Date-A-Scientist!

Machine Learning Fundamentals

Final project

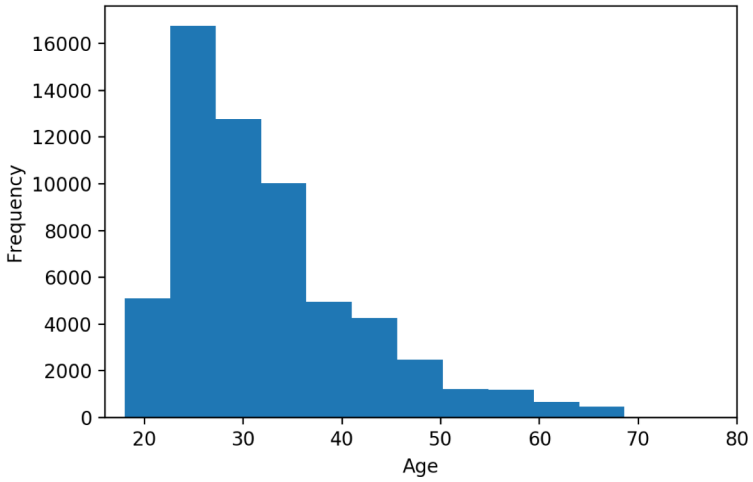
Zineb ALY – March 2019

Outline

- Exploration of the dataset
- Questions to answer throughout the project
- Augmenting data
- Comparison between two classification approaches
- Comparison between two regression approaches
- Conclusions

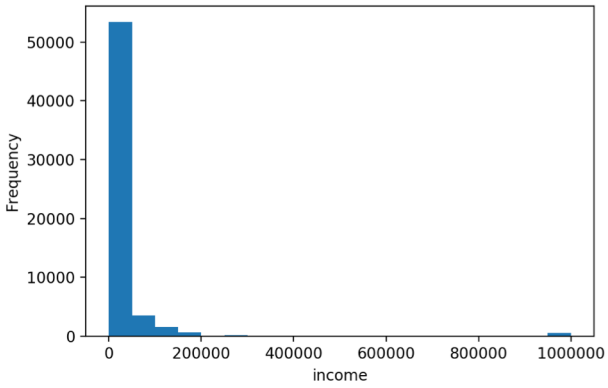
Let’s explore our dataset

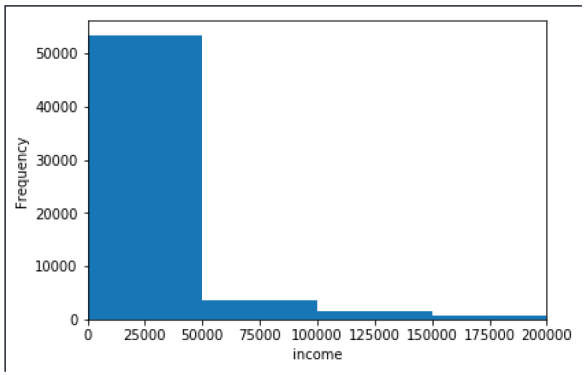
- Data provided has the following features: body type, diet, drinks, drugs, education, ethnicity, height, income, job, offspring, orientation, pets, religion, sex, sign, smokes, speaks, status and some essays written by the dating app users!

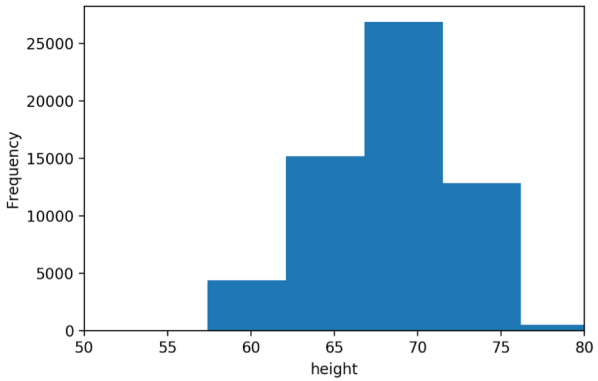


- Most users of the dating app are aged between 25 years old and 35 years old. We observe from the plot that the older people get, the less they are tempted to use the dating app!

Let’s explore our dataset







- Both figures above show that most users have an income between 0 to 50K.
- Most users’ height is around 70.

Questions to answer throughout the project

- Can we predict users gender/sex with drinks, drugs, smokes and the length of the essay the user wrote?

- Can we predict income with length of essays, drinks, drugs and smokes?

Augmenting data

- Most of the data provided is categorical, therefore, we need to create some numerical data based on the information given in our dataset. I first explored the possible answers in the column I was interested in, then I associated a label with each category of answers. Below is the code I used to label the column containing gender information and adding it to the dataset:

```
sex_mapping = {'m':0, 'f':1}
all_data["sex_code"] = all_data.sex.map(sex_mapping)
```

- The new column formed is named sex_code and contains 0 if the user is a male and 1 if the user is a female.

Augmenting data

- I also added multiple columns following the same procedures to create the following columns:

drinks_code

drugs_code

religion_code

diet_code

Smokes_code

- After creating the new columns, I normalized data and I removed all the rows where “NaN” was present.

Augmenting data

- Similarly, I combined all the essays written by the users in their profiles into one essay and I calculated the length of the essay, which I added as a new column to the data set:

```
## combining the essays
essay_cols = ["essay0","essay1","essay2","essay3","essay4","essay5","essay6","essay7","essay8","essay9"]
# Removing the NaNs
all_essays = all_data[essay_cols].replace(np.nan, '', regex=True)
# Combining the essays
all_essays = all_essays[essay_cols].apply(lambda x: ' '.join(x), axis=1)
all_data["essay_len"] = all_essays.apply(lambda x: len(x))
```

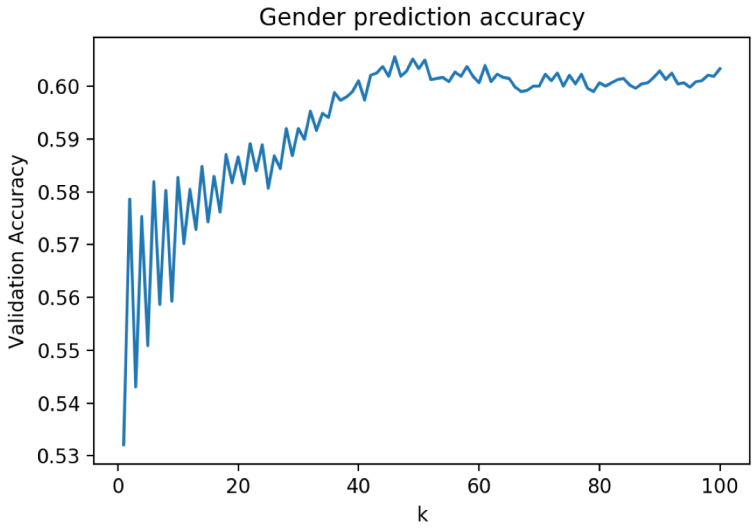
Comparison between two classification approaches

- Lets compare the performance of Support victor machine classifier, Naive Bayes classifier and k-Neighbors Classifier to answer the following question: Can we predict users gender/sex with drinks, drugs, smokes and the length of the essay the user wrote?

	Support victor machine	Naive Bayes classifier	k-Neighbors Classifier
Accuracy score	0.602	0.603	0.62
Recall score	0.001	0.0	-

Comparison between two classification approaches

- The accuracy of the k-Neighbors Classifier depend on the number of the neighbors that we set, the figure shows that the accuracy is stable around k=40, and the accuracy is about 0.62.
- K-Neighbors classifier has slightly a better performance than the other classifiers tested. However, overall, an accuracy of 0.60 is still not very satisfying!



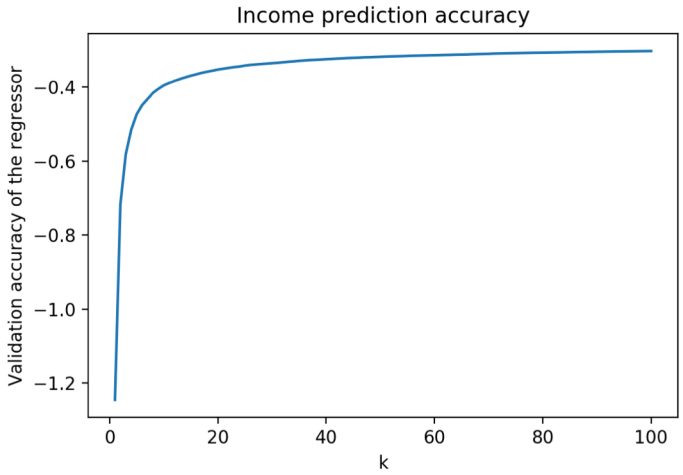
Comparison between two regression approaches

- Lest compare the performance of two regression approaches: Linear regression and k-neighbors regressor to answer the following question: Can we predict income with length of essays, drinks, drugs and smokes?

- Linear regression appears to be faster and less sophisticated than the k-neighbors regressor, however, in both cases, both regressors do not have a good accuracy in predicting the income based on the features chosen!

Comparison between two regression approaches

- The R_square factor for the linear regression is approximatively 0.01, which is clearly a poor linear fit as the closer the R_square to 1, the better the linear fit.
- The k-neighbors regressor also provides a poor model as the accuracy score is equal to -0.45!
- Clearly the results are not impressive but we can still study the accuracy of k-neighbors regressor as function of the number of neighbors, as shown in the figure



Conclusions

- Overall, we couldn’t find impressive predictions for the questions I was interested in! However, we were able to understand the main procedures to follow in order to make good use of the fundamentals of machine learning used in this course.

- Clearly, there wasn’t a strong correlation between the features I used and the type of answer/output I was looking for. Therefore, to better answer my question, I would think about other important features like education, degree, type of job, culture etc!