Final Report

Predicting Chemical Species with Chemical Sensor Data

Contributors: Samuel Scovronski

Introduction

Data Source:

The data for this project was provided by a group of researchers working in the BioCircuits Laboratory at the University of California San Diego. The data consists of over 13,000 experimental runs conducted on a chemical sensor array and recording their electrical signals produced by exposing them to various gases at varying concentrations.

Who cares about this data and why?

This dataset is of particular interest to companies that use industrial processes that alter the chemistry of their products. For example, pharmaceutical, semiconductor, and chemical producer companies care about having the right chemical and concentration to ensure that the desired chemical reaction occurs at the required rate. Hence, their product has consistent quality and performs as intended without harming the end-user. However, most of these companies still rely on human intervention to select the chemical and its concentration. A sensor array system capable of identifying the current chemical being used and its concentration could act as a backup quality/safety protocol.

Problem Statement

What researchers/ suppliers care about:

Researchers and potential suppliers of sensors like these want to spend as little time as possible performing product validation and calibration. Performing experimental runs to create a valid model that works in the end user's system is financially and temporally costly. Their goal is to create a model capable of determining which chemical is being used and its concentration with high accuracy in as few experimental runs as possible

What Industries care about

Companies that create products by altering the chemistry of raw materials with industrial processes care about two aspects of a chemical analysis sensor array: how will it impact the bottom line, and how accurate is it? The factors that determine the bottom line are implementation costs and potential savings (improved product consistency or reduced product liability). The potential savings a company could realize is circumstantial, so it is difficult to provide a value. However, there is a clear relationship between implementation costs and the number of sensors used since individual chemical sensors can reach up to $5,500. The accuracy is important because the company will want to realize the gains before the sensors fail, and they need to be able to trust that it is working as intended.

Data Wrangling

The dataset provided by the BioCircuits Lab contained 13,910 rows and 129 columns in a dtl format from experimental runs tested in batches. It includes 128 predictor features extracted from the sensor electrical signals and two target variables: the processed chemical code and the concentration of that chemical species. Below are some of the problems that were found with the dataset and actions taken to address the problems

Problem 1: The dataset had a lot of outliers if we use the standard definition of an outlier being greater than 1.5 times the interquartile range (IQR) outside of the IQR. In fact, 35.3% of the experimental runs had at least one predictor feature that was deemed an outlier

Solution: Although there were a lot of outliers following the standard definition, it was found that the values were most likely not erroneous, which is what we are looking for. All experimental runs with an outlier were kept as part of the dataset.

Problem 2: With 128 predictor features, the dimensionality of the data is large, and it is difficult to visualize the interactions between features.

Solution: Because all dimensions are useful and cannot be reasonably deleted, algorithmic methods (i.e. PCA, elimination of highly correlated features, or simple models with a model complexity penalty) are needed. These various methods will be used as part of the data pipeline.

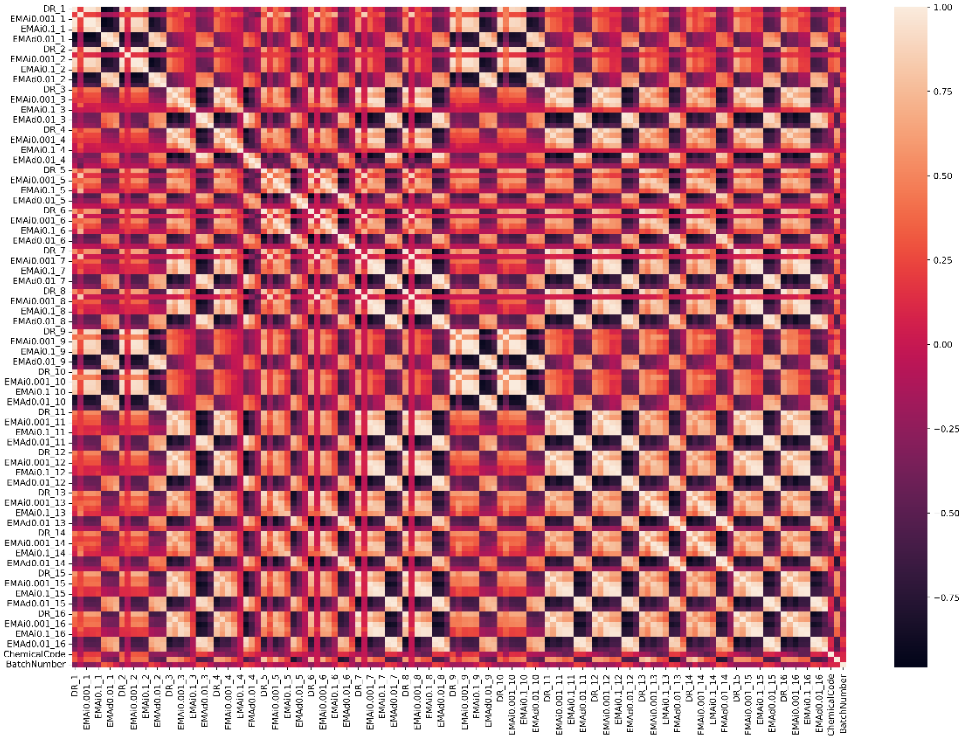
Problem 3: The predictor features have different units and significantly different scales, which could cause our model to over-emphasize the features with a larger scale.

Solution: All features will be standardized such that the mean will be 0 and the standard deviation will be 1.

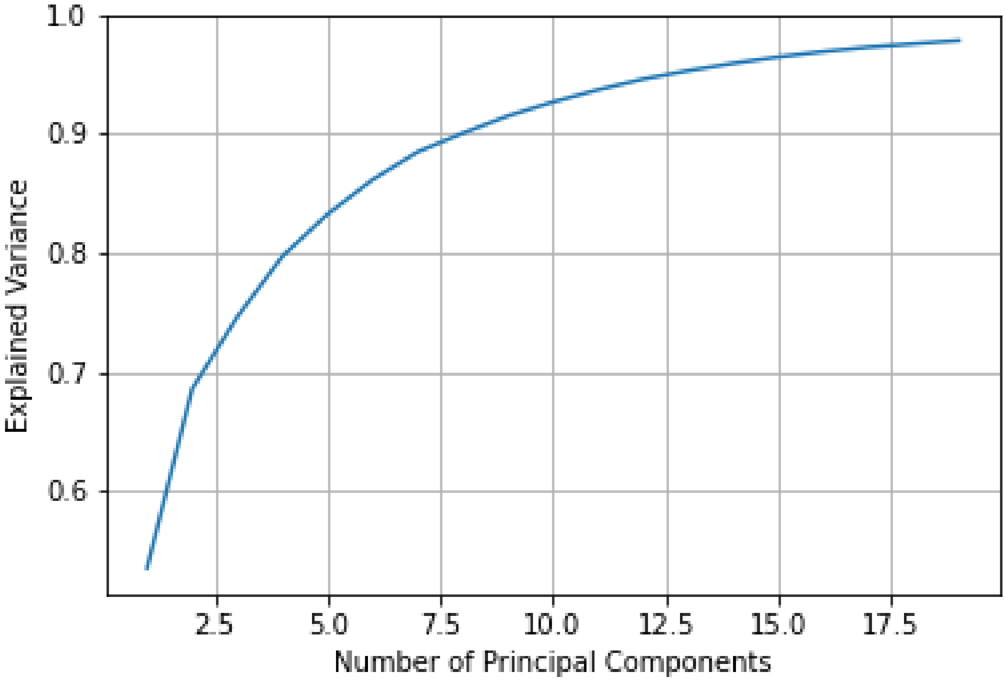
Exploratory Data Analysis

During EDA, four main things were observed: A high correlation among the predictor features. The number of dimensions can be reduced dramatically while retaining most of the explained variability. The predictor features displayed high skewness and the association of outliers with concentration levels.

This heat map shows the correlation between all features. The scale indicates that very dark and light regions indicate high correlation (either positive or negative), while the purple, orange, and red regions indicate lower correlation. As can be seen, large portions of the heat map show a strong correlation. However, no predictor features are highly correlated to the chemical code, and only three predictor features are correlated to the chemical concentration.

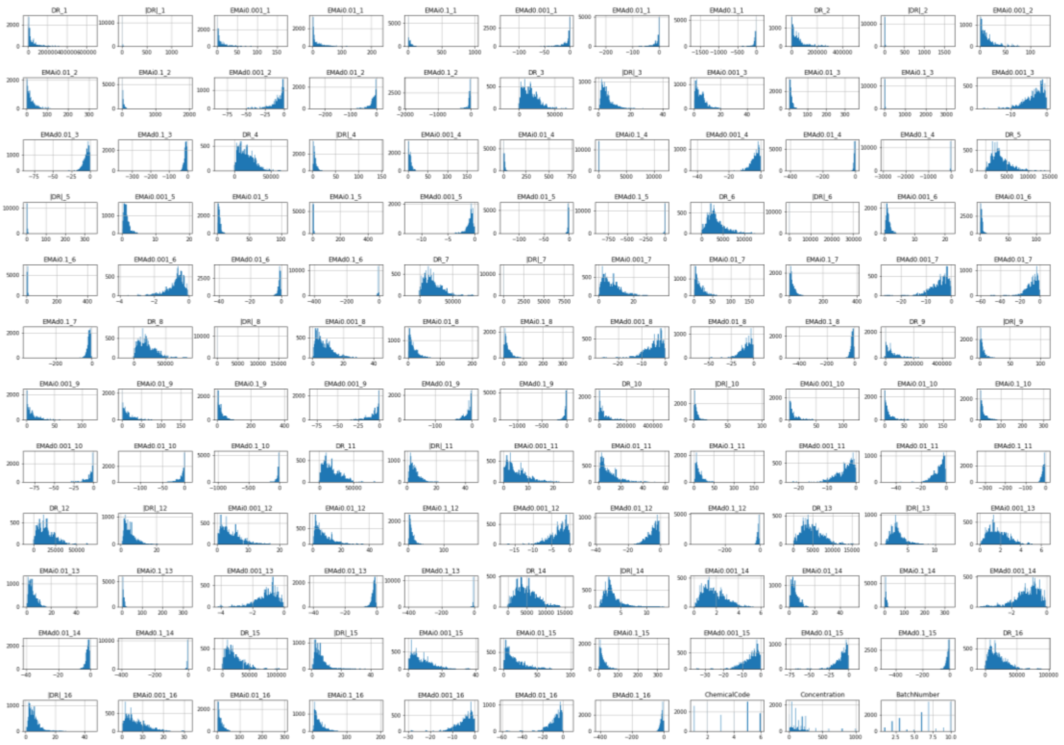


Because there is a high correlation among features and we have high dimensionality, the dataset is a good candidate for dimensional reduction techniques such as principal component analysis (PCA). The number of dimensions can be dramatically reduced by looking at the tradeoff between the number of components and the percent of the variance in the original dataset that is still explained with these components. With just eight features, 90% of the variance can be explained, and 98% can be explained with 18 features.



Below is a grid of the histogram curves for each feature, in which we can see just that there are large white spaces with a couple of tall bars towards one side indicating high kurtosis values. Overall, 19 predictor features had a Kurtosis value greater than 1000, with the highest value being 13,909 for feature |DR|_6.

It was found that outliers were associated with higher concentration runs, certain chemicals, and specific experimental batches using bootstrap hypothesis testing. Outliers are now defined as greater than five standard deviations away from the mean. For all concentrations 300 ppmv or greater, the proportion of the number of runs deemed as outliers was significantly different than the proportion of the runs from the population that had a concentration of that magnitude. The outliers were also associated with experimental runs flowing chemical number 5, Acetone. Lastly, outliers were also associated with batch numbers 1, 10, and 4. At this point, interactions or confounding of these associations have not been tested.



Modeling Setup & Results

Model Pipeline

The data was split into three groups to address potential companies' concerns about accuracy: a training set, a test set, and a future set. The training and test set values were randomly chosen from the data acquired in the first five batches. In contrast, the future group consisted of all experimental runs after batch five, so we could determine the long-term model accuracy. A function was created to automatically split the data into these three groups based on test size and feature sets as inputs. This allowed us to create multiple models of various training sizes (to determine the minimum number of experimental runs required to create an accurate model) and various feature sets (to determine the minimum number of sensors needed. Due to the number of sensors used, every combination of sensors was not evaluated. Instead, a basic Decision Tree model was created to determine which features were the most important, and then the sensors used in that list were combined in various ways.

Baseline Model

Before setting off to create complex models, a couple of dumb or simplistic models were created first. Two models were made: random selection of chemical code and most frequent chemical code. The random selection model had an accuracy of 15% on the test data, which is only slightly off from the theoretical of 16.7%. The frequentist model performed better with an accuracy of 28% on the test data, so this will be the baseline score to beat with the created model.

Model Selection

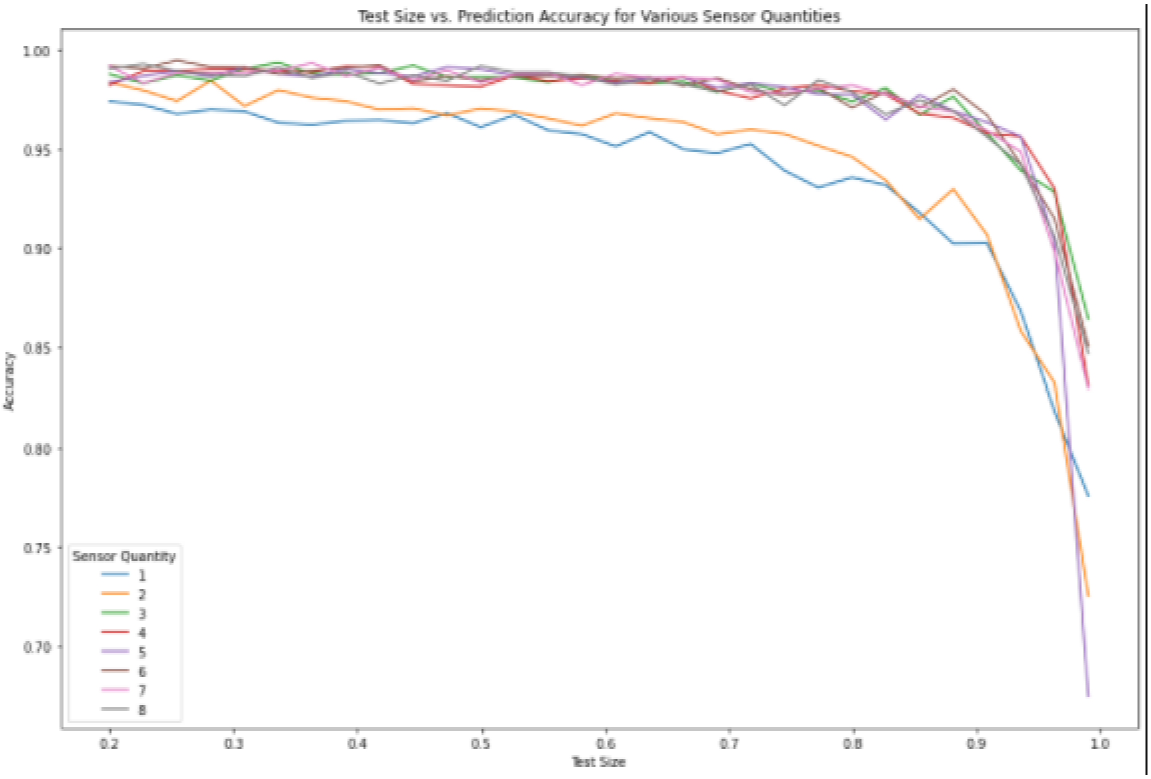
There are many classifier algorithms available for us to use, all with pros and cons. A cross-fold evaluation was performed with the same dataset to evaluate which models would work better for the dataset. Three algorithms were selected to continue evaluating (See results in Table 1). Linear SVC, K Nearest Neighbors, and Random Forest. Next, these three algorithms were tested again except using various test sizes and features sets to see which model had the best accuracy among all tests. It should be noted here that the Linear SVC models had difficulty converging, so the accuracy values are to be taken with a grain of salt. The Random Forest classifier algorithm performed the best, so it was chosen as the model type to address the concerns of both parties mentioned in the problem statement section (suppliers and customers).

Table 1: Model Comparison

Model Type	Fit Time	Score Time	Accuracy
Decision Tree	0.2933	0.0017	0.966
Random Forest	1.6627	0.0232	0.989
kNN	0.0848	0.1237	0.978
Linear SVC	0.6342	0.0055	0.983
Gaussian Naïve Bayes	0.0093	0.0032	0.803

Test Size and Sensor Quantity Tradeoff

A tradeoff table was created by creating models for every sensor quantity and test size combination. The table allows for the selection of a minimum desired accuracy and determining which mix of test size and sensor quantity is capable of delivering those results. The first thing to note is using less than three sensors reduces the model accuracy. The second thing to note is that model accuracy stays relatively stable until a test size greater than 0.8 is used (only 727 experimental runs are used to create the models). Based on these results, two potential solutions balance the needs of the suppliers and consumers while achieving a test accuracy of >99%: 3 sensors requiring 2,400 experimental runs or five sensors requiring 1,820 experimental runs.



Long Term Accuracy Assessment

Using a model with five sensors and a test size of 0.5, the long-term performance of the model can now be evaluated using the third bucket of data we kept aside. Quickly, it is seen that the model does not continue to perform well. Even though the model metrics are great for the test data with an overall accuracy of 99% (Table 2), the metrics for the ongoing data are only slightly better than our baseline model with an overall accuracy of 40% (Table 3). At this point, it is unknown whether or not the inaccuracy is due to poor model creation or a change within the testing system.

Table 2: Test Performance

Chemical Code	Precision	Recall	F1 score	Quantity
1	.99	.99	.99	356
2	.99	1	1	503
3	.99	.99	.99	215
4	.98	.97	.98	228
5	.99	.99	.99	476
6	.97	1	.99	39

Table 3: Future Performance

Chemical Code	Precision	Recall	F1 score	Quantity
1	.49	.36	.41	1854
2	.61	.71	.65	1921
3	.18	.51	.26	1210
4	.16	.2	.18	1481
5	.93	.58	.71	2057
6	0	0	0	1754

Future Improvements

Sadly, there was not enough time to create a model to predict the chemical concentration. This prediction would also be critical to potential customers looking to use the sensors for quality assurance.

Further investigation into why model accuracy is excellent for the first five batches, even when the training set is minimal but then works poorly on the remainder of the experimental data, needs to be investigated further. Specifically, determine if the lower accuracy is due to the model or external factors such as sensor reading errors or test system changes. We could also explore compensating methods if the inaccuracy is due to sensor drift/ read errors.

Citations

A Vergara, S Vembu, T Ayhan, M Ryan, M Homer, R Huerta. "Chemical gas sensor drift compensation using classifier ensembles." Sensors and Actuators B: Chemical 166 (2012): 320-329. https://www.researchgate.net/publication/216301619_Gas_sensor_drift_mitigation_using_classifier_ensembles

I Rodriguez-Lujan, J Fonollosa, A Vergara, M Homer, R Huerta. "On the calibration of sensor arrays for pattern recognition using the minimal number of experiments." Chemometrics and Intelligent Laboratory Systems 130 (2014): 123-134.

MKS. “Residual Gas Analyzer 1-200 amu, Faraday Detector, Open Ion Source.” Source: https://www.newport.com/p/EV2-210-000FT?xcid=goog-pla-EV2-210-000FT