Using Machine Learning Models to Complete OkCupid Demographics Data Sets: An Initial Investigation

Machine Learning Fundamentals

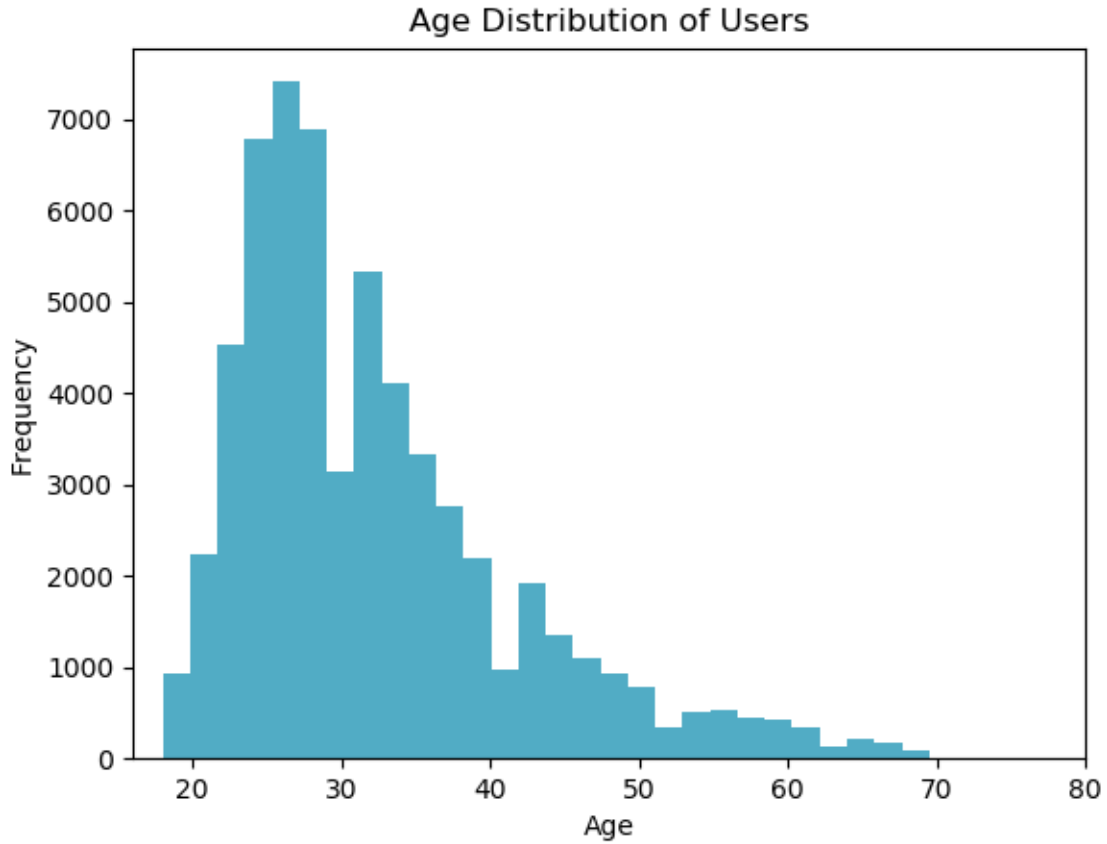
Codecademy Intensive Capstone Project

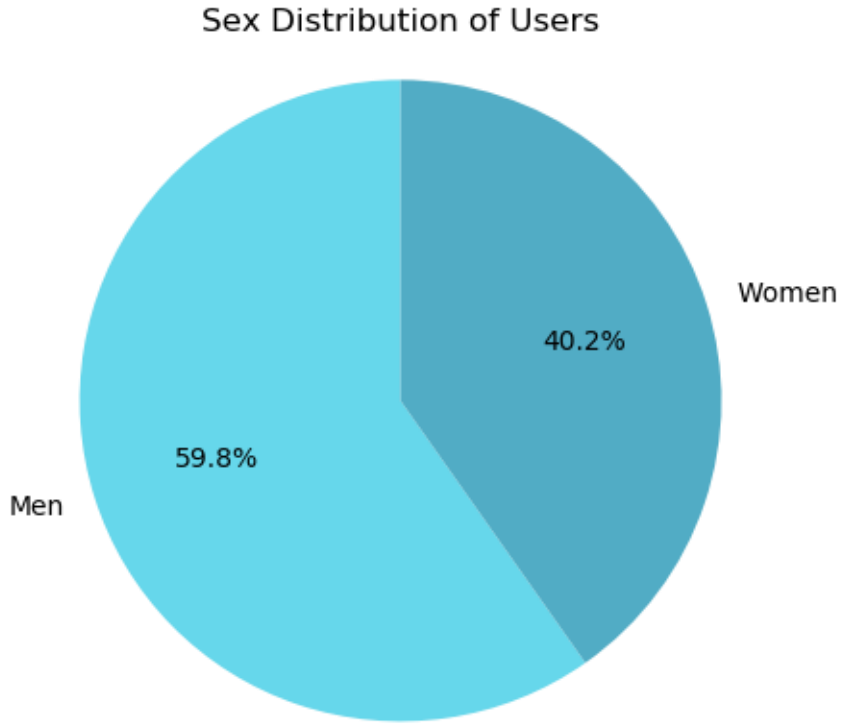
Jeffrey Egan / December 2018

Project Overview

- For this project we were provided user profile data from the dating website OkCupid.
- The user profile data provides a self-reported census that contains sex, age, level of education along with many other attributes.
- The data however has several user records with incomplete profile data.
- While most dating websites rely on user subscription fees for revenue, OkCupid is a free service that relies primarily on income from selling advertisements.
- We understand that advertising is most effective and can demand more per view or click when it is targeted based on a viewer's demographics and interests.
- Can we create machine learning models to predict and complete missing fields in a user's profile so that we can improve the advertising and thus revenue for OkCupid?

Inspecting Basic Demographics User Data





User data for age skews young with the data set having a median age of 30 years, and a 75th percentile at 37 years old.

dragonaur

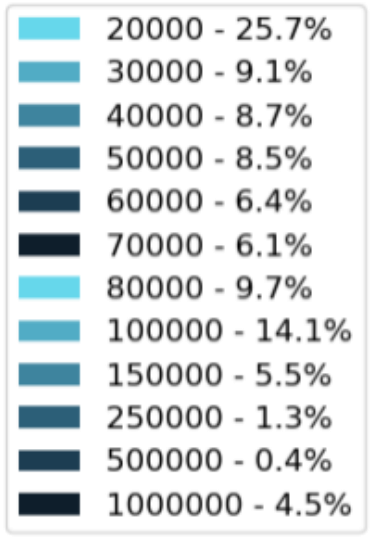
okcupid

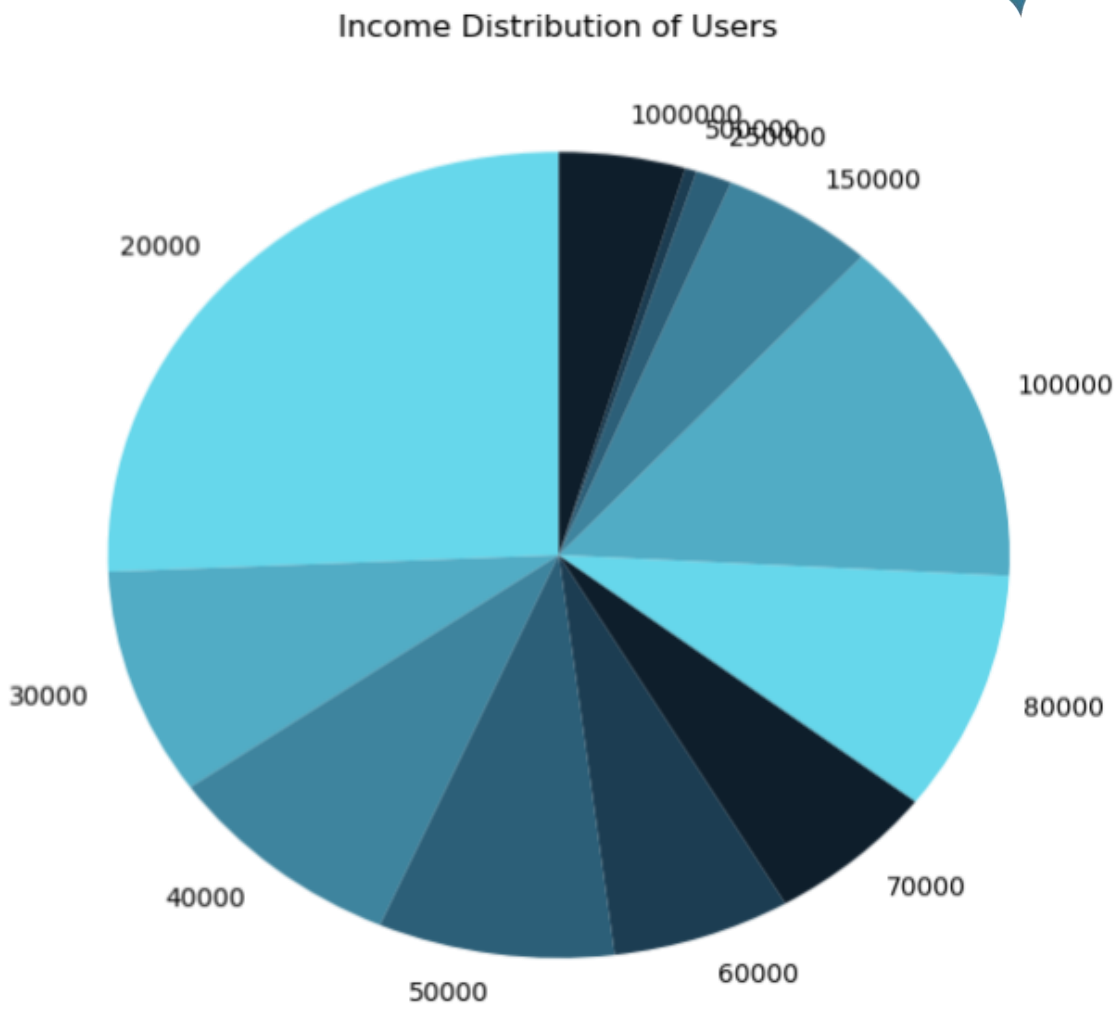
Inspecting User Income Data

The provided data set includes data for 59,946 users.

Unfortunately, inspection revealed that 80.8% of users provided no income data!

Income distribution for the remaining 11,504 users in the data set is presented along with a Five Number Summary.





Minimum Value	$20k
First Quartile	$20k
Median	$50k
Third Quartile	$100k
Maximum Value	$1,000k

Interesting Observation: 6.2% of users report an income in excess of $200k

Questions to Provide Insights

- Perspective advertisers are interested in income levels of OkCupid's user base. That said, the provided data set has many entries with missing Income information. Can we use other features such as education, age, and field of employment to generate regression models to accurately predict a user's income and populate the omitted data?

- There is also an opportunity to take OkCupid data and gain insights that may appeal more broadly to advertisers. For example, advertisers have browser history metadata that allows them to deduce a user's field of employment and whether or not they have dietary preferences like vegetarianism. Using OkCupid data, can we generate a classification models to predict a user's sex so that advertisers may augment their own data sets and develop even more targeted ads?

Augmenting Data with New Columns

- education_code
  - To use education level in the machine learning models, education level strings must be mapped to numbers. Further more, education level is a progressive trait so efforts were made to group similar education levels and linearize the data as shown in the python snippet below.

```
education_mapping = {"dropped out of space camp ": 0, "working on space camp": 0, "space camp": 0,
                     "graduated from space camp": 0, "dropped out of high school": 1, "working on high school": 2,
                     "high school": 2, "graduated from high school": 3, "dropped out of two-year college": 3,
                     "dropped out of college/university": 3, "working on two-year college": 4, "two-year college": 4,
                     "working on college/university": 5, "college/university": 5, "graduated from two-year college": 6,
                     "graduated from college/university": 7, "dropped out of masters program": 7,
                     "dropped out of law school": 7, "dropped out of med school": 7, "working on masters program": 8,
                     "working on med school": 8, "med school": 8, "working on law school": 8, "law school": 8,
                     "masters program": 8, "graduated from masters program": 9, "dropped out of ph.d program": 9,
                     "working on ph.d program": 10, "ph.d program": 10, "graduated from law school": 11,
                     "graduated from ph.d program": 11, "graduated from med school": 11}
df['education_code'] = df['education'].map(education_mapping)
```

- diet_code:
  - To use diet in the machine learning models, strings again must be mapped to numbers. While this data is not linear, grouping dietary preferences into similar categories was completed: anything, vegetarian/vegan, influenced by religion, and other. Later, this grouping allowed exclusion of certain data (e.g. religious preferences and other) for testing the hypothesis).

```
diet_mapping = {"mostly anything": 0, "anything": 0, "strictly anything": 0, "mostly vegetarian": 1, "mostly other": 3,
                "strictly vegetarian": 1, "vegetarian": 1, "strictly other": 3, "mostly vegan": 1, "other": 3,
                "strictly vegan": 1, "vegan": 1, "mostly kosher": 2, "mostly halal": 2, "strictly kosher": 2,
                "strictly halal": 2, "halal": 2, "kosher": 2}
df['diet_code'] = df['diet'].map(diet_mapping)
```

Augmenting Data with New Columns

- income_z: column of Z-Score Normalized income
  - Z-Score normalization was chosen because while most reported incomes were below $200k, a small percentage of users reported income as high as $1M. A simple Min-Max normalization process would have removed much of the fidelity for the majority of our data set.

```
df = df[df.income != -1]
df['income_z'] = (df['income'] - df['income'].mean()) / df['income'].std(ddof=0)
```

  - The income_z column of data was not ultimately required for any machine learning models discussed in this report but may be useful in future analyses.

Regression Models Surveyed

Multiple Linear Regression

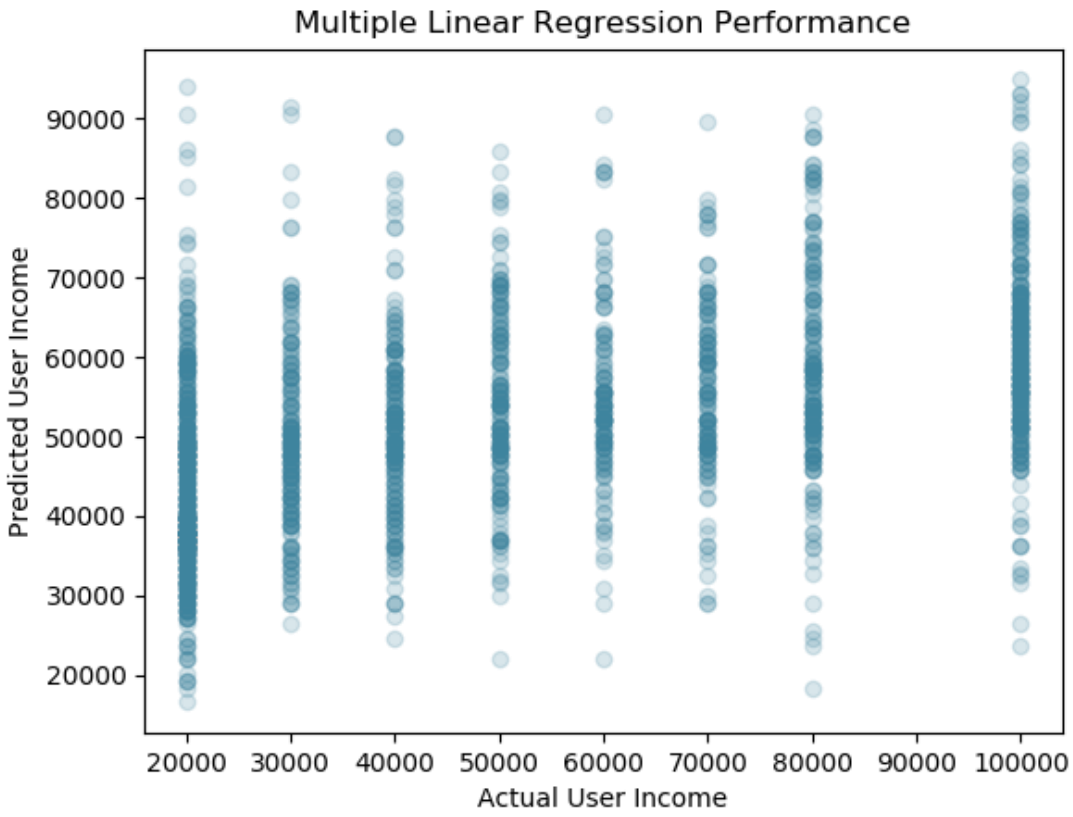
- Input Features must be Linearized
  - Works well for some inputs like age or level of education but stumbles with data that can't be linearized like job field.
- Input Feature data limited to:
  - `education_code` and `age`
- Input Labels data:
  - `income` where income <$120k, and != -1
  - Both models performed better after culling high income outliers from the training/testing data.
- Model initialization, fitting, and scoring time: 0.002 seconds / trial

K-Nearest Neighbors Regressor

- Model performs a k-NN assessment before running a weighted regression against chosen neighbors so all desired inputs are viable.
  - Value k chosen from Monte Carlo analysis.
- Input Feature data:
  - `education_code`, `job_code`, and `age`
- Input Labels data:
  - `income` where income <$120k, and != -1
- Model initialization, fitting, and scoring time: 0.03 seconds / trial

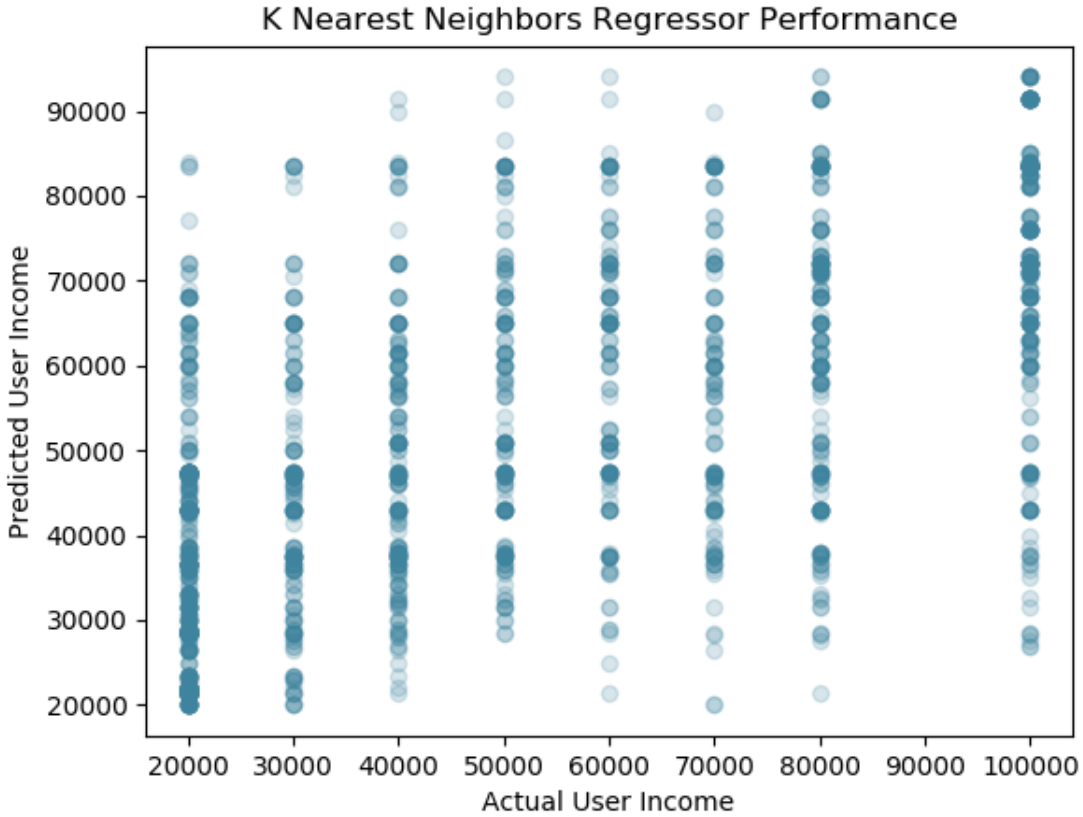
Regression Model Results

Multiple Linear Regression



Limiting input features to only education and age negatively impacts performance. R^2 Score: 0.220904

K-Nearest Neighbors Regressor



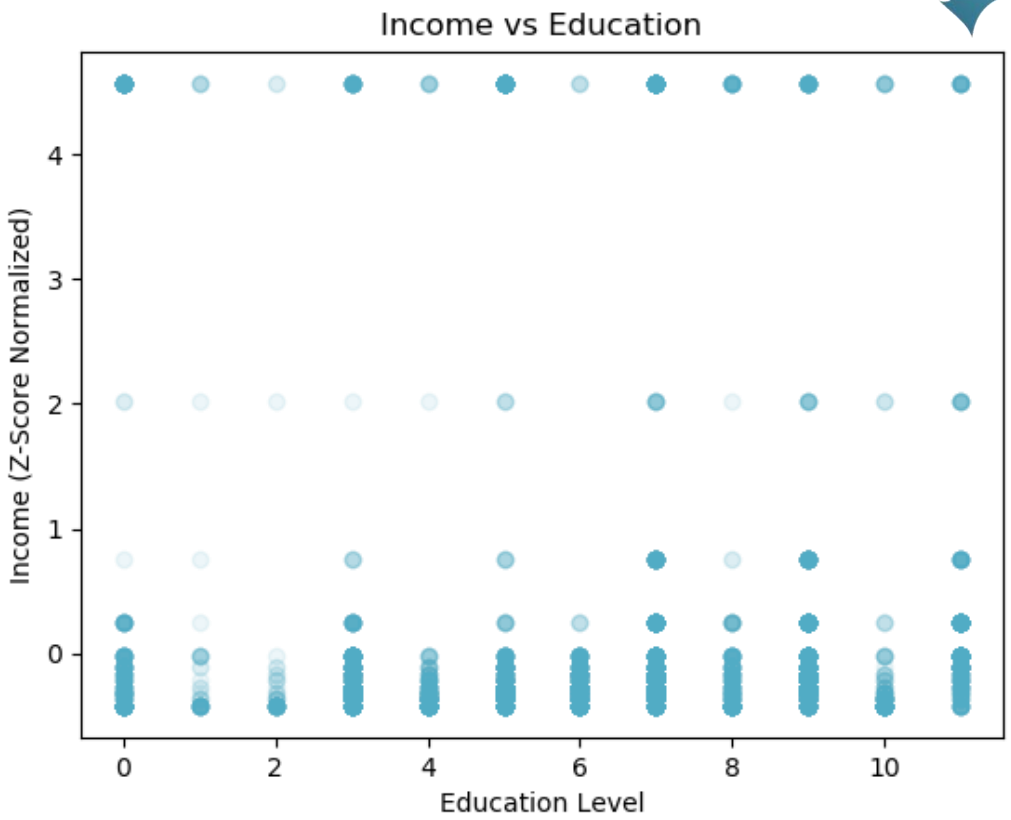
Ability to use all three desired input features improves performance. R^2 Score: 0.421187

dragonaur

okcupid

Insights from Employing Regression Models

- In the end both models failed to meet performance goals for $R^2 > 0.7$.
- For our test question, a multiple linear regression model is not ideal since some of our feature data is not highly linear or not linear at all.
- In the end though it just appears that there is not a strong enough relationship between features and labels as one would hope (from data supplied, it appears that education level has little or no impact on income). Suggestions to improve this are covered later.



Education level alone does not correlate strongly enough with income in the OkCupid data provided.

- Both models take little processing time to initialize, fit, and score and may be quickly assessed.

dragonaur

okcupid

Classification Models Surveyed

K Nearest Neighbors Classification

- Input Feature data:
  - Features `job_code` and `diet_code`
  - Input data cleaned to remove dietary preferences other than “*anything* and *vegan*/*vegetarian*.
- Input Labels data:
  - `sex_code` - the binary representation of male/female
- Model initialization, fitting, and scoring time: 0.432 seconds / trial

Support Vector Machine Classification

- Input Feature data:
  - Features `job_code` and `diet_code`
  - Input data cleaned to remove dietary preferences other than “*anything* and *vegan*/*vegetarian*.
- Input Labels data:
  - `sex_code` - the binary representation of male/female
- Uses the Radial Basis Function kernel
- Monte Carlo trials reveals performance is indifferent to gamma. (0.05 to 100)
- Model initialization, fitting, and scoring takes significant time: 11.833 sec / trial

K Nearest Neighbors Classifier Setup

For our classification model we are seeking to identify the sex of a user based on career field and dietary preferences.

Since both sexes appear with roughly equal measure, the model's accuracy score should be the dominant driver for choosing k.

That stated, there is no reason to choose k such that the secondary metrics of precision and recall suffer.

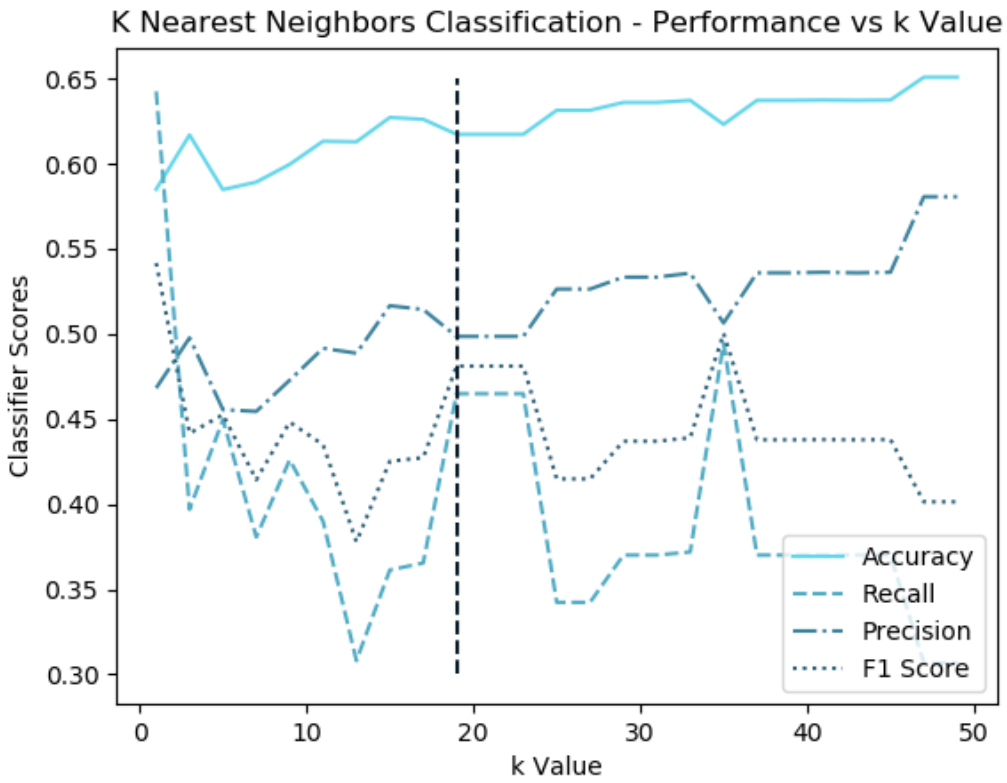
With these considerations, trials reveal k=19 as the optimal number of nearest neighbors for the classifier.

dragonaur

okcupid

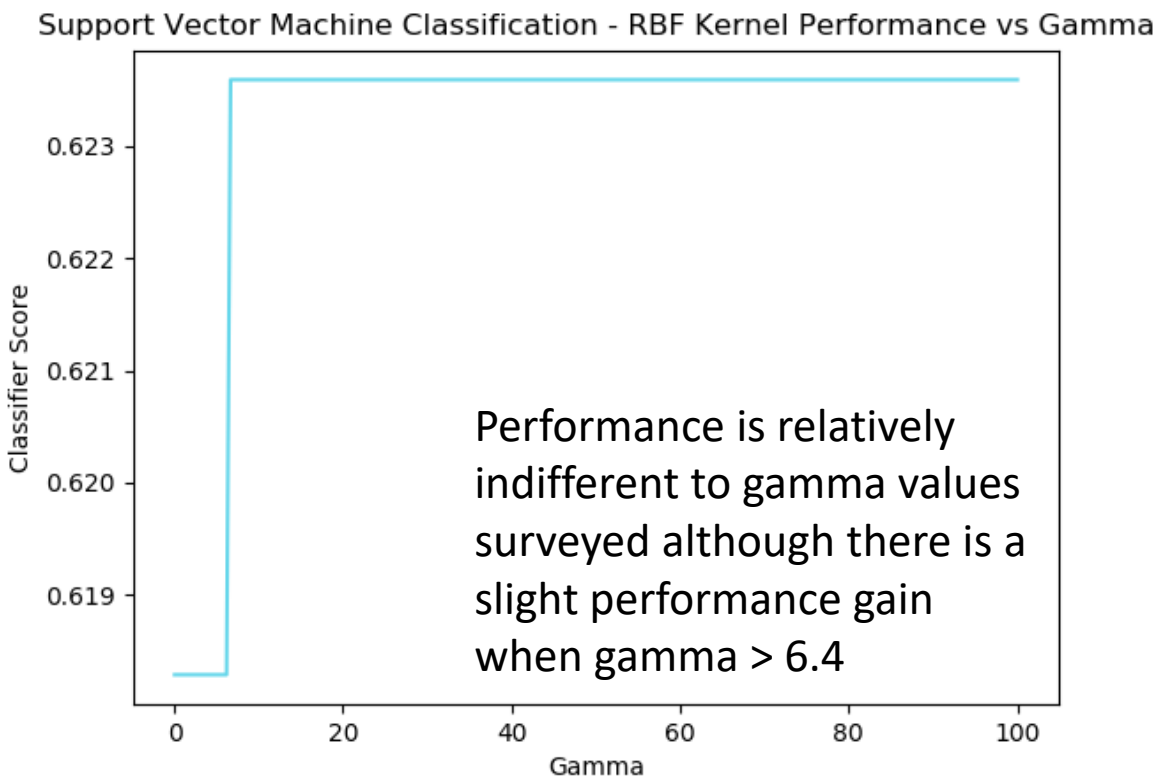
Classification Model Results

K Nearest Neighbors Classification



Accuracy Score: 0.617232

Support Vector Machine Classification



Accuracy Score: 0.623588

dragonaur

okcupid

Insights from Employing Classification Models

- Both models performed better than the regression models produced, however, both failed to meet performance goals for accuracy > 0.7 with both the k-NN and SVM models scoring 0.62.
- Efforts to cleanse training data and improve model performance were exhausted.
- The largest improvement (scores from ~0.4 to ~0.6) came from cleansing diet data to exclude religious dietary preferences from the input training data for example.
- Some data cleansing options were evaluated but ultimately discarded because while improving performance score on paper, restricting the viable training data beyond a certain point negates the practical usefulness of the model in solving our stated problem.

Insights from Employing Classification Models

- Like the regression models, the classifier models would also benefit greatly from finer resolution information within some features. Some suggestions for data augmentation is presented on the following slide.

- On a final note, despite nearly equivalent accuracy performance, the SVM takes substantially longer to implement, train, and run than the kNN. Thus when iterating on input feature subsets and data cleansing choices, it may be more prudent to prototype and iterate with data cleansing decisions while using the K Nearest Neighbors classifier.

Enabling More Accurate ML Models

- Augmenting at least two of the primary data features used with new information would enable the development of more accurate machine learning models.

  - The Job Field feature as it stands right now is not sufficient enough to enable regression models to predict income. For example, the “Medicine / Health” field could conceivably include users with diverse positions ranging from Pharmacy Technician to Neurosurgeon. Thus it is also safe to assume a wide variance in reported incomes, even within the same field. Similarly with regard to our classification model, while the “Medicine / Health” field may have users with similar portions of men and women, users with job title like “Labor and Delivery Nurse” may skew female. Augmenting the Job Field data with Profession or Job Title is recommended.

  - Similar arguments can be made to augment Education with details on the field of study. For example, a user with a MS. in Electrical Engineering may in general earn more than a user with a M.A. in Literature but with current data, regression models treat both of these users as the same as they have an equivalent education level. Augmenting the Education data with Field of Study or Degree Title is recommended.

Overall Conclusion and Next Steps

- Supervised machine learning regression and classification models were created to predict and complete missing fields in user data profiles with the aim of improving OkCupid's ability to target advertisements.

- Unfortunately, many input features one would initially think viable for training machine learning models weren't specific enough or simply didn't correlate well enough to inform sufficiently accureate predictions.

- It is this report's recommendation that OkCupid allows users to augment job field and education data as described with more granular information fields (e.g. job title and field of study) in order to improve models and ultimately improve OkCupid's ability to effectively deliver advertisements.