1. Introduction

Background.

The consequences of a changing climate are many and far-reaching. One consequence that has gained significant attention over the past decade is ocean deoxygenation, a decrease in levels of dissolved O_2. The impacts of ocean deoxygenation are only now becoming fully apparent: e.g. changes to ocean biogeochemistry, macro- and microorganism death due to hypoxia, and increased ocean production of nitrous oxide (N_2O), a greenhouse gas (Keeling et al., 2010).

Thus, we wanted to answer the questions: what factors predict ocean deoxygenation over time? Given these factors, are we able to accurately predict ocean oxygenation in new data? Oceanic levels of O_2 determine global populations' access to food and other resources, and the ability to predict this information may guide policies by (1) generating potentially-useful information on factors important to preventing or slowing deoxygenation, and (2) demonstrating potential future trends.

We are using data from the California Cooperative Oceanic Fisheries Investigations (CalCOFI), which we downloaded from Kaggle. CalCOFI's data represent the longest oceanographic time series in the world, spanning the years 1949 to 2016. Measured off the coast of California, the dataset includes larval and egg abundance data on 250+ species, as well as oceanographic data such as water temperature, salinity, chlorophyll, and oxygenation.

Our primary outcome for these classification models is dissolved seawater oxygen in µmol/Kg, and whether a given observation is **above/below the median value for our dataset**.. A list of 14 predictors can be found in the Models section.

Data cleaning and preparation.

After downloading and importing the data, we cleaned and prepared it. We restricted the dataset to observations after 2008. The original timeseries was quite long (1949-2016, with 800k+ observations) and would have significantly increased computational effort; this also provides a more current (i.e. approximately prior-decade) interpretation of results. Then, since there was class imbalance in our outcome (above/below median oxygen saturation), we used `caret::downsample()` to randomly sample our data so that all classes have the same frequency as the minority class. Finally, for increased computational efficiency, we took a 20% random sample from the resulting downsampled dataset of ~25k, leaving us with ~5,000 observations.

Next, we removed from consideration features with large amounts of missing data. We had selected a preliminary list of predictors, but several of these had large proportions of data missing. Finally, for the 14 features that remained (see Models for the final list), we removed any missing values for predictors.

After a final dataset was created for analytic purposes, training and testing datasets were created using `caret::createDataPartition()`, with 80 percent of data dedicated to training and 20 percent to testing. The training set was used for all subsequent analyses, including comparison of candidate models, with the testing set only used to evaluate the final chosen model.

Using the training set, we explored outliers for all variables contained in the dataset. While some variables had extremely skewed distributions (e.g. right-skewed phaeophytin, ammonium, and nitrite concentrations), none had outliers visible from boxplots that warranted consideration of correction. Exploratory outlier visualizations are not shown in this document, but code for the boxplots is included in an appendix at the bottom of the submitted RMarkdown file.

```
## # A tibble: 2 x 2
##   outcome     n
##   <fct>   <int>
## 1 Above   32903
## 2 Below   12718
```

```
## # A tibble: 2 x 2
```

```
##   outcome     n
##   <fct>   <int>
## 1 Above   12718
## 2 Below   12718
```

```
## # A tibble: 2 x 2
##   outcome     n
##   <fct>   <int>
## 1 Above    2524
## 2 Below    2563
```

2. Exploratory analysis/visualization

2.a. Correlation.

A correlation matrix was created (see Figure 1) to estimate linear (Pearson's) correlation values among all features. Several features were highly correlated, particularly the feature pairs temperature/potential density, SiO_3/PO_4, and PO_4/potential density. Since correlation amongst predictors can cause problems, we were careful to consider these highly-correlated predictors in both model selection and interpretation.

2.b. Cluster analysis.

[. . .]

3. Models

3.a. Candidate models.

(i) Logistic Regression

[. . .]

3.b. Tuning.

[. . .]

3.c. Final Model.

[. . .]

3.c.i. Test error rate.

3.c.ii. Important variables.

3.c.iii. Interpretation and limitations.

4. Conclusions

4.a. Summary of findings.

. . .

4.b. Limitations of findings.

These data were exclusively gathered in the Pacific Ocean off the coast of California; thus, these prediction models and any interpretations may not be generalizable to data outside of this geographic area. Additionally, we limited analyses to data from 2008-2016, which may limit temporal interpretation.

References

Keeling RF, Kortzinger A, Gruber N (2010). Ocean Deoxygenation in a Warming World. Annu Rev Marine Sci, 2:463–93.

Watson AJ, Lenton TM, Mills BJW (2017). Ocean deoxygenation, the global phosphorus cycle and the possibility of human-caused large-scale ocean anoxia. Philos Trans A Math Phys Eng Sci, 375(2102).

Appendix: Figures & Tables

Figure 1. Correlation matrix of features