In recent years, there has been a massive rise in the usage of dating apps to find love. Many of these apps use sophisticated data science techniques to recommend possible matches to users and to optimize the user experience. These apps give us access to a wealth of information that we've never had before about how different people experience romance.

In this capstone, you will analyze some data from OKCupid, an app that focuses on using multiple choice and short answers to match users.

The dataset provided has the following columns of multiple-choice data:

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

And a set of open short-answer responses to :

- 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 - Favorite 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...

Introduction

In this capstone, you will create a presentation about your findings in this OkCupid dataset.

The purpose of this capstone is to practice formulating questions and implementing Machine Learning techniques to answer those questions. We will give you guidance about the kinds of questions we asked, and the kinds of methods we used to answer those questions. But the questions you ask and how you answer them are entirely up to you. We're excited to see what kinds of different things you explore. Compared to the other projects you have completed this far, we are requiring few restrictions on how you structure your code. The

project is far more open-ended, and you should use your creativity. In addition, much of the code you write for later parts of this project will depend on how you decided to implement earlier parts. **Therefore, we strongly encourage you to read through the entire assignment before writing any code.**

Load in the DataFrame

The data is stored in **profiles.csv**. We can start to work with it in **dating.py** by using Pandas, which we have imported for you with the line:

```
import pandas as pd
```

and then loading the csv into a DataFrame:

```
df = pd.read_csv("profiles.csv")
```

Explore the Data

Let's make sure we understand what these columns represent!

Pick some columns and call `.head()` on them to see the first five rows of data. For example, we were curious about `job`, so we called:

```
df.job.head()
```

You can also call `value_counts()` on a column to figure out what possible responses there are, and how many of each response there was.

Visualize some of the Data

We can start to build graphs from the data by first importing Matplotlib:

```
from matplotlib import pyplot as plt
```

and then making some plots!

For example, we were curious about the distribution of ages on the site, so we made a histogram of the `age` column:

```
plt.hist(df.age, bins=20)
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.xlim(16, 80)
plt.show()
```

Try this code in your own file and take a look at the histogram it produces!

Formulate a Question

As we started to look at this data, we started to get more and more curious about Zodiac signs. First, we looked at all of the possible values for Zodiac signs:

```
df.sign.value_counts()
```

We started to wonder if there was a way to predict a user's Zodiac sign from the information in their profile. Thinking about the columns we had already explored, we thought that maybe we could classify Zodiac signs using drinking, smoking, drugs, and essays as our features.

Augment your Data

In order to answer the question you've formulated, you will probably need to create some new columns in the DataFrame. This is especially true because so much of our data here is categorical (i.e. `diet` consists of the options `vegan`, `vegetarian`, `anything`, etc. instead of numerical values).

Categorical data is great to use as labels, but we want to create some numerical data as well to use for features.

For our question about Zodiac signs, we wanted to transform the `drinks` column into numerical data. We used:

```
drink_mapping = {"not at all": 0, "rarely": 1, "socially": 2, "often": 3, "very often": 4, "d
```

```
all_data["drinks_code"] = all_data.drinks.map(drink_mapping)
```

These lines of code created a new column called 'drinks_code' that mapped the following `drinks` values to these numbers:

drinks	drinks_code
not at all	0
rarely	1
socially	2
often	3
very often	4
desperately	5

We did the same for `smokes` and `drugs`.

We also wanted some numerical data about the short answer essays. We combined them all into one string, took out the `NaN`s, and then created a new column called `essay_len`:

```
essay_cols = ["essay0","essay1","essay2","essay3","essay4","essay5","essay6","essay7","essay8
```

```
# 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))
```

We also created a column with average word length and a column with the frequency of the words "I" or "me" appearing in the essays.

Normalize your Data!

In order to get accurate results, we should make sure our numerical data all has the same weight.

For our Zodiac features, we used:

```
feature_data = all_data[['smokes_code', 'drinks_code', 'drugs_code', 'essay_len', 'avg_word_l
```

```
x = feature_data.values
min_max_scaler = preprocessing.MinMaxScaler()
x_scaled = min_max_scaler.fit_transform(x)
```

```
feature_data = pd.DataFrame(x_scaled, columns=feature_data.columns)
```

Use Classification Techniques

We have learned how to perform classification in a few different ways.

- We learned about K-Nearest Neighbors by exploring IMDB ratings of popular movies
- We learned about Support Vector Machines by exploring baseball statistics
- We learned about Naive Bayes by exploring Amazon Reviews

Some questions we used classification to tackle were:

- Can we predict sex with education level and income??
- Can we predict education level with essay text word counts?

Use Regression Techniques

We have learned how to perform Multiple Linear Regression by playing with StreetEasy apartment data. Is there a way we can apply the techniques we learned to this dataset?

Some questions we used regression to tackle were:

- Predict income with length of essays and average word length?
- Predict age with the frequency of "I" or "me" in essays?

We also learned about K-Nearest Neighbors Regression. Which form of regression works better to answer your question?

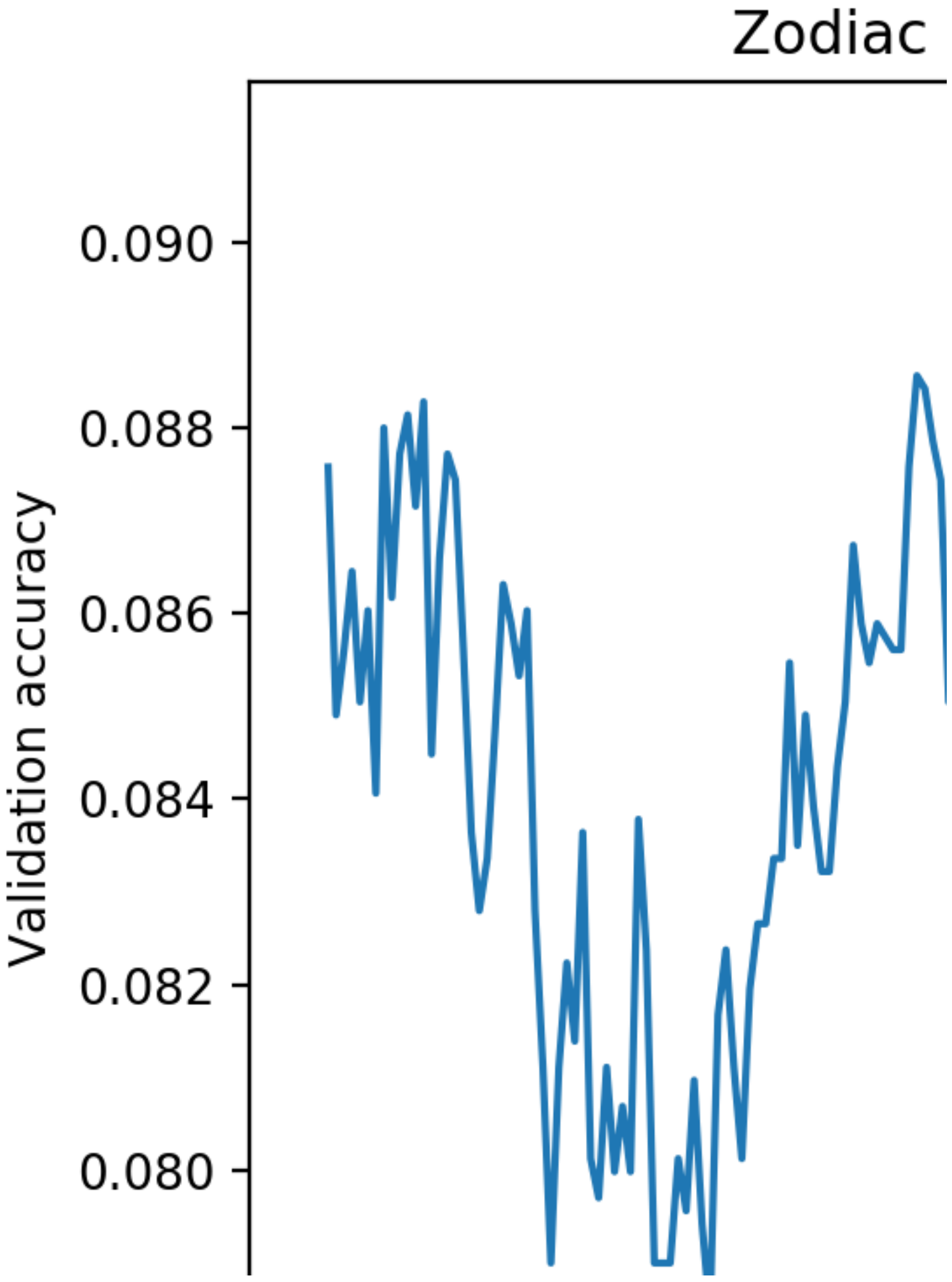
Analyze the Accuracy, Precision and Recall

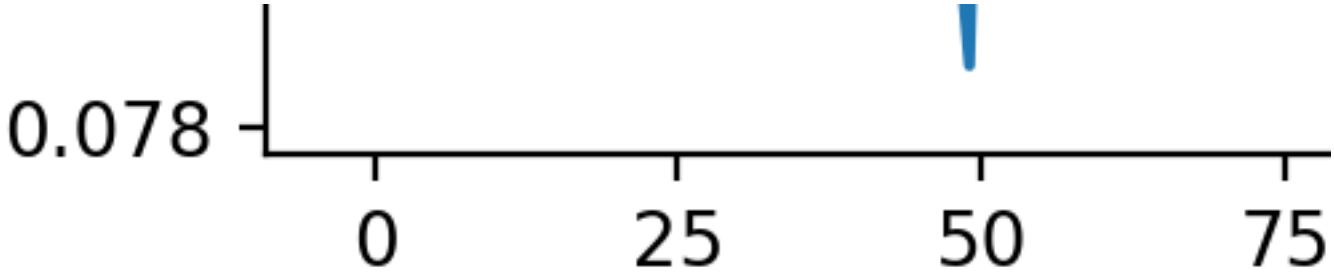
After you have trained your model and run it, you will probably be curious about how well it did.

Find the accuracy, precision, and recall of each model you used, and create graphs showing how they changed.

For our question of classifying Zodiac signs, one graph we produced showed classification accuracy versus k (for

K-Nearest Neighbors):





The accuracy we would expect from predicting a Zodiac sign by randomly selecting one would be 1/12, or 0.0833. Our model did not significantly outperform this number. We were unimpressed.

Create your Presentation

We want to see:

- at least two graphs containing exploration of the dataset
- a statement of your question (or questions!) and how you arrived there
- the explanation of at least two new columns you created and how you did it
- the comparison between two classification approaches, including a qualitative discussion of simplicity, time to run the model, and accuracy, precision, and/or recall
- the comparison between two regression approaches, including a qualitative discussion of simplicity, time to run the model, and accuracy, precision, and/or recall
- an overall conclusion, with a preliminary answer to your initial question(s), next steps, and what other data you would like to have in order to better answer your question(s)

Good luck!