DATE-A-SCIENTIST

Machine Learning Fundamentals

Justin Haut

November 11, 2018

TABLE OF CONTENTS

QUESTIONS TO ANSWER

Classification Question

- Can we predict which religion a person may be based on how much they— drink, smoke, and do drugs

  ***I parsed out only the religion piece ignoring how serious they were about it. (not the best idea in hindsight)**

Regression Question

- Can we predict income based on the amount a person drinks and smokes?

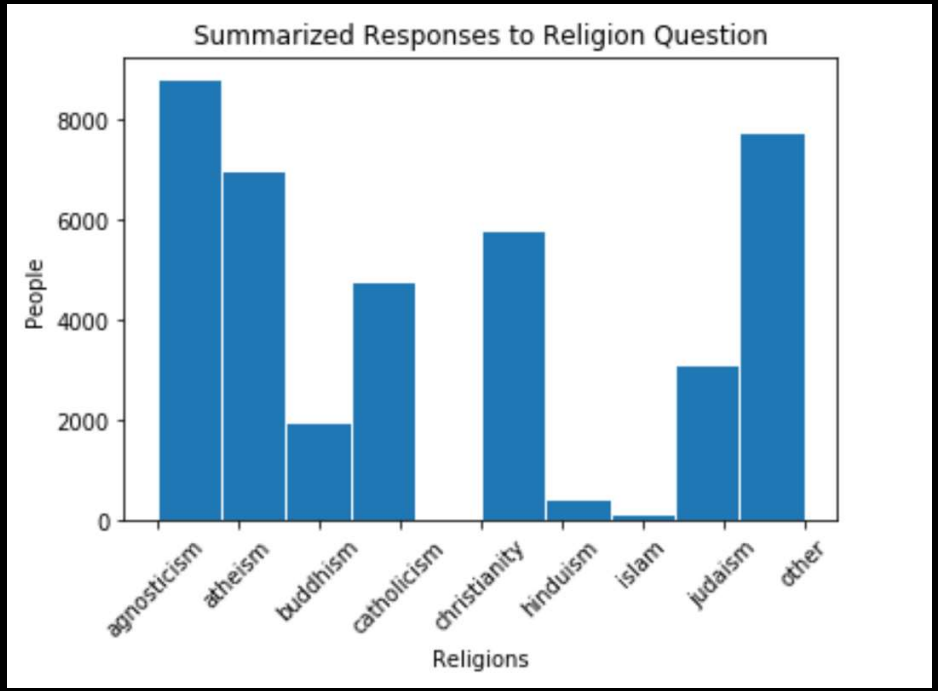
  ***I removed the people who gave a -1 answer for income.**

*****I converted NaNs under the following categories as follows… I realize this is introducing bias ☹:**

- Drink –NaN's converted to: 'maybe so and maybe not'
- Smoke – NaN's converted to: "what momma don't know don't hurt her"
- Use drugs – NaN's converted to: "experimented"

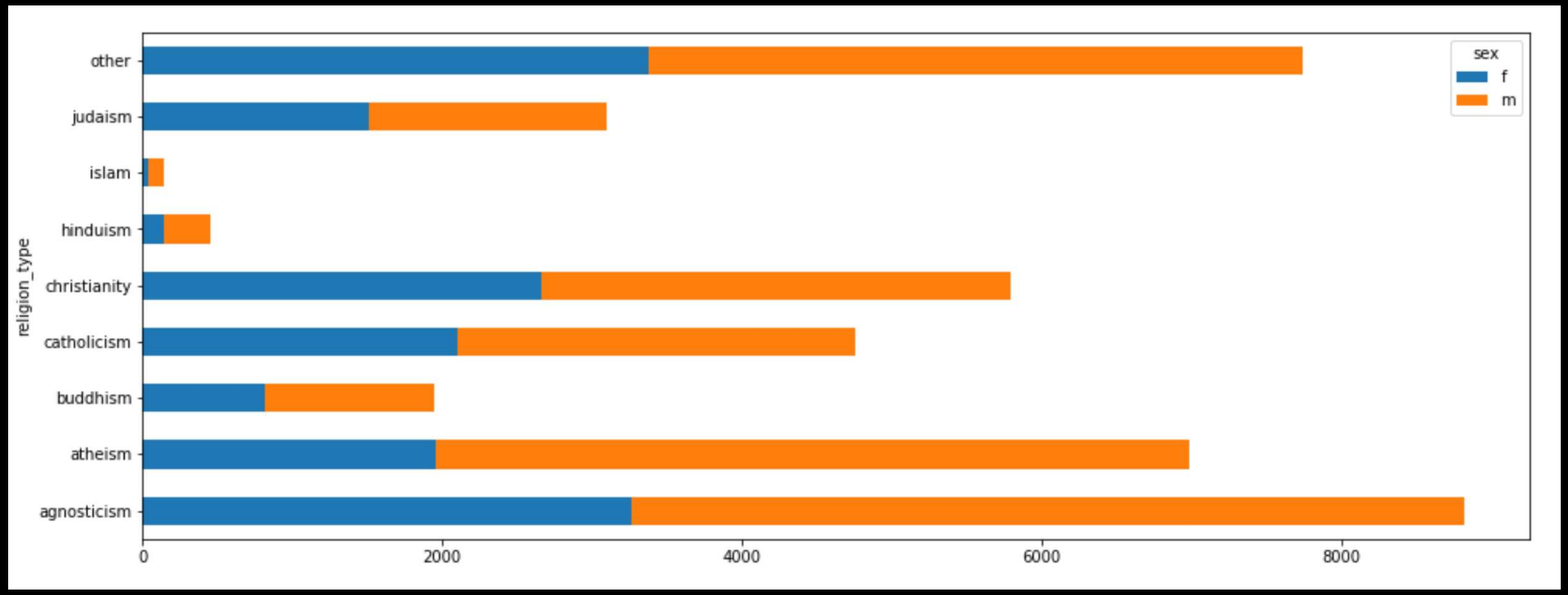
DATA EXPLORATION - RELIGION CONDENSED

- This graph depicts the spread of responses users chose as their religion.
- I ignored intensity/seriousness of practice, which, in hindsight was maybe not the best idea.

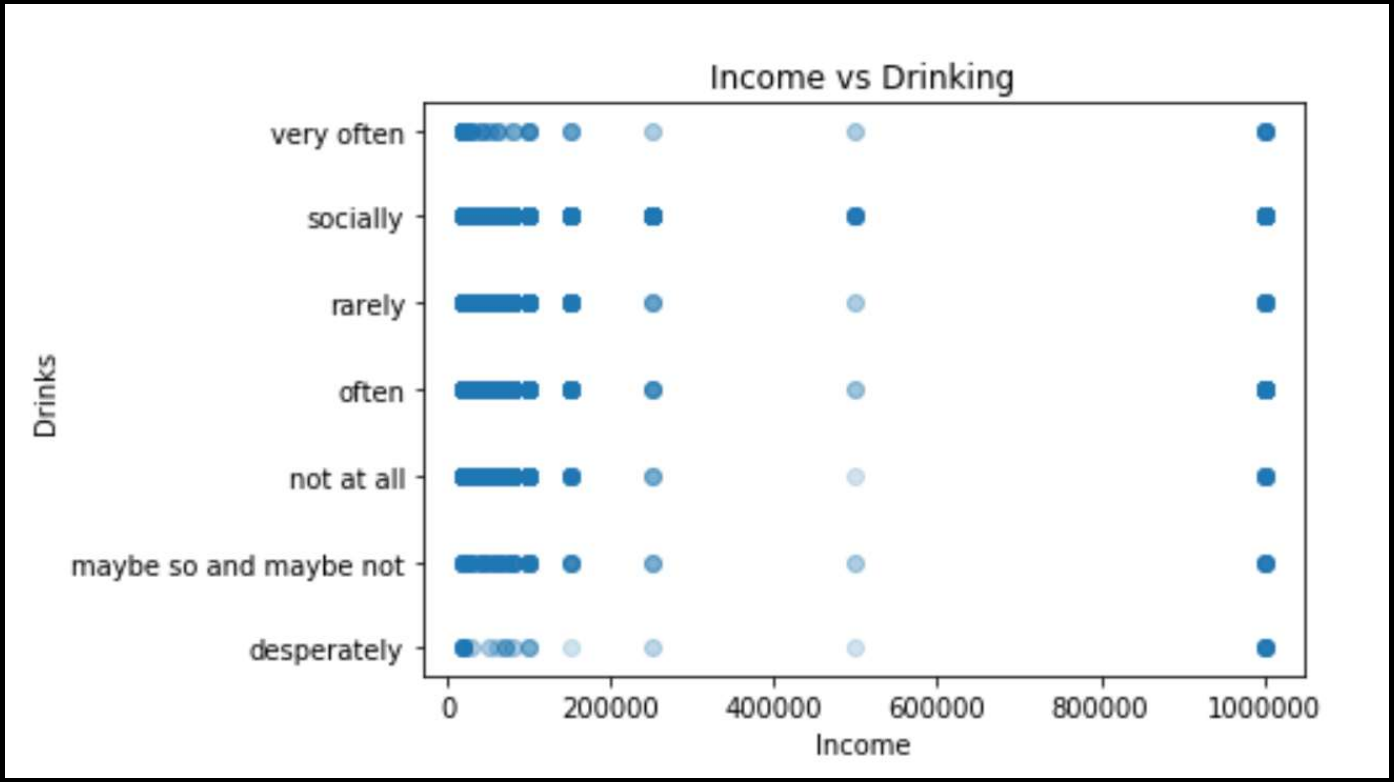


DATA EXPLORATION - RELIGION BY SEX

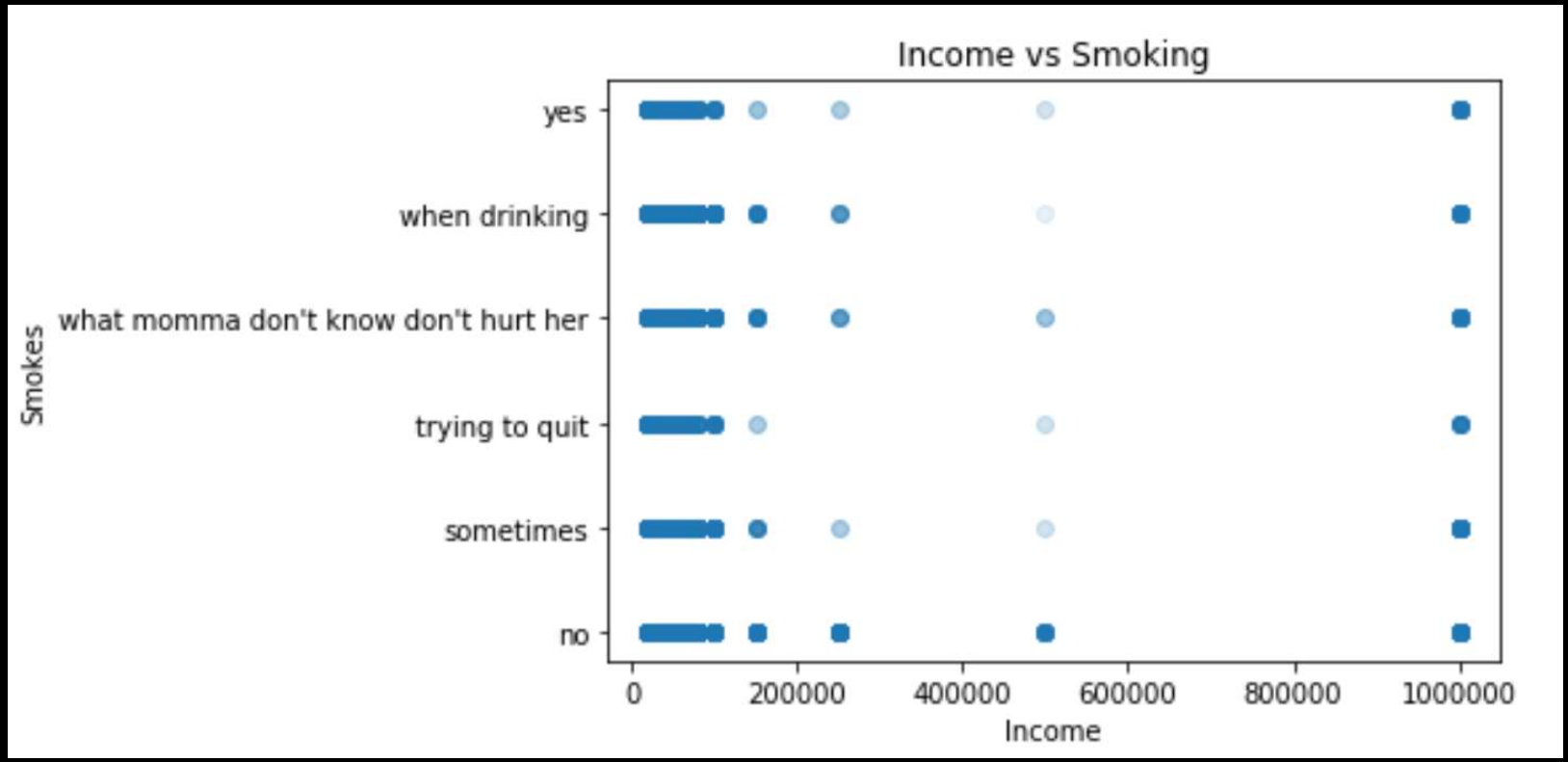
- I was curious about the break out between male and female users and their choice of religion.



DATA EXPLORATION – INCOME TO DRINKING FREQUENCY



DATA EXPLORATION – INCOME TO SMOKING FREQUENCY



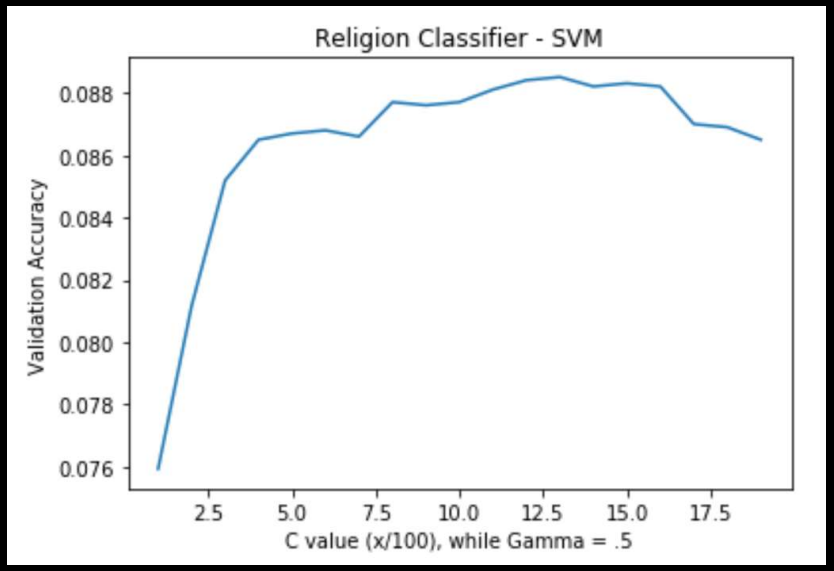
NEW COLUMNS CREATED

- Religion question columns:
  1. Created 'religion_type' by taking the first word from each answer choice.
     - df.religion.str.split(n=1).str[0]
  2. Created 'religion_vals' by mapping each unique item from religion type to an arbitrary number. In this case, in order of popularity rank.
     - df.religion_type.map({'agnosticism':10,'other':9,'atheism':8,'christianity':7,'catholicism':6,'catholicism':5,'judaism':4,'buddhism':3,'hinduism':2,'islam':1})
  3. I also created values column mappings for Drugs, Drinks, and Smokes.

RELIGION – CLASSIFICATION MODEL GRAPHS

Support Vector Machine

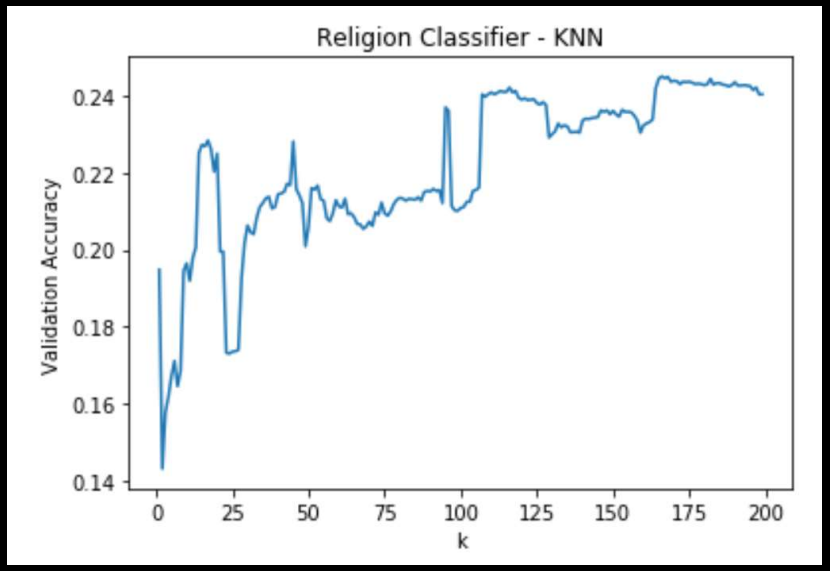
C = .125, Gamma = .5




```
SVM Validation Accuracy: 0.08862034239677745
Time to run SVM: 40.049803277600404
```

K-Nearest Neighbors

K = 166




```
KNN validation accuracy: 0.24496475327291037
Time to run KNN: 0.5987632508190472
```

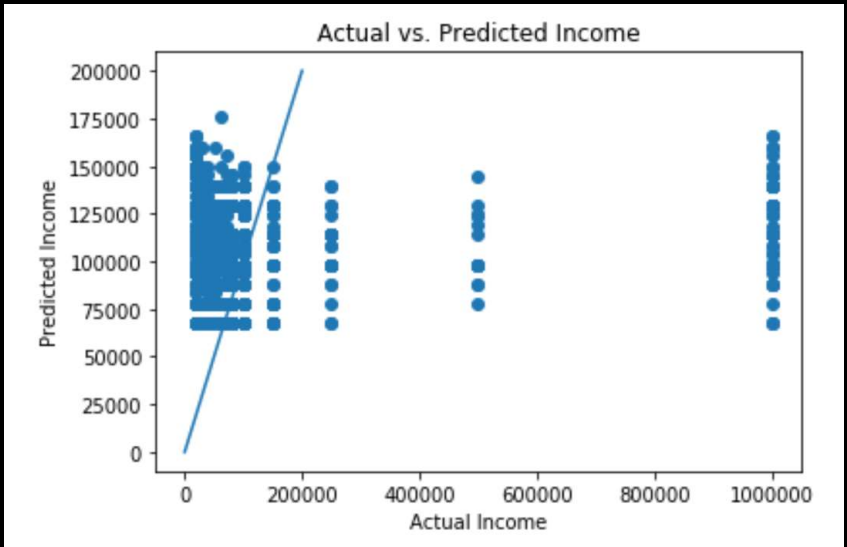
CLASSIFICATION CONCLUSION

Classification Question – Religion and Temptations

- Looking at the results between SVM and KNN it seems that while KNN has a 24% accuracy rate, I have a feeling this is because there are 5 top religions selected. 1/5.
- SVM on the other hand has an 8.8% accuracy, which seems more realistic.
- It would be interesting to dig further into this question by:
  - Breaking the dataset into age groups
  - Breaking the dataset into male and female and then looking for classification
  - Only using those users who are strict about their religion.

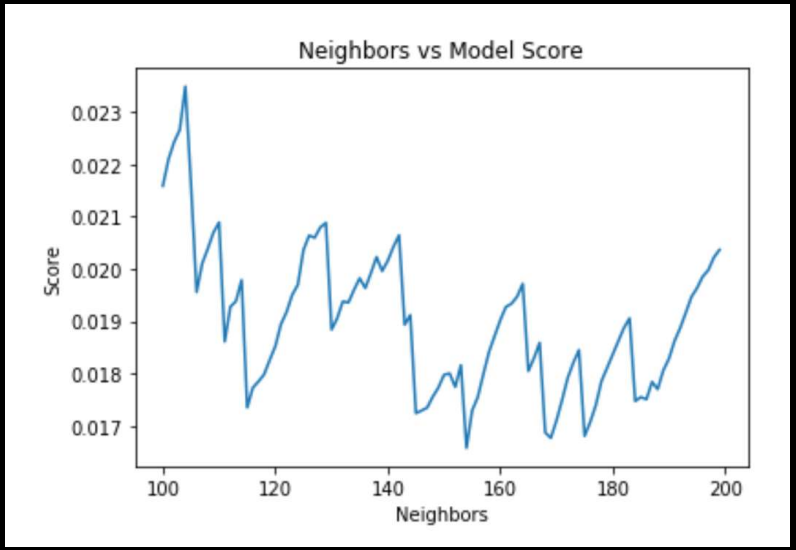
INCOME – REGRESSION MODEL GRAPHS

Multiple Linear Regression




```
Time to run MLR: 0.0031150296350688222
MLR Score: 0.009239106113491546
MLR Score: 0.014302392556309718
MLR Coefficients: [10134.14373695 15799.92667442]
```

K-Nearest Neighbors Regressor




```
Time to run KNN Regressor: 0.06612162764872664
KNN Regressor Score: 0.023491905152347337
```

REGRESSION CONCLUSION

Regression Question – Income, Cigarettes, and Booze

- Looking at the results between MLR and KNN Regressor, both scored very similarly low, .0143 vs .0234
- MLR ran much faster than KNN Regressor.
- It would be interesting to dig further into this question by:
  - Getting more data for the upper income brackets
  - Breaking the data into age groups

GOING FORWARD

- I plan to:
  - Get better with pandas by playing with public datasets
  - Read up on Scikit-Learn's documentation
  - Sign up for a Codecademy Pro membership!

THANKS FOR YOUR TIME AND FOR CREATING THIS COURSE! ☺

THE END