "buddhism and laughing about it": "none" ,
                "buddhism": "some" ,
                "buddhism and somewhat serious about it": "somewhat",
                "hinduism but not too serious about it": "little",
                "hinduism": "some",
                "buddhism and very serious about it": "very",
                "hinduism and somewhat serious about it": "somewhat",
                "hinduism and laughing about it": "none",
                "hinduism and very serious about it": "very"
}
religion_mappings_to_numbers = {"none": 0, "little": 1, "some": 2, "somewhat": 3, "very": 4}
female["eastern"] = female.religion.map(eastern_mappings).map(religion_mappings_to_numbers)
female_eastern = female[["orientation", "eastern", "drinks", "smokes", "drugs"]].dropna()
```

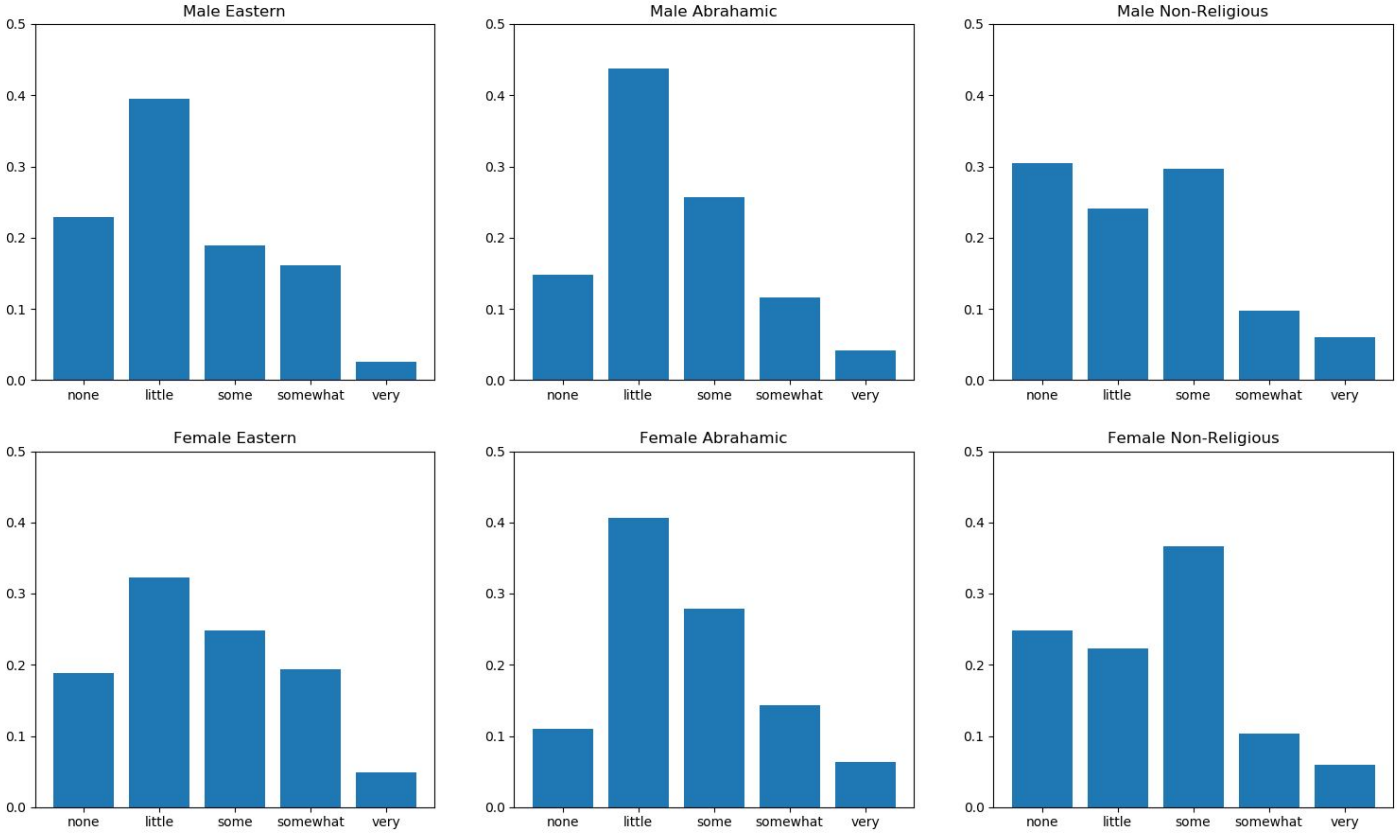
and similarly for Abrahamic (christianity, catholicism, judaism, islam) and non-religous (atheism, other, agnosticism).


Abrahamic
straight
92.8%
3.5%
gay
3.7%
bisexual
Eastern
straight
80.9%
9.3%
gay
9.8%
bisexual
Non-Religous
straight
80.0%
7.3%
gay
12.6%
bisexual

Religion

For the sake of exploring the dataset, the next slide shows bar plots describing the distribution of religious adherence between the sexes. Three categories are the same as before: Abrahamic religions (Christianity, Islam, Judaism and Catholicism), Eastern (Buddhism and Hinduism) and Non-Religious (Atheism, Agnosticism and Other). As expected, most people say that the traditional religions have little influence on their lives, and in the category non-religious, which includes other belief systems, more people say that religion has more of an influence. Moreover, we see that women in general seem to be more influenced by religion.

The categories are as described on the left, and the results were filtered by sex. The y-axes are given in percentages to make comparison easier.


Male Eastern
Male Abrahamic
Male Non-Religious
Female Eastern
Female Abrahamic
Female Non-Religious
0.0
0.1
0.2
0.3
0.4
0.5
none
little
some
somewhat
very

Questions

The rationale behind grouping the religions the way it was done, is that the Abrahamic religions have stricter rules around morality relating to sexuality, so that stricter adherence to these religions might imply less self-identification as sexualities other than straight. Similarly, since most respondents are presumably western, openness towards eastern religions with less of an emphasis on restricting sexuality, might imply a higher degree of self-identification as sexualities other than straight.

Another question is to see if there is a correlation between religious adherence and vices such as drinking, smoking and doing drugs.

Now we will proceed to see if a K Nearest Neighbor algorithm or Support Vector Machine can learn to predict sexuality based on religious beliefs and vices, and if Multilinear Regression and K Means Regression can predict religious adherence based on vices.

- Can we predict adherence to religion with health-risking activities?
- Can we predict bi-sexuality from religious beliefs?

Multilinear Regression

The data were prepared as (and similarly for the other religious groups).

```
min_max_scaler = preprocessing.MinMaxScaler()

y_abrahamic_scaled =
min_max_scaler.fit_transform(female_abrahamic.abrahamic.values.reshape(-1, 1))

x_abrahamic_scaled = min_max_scaler.fit_transform(female_abrahamic[["drinks", "smokes",
"drugs"]].values)
```

The independent values are the vices (“drinks”, “smokes”, “drugs”) and the dependant variable is the religion (“abrahamic”, “eastern” or “non_religious”).

The results of this regression became

```
|-------------------------------Abrahamic-----------------------------------|
coef_ = [[-0.21171999  0.01233301 -0.19505979]]
intercept_ = [0.51021295]
score train: 0.0258930012315
score test: 0.0397389031594
|-------------------------------Eastern-------------------------------------|
coef_ = [[-0.26800835 -0.00577933 -0.09428763]]
intercept_ = [0.49611984]
score train: 0.0329805902261
score test: -0.0694279221229
|-----------------------------Non-Religious---------------------------------|
coef_ = [[-0.08262075 -0.0058954  -0.00905037]]
intercept_ = [0.4140075]
score train: 0.00251167926341
score test: 0.00331405851064
|---------------------------------------------------------------------------|
```

We see that most coefficients are negative, indicating that higher degrees of vices implies less adherence to religion. However, the correlation is very weak, but the Abrahamic religions show a slightly stronger correlation.

K-Means Regression

For K-Means Regression we prepared the data as in the previous slide. The next slide shows the accuracies of the three models with respect to the k. The most accurate models were created with k's equal to

K = 88 for Abrahamic

K = 88 for Eastern

K = 97 for Non-Religious

The results are

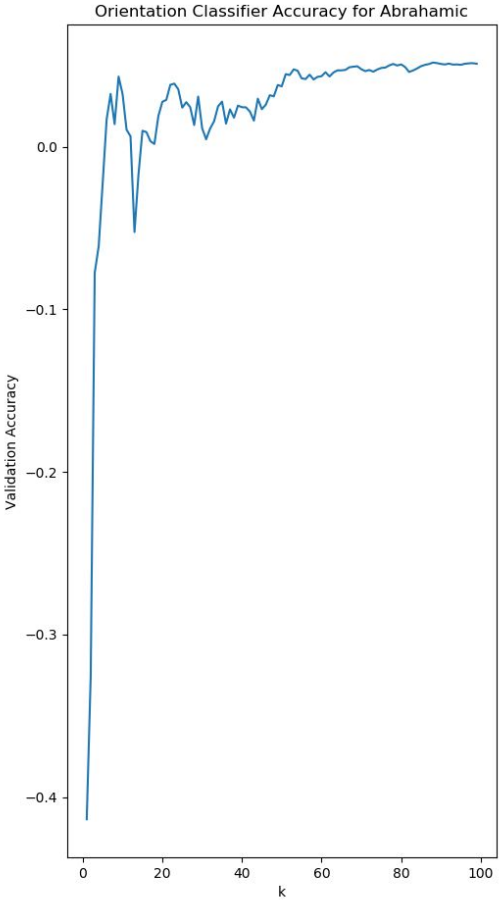
k_maxs = [88, 88, 97]

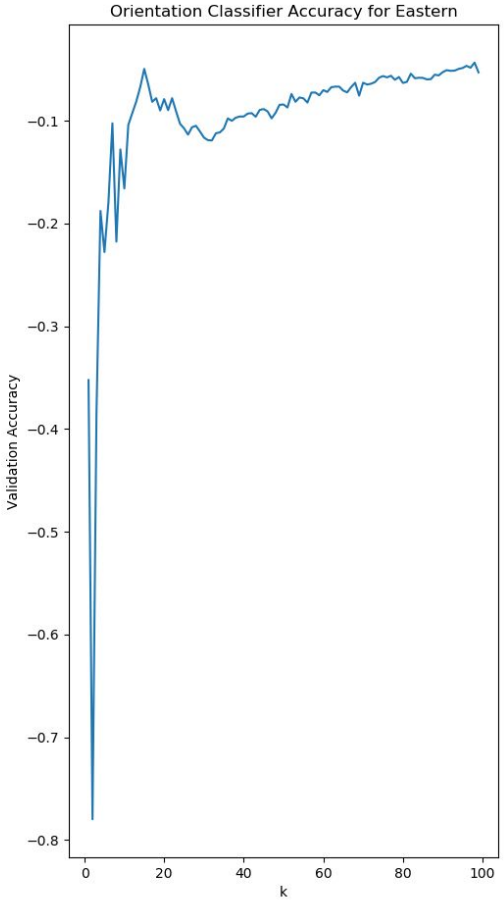
Score for Abrahamic: 0.0517334646793

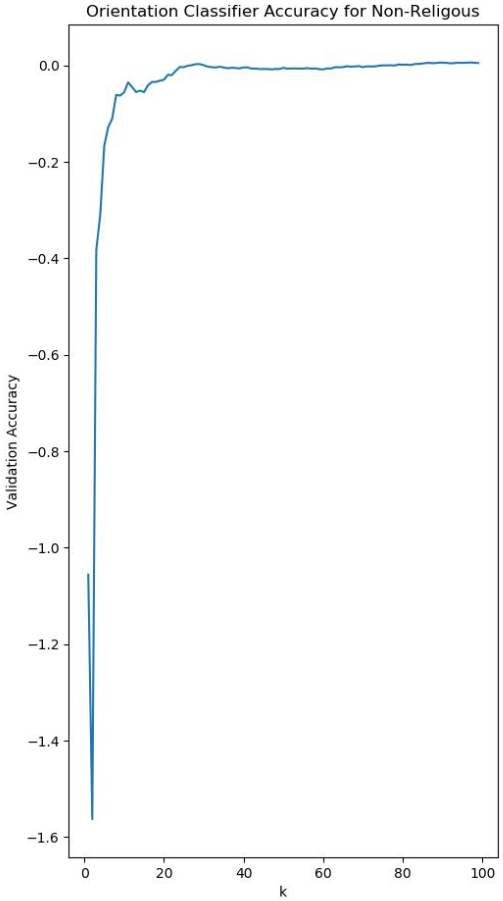
Score for Eastern: -0.0552390791027

Score for Non-religious: 0.00585402354818

Again, the scores show a weak correlation, but it is perhaps worth noting that for the non-religious group the R^2 is an order of magnitude lower, indicating even less of a correlation for non-religious women.


Orientation Classifier Accuracy for Abrahamic
Validation Accuracy
k
0.0
−0.1
−0.2
−0.3
−0.4
0
20
40
60
80
100


Orientation Classifier Accuracy for Eastern
Validation Accuracy
k
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
0
20
40
60
80
100


Orientation Classifier Accuracy for Non-Religous
Validation Accuracy
k
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
−1.6
0
20
40
60
80
100

Classification - K-Nearest Neighbors

For K-Nearest Neighbors classification, we want to predict sexual orientation, based on religious adherence and tendency to vices. The next slide shows the accuracies of the three models with respect to the k. The most accurate models were created with k's equal to

K = 7 for Abrahamic

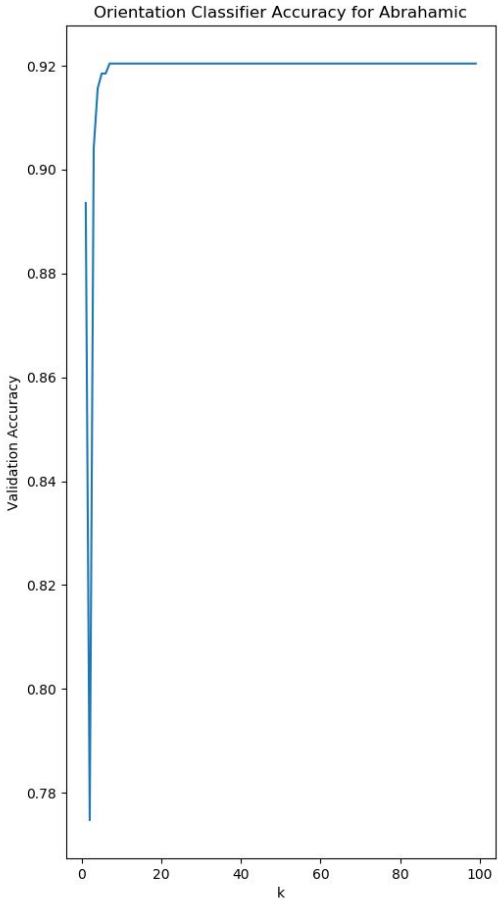
K = 16 for Eastern

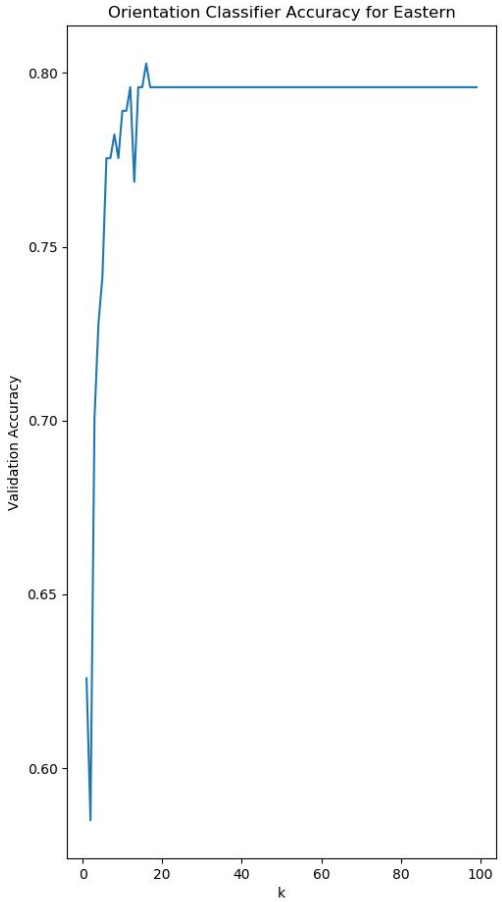
K = 53 for Non-Religious

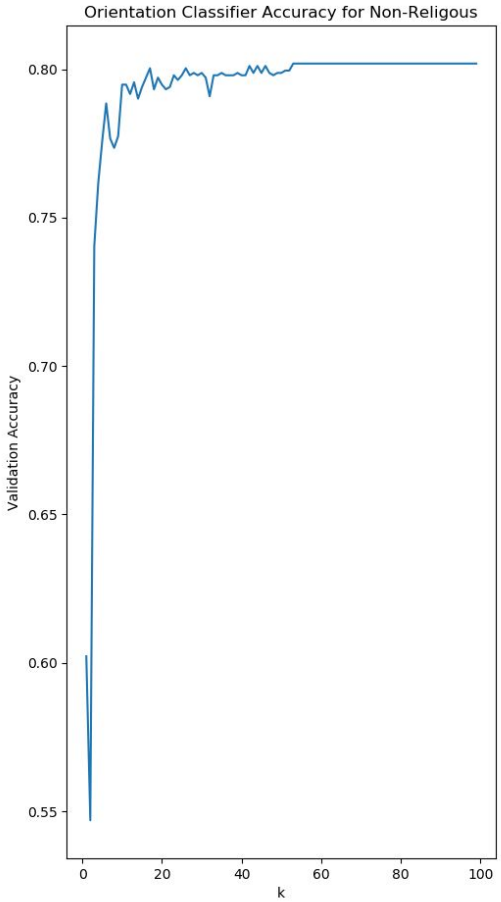
The results are

```
k_maxs = [7, 16, 53]
Score for Abrahamic: 0.920421860019
Score for Eastern: 0.802721088435
Score for Non-Religious: 0.801894238358
```

These results look very close to the percentages displayed in the pie charts a few slides back, so this model has not managed to be better at predicting orientation


Orientation Classifier Accuracy for Abrahamic
Validation Accuracy
k
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0
20
40
60
80
100


Orientation Classifier Accuracy for Eastern
Validation Accuracy
k
0.80
0.75
0.70
0.65
0.60
0
20
40
60
80
100


Orientation Classifier Accuracy for Non-Religous
Validation Accuracy
k
0.80
0.75
0.70
0.65
0.60
0.55
0
20
40
60
80
100

Classification - Support Vector Machine

For Support Vector Machine classification, we again want to predict sexual orientation, based on religious adherence and tendency to vices. The most accurate models were created with parameters $C = 1$ and $k = 9$ and all three cases.

The results are shown on the right, and again, the algorithm didn't seem to outperform randomly picking a label. In fact, for Eastern and Non-Religious is did slightly worse, and the whole process took a lot longer, because two parameters had to be found for the most accurate models.

The results are

Score for Abrahamic: 0.920421860019
Score for Eastern: 0.78231292517
Score for Non-Religious: 0.797158642463

Conclusions

It appears that little or no correlation was found between vices and religious adherence or vices and religious adherence with sexual orientation. Both classification algorithms did not perform better than randomly picking a female respondent from the survey. For the regression techniques, there did seem to be a small indication that more frequent drinking, smoking and drug habits correlated negatively with being devoutly religious, with the non-religious group being less dependent on frequency of vices. It should be noted in this context that the group non-religious is somewhat ambiguous since it includes both atheism and agnosticism as well as any other type of religion, not included in the other two groups.

Perhaps the most interesting finding in this analysis was that as much as 92% of the female respondents adhering to Abrahamic religions, identified as straight, while that number was only 80% in the other two groups. Also, for the non-religious group, bisexuality was more common than being gay, which could indicate less restricted sexuality for women not subscribing to the conventional religions.

Conclusions

Instead of analysing a person's vices and religion, we could have run a Naive Bayes Classification algorithm on the personal essays. Perhaps using certain words or bringing up certain topics is a better indicator of something as personal as sexual orientation.