Lifestyle Choices: Can We Predict Them?

An analysis of OK Cupid Data

Overview: Research Questions

- Can you predict **offspring from smoking, drinking, and drug choices**?

  (In this case offspring means not just whether someone has kids or not, but what their actual preferences are for having children at all)

- Can you predict **if someone is male or female based on smoking, drugs, and drinking choices**?

Predicting Lifestyle Choices: New Columns

```
drink_mapping = {"not at all": 0, "rarely": 1,
"socially": 2, "often": 3, "very often": 4,
"desperately": 5}

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

Likewise:

```
df["smokes_code"] =
df.smokes.map(smokes_mapping)

df['drugs_code'] = df.drugs.map(drugs_mapping)

df["offspring_code"] =
df.offspring.map(offspring_maping)

sex_mapping = {"m": 2, "f": 1}
```

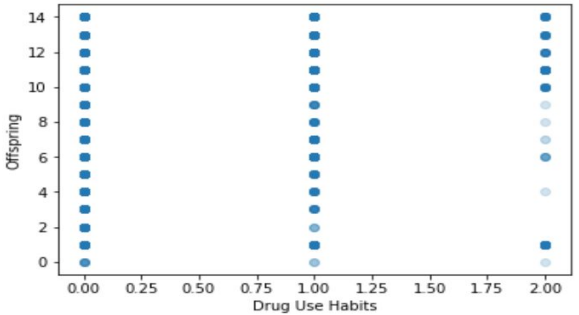
- Preparing the data for analysis involved the creation of new columns, which was done by the simple process of mapping.
- The offspring code featured a range of numbers from 0 to 14, with 0 marked as “has kids and wants more” and 14 “doesn’t have kids and doesn’t want them”.
- We also had to fill in any NaN with reasonable values; we knew our data were ready if our columns registered “False” to this query:

```
df.isna().any()
```

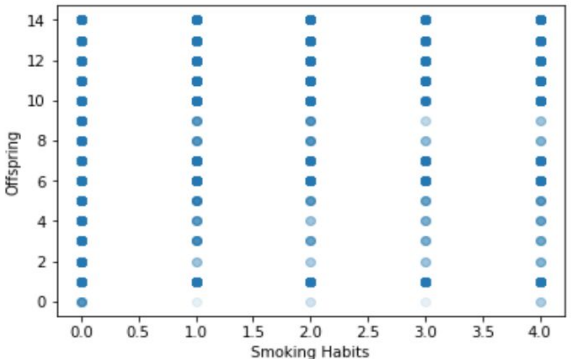
Our Data

Our data comes from OKCupid and was provided in CSV form. Here is some basic data exploration.

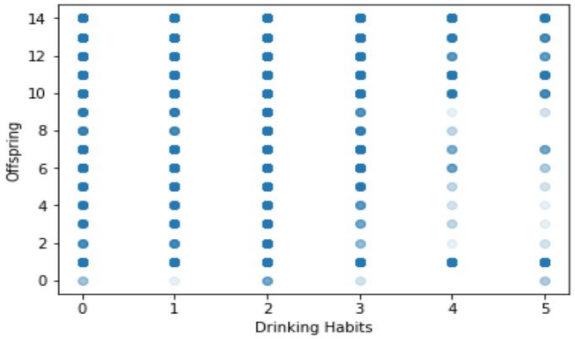
Not a lot of people who use drugs have children and/or want more!



This graph may only really tell us that most people don't smoke.



Not a lot of people who drink "desperately" have children and/or want more!



Predicting Lifestyle Choices: Regression

- We used two different machine learning regressions to tackle the same question: Can you predict **offspring from smoking, drinking, and drug choices**?

- The two regression implementations we used were: multilinear regression (MLR) and K-Nearest Neighbor Reggressor (K-Nearest)

MultiLinear Regression

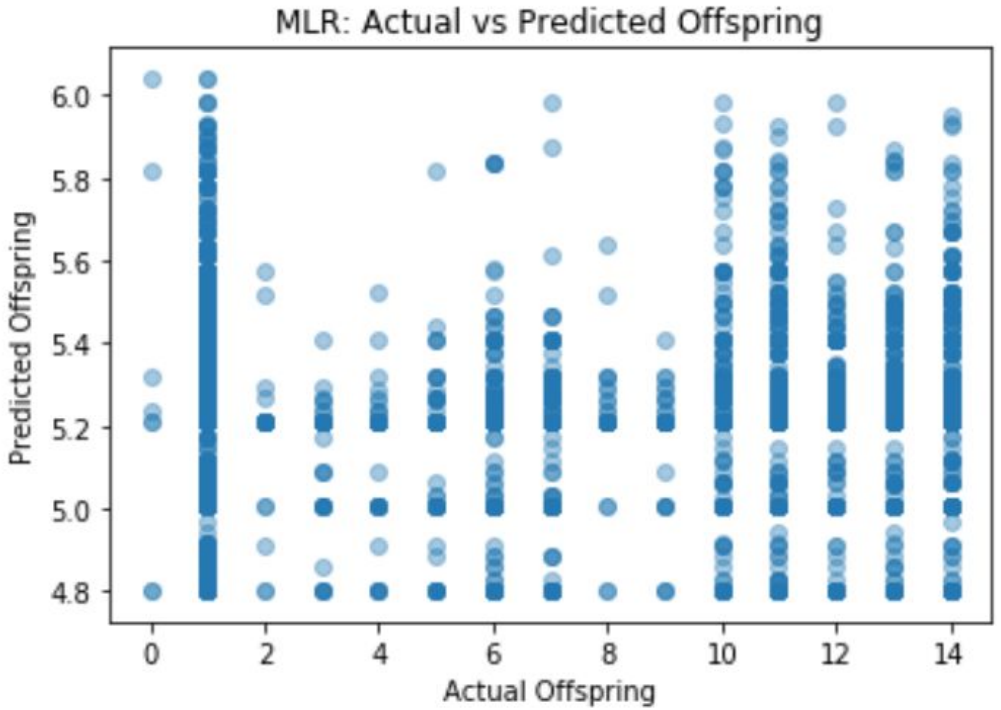
Our MLR score was not impressive:

mlr.score(x_test,y_test):
0.0003515942822691631

mlr.score(x_train,y_train):
0.001032802654101328

But we did at least learn which of our coefficients carries the most predictive power…

[('smokes_code', array([0.02709139, 0.05706273, 0.2022135]))]
(It's Drinking!)



K-Nearest Neighbor Regressor

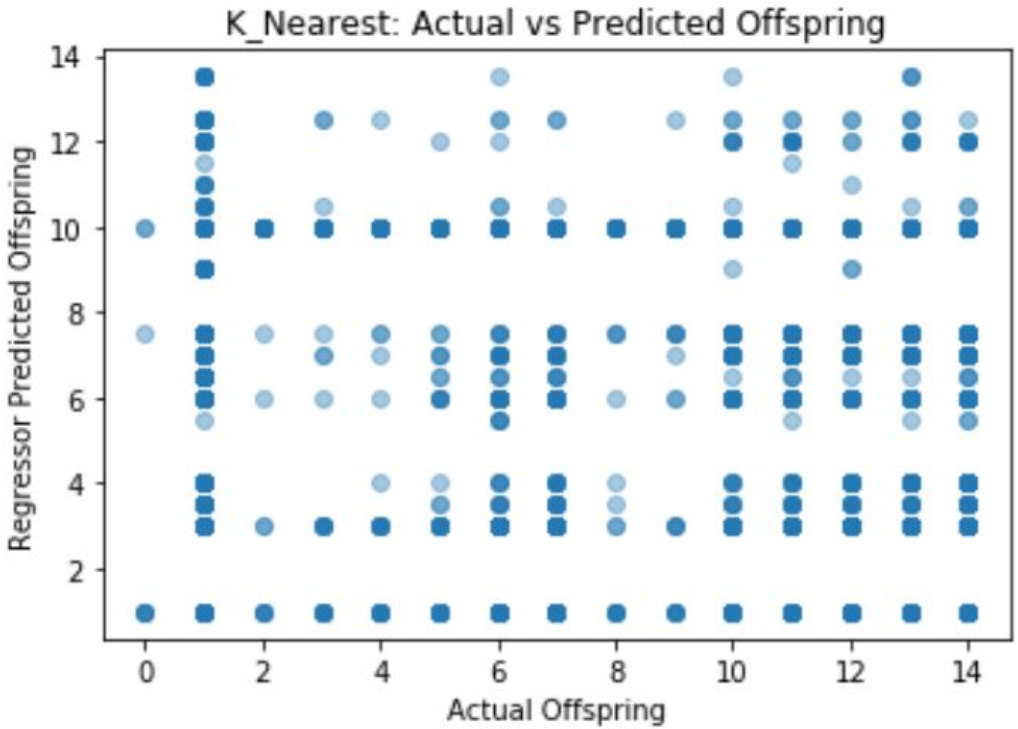
This regression scored much better!

regressor.score(x_test,y_test):
-0.5821071653137004

regressor.score(x_train,y_train):

-0.5562999586593569

And it was simple & efficient to complete!



Predicting Lifestyle Choices: Classifier

- We used two different machine learning classifiers to see if we could **predict if someone was male or female based on smoking, drugs, and drinking choices**.

- The two classifier implementations we used were: Support Vector Machine (SVM) and K-Nearest Neighbor (K-Nearest)

Predicting Lifestyle Choices: SVM

Accuracy is calculated by finding the total number of correctly classified points and dividing by the total number of points. Accuracy_score: **0.5962468723936614**

Recall measures the percentage of relevant items that your classifier found. Recall_score: **0.5962468723936614**

Precision is the number of true positives the algorithm correctly predicted divided by the number of times it predicted positives. Precision_score: **0.3555103328392231**

F1 score is the harmonic mean of precision and recall. F1_score: **0.4454327698147536**

Predicting Lifestyle Choices: K-Nearest

Accuracy is calculated by finding the total number of correctly classified points and dividing by the total number of points. Accuracy_score: **0.5756463719766473**

Recall measures the percentage of relevant items that your classifier found. Recall_score: **0.5756463719766473**

Precision is the number of true positives the algorithm correctly predicted divided by the number of times it predicted positives. Precision_score: **0.49380054287564057**

F1 score is the harmonic mean of precision and recall. F1_score: **0.47123445599445146**

Conclusions

- Our multilinear regression was not very impressive, but at least we did learn which of our coefficients has the most power.
- Our K-Nearest Neighbor regression performed better, and showed that those people with children are more predictable in their other lifestyle choices.

- In terms of classifiers, our SVM model was more accurate, but our K-Nearest Neighbor classifier was far more precise, making it the superior choice, as reflected in the F1 score.
- However, neither model was terribly impressive. Using the F1 score, we can see a below 50% rating. Not good! You really can't tell if someone is male or female based on the drug, smoking, and drinking habits!

- Don't judge a book by its cover!
- Next steps: Are there other lifestyle habits that people with kids engage in more predictably?
- Does smoking predict drug choices or drinking, or vice versa?