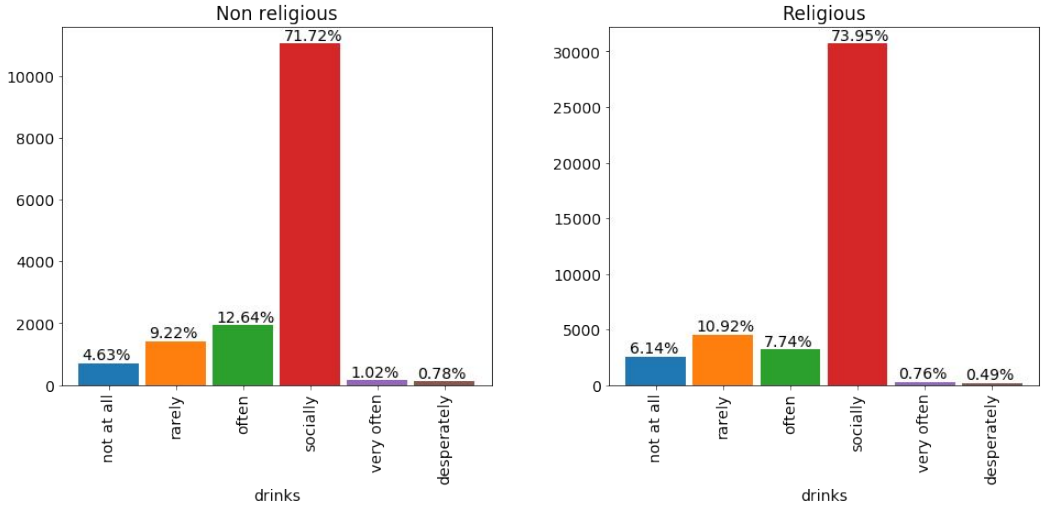
Date-A-Scientist Project

Adventures in Machine Learning

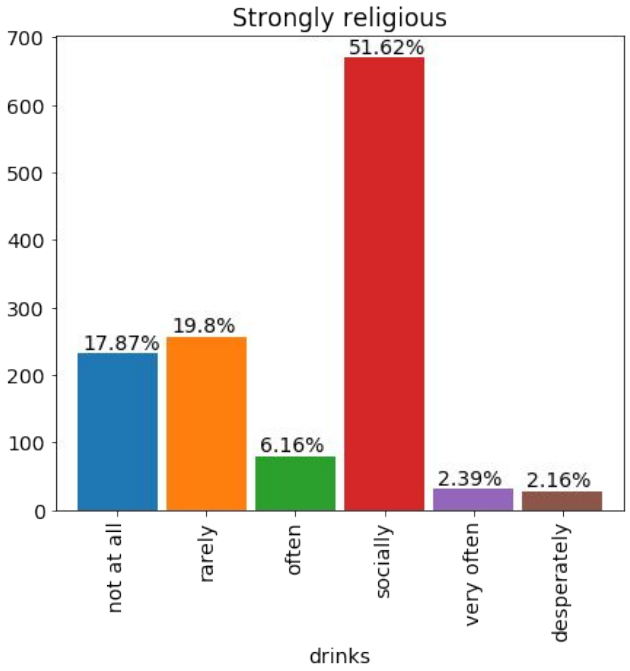
Initial Exploration - Part 1

After examining a small sample of the data I retrieved all the value counts for the non-essay questions. This led me to take a further look at how the religion and drinks columns were related. I knew that some religions advocated abstinence, but was interested to see if this held true with real world data and produced the following graphs, breaking up people in religious and non-religious.



Initial Exploration - Part 2

From my analysis it appeared that religious people were more likely to not drink at all, with the proportion of people with the drinks answer “not at all” higher amongst religious people (6.14% vs 4.63%). I then decided to see if this affect was further amplified in more devout people, that answered they were “very serious about” their religion. This effect was present with an even greater proportion of people not drinking in this group (17.87%). Interestingly it was also the case that this group proportionally contained more people that answered they drank "very often" or "desperately".



Question Time

After gaining a bit more insight into the data I decided I dig deeper and try to answer the following questions:

- Classification - **Can we predict how strongly a person believes in their religion?**
  Just as religion can possibly influence a person's view of the world, it can also influence the choices they make. I was interested in seeing how strongly this influence was filtered through to lifestyle and subsequently if it allows us to predict how strongly they held their religious beliefs.

- Regression - **Does someone being religious, and their strength of belief, affect the sentiment of their essay questions? How does this compare to other factors in their life?**
  As religion can have an impact on how people view the world I was interested to see if this influenced how positive or negative their essay questions were, how strong the effect was and how it compared to other factors.

Data Preparation

With my interest piqued I created the following new columns required to answer my questions:

- **religion_code** and **religion_strength** - Using the religion column I extracted either the specific religion (eg. 'atheism', 'buddhism', etc) or how strongly they held their religion (eg. 'but not too serious about it', etc). I then mapped these text values to numbers.
- **religiosity** - This is a number between -4 and +4. A negative number represents if they're atheist or agnostic; a positive number indicates they're christian, jewish, etc. The scale of the number is derived from **religion_strength**.
- **essays_sentiment** - To create this I merged all essay questions together. Then I used the TextBlob library (https://textblob.readthedocs.io/en/dev/) to extract the sentiment.
- **Speaks_count** - I derived this by splitting the speaks column and get a count of the values.
- I also created a number of other columns that mapped the text values to numbers, Eg. drinks_code, smoke_code, drugs_code, status_code, job_code.

Classification - Overview

My classification question was:

Can we predict how strongly a person believes in their religion?

In order to answer this I used the following features:

- age
- drinks_code
- drugs_code
- orientation_code
- sex_code
- smokes_code
- status_code
- religion_code

Classification - K-Nearest Neighbor

After creating multiple KNN classifier modifiers with an increasing number of neighbors, I ended up with the following, best accuracy:

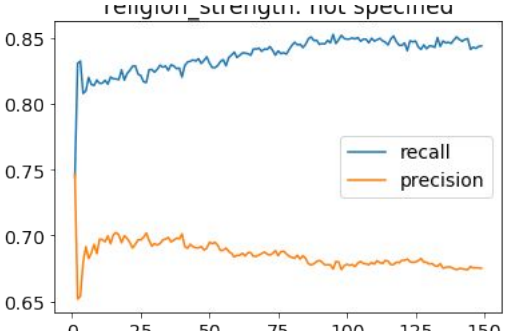
- Max Accuracy: 0.5876219868212528
- Best number of neighbors: 101

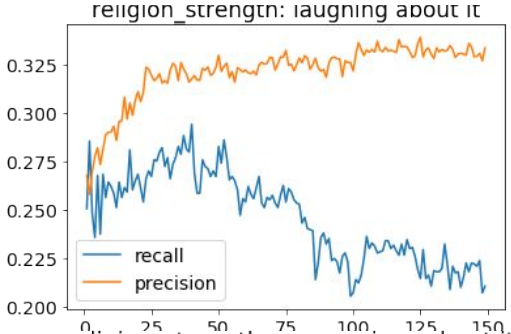
However the model was quite poor at predicting labels with fewer items, as demonstrated in the following classification report for the most accurate model:

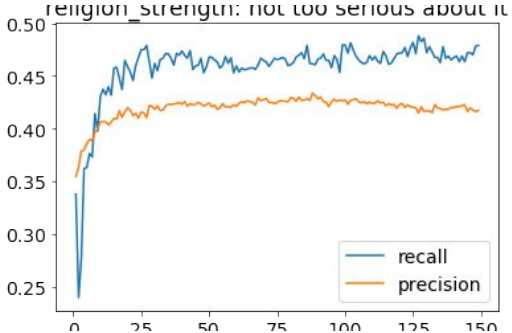
	Precision	Recall	f1-score	support
0	0.68	0.85	0.75	6452
1	0.34	0.21	0.26	1749
2	0.43	0.48	0.45	2458
3	0.00	0.00	0.00	876
4	0.31	0.02	0.03	454
avg / total	0.55	0.58	0.53	11989

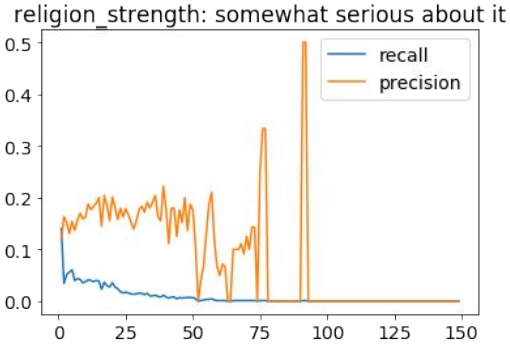
Classification - KNN Precision + Recall

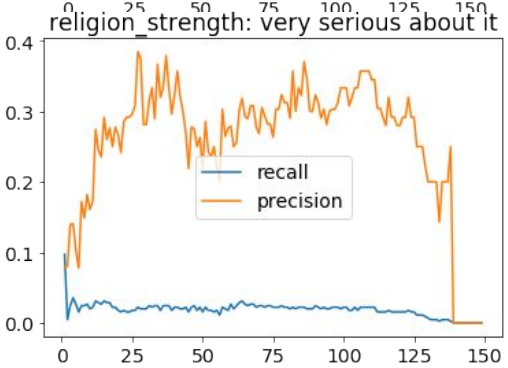
Below are the progression in precision and recall for each class as n_neighbors increased.











Classification - Support Vector Machine

Due to my SVC model taking an extremely long time to run, I was only able to evaluate a few permutations with different **C** and **gamma** values. I ended up with the following best accuracy:

- Max Accuracy: 0.5875385770289432
- Best C/gamma combination: C = 10, gamma = 1

As with the KNN classifier it had low precision and recall for underrepresented classed:

	precision	recall	f1-score	support
0	0.69	0.84	0.76	6452
1	0.31	0.22	0.26	1749
2	0.43	0.47	0.45	2458
3	0.15	0.02	0.04	876
4	0.13	0.01	0.02	454
avg / total	0.52	0.59	0.54	11989

Classification Comparison

After running my 2 classification models I can to the following conclusions:

- Predicting multi-class data, where the amount of data per class is skewed for one class, can result in a model which simply predicts that everything is the most prevalent class. This makes sense in the case of a KNN classifier, as the abundance of neighbors of a different class can overwhelm the data points which are less numerous; the graphs for KNN precision/recall showed this with the precision and recall dropping to 0 for the 2 classes with the least amount of data, as K increased.
- Counter to what I learnt, Scikit's SVC classifier ran extremely slowly; the fact I was predicting a multiple classes and may have added to the time it took for my model to fit the data.
- Due to the speed issues I had with the SVC classifier, it is hard to make a direct comparison between the SVC and KNN classifiers. Accuracy/Precision/Recall results for both showed that a high accuracy is a poor indicator of how "good" the model is. If aiming for a higher recall a lower accuracy may be an acceptable result.

Regression - Overview

My regression question was:

> **Does someone being religious, and their strength of belief, affect the sentiment of their essay questions? How does this compare to other factors in their life?**

To test this I used the following features:

- religiosity
- age
- drinks_code
- speaks_count

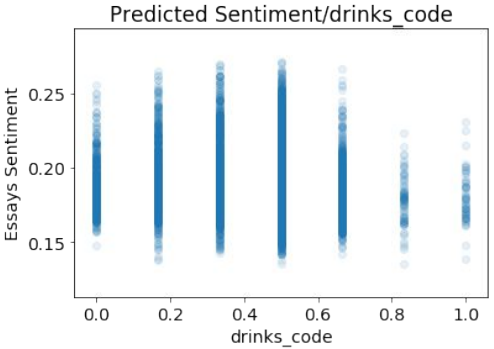
Regression - Linear Regression

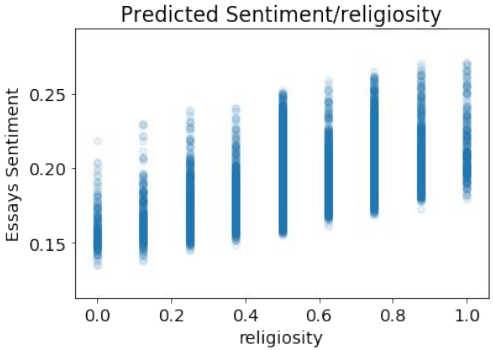
The linear regression model returned the following results:

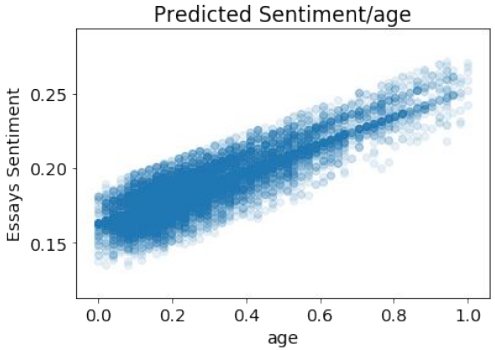
- Score using training data: 0.0319380559874195
- Score using test data: 0.035876844882340664
- Coefficients -
  - age: 0.09063423177031563
  - religiosity: 0.04888839881327408
  - speaks_count: -0.0110209944109511 37
  - drinks_code: -0.0009614402935981833

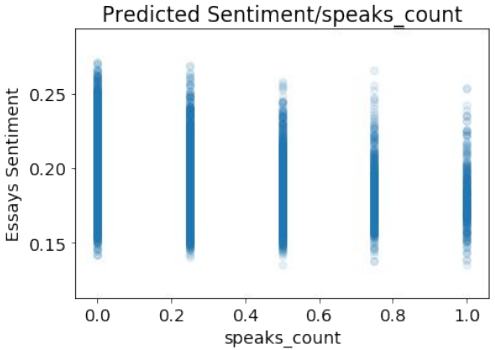
Regression - Linear Regression

Using the most accurate linear regression model here are the predicted values of essays sentiment when plotted against specific features. From these it becomes clear what correlations they have.









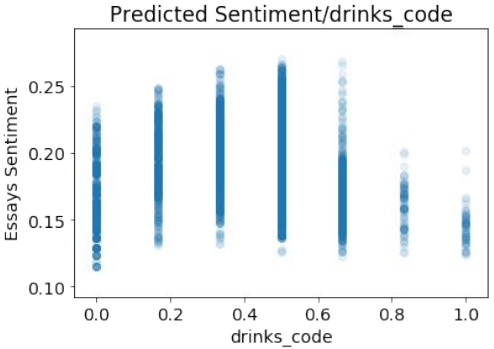
Regression - K-Nearest Neighbor

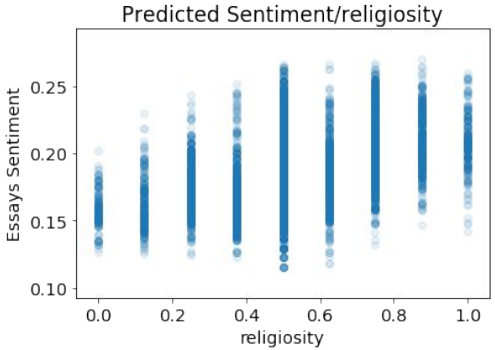
When I used the k-nearest neighbors regression model on the same data set I received the following scores:

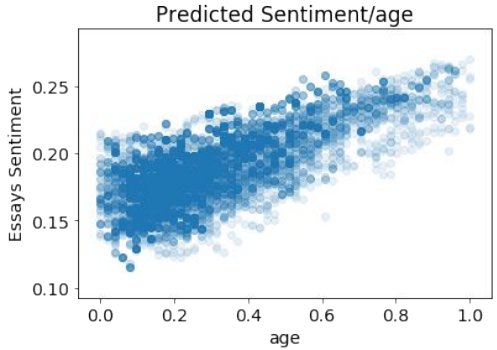
- Score using training data: 0.047150790976623
- Score using test data: 0.0322627439824229

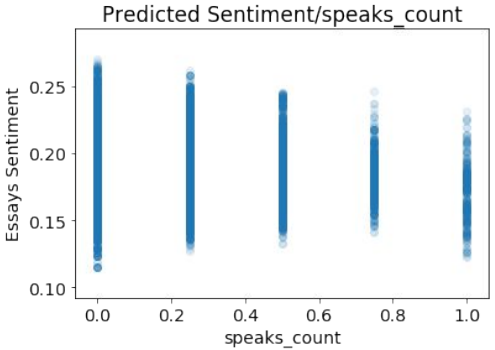
Regression - K-Nearest Neighbor

Using the most accurate KNN model here are the predicted values of essays sentiment when plotted against specific features. From these it becomes clear what correlations they have.









Regression - Comparison

After running my 2 models I came to the following conclusions:

- The correct **weights** parameter for the KNN regression is essential. Linear regression handles data with fewer discrete values better than KNN regression, if the **weights** for the KNN regressor is set to “distance” .
    - If for example x has 5 discrete value (scaled between 0 and 1), but y has thousand of discrete values, for any given y value, the KNN distance calculation will mostly find a neighbor with the same x value, unless the number of neighbors is extremely high. Linear regression is less affected by this clustering of data, as the loss of each data point is calculated independently of others.
- Linear regression, in the case of my dataset, is quicker to run. This is because, for KNN, we have to run the model multiple times to discover the optimum number of neighbors.
- The Linear Regression model also had a higher accuracy. This could possible have been due to the how the data is clustered.

Conclusion

In attempting to answer my questions I had mixed outcomes. Regarding the question,

> **Can we predict how strongly a person believes in their religion?**

I ended up learning a lot about multi-class classification and what to do when datasets have uneven number of rows per class. Whilst my models were able to predict some of the classes, better than chance, I couldn't say with confidence if this was just an artifact of bad initial data or me picking the wrong approach.

If I was to do it over again, I would:

- Have a classifier per each class and try and predict a binary label.
- Try and generate more data for the under represented classes.

Conclusion

Regarding the question,

> **Does someone being religious, and their strength of belief, affect the sentiment of their essay questions? How does this compare to other factors in their life?**

I feel I was slightly more successful in gaining some insights. The linear regression model showed that, yes, how religious someone was did have a (positive) impact on their essay sentiment. Age had a greater positive impact, whilst the amount of drinking had a negative impact. This appears to tally with other research (see https://medium.economist.com/why-people-get-happier-as-they-get-older-b5e412e471ed) suggesting there is some validity to this finding.

Conclusion - What next

If I was to ask for more data to better help answer my questions, I'd like to know about people's habits of worship, for example:

- How often people go to a church, temple, etc
- How often they pray, meditate, etc

These might provide a better indicator of how strongly people believe in their religion.

I'd also like to investigate how other factors influence people's essay sentiment and see if there's correlation between essay sentiment and self reported happiness - is essay sentiment a good proxy for happiness? Could it be used to spot depression, for example?