DATE-A-SCIENTIST CAPSTONE

Machine Learning Fundamentals
Troy Beckett
12/2/2018

codecademy

Table of Contents

- Questions to Answer
- Exploration of the Dataset
- Augmenting the Dataset
- Classification Approach
- Regression Approach
- Conclusions/Next steps

Questions to Answer

There were many questions that we could have come up and some of them were pretty interesting considering that many drinks were had. Here are the top three questions that we wanted to answer:

#1. Can we predict sex by essay? (Using Classification Techniques)

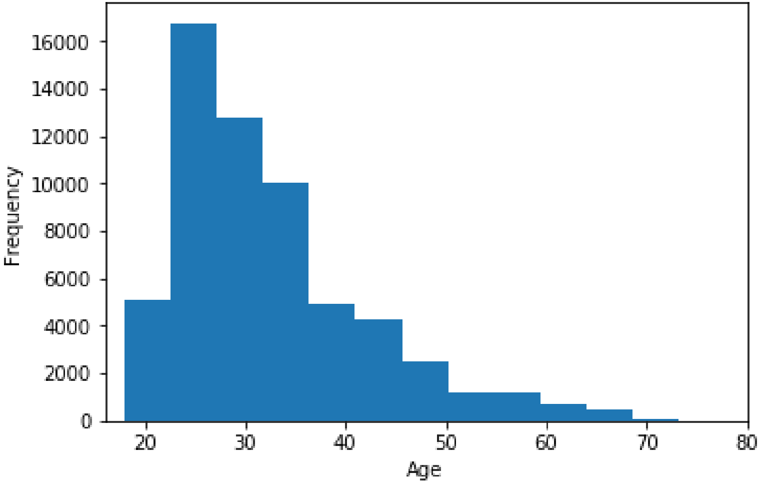
- I’ve always heard that women had more to say than men so I thought I could prove it by analyzing the essays.

#2. Can we predict height by income level? (Using Regression Techniques)

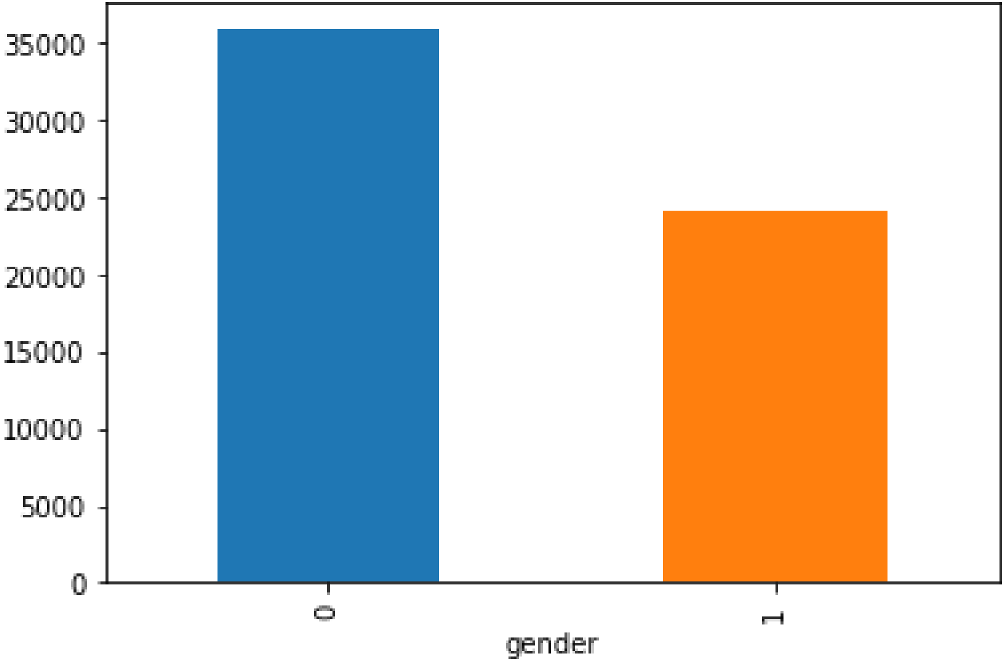
- Again, this may be a myth but I have heard that taller people make more money. Can I prove it?

Exploration of the Dataset

Exploring the dataset started with first trying to determine if I could separate the essays by age. Next, I wanted to explore if I was able to group number of words in each essay by age.

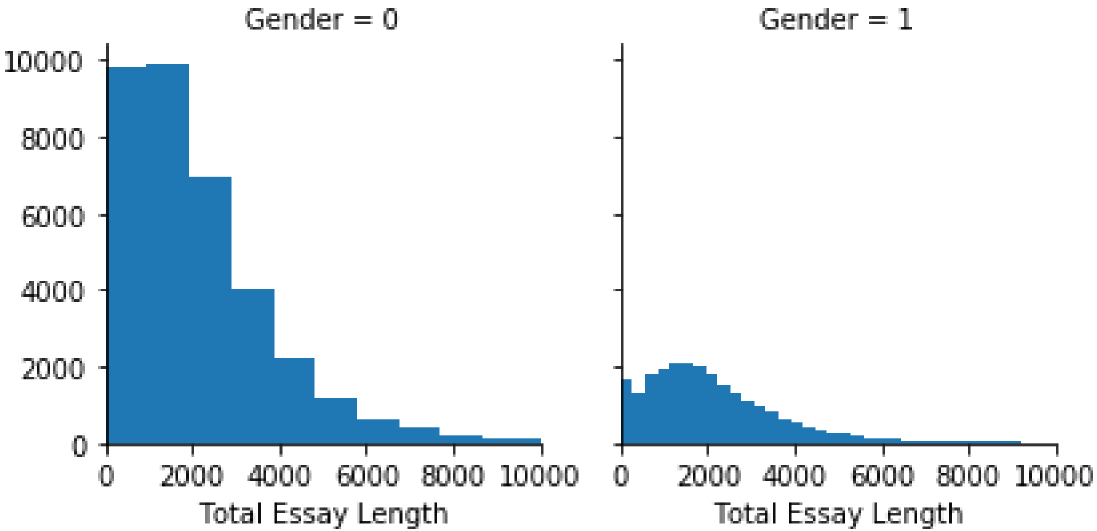


Augmenting the Dataset



Next, I needed to change some columns to create categories of gender.

Classification Approach



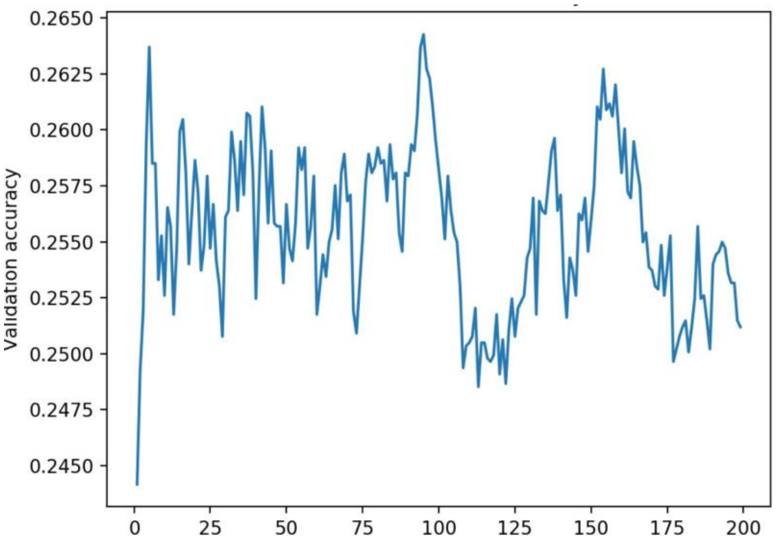
Once I was able to figure out gender, I needed to calculate total essay length by gender so I created another column to capture essay length.

My code example

```
df["Gender"]=np.where(df["sex"]=="m",0,1)
# Clean essay columns - strings only
df.essay0.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay1.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay2.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay3.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay4.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay5.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay6.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay7.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay8.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.essay9.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
essay_cols =
["essay0","essay1","essay2","essay3","essay4","essay5","essay6","essay
7","essay8","essay9"]
```

```
# Removing the NaNs
all_essays = df[essay_cols].replace(np.nan, '', regex=True)
# Combining the essays
all_essays = all_essays[essay_cols].apply(lambda x: ' '.join(x), axis=1)
df["Total Essay Length"] = all_essays.apply(lambda x: len(x))
print(df.head())
g = sns.FacetGrid(data=df, col='Gender')
g.set(xlim=(0, 10000))
g.map(plt.hist, 'Total Essay Length', bins=100)
gender_class = df[(df['Gender'] == 0) | (df['Gender'] == 1)]
print(gender_class.shape)
X = all_essays#gender_class[essay_cols]
y = gender_class['Gender']
#print(X[0])
bow_transformer = CountVectorizer(analyzer=text_process).fit(X)
len(bow_transformer.vocabulary_)
```

Regression Approach



While I was trying to create regression to show age by income. The code proved too difficult to create anything worth showing. I guess I should have come up with better questions. While I have always heard that taller people make more money, I could not prove it.

Conclusion & Final Thoughts

Course Overview

I think the course was well designed and opened my eyes to some of the concepts that used in machine learning. I definitely think using real-world data and putting concepts learned into practice is harder than expected. For example, this capstone was much harder than anticipated and almost gave up on finishing the work due to not getting the code to function properly. I'm glad I at least attempted the process and now know that the work is more involved and effort should definitely not be taken lightly.

- **Conclusion:**
    - **While the data was difficult to extract, at least at my current skill level, the results indicate that there is no difference in trying to determine sex by length of essays. I also could not prove any correlation between height and income.**
- **Final Thoughts & Next steps:**
    - **Go back and review challenge lessons that used real-world data. Find challenges that allow me to continue to put to use the knowledge gained from the course.**