Capstone Project

Machine Learning Fundamentals

Pablo Gancharov
March 14 2019

Table of Contents

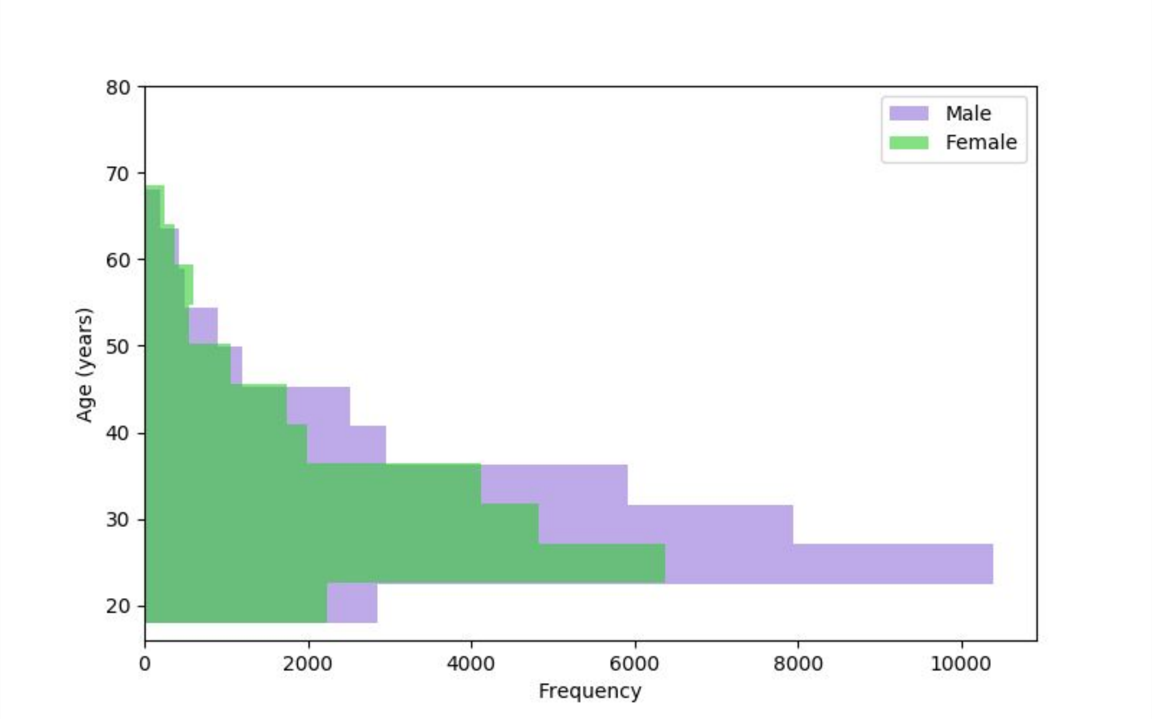
Exploration of the OkCupid dataset

Dataset general numbers:

- There are 59946 records
- 31 columns
- Only 3 numerical columns
- 10 columns contain free text
- Some columns contain labels separated by comma
- 24 columns contain NAN values

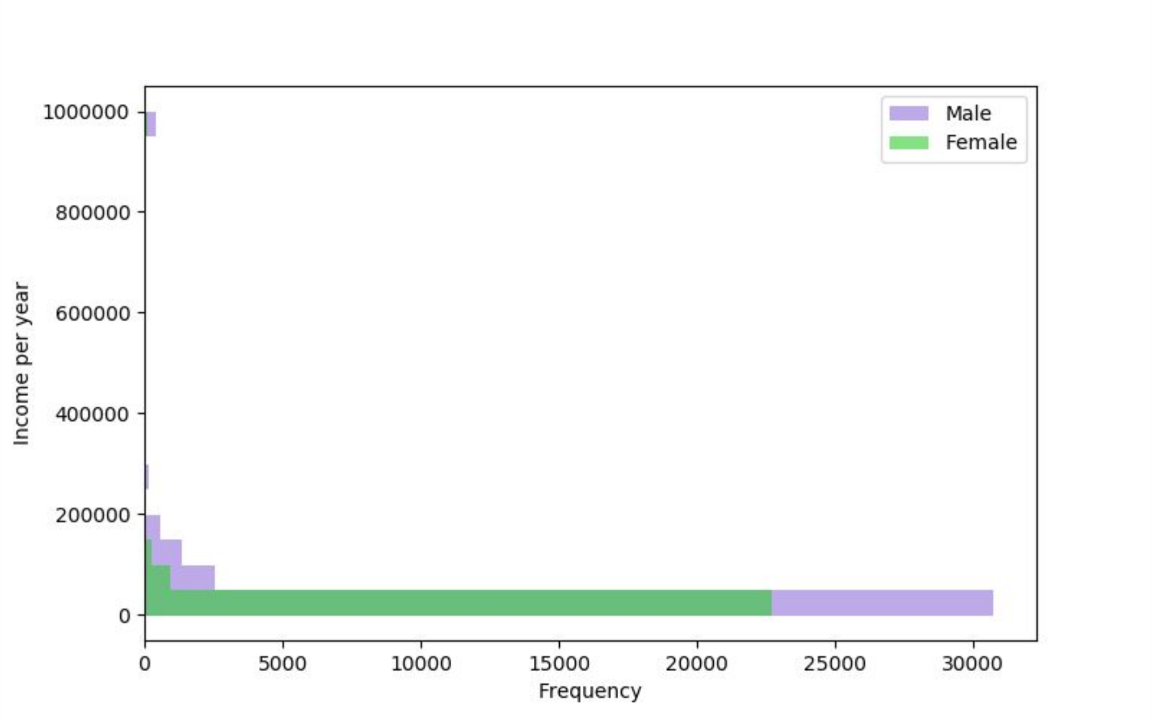
Exploration of the OkCupid dataset

Dataset overview on images:



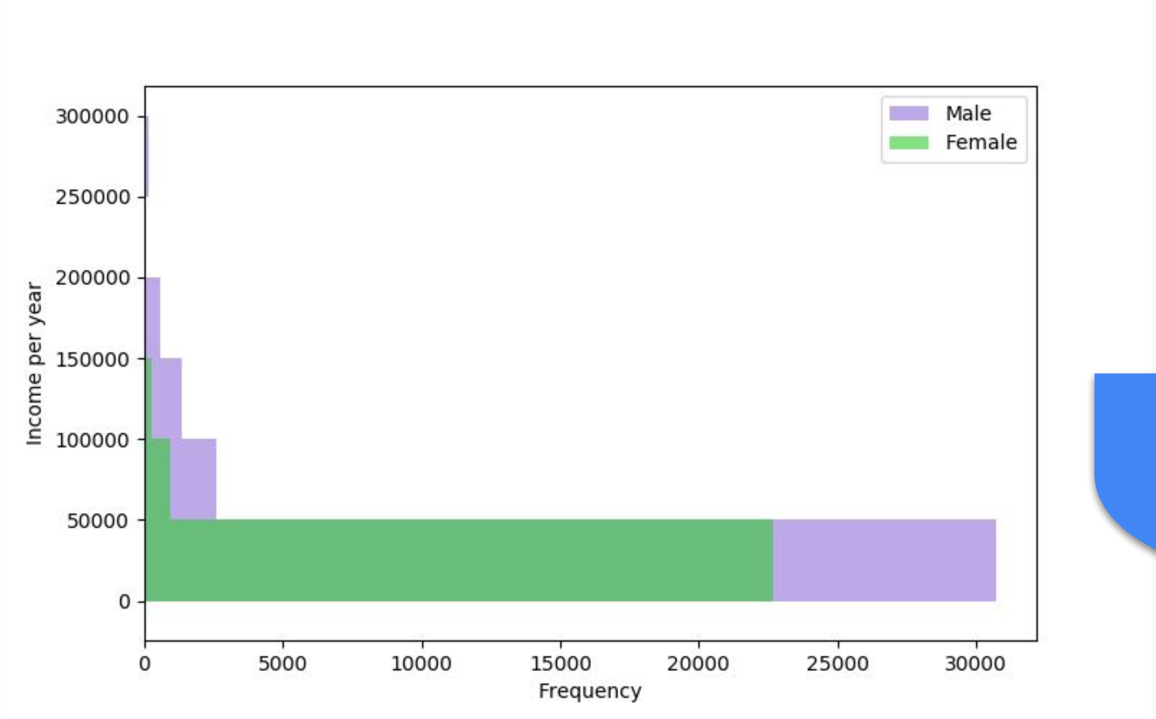
Exploration of the OkCupid dataset

Dataset overview on images:



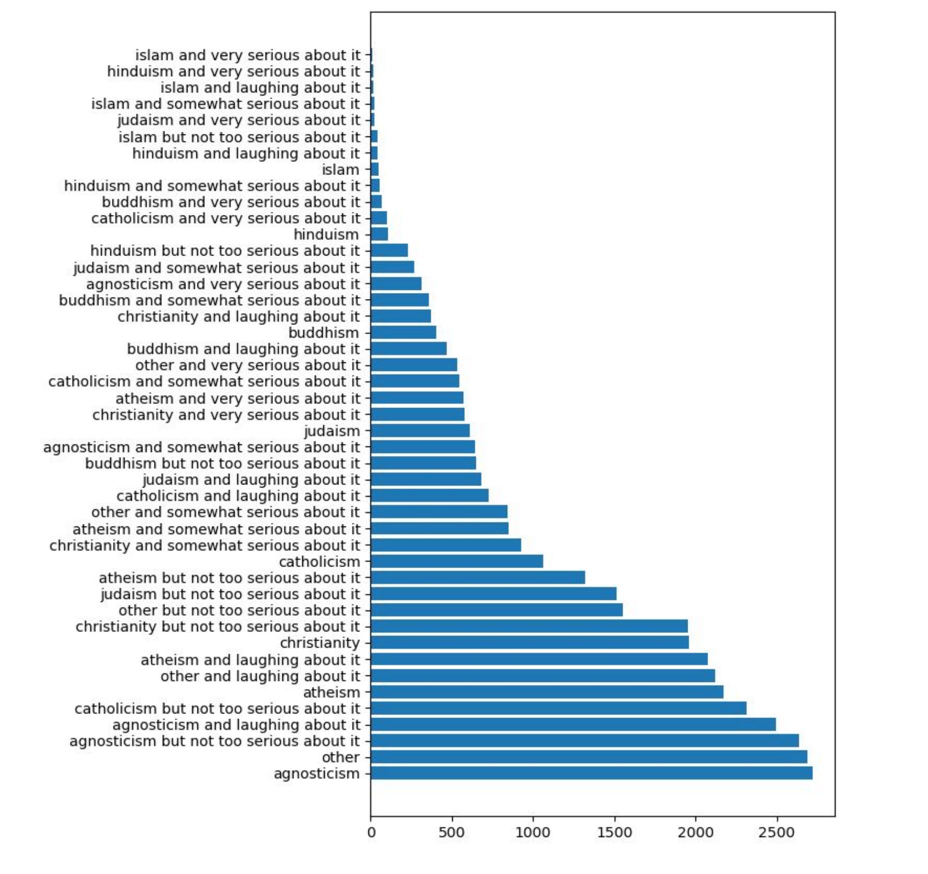
Exploration of the OkCupid dataset

Dataset overview on images:

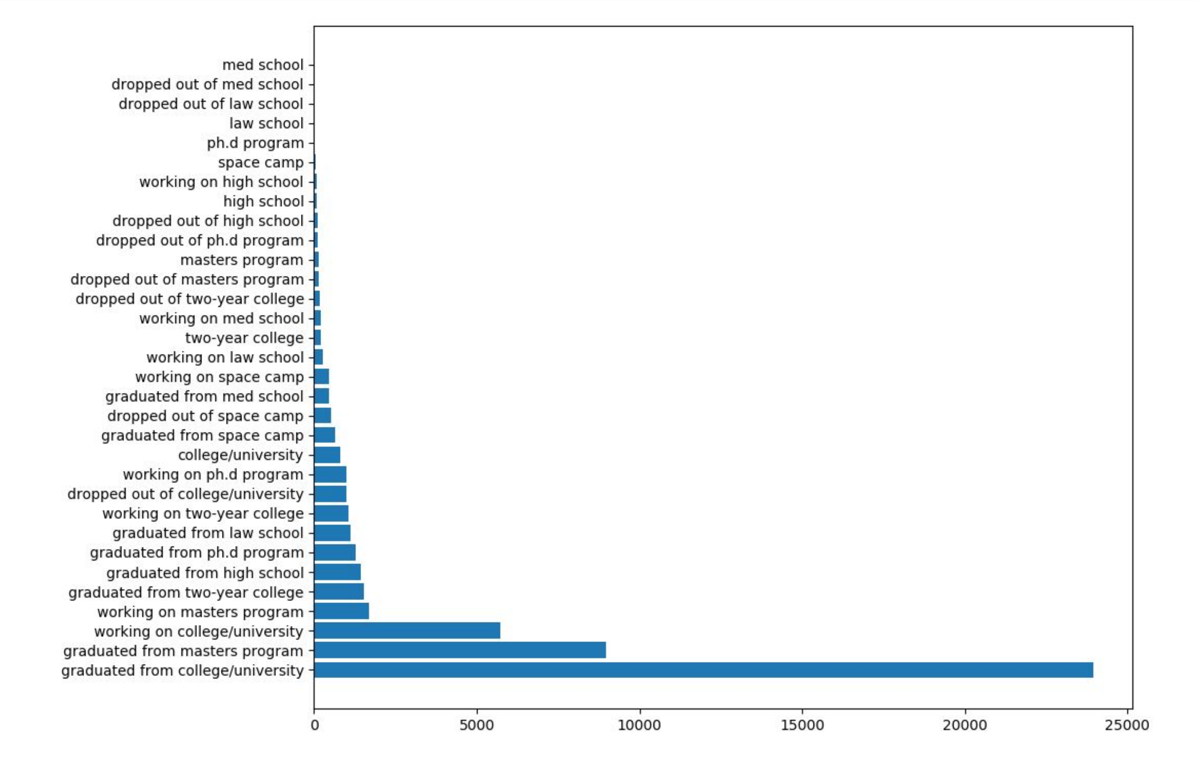


This is a close up of previous chart, excluding high income values

Exploration of the OkCupid dataset



Exploration of the OkCupid dataset



Question(s) to Answer

Q1: Are there clusters clearly defined?

I feel curiosity to understand the shape or meta-shape of this set of data by using the power of unsupervised learning techniques.

Question(s) to Answer

Q2: Can I predict ‘sex’ by sentiment analysis score?

I will try to predict the category ‘sex’ based on the sentiment analysis of the essays

Question(s) to Answer

Q3: Can I predict the income?

I´m going to try to predict the salary based on all the other attributes. So you can know if you deserve a raise or not

Augmenting the Dataset

The comma separated values were separated in new columns in the case of ‘speaks’ and ‘ethnicity’ columns

```
Index([u'age', u'body_type', u'diet', u'drinks', u'drugs', u'education',
       u'essay0', u'essay1', u'essay2', u'essay3', u'essay4', u'essay5',
       u'essay6', u'essay7', u'essay8', u'essay9', u'ethnicity', u'height',
       u'income', u'job', u'last_online', u'location', u'offspring',
       u'orientation', u'pets', u'religion', u'sex', u'sign', u'smokes',
       u'speaks', u'status'],
       dtype='object')
```

Before

Augmenting the Dataset

The comma separated values were separated in new columns in the case of 'Speaks' and 'ethnicity' columns

```
Index([u'age', u'body_type', u'diet', u'drinks', u'drugs', u'education',
       u'essay0', u'essay1', u'essay2', u'essay3', u'essay4', u'essay5',
       u'essay6', u'essay7', u'essay8', u'essay9', u'ethnicity', u'height',
       u'income', u'job', u'last_online', u'location', u'offspring',
       u'orientation', u'pets', u'religion', u'sex', u'sign', u'smokes',
       u'speaks', u'status', u'speaks_afrikaans',
       u'speaks_afrikaans (fluently)', u'speaks_afrikaans (okay)',
       u'speaks_afrikaans (poorly)', u'speaks_albanian',
       u'speaks_albanian (fluently)', u'speaks_albanian (okay)',
       u'speaks_albanian (poorly)', u'speaks_ancient greek',
       u'speaks_ancient greek (fluently)', u'speaks_ancient greek (okay)',
       u'speaks_ancient greek (poorly)', u'speaks_arabic',
       u'speaks_arabic (fluently)', u'speaks_arabic (okay)',
       u'speaks_arabic (poorly)', u'speaks_armenian (fluently)',
       u'speaks_armenian (okay)', u'speaks_armenian (poorly)',
       u'speaks_basque', u'speaks_basque (fluently)', u'speaks_basque (okay)',
       ...
       u'speaks_yiddish (fluently)', u'speaks_yiddish (okay)',
       u'speaks_yiddish (poorly)', u'ethnicity_asian', u'ethnicity_black',
       u'ethnicity_hispanic / latin', u'ethnicity_indian',
       u'ethnicity_middle eastern', u'ethnicity_native american',
       u'ethnicity_other', u'ethnicity_pacific islander', u'ethnicity_white'],
       dtype='object')
```

After

Augmenting the Dataset

Also I mapped into numeric values the following categories:

'drinks'	'drinks_code'
'drugs'	'drugs_code'
'smokes'	'smokes_code'
'pets'	'pets_code'
'education'	'education_code'

I used my own criteria for these two, and that can be a source of bias in the models

```
# EXAMPLE:
# map 'pet' into codes
pets_mapping = {"likes dogs and likes cats" : 2, "likes dogs" : 1, "likes dogs and has cats":3, "has dogs"
: 2, "has dogs and likes cats" : 3, "likes dogs and dislikes cats": 0, "has dogs and has cats": 4, "has
cats": 2, "likes cats": 1, "has dogs and dislikes cats": 1, "dislikes dogs and likes cats": 0, "dislikes
dogs and dislikes cats": -2, "dislikes cats": -1, "dislikes dogs and has cats": 1, "dislikes dogs":-1 }
df["pets_code"] = df.pets.map(pets_mapping)
```

Augmenting the Dataset

Also I mapped into numeric values the following categories:

```
# EXAMPLE: map 'pet' into codes

pets_mapping = {"likes dogs and likes cats" : 2, "likes dogs" : 1,
"likes dogs and has cats":3, "has dogs" : 2, "has dogs and likes
cats" : 3, "likes dogs and dislikes cats": 0, "has dogs and has
cats": 4, "has cats": 2, "likes cats": 1, "has dogs and dislikes
cats": 1, "dislikes dogs and likes cats": 0, "dislikes dogs and
dislikes cats": -2, "dislikes cats": -1, "dislikes dogs and has
cats": 1, "dislikes dogs":-1 }


df["pets_code"] = df.pets.map(pets_mapping)
```

Augmenting the Dataset

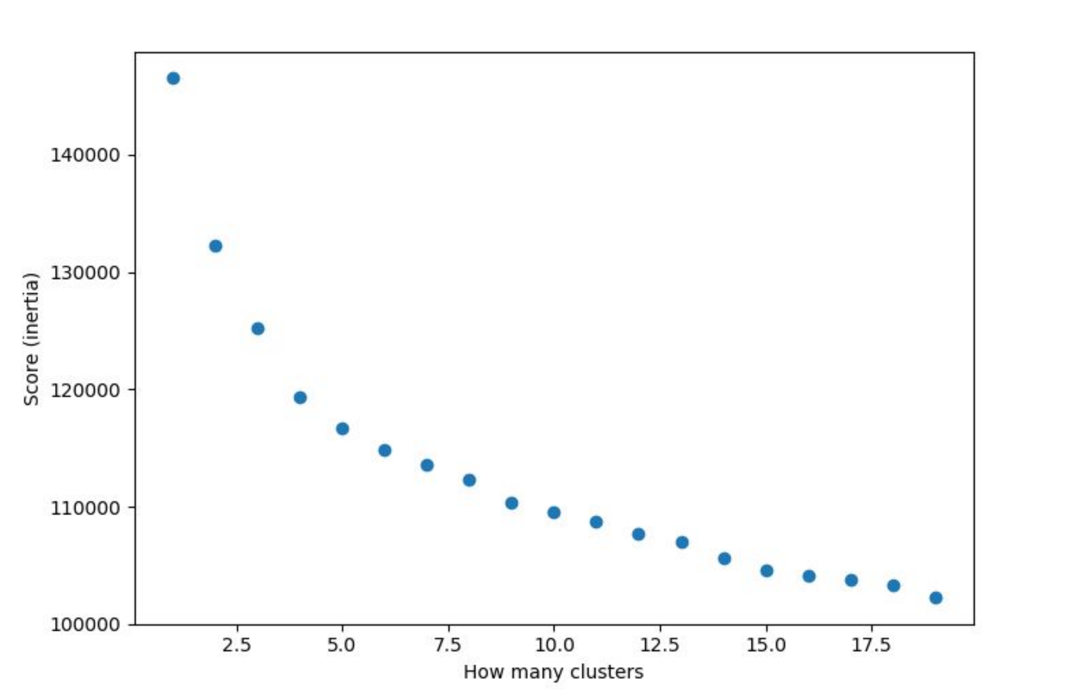
Regarding the Essays

- I created one column with the length of all the essays combined
- I applied VADER sentiment analysis (https://github.com/cjhutto/vaderSentiment) to each essay.

```
# EXAMPLE:
analyser = SentimentIntensityAnalyzer()
df[essay_cols] = df[essay_cols].astype(str)
df["essay0_sentiment_score"] = df["essay0"].map(lambda x: analyser.polarity_scores(x)["compound"])
```

Clustering Approaches

Let’s find how many clusters there are:



Q1: Are there clusters clearly defined?

The answer is **no**, there isn’t any clear “elbow” in the chart. That means that there is no special “K” number of clusters that group data in a compact way.

Classification Approaches

In order to be able to understand the correlations between labels, I had to represent them as 0s or 1s instead of strings using ‘get_dummies’:

```
# augment categorical data: diet, body_type,  'job', 'sex'
df = pd.concat([df, (df['diet'].str.get_dummies(sep=', ').add_prefix('diet_')) ], axis=1)
df = pd.concat([df, (df['body_type'].str.get_dummies(sep=', ').add_prefix('body_type_')) ], axis=1)
df = pd.concat([df, (df['job'].str.get_dummies(sep=', ').add_prefix('job_')) ], axis=1)
df = pd.concat([df, (df['sex'].str.get_dummies(sep=', ').add_prefix('sex_')) ], axis=1)
```

Classification Approaches

Interesting correlations found:

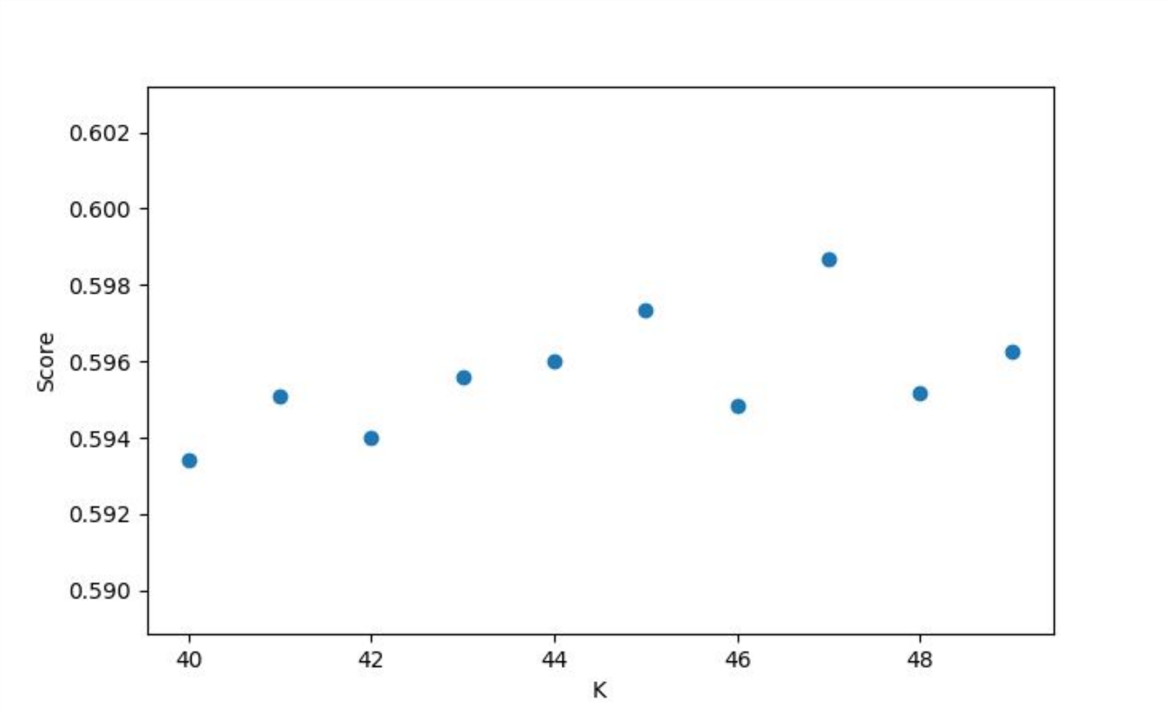
```
Top 20 highest absolute correlations:
        height                          sex_m                                   0.649790
                                        sex_f                                   0.649790
        ethnicity_asian                 ethnicity_white                         0.440331
        drugs_code                      smokes_code                             0.328228
        body_type_curvy                 sex_m                                   0.316167
                                        sex_f                                   0.316167
        essay_len                       essay9_sentiment_score                  0.302298
        body_type_average               body_type_fit                           0.298922
        body_type_athletic              body_type_average                       0.282239
        essay_len                       essay1_sentiment_score                  0.277380
        age                             job_student                             0.263225
        essay_len                       essay0_sentiment_score                  0.259825
        body_type_athletic              body_type_fit                           0.244669
        diet_anything                   diet_mostly anything                    0.233621
        essay1_sentiment_score          essay2_sentiment_score                  0.231207
        ethnicity_black                 ethnicity_white                         0.228297
        essay0_sentiment_score          essay9_sentiment_score                  0.224379
        drinks_code                     drugs_code                              0.221892
        essay_len                       essay2_sentiment_score                  0.221497
        essay0_sentiment_score          essay1_sentiment_score                  0.221490
```

I filtered out the correlations about language, because they feel unimportant to me

Classification Approaches

Q2: Can I predict 'sex' by sentiment analysis score?

Using KNN I never get better than 0.6, which is a bad score.



Classification Approaches

Q2: Can I predict 'sex' by sentiment analysis score?

Anyway by using Support Vector Machines (SVC) I only got score = 0.5988323603

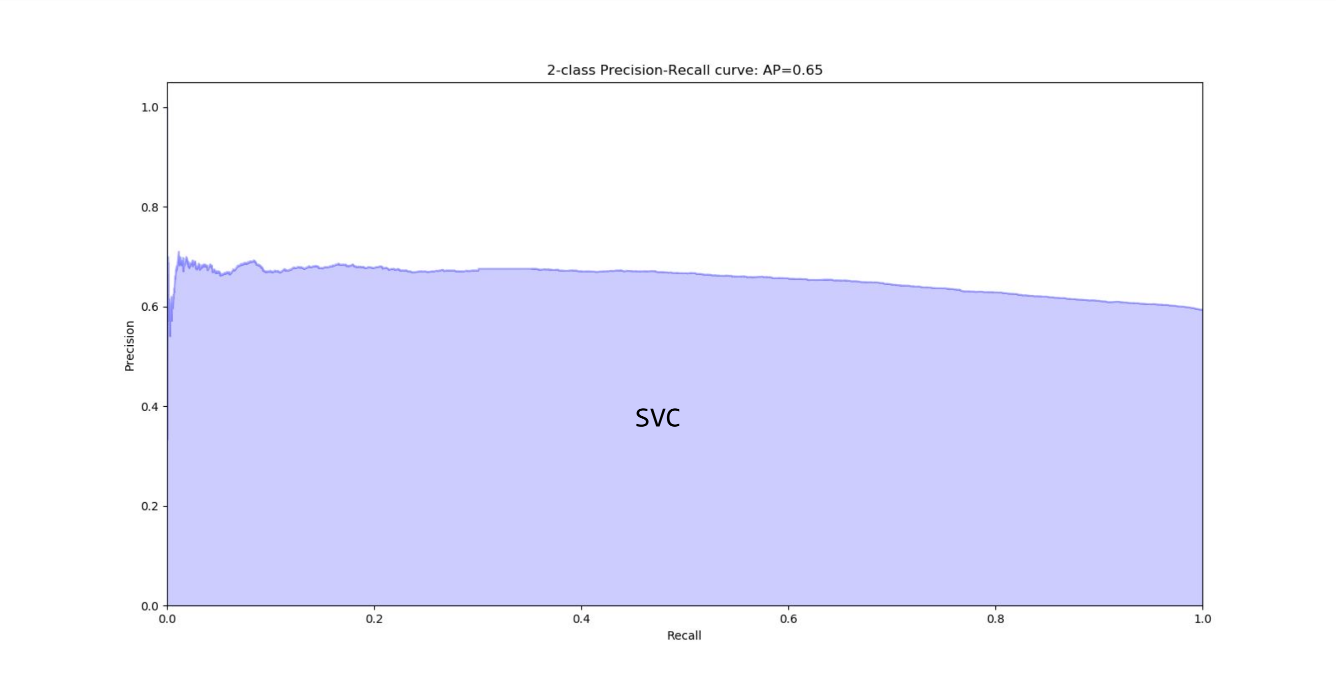
Conclusion:

Having in mind that the 2 approaches generate similar results, I prefer SVC, because it is much faster and direct.

Q2 Answer: You can predict, but with low accuracy (~60%).

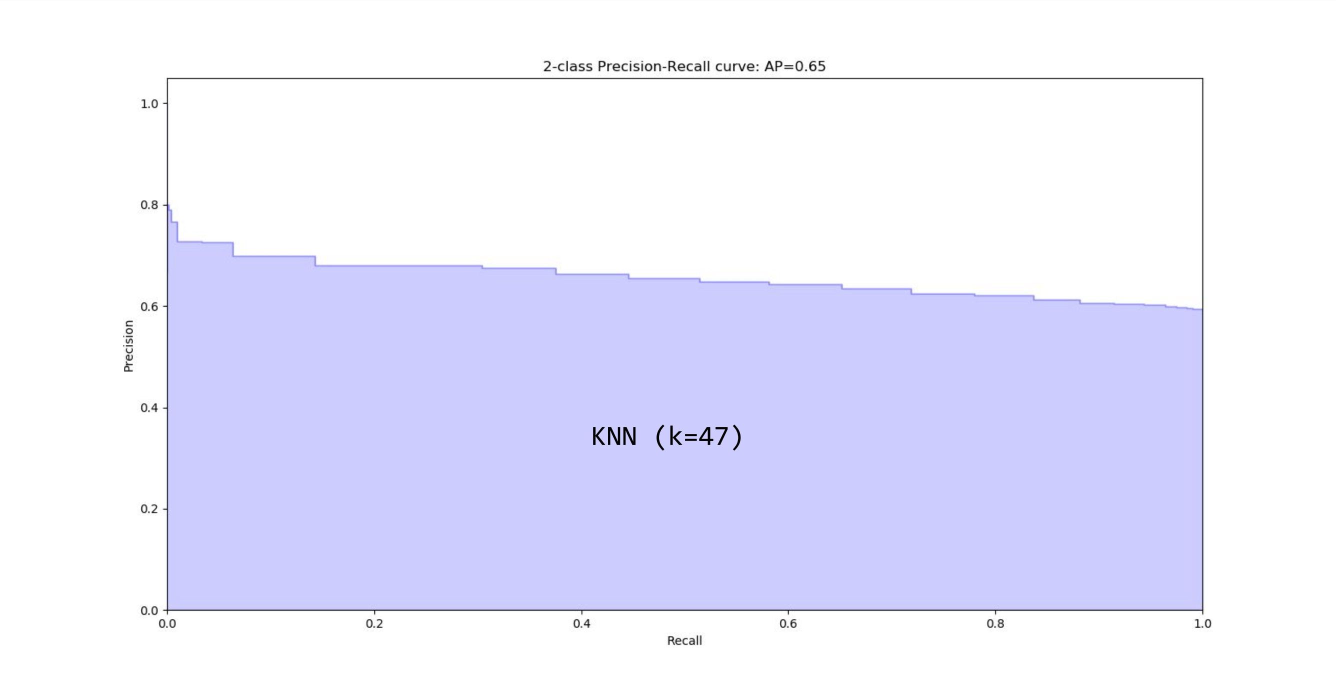
Classification Approaches

Precision and recall analysis



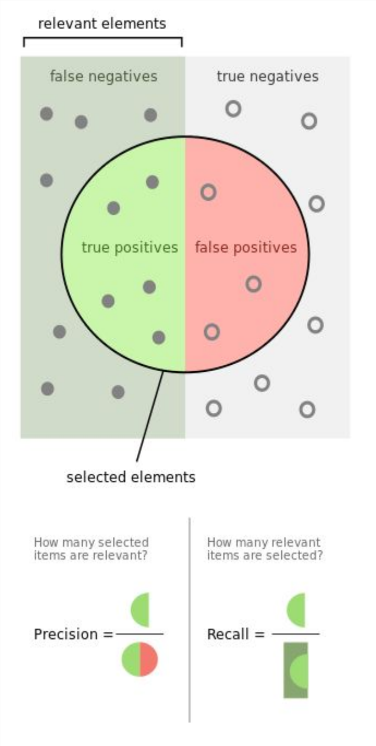
Classification Approaches

Precision and recall analysis



Classification Approaches

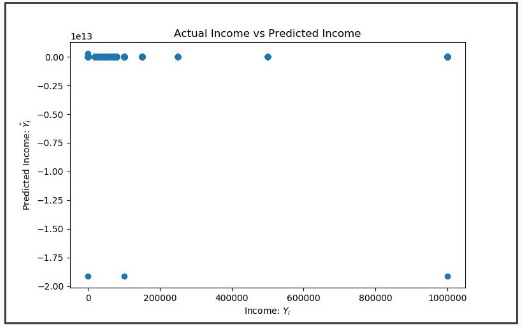
Precision and recall analysis

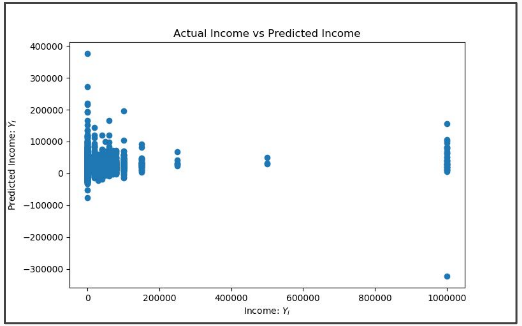


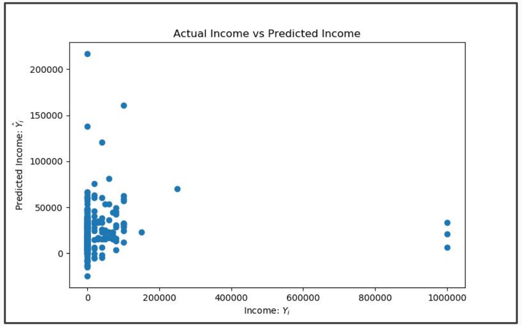
In both cases, the trade of between precision and recall was pretty stable.

Regression Approaches

Q3: If you fill the survey, can I predict your income?







`test_size=0.4`

```
Train score:
0.0820010966187
Test score:
-0.0264339845472
```

`test_size=0.1`

```
Train score:
0.118094465003
Test score:
-8.66289513941e+12
```

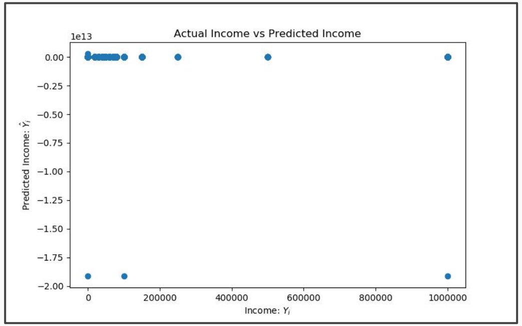
`test_size=0.01`

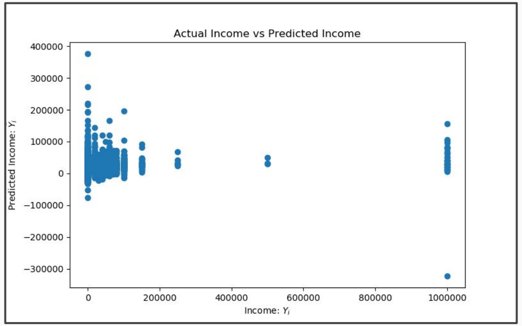
```
Train score:
0.0820010966187
Test score:
-0.0264339845472
```

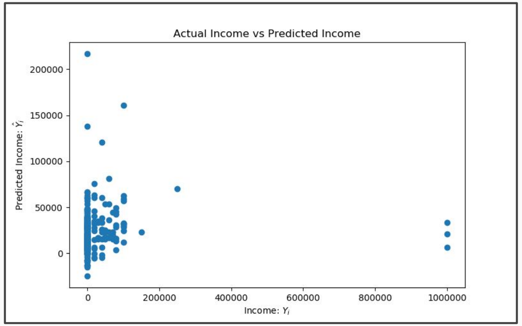
Regression Approaches

If you fill the survey, can I predict your income?

`Linear Regression: seems not to work`







`test_size=0.4`

```
Train score:
0.0820010966187
Test score:
-0.0264339845472
```

`test_size=0.1`

```
Train score:
0.118094465003
Test score:
-8.66289513941e+12
```

`test_size=0.01`

```
Train score:
0.0820010966187
Test score:
-0.0264339845472
```

Not even close

Conclusions/Next steps

- Q1: Are there clusters clearly defined?
    - No, there are not clear groups
- Q2: Can I predict ‘sex’ by sentiment analysis score?
    - Yes, but not in high accuracy
- Q3: Can I predict the income?
    - No, data available is not enough to do it

Conclusions/Next steps

- I´d like to try removing the high income individuals from the set (outliers), if we can get better clustering results.
- I know there´s much more information to extract from the essays, I suggest to continue that path.
- When trying to predict the income, numbers seem to improve when I assign 99% to train and 1% to test. I´d like to try with a bigger dataset.
- Also the high dimensionality of the final set can be a problem and may deserve more research.

Thank You