Codecademy Machine Learning Course: Date-A-Scientist

Machine Learning Fundamentals

Seb Galer

October 2018

Outline

- Background
- Exploration of the data
- Question 1
- Question 2
- Further Work

Background

The data analysed here was provided by OKCupid and consists of the following sections:

- body_type
- diet
- drinks
- drugs
- education
- ethnicity
- height
- income
- job
- offspring
- orientation
- pets
- religion
- sex
- sign
- smokes
- speaks
- status

- essay0 - My self summary
- essay1 - What I'm doing with my life
- essay2 - I'm really good at
- essay3 - The first thing people usually notice about me
- essay4 - Favourite books, movies, show, music, and food
- essay5 - The six things I could never do without
- essay6 - I spend a lot of time thinking about
- essay7 - On a typical Friday night I am
- essay8 - The most private thing I am willing to admit
- essay9 - You should message me if…

Exploration of the data

The dataset contains results for 59946 users comprised of 35829 males and 24117 females.

For this work I will be examining the information stored in the **essays** as well as **education**, **religion**, **income** and **speaks** to look for trends and correlations.

As is often the case for real world data there are many sections that are incomplete and a balance must be struck between dropping data and bias due to incompleteness. For this purpose any blank essay questions were filled with blank comments (‘ ’) whilst records with missing data in the other columns being considered were dropped leaving 36950 records (22097/14853 male/female).

Exploration of the data

The education and religion sections contain the following breakdown and labels.

Level	Counts	Level	Counts
graduated from college/university	16013	working on law school	167
graduated from masters program	6139	dropped out of two-year college	156
working on college/university	4119	working on med school	145
graduated from two-year college	1171	two-year college	136
working on masters program	1155	dropped out of masters program	113
graduated from high school	1085	dropped out of ph.d program	101
graduated from ph.d program	913	dropped out of high school	84
dropped out of college/university	846	working on high school	67
working on two-year college	805	masters program	59
graduated from law school	781	high school	57
working on ph.d program	685	space camp	38
graduated from space camp	505	ph.d program	17
college/university	465	dropped out of law school	14
dropped out of space camp	420	law school	14
working on space camp	347	dropped out of med school	8
graduated from med school	320	med school	5

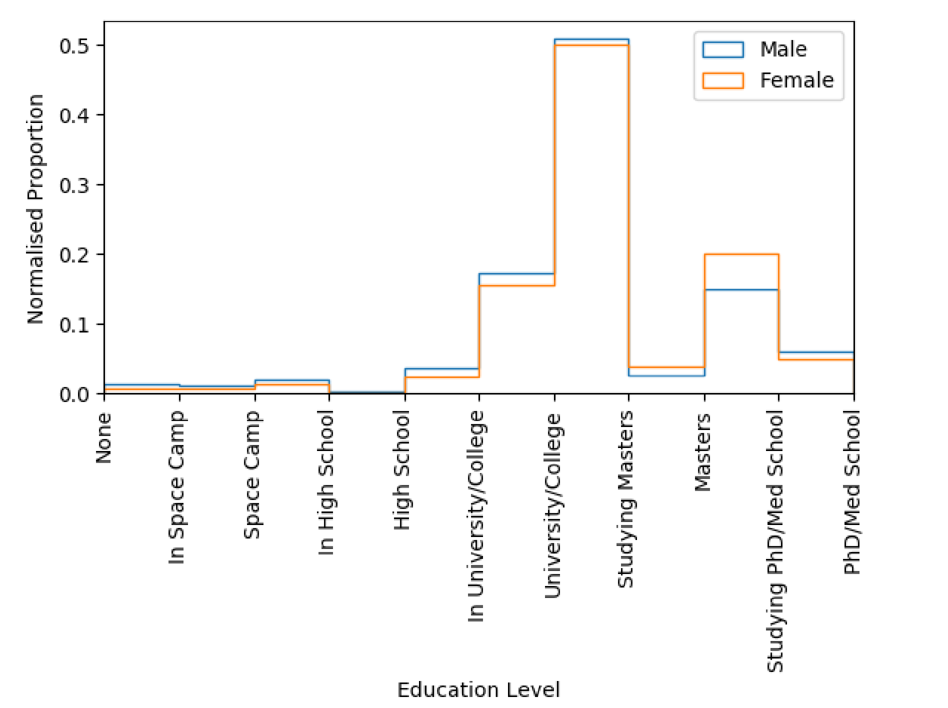
```
# Generated with
print(df_cleaned.education.value_counts())
print(df_cleaned.religion.value_counts())
```

Label	Counts	Label	Counts
agnosticism but not too serious about it	2531	atheism and very serious about it	525
agnosticism	2521	catholicism and somewhat serious about it	518
other	2376	other and very serious about it	484
agnosticism and laughing about it	2353	buddhism and laughing about it	443
catholicism but not too serious about it	2154	buddhism and somewhat serious about it	351
atheism	1989	christianity and laughing about it	348
other and laughing about it	1954	buddhism	345
atheism and laughing about it	1936	agnosticism and very serious about it	294
christianity but not too serious about it	1838	judaism and somewhat serious about it	255
christianity	1713	hinduism but not too serious about it	224
judaism but not too serious about it	1481	hinduism	100
other but not too serious about it	1472	catholicism and very serious about it	96
atheism but not too serious about it	1253	buddhism and very serious about it	64
catholicism	926	hinduism and somewhat serious about it	58
christianity and somewhat serious about it	864	hinduism and laughing about it	43
atheism and somewhat serious about it	810	islam but not too serious about it	40
other and somewhat serious about it	791	islam	39
catholicism and laughing about it	686	judaism and very serious about it	22
judaism and laughing about it	644	islam and somewhat serious about it	19
buddhism but not too serious about it	622	islam and laughing about it	14
agnosticism and somewhat serious about it	609	hinduism and very serious about it	14
judaism	580	islam and very serious about it	12
christianity and very serious about it	539		

Exploration of the data - Education

To reduce the number of categories these were mapped as a new column as follows where a drop out is considered the level below the course dropped out from. The graph to the right shows the split by gender showing very similar education levels at most levels except Phd/Med school in progress.

Level	Map
PhD/Med School completed	10
PhD/Med School in progress	9
Masters completed/graduate	8
Masters in progress	7
College/University/Law School completed	6
College/University/Law School in progress	5
High School completed	4
High School in progress	3
Space Camp	2
Space Camp in progress	1
Space Camp drop out	0



Note that due to lack of familiarity with the American education system this may not be a truly accurate reflection of education levels. This may lead to bias/incorrect results later.

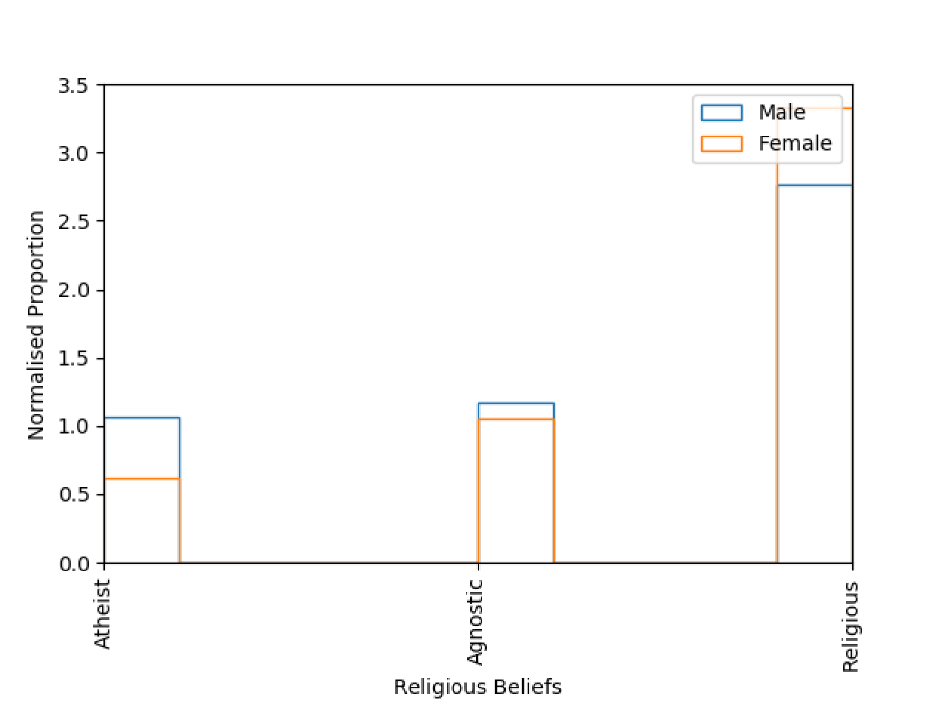
```
# Generated with
df_cleaned['mapped_education'] = df_cleaned.education.map(education_map)
```

```
plt.hist(male_data.mapped_education, bins=10, histtype='step', density=True, fill=False)
plt.hist(female_data.mapped_education, bins=10, histtype='step', density=True, fill=False)
plt.xlabel("Education Level")
plt.ylabel("Normalised Proportion")
plt.legend(['Male', 'Female'])
plt.xlim(0, 5)
plt.xticks(np.arange(11), ('None', 'In Space Camp', 'Space Camp', 'In High School', 'High School',
                           'In University/College', 'University/College', 'Studying Masters',
'Masters',
                           'Studying PhD/Med School', 'PhD/Med School'), rotation=90)
plt.show()
```

Exploration of the data - Religion

To reduce the number of categories these were mapped as a new column as follows where the additional comments indicating strength of faith were dropped to reduce the number of categories.

Level	Map
Religious (all categories)	2
Agnostic	1
Atheist	0



The lack of strength of faith may add a skew to the data and should be considered when interpreting the results.

```
# Generated with
df_cleaned['mapped_religion'] = df_cleaned.religion.map(religion_map)
```

```
plt.hist(male_data.mapped_religion, bins=10, histtype='step', density=True, fill=False)
plt.hist(female_data.mapped_religion, bins=10, histtype='step', density=True, fill=False)
plt.xlabel("Religious Beliefs")
plt.ylabel("Normalised Proportion")
plt.legend(['Male', 'Female'])
plt.xlim(0, 2)
plt.xticks(np.arange(3), ('Atheist', 'Agnostic', 'Religious'), rotation=90)
plt.show()
```

Question 1

Does positivity increase with education level, religious belief and income? Can these factors be used to predict a user’s level of positivity?

Approach

Train a Naïve Bayes algorithm on sentiment to measure the sentiment of the essay answers. Use classification type methods to predict sentiment.

Limitations

Limited labelling accuracy for religion and education, NB model accuracy, how truthful people are in answering the questions (especially the essay questions), the length of essay questions may skew sentiment scoring.

Question 1

1. The Naïve Bayes MultinomialNB model was selected from the sklearn library
2. It was then trained with a 1.6 Million sentiment labelled tweet dataset available from: https://www.kaggle.com/kazanova/sentiment140
3. The data was prepared with the sklearn.feature_extraction.text library CountVectorizer and sklearn's train_test_split.
4. This model was then used to calculate a sentiment score for each essay answer and a total overall score.

The model accuracy, score, recall and precision were all 0.782 which is fair to good. It took 68s to train the Naïve Bayes model (several minutes on i5 6400 windows PC) and required ~1.5GB of RAM.

Better algorithms can score much higher than this.

```python
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB

sentiment_header = ['sentiment', 'number', 'date', 'query', 'user', 'tweet']
sentiment_data = pd.read_csv('training.1600000.processed.noemoticon.csv',
                             encoding="ISO-8859-1", names=sentiment_header)
sentiment_data.drop(['user', 'number', 'date', 'query'], axis=1, inplace=True)

counter = CountVectorizer()
counter.fit(sentiment_data.tweet)
sentiment_counts = counter.transform(sentiment_data.tweet)

X_train, X_test, y_train, y_test = train_test_split(sentiment_counts,
sentiment_data.sentiment, test_size=0.2,
                                                    random_state=1)

classifier = MultinomialNB()
classifier.fit(X_train, y_train)
print('Naive Bayes prediction score is %.1f%%' % (classifier.score(X_test,
y_test)*100))

for column in sub_columns[3:]:
  new_col_name = column + '_sentiment'
  sentiment_list.append(new_col_name)
  df_cleaned[new_col_name] =
classifier.predict(counter.transform(df_cleaned[column]))

df_cleaned['overall_sentiment'] = df_cleaned[sentiment_list].mean(axis=1)  #
aggregate essay sentiments
```

Question 1 - predictions

The Linear Regression and KNN regressor models from sklearn were used to investigate correlation between essay sentiment and income, education level and religious belief.

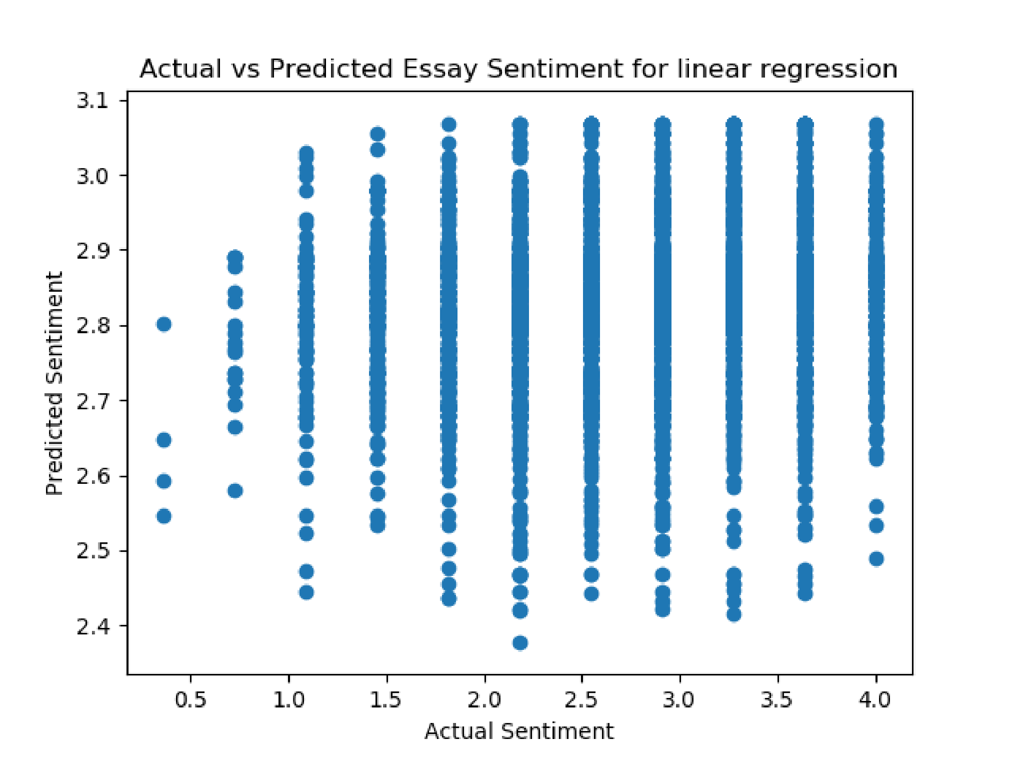
	Linear Regressor	KNN Regressor
Score	0.024	0.013
Mean Squared Error	0.417	0.540
Time (s)	0.3	64

These results indicate little to no correlation and whilst KNN scored lower both scores are too low to draw a meaningful conclusion regarding model performance.

```
# Relevant code
from sklearn.linear_model import LinearaRegression

features = ['mapped_education', 'mapped_religion', 'scaled_income',
'normalised_income', 'income', 'no_of_languages']
XC_train, XC_test, yc_train, yc_test = train_test_split(df_cleaned[features],
df_cleaned['overall_sentiment'],
                                                        test_size=0.2,
random_state=1)
lm = LinearRegression()
lm.fit(XC_train, yc_train)
yc_predict = lm.predict(XC_test)

print("Train score: %f" % lm.score(XC_train, yc_train))
print("Test score: %f" % lm.score(XC_test, yc_test))
```




```
plt.scatter(yc_test, yc_predict)
plt.xlabel("Actual Sentiment")
plt.ylabel("Predicted Sentiment")
plt.title("Actual vs Predicted Essay Sentiment for linear regression")
plt.show()
```

Question 2

Can you predict the number of languages someone speaks from their education level?

Approach

Given the number of languages spoken in the speaks column determine what, if any, correlation there is to education level.

Limitations

Limited labelling accuracy for number of languages and education, how truthful people are in answering the questions.

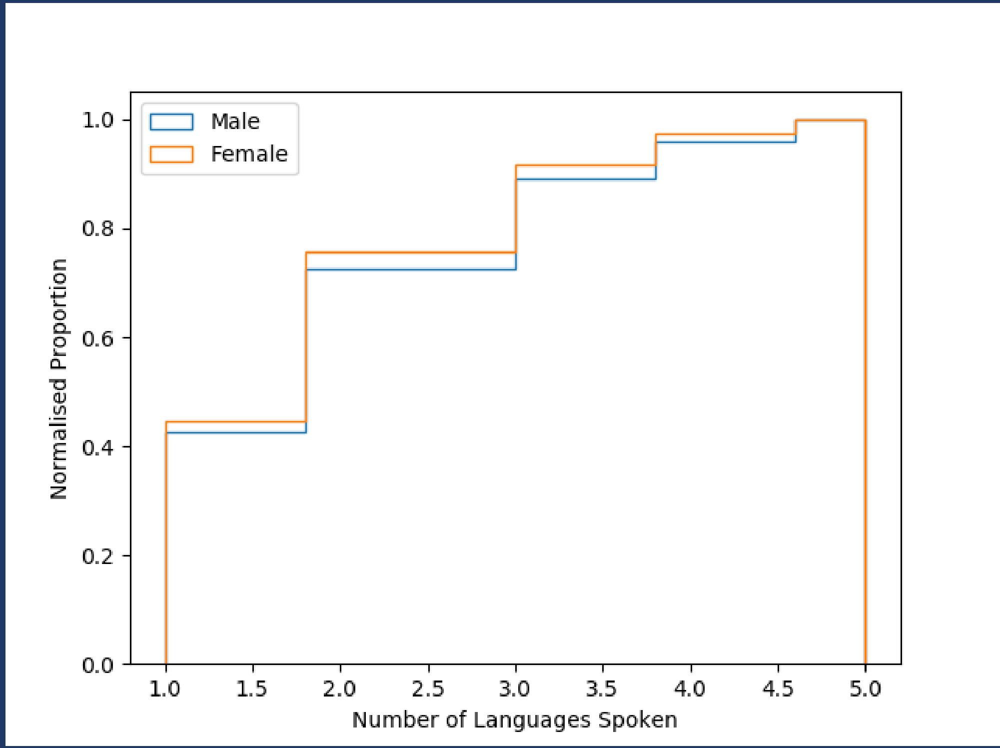
Question 2

The data for the number of languages was prepared by splitting the **speaks** data into number of languages.

```
print(df_cleaned['speaks'].head(10))

## OUTCOME
0                                                 english
1     english (fluently), spanish (poorly), french (...
5                  english (fluently), chinese (okay)
7                             english, spanish (okay)
8                                                 english
9                                      english (fluently)
11           english (fluently), sign language (poorly)
13                                                english
14                                                english
15                 english (fluently), spanish (okay)

df_cleaned['no_of_languages'] = df_cleaned['speaks'].str.split(',').str.len()
```



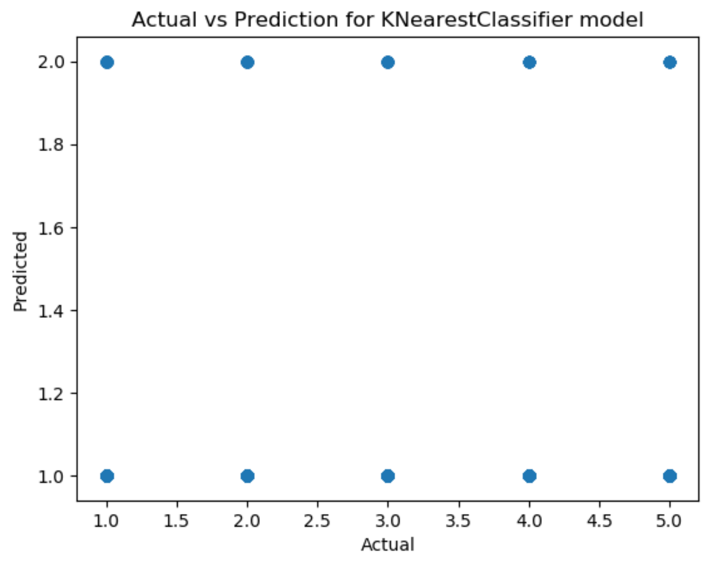

```
plt.hist(male_data.no_of_languages, bins=10, histtype='step', density=True, fill=False, cumulative=True)
plt.hist(female_data.no_of_languages, bins=10, histtype='step', density=True, fill=False,
cumulative=True)
plt.xlabel("Number of Languages Spoken")
plt.ylabel("Normalised Proportion")
plt.legend(['Male', 'Female'], loc='upper left')
plt.show()
```

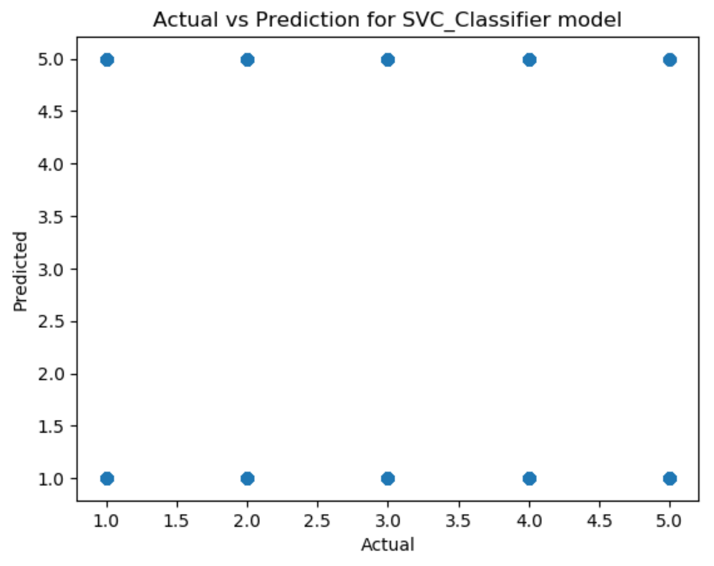
Question 2

The data was fit with sklearn’s K-Nearest Classifier and Support Vector Machine Classifier algorithms. A linear regressor was also run however it scored 0.007 with an mean squared error of 1.127.

Both models are fairly simple although SVM required considerably more time to run. The model metrics were:

	KNN	SVM
Accuracy	0.434	0.349
Recall	0.434	0.349
Precision	0.434	0.349
Time (s)	91	209





Whilst the scores for KNN were higher the plot shows it only guessed two classes. This is likely due to the class imbalance where ~70% of entries spoke 1 or 2 languages. This was accounted for in the SVM model using class weights.

```
plt.hist(male_data.no_of_languages, bins=10, histtype='step', density=True, fill=False, cumulative=True)
plt.hist(female_data.no_of_languages, bins=10, histtype='step', density=True, fill=False,
cumulative=True)
plt.xlabel("Number of Languages Spoken")
plt.ylabel("Normalised Proportion")
plt.legend(['Male', 'Female'], loc='upper left')
plt.show()
```

Conclusions

Question 1

The sentiment of a tweet or sentence can be measured, with some confidence, using the Naïve Bayes algorithm trained with a sentiment based training set. However, there is virtually no correlation between this and the individual's education level, religious belief and income.

Question 2

Some level of prediction can be made based on education but the models performed generally quite badly. KNN because it only predicted two of the five classes and SVM, despite predicting all classes scored lower. This suggest that the correlation is very weak as found with the linear regressor model which scored 0.007 with a mean squared error of 1.127.

Future Work

The first thing to investigate would be to refine the categorization of the education, religion and speaks data. For religion this would include further splitting the data by conviction of belief for speaks this would look at fluency in languages not just number.

Improve sentiment classification score using improved/alternative algorithms. Investigate whether a twitter training set is the most appropriate (use of words may well be different). Improve model optimization (sklearn's gridsearchCV or similar) as only a single hyper parameter was investigated and only for a limited range of values.

Finally, work needs to be done to account for the class imbalance for KNN (either by balancing the set (removal or additional of entries) or by weighting.