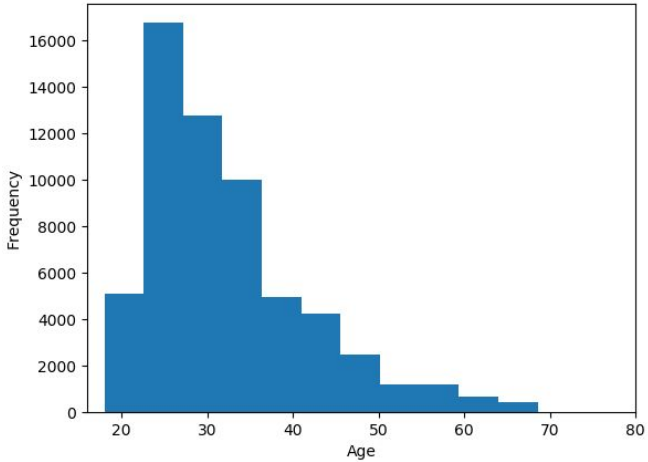
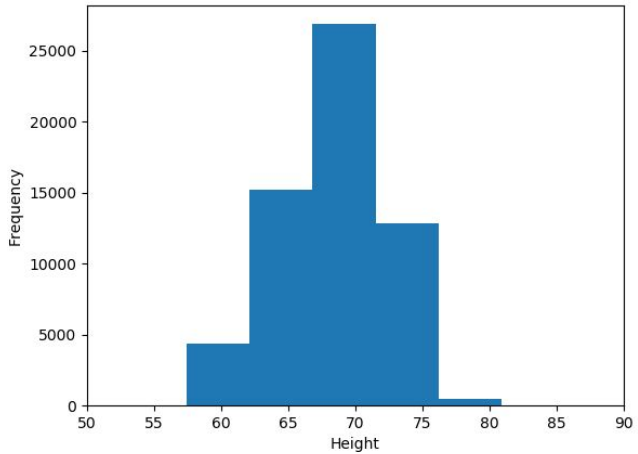

Date-A-Scientist
Capstone Project - Machine Learning Fundamentals
Clay Thomas
January 29, 2019

Presentation Contents





Exploring the Data

Two of the initial graphs I used to explore the data were:

1. A histogram showing the frequency of each age
2. A histogram showing the frequency of each height

I also took a look at the contents of the ‘religion’ feature, using .head() and .value_counts().

The Question

After exploring the dataset, I decided I wanted to focus on predicting people's religious beliefs.

Specifically, **can dataset features help us predict whether or not a person is traditionally religious?**

For this question, I took 'atheism', 'agnosticism', and 'other' answers to indicate that a person is not traditionally religious.

After exploring the data with my initial question in mind, I decided to narrow my focus to the following 6 features:

1. Drinks
2. Smokes
3. Drugs
4. Diet
5. Age
6. Education
7. Religion (target)

Augmenting the Data

I wanted to be able to easily move between categorical and numerical values for the features I was interested in. Therefore, I created new columns containing numerical codes for each of the relevant categorical features (drinks/smokes/drugs/diet/ethnicity/education), by following the directions in capstone_instructions.md. For instance, to create a column containing numerical codes for diet, I used:

```
diet_mapping = {"mostly anything": 0, "anything": 1, "strictly anything": 2, "mostly vegetarian": 3, "mostly other": 4, "strictly vegetarian": 5, "vegetarian": 6, "strictly other": 7, "mostly vegan": 8, "other": 9, "strictly vegan": 10, "vegan": 11, "mostly kosher": 12, "mostly halal": 13, "strictly halal": 14, "strictly kosher": 15, "halal": 16, "kosher": 17}

df["diet_code"] = df.diet.map(diet_mapping)
```

Classification: Naive Bayes Classifier

I used the NBC to predict religiosity based on a binary metric: not-traditionally religious (1) or traditionally religious (0).

Simplicity: For this model, I decided to include only the features my priors led me to *expect* were correlated with religious belief: drug use, age, and education.

Time to Run: 0:00:00.014564

Accuracy/Precision/Recall:

Accuracy: 0.5973333333333334
Precision: 0.5973333333333334
Recall: 1.0
F1: 0.7479131886477464

Comparison/Analysis: The model didn't improve on chance, as it assigned every row a '1'.

Classification: K-Nearest Neighbors

I also ran the K-Nearest Neighbors Classifier to predict religiosity based on the same binary metric I used with the Naive Bayes Classifier: not-traditionally-religious (1) or traditionally-religious (0).

Simplicity: I used the same features I used for the NBC, in order to make a direct comparison of the two models.

Time to Run: 0:00:00.200215 (at k=63)

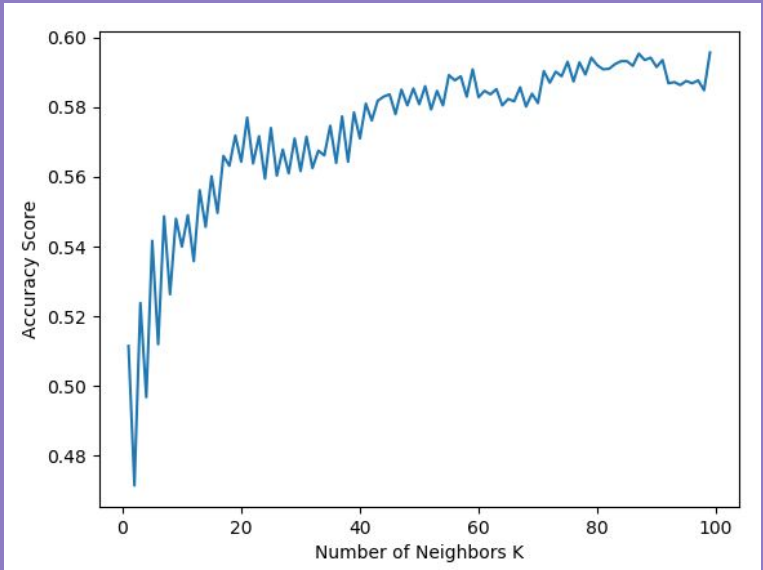
Accuracy/Precision/Recall (k=63):

Accuracy: 0.5851666666666666
Precision: 0.5961707359915686
Recall: 0.9469866071428571
F1: 0.7317020588552334

Comparison/Analysis: The model came back with slightly different (worse!) results compared to the NBC.

Classification: K-Nearest Neighbors

Accuracy went up as k increased, with an optimal k around 63 - but it still never beat random chance, which would predict 0.5926 accuracy (the model topped out around 0.5852).



Regression: K-Nearest Regressor

I ran a K-Nearest regression that would return a “Religiosity Score”, giving me the likelihood that a person was traditionally religious (closer to 1) or not (closer to 0). I

Simplicity: I added in the remaining features I’d coded in the data (drinks/smokes/diet), but otherwise kept the set-up consistent to the classification models.

Time to Run: 0:00:00.466201 (at k=63)

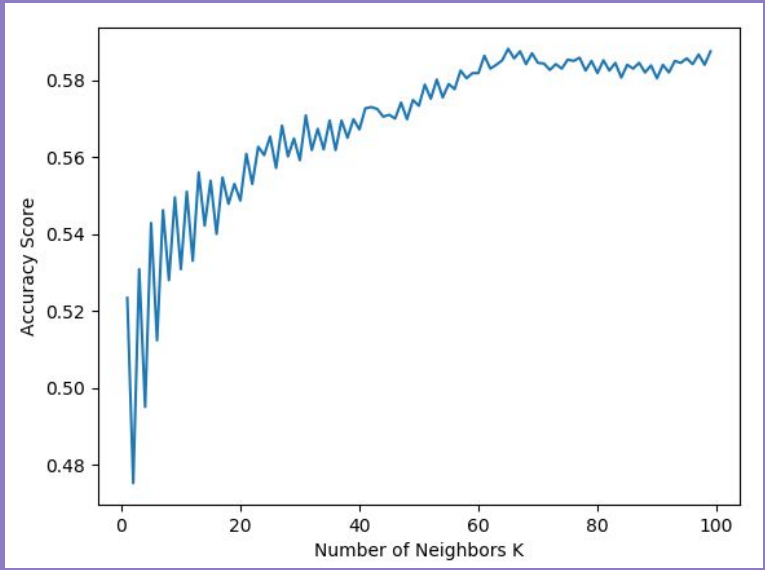
Accuracy/Precision/Recall (at k=63):

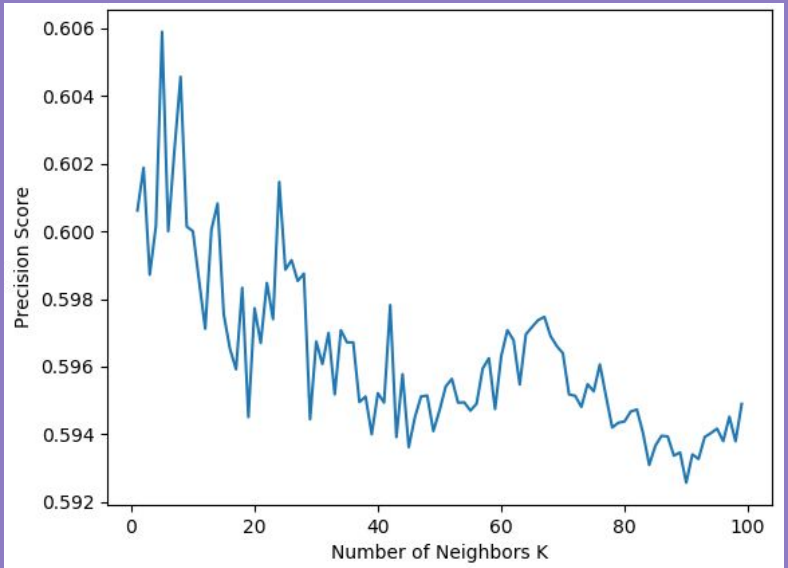
Accuracy: 0.5521666666666667
Precision: 0.5968891769280622
Recall: 0.7709263392857143
F1: 0.6728357482040668

Comparison/Analysis: The model performed still worse than the others! Worse than chance on such a big dataset - that’s kind of hard for me to wrap my head around, and really makes me feel like I’m missing something.

Regression: K-Nearest Regressor

Accuracy went up as K increased, but this model still didn't improve on chance.
Precision got slightly worse as k increased.





Regression: Multiple Linear Regression

Finally, I ran an MLR model to compare to the linear regression I'd done.

Simplicity: Identical set-up to the K-Nearest regression model.

Time to Run: 0:00:00.013372

Accuracy/Precision/Recall:

Accuracy: 0.5973333333333334
Precision: 0.5973333333333334
Recall: 1.0
F1: 0.7479131886477464

Comparison/Analysis: The validation scores were identical to the NBC because I used .round() after receiving a continuous/binary value error. I know there's got to be a better way to do this. Unfortunately am out of time and need to submit! Planning to work on this issue over the next couple of days.

Preliminary Answers and Analysis

Overall, the models I created proved to be no better than chance at predicting a person's religiosity/agnosticism. If I trusted my methods, this might be an interesting result - are people's beliefs really unconnected to the features I looked at?? However, I'm more inclined to think that I'm doing something wrong - I haven't had much of a chance to revisit/review the model, and I'm hoping to improve on this initial effort after receiving feedback.

I intend to continue with Python/ML resources in Codecademy to keep adding to my skillset - this is the first coding class I've ever taken, and I've really enjoyed it. I know I have a long way to go, though - and that will be clear from my code.

As my ML/Python skills improve, I would be interested to keep working on this model and add in additional features from this dataset, e.g. counting uses of the word 'God' in the essay answers and adding that to the features I'm using.

Additional Data Desired: I would love to see some data on political beliefs added to the dataset, and use that both as a feature to help predict religiosity, and as a predictive target in its own right.