Unit 10 - Capstone 1 - In-depth Analysis

> *You've learned several techniques to use supervised and unsupervised learning to help build your predictive models... The techniques you'll use in this project depend on your dataset. (Springboard)*

Data File and Format

File

- SFBayWaterQualityCombined.csv - all water quality parameters, original combined file

Format

- CSV format; one header row; 27 columns

Columns

```
Date, Time, Station Number, Distance from 36, Depth, Discrete Chlorophyll,
Chlorophyll a/a+PHA, Fluorescence, Calculated Chlorophyll, Discrete Oxygen,
Oxygen Electrode Output, Oxygen Saturation %, Calculated Oxygen, Discrete SPM,
Optical Backscatter, Calculated SPM, Measured Extinction Coefficient, Calculated
Extinction Coefficient, Salinity, Temperature, Sigma-t, Nitrite, Nitrate +
Nitrite, Ammonium, Phosphate, Silicate, DateTime
```

Linear Regression

> Regression analysis is a subfield of supervised machine learning. It aims to model the relationship between a certain number of features and a continuous target variable. (https://towardsdatascience.com)

Several columns in the dataset were calculated based on a linear regression between two measurements. I thought it could be interesting to try to replicate these results.

Background

/via T. Schraga, USGS:

> You can not replicate our Calculated or Measured Light Extinction Coeff as we do not save the PAR data, or the pre-1988 Secci data. Note the difference between our Calculated and Measured parameters; I updated the documentation for you.
>
> Your best bet for matching our calculated results is to conduct a linear regression with Discrete Oxygen v. Oxygen electrode output. Again, we occasionally dropped outliers (>2 x st dev), but it was less often. You can only do this for any individual cruises up until August 2016; at that point we stopped collecting discrete oxygen samples because the sensors are so good at holding the calibration.
>
> You can try to replicate our regressions between discrete chl+fluorescence and discrete SPM+Optical Backscatter volts but you will usually not match perfectly. The regressions are a process; often there are regressions for each sub embayment (e.g. stn 36-22 = south bay, or south bay is split 36-25 and 24-22) - this is because of different phytoplankton populations, tide changing (flood/ebb), different sediments in the embayments, and more. Note, every regression we do for every parameter is cruise specific.

Features Used

For purposes of this investigation, there are three interesting sets of features.

Oxygen

- Discrete Oxygen - Concentration of dissolved oxygen in water samples measured in the laboratory
- Oxygen Electrode Output [volts] - Voltage output of the oxygen electrode, a relative measure of the concentration of oxygen dissolved in the water.
- Calculated Oxygen [mg/L] - Estimated concentration of dissolved oxygen, calculated from the oxygen electrode voltage output which is (calibrated) using linear regression with the discrete measures of the dissolved oxygen.

Suspended Particulate Matter (SPM)

- Discrete SPM [mg/L] - Concentration of suspended sediments (or suspended solids) in bay water at the corresponding station and depth, measured by weighing the mass of solids collected onto filters after drying.
- Optical Backscatter [volts] - OBS Raw voltage-based output of the optical backscatter sensor, a relative measure of the concentration of suspended sediments or solids in the water.
- Calculated SPM [mg/L] - The estimated concentration of suspended sediments, calculated from the OBS voltage output and linear regression (calibration) between the discrete measures of suspended solids and the OBS voltage.

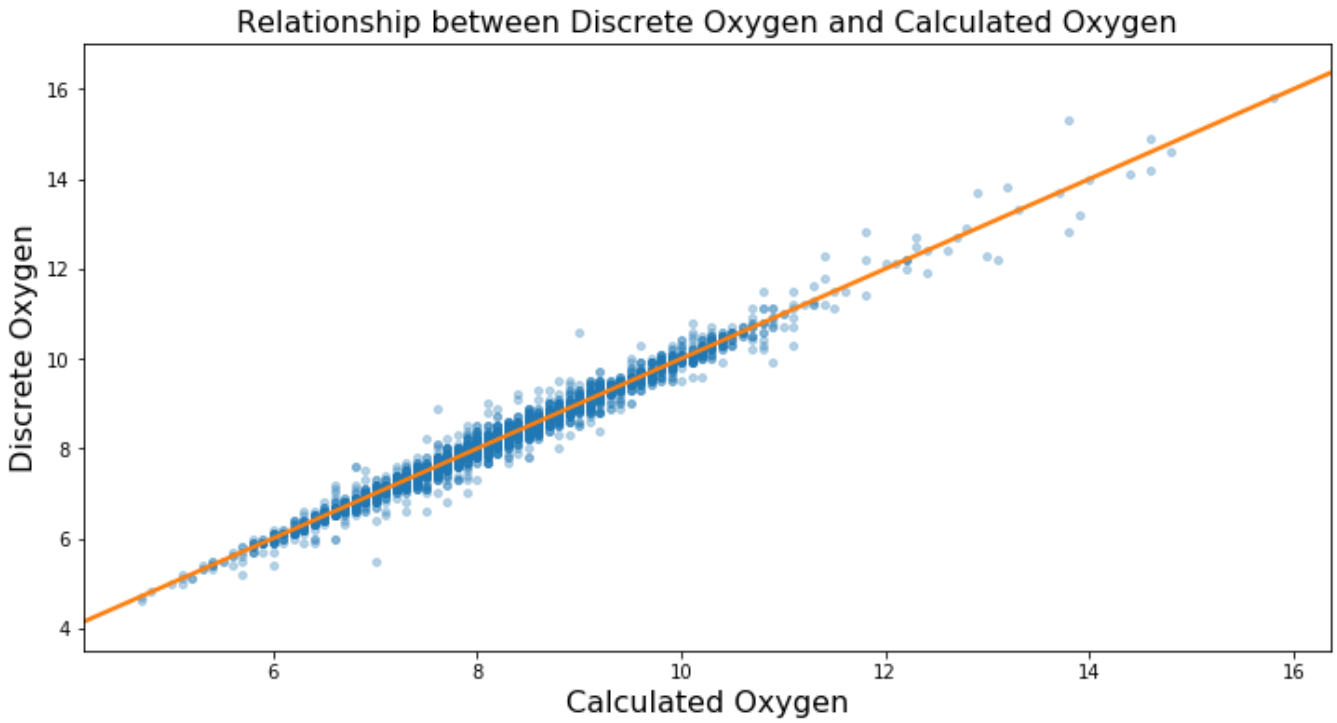
Chlorophyll

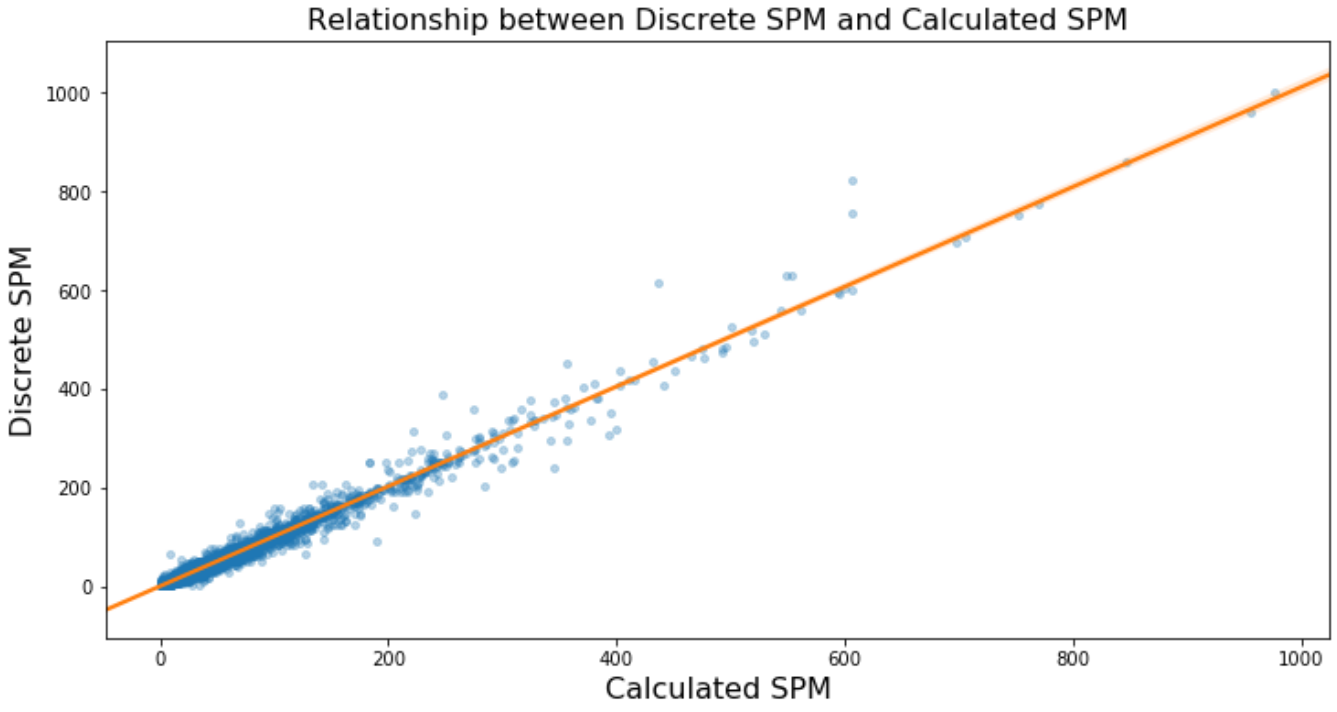
- Discrete Chlorophyll [mg/m3] - Concentration of chlorophyll content, measured by laboratory analysis of a water sample collected onto a filter.
- Fluorescence [volts] - Raw voltage-based output of the fluorometer, a relative measure of the concentration of chlorophyll-a in the water.
- Calculated Chlorophyll [mg/m3] - Estimated concentration of chlorophyll-a in water samples from in-vivo fluorescence measured with a ship-board fluorometer. The calculations are based on linear regressions of fluorescence and Discrete Chlorophyll.

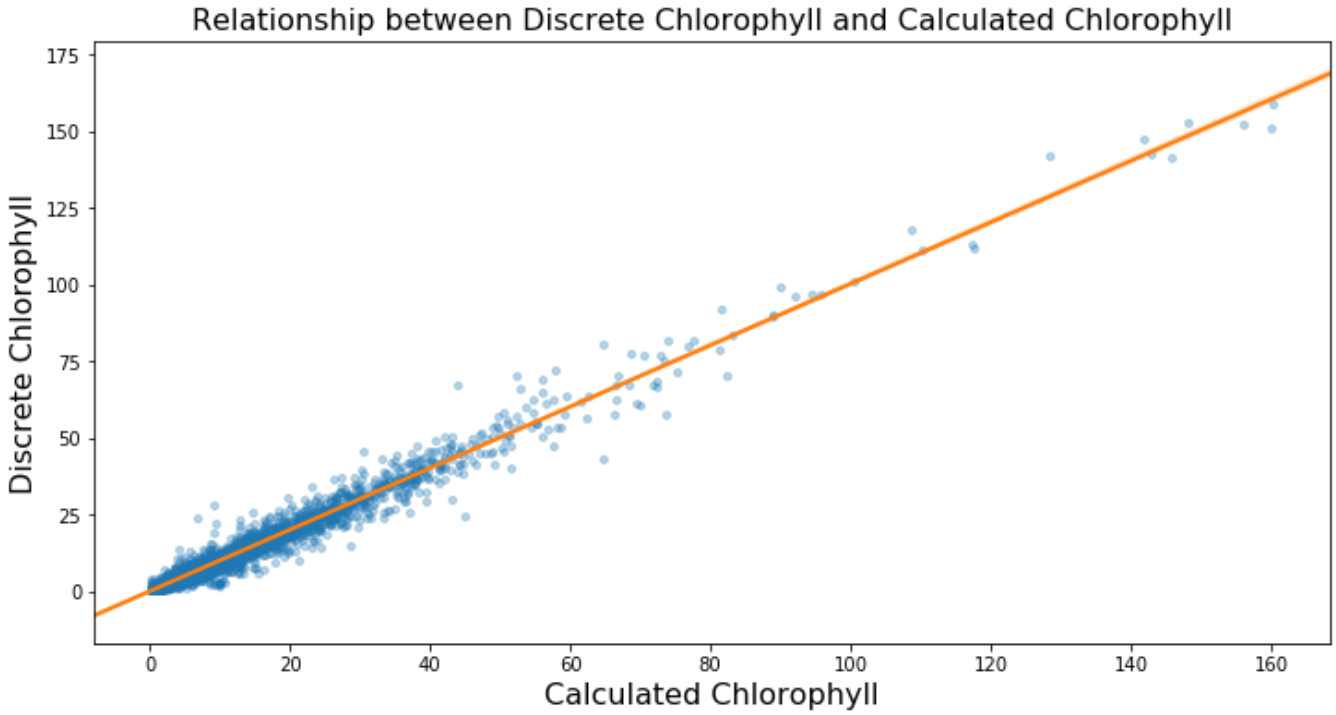
Confirm Correlation of Data Values

I first confirmed the correlation between "Discrete" and "Calculated" values in the data, by graphing.

- Oxygen - discrete vs calculated
- SPM - discrete vs calculated
- Chlorophyll - discrete vs calculated


Relationship between Discrete Oxygen and Calculated Oxygen
Discrete Oxygen
Calculated Oxygen
4
6
8
10
12
14
16


Relationship between Discrete SPM and Calculated SPM
Discrete SPM
Calculated SPM
0
200
400
600
800
1000



Replicate Linear Regression

For each parameter, (Oxygen, SPM, Chlorophyll), I grouped the dataset by cruise date and sorted by "Discrete" measurement. I then looked for cruise data where the number of discrete measurements, paired with the other value required for the regression, was large.

For example:

Date	Station Number	Discrete_Oxygen	Oxygen_Electrode_Output	Calculated_Oxygen
2019-06-04	479	0	479	479
...	...	...	...	...
2012-01-10	549	15	549	549
2015-10-15	540	15	526	526
2011-06-14	532	16	532	532
2012-02-07	296	26	0	0
2011-12-13	259	29	0	0

The cruise for 2011-06-14 looked the most promising. There were more Discrete Oxygen values for two other cruises, but no Oxygen Electrode Output values were measured on those dates.

Once I had selected a cruise, I extracted only the relevant data for that date, then fitted a regression model using `ols` from Python's `statsmodels.formula.api`.

Oxygen

e.g.:

```
model = ols('Discrete_Oxygen ~ Oxygen_Electrode_Output',
            data = df_tmp).fit()


df_tmp['Oxygen_prediction'] = round(model.predict(df_tmp),2)
```

O2 linear regression output:

```
In [15]: df_tmp
```

executed in 17ms, finished 19:36:28 2019-12-15

Out[15]:

	Discrete_Oxygen	Oxygen_Electrode_Output	Calculated_Oxygen	Oxygen_prediction
172413	5.9	5.8	5.8	5.87
172420	6.4	6.4	6.5	6.45
172436	6.7	6.7	6.8	6.74
172456	7.1	7.1	7.2	7.13
172505	7.3	7.3	7.3	7.32
172527	8.0	7.9	7.9	7.91
172587	7.5	7.5	7.5	7.52
172639	6.9	6.8	6.8	6.84
172654	7.0	7.0	7.0	7.03
172718	8.1	8.1	8.1	8.10

As suggested by T. Schraga of USGS, the predicted results were a good match for the "calculated" values in the data set.

SPM and Chlorophyll

I used the same method to choose cruises and fit models for SPM and Chlorophyll.

The predicted values for SPM were pretty good across all stations.

	Station Number	Discrete_SPM	Optical_Backscatter	Calculated_SPM	SPM_prediction
207735	36.0	NaN	3.48	109.0	110.961
207736	36.0	147.8	5.10	156.0	160.996
207737	36.0	NaN	4.13	128.0	131.037
207738	36.0	43.8	1.09	39.0	37.144
207739	36.0	NaN	4.34	134.0	137.523
...	...	...	...	...	...
208277	657.0	NaN	0.64	19.0	23.246
208278	657.0	20.0	0.73	25.0	26.026
208279	657.0	NaN	0.68	22.0	24.481
208280	657.0	NaN	0.75	26.0	26.643
208281	657.0	NaN	0.65	19.0	23.555

However, as suggested by T. Schraga, the Chlorophyll prediction was very poor.

	Station Number	Discrete_Chlorophyll	Fluorescence	Calculated_Chlorophyll
213931	36.0	3.6	0.43	3.8
213932	36.0	NaN	0.46	4.7
213933	36.0	NaN	0.43	3.9
213934	36.0	4.3	0.46	4.7
213935	36.0	NaN	0.43	3.8
...	...	...	...	...
214426	657.0	10.1	0.33	12.5
214427	657.0	12.0	0.30	11.5
214428	657.0	NaN	0.28	10.7
214429	657.0	NaN	0.30	11.2
214430	657.0	NaN	0.33	12.7

I next separated my "one cruise" data into chunks, based on groups of stations:

- Group 0: South Bay
- Group 1: Central Bay
- Group 2: Golden Gate
- Group 3: San Pablo Bay
- Group 4: Suisun Bay
- Group 5: Lower Sacramento River
- Group 6: Lower Sacramento River

When I fitted the model by station group, the Chlorophyll prediction results were much better.

Chlorophyll prediction for station group 0:

```
station group 0
         Calculated_Chlorophyll  Chlorophyll_prediction
213931                      3.8                   4.700
213932                      4.7                   4.767
213933                      3.9                   4.700
213934                      4.7                   4.767
213935                      3.8                   4.700
...                         ...                     ...
214015                      4.1                   4.432
214016                      4.1                   4.432
214017                      4.0                   4.410
214018                      4.0                   4.410
214019                      4.1                   4.432

[89 rows x 2 columns]
```

Chlorophyll prediction for station group 1:

```
station group 1
         Calculated_Chlorophyll  Chlorophyll_prediction
214020                      3.7                   3.718
214021                      3.9                   3.778
214022                      3.7                   3.718
214023                      3.8                   3.718
214024                      3.6                   3.718
...                         ...                     ...
214107                      4.9                   3.778
214108                      3.9                   3.537
214109                      4.7                   3.718
214110                      4.1                   3.597
214111                      4.8                   3.718

[92 rows x 2 columns]
```

(station groups 2 - 6 not shown)

I tried tweaking the station groups to get better results. This actually got worse results, so I desisted.

For completeness, I also re-fitted the regression model for SPM using station groups. I can't say that this was necessarily an improvement. The results for the entire Bay are very similar.

SPM prediction for station group 0:

```
station group 0
         Calculated_SPM  SPM_prediction
207735            109.0         107.223
207736            156.0         154.597
207737            128.0         126.231
207738             39.0          37.332
207739            134.0         132.372
...                 ...             ...
207818             31.0          29.144
207819             23.0          21.541
207820             35.0          33.823
207821             29.0          27.097
207822             48.0          46.105

[88 rows x 2 columns]
```

SPM prediction for station group 1:

```
station group 1
         Calculated_SPM  SPM_prediction
207823            27.0          28.806
207824            43.0          53.302
207825            32.0          36.243
207826            23.0          23.120
207827            38.0          44.554
...                ...             ...
207916            13.0           8.247
207917            17.0          13.934
207918            14.0           8.685
207919            18.0          16.121
207920            13.0           7.372

[98 rows x 2 columns]
```

(station groups 2 - 6 not shown)

K Nearest Neighbors

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. …

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. Towards-data-science

Features Used

For purposes of this investigation, I chose to examine Salinity, Temperature, and Calculated Chlorophyll, classified by station group. I chose these features because I have examined them previously by station group.

I initially separated the stations into six groups, as before. However, after running the KNN algorithm and viewing the results, I combined groups 5 and 6 into one "Sacramento River" group.

I borrowed the KNN code from the tutorial posted at benalexkeen.com.

Steps:

- Add Station group column to DataFrame.
- Extract data for on cruise (one date), for simplicity.
- Create design matrix X and target vector y (where X contains Salinity, Temperature, and Calculated Chlorophyll and y is Station group.
- Scale X values.
- Split data into **test** and **train** sections.
- Fit a model.
- Test the model.
- Evaluate the accuracy of the model.
- Plot the results.

Predicted accuracy was 0.966. The plot implies that I might get better results if I adjust station groups 2, 3, and 4 slightly, but I think it's pretty good.