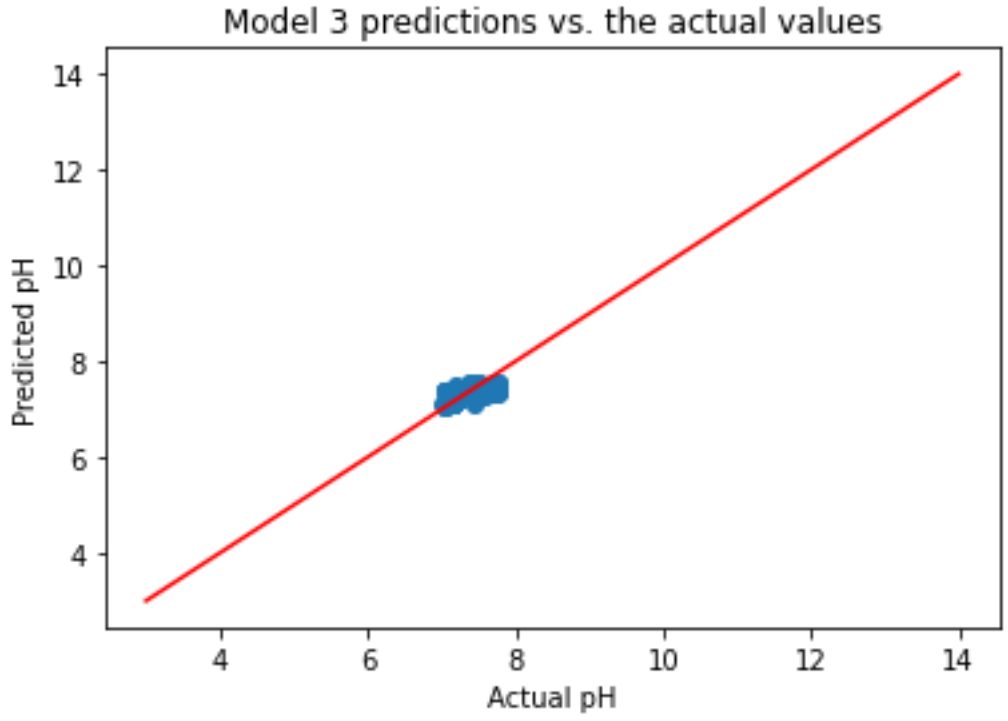
PoolPal Modeling Proposal

For the current modeling phase of this project, we went through a few different models with the goal to make a model which could predict pH with a reasonable amount of accuracy. The hope was to understand the data and what may be feasible. What we discovered was that although we are able to predict pH, it could be challenging to create a forecasting model which predicts pH changes given other sensor data. Please see below our findings with the various models we tested. Ultimately, we have discovered that an LightGBM model works best for predicting pH given other sensor data and propose the potential exploration into utilizing time series regression modeling for future development to create some predictive power.

We will go through each model and share the metrics at the end as well. Generally speaking, we had constrained the model to focus on a smaller pH range to remove bad sensor data and we selected features which might be helpful. For most models we decided on a range between 7pH and 8pH. We also removed features that were deemed obsolete based on the features document and also removed features with collinearity (as some features were calculated based on others).

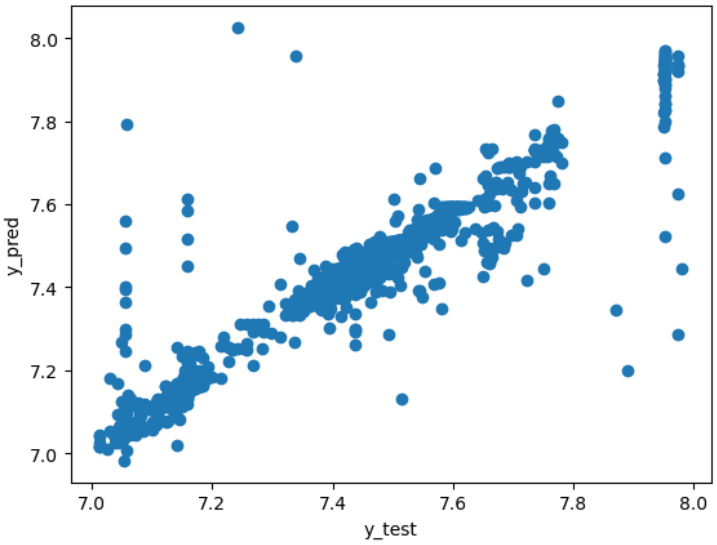
Linear Regression

We found that Linear Regression struggles when trying to model out the data. We constrained the data to only those points where the pH is within 7.2-7.8. The RMSE was 0.042 and the R^2 was 0.542. Observing the chart where we compared predicted pH to actual pH, there is a tight cluster. The features used in this model were salinity, turbidity, ORP, TDS, pressure in/out, pump current, fresh water, drain, human counter, temperature, and water level. Overall, a strict OLS model did not perform well.



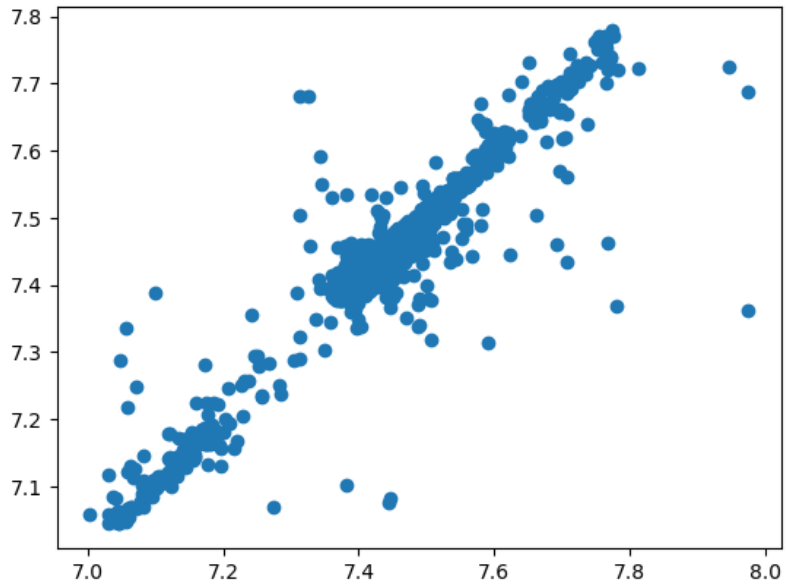
XGBoost

XGBoost performed somewhat better with the data. WIth this we constrained it from 7.2 pH to 8pH. The MSE was around 0.001 with an R^2 of approximately 0.90. The features selected were salinity, turbidity, ORP, TDS, Pressure in/out, pump current, human counter, temperature, and water level. We note that while for most of the pH range we get a reasonable correlation, at pH above 7.8 the model struggles. This may be due to outliers leading to an elevated pH or sensor data not being clear.

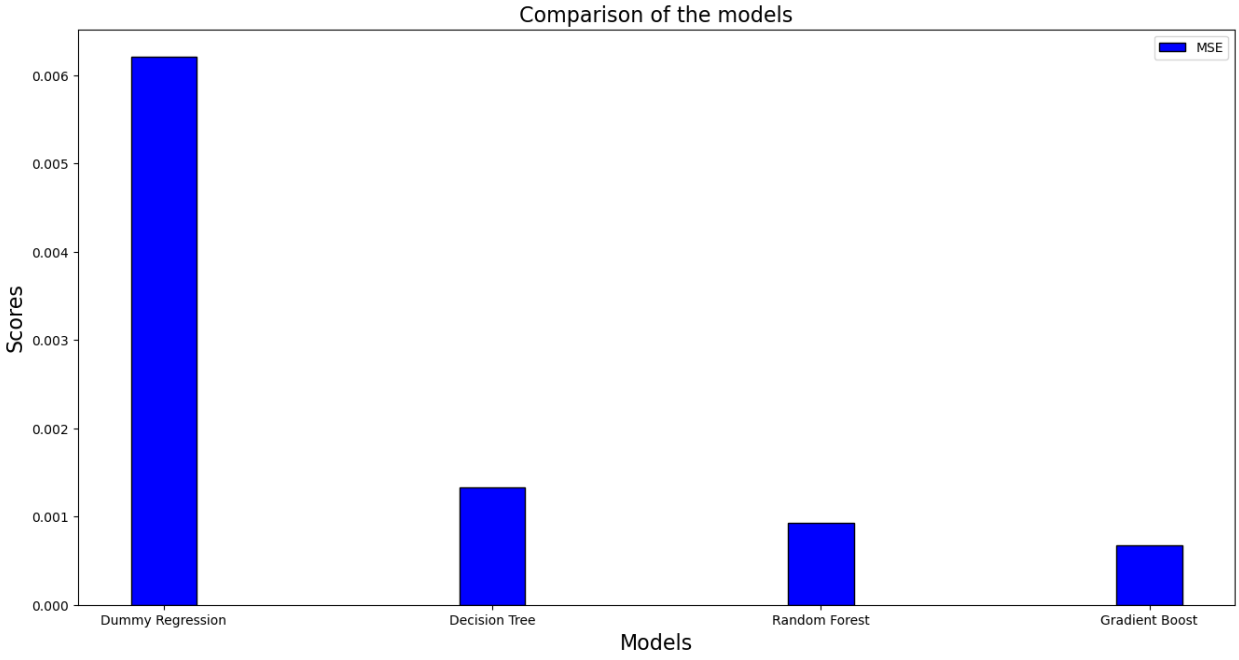


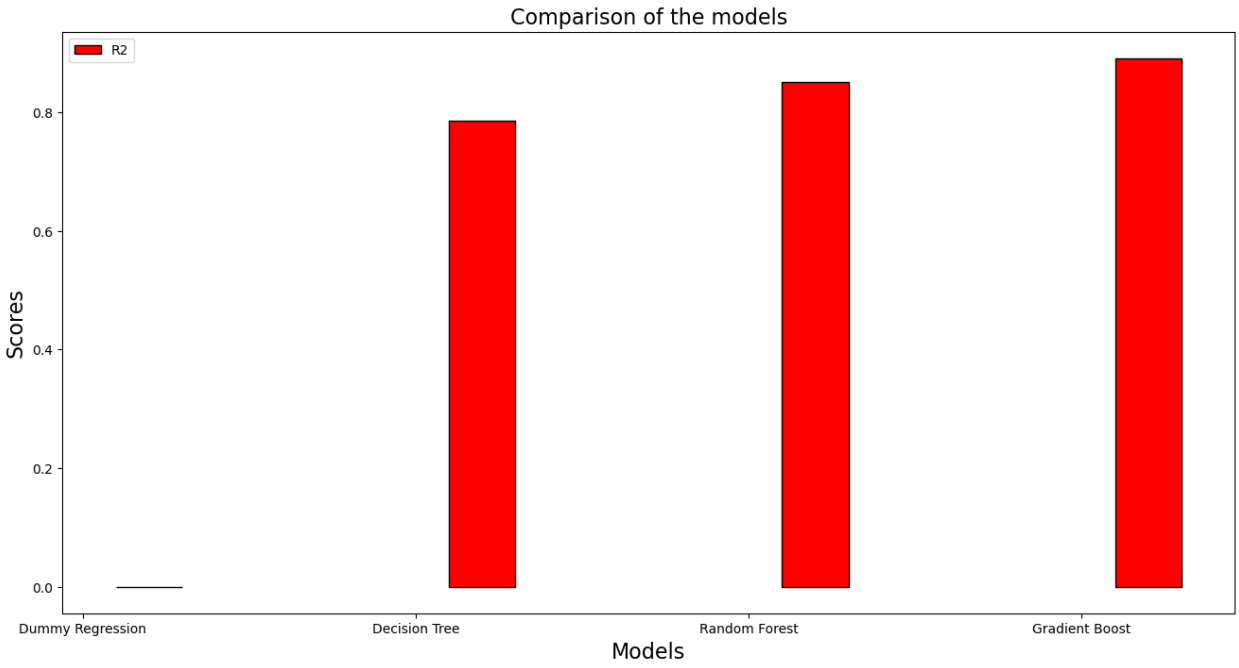
LightGBM

With LightGBM we have similar results to XGBoost with a slight edge. The R^2 was 0.92 with a MSE of 0.0004. The features chosen were turbidity, ORP, TDS, pressure in/out, pump current, human counter, and water level. Of the three, this model performed the best.

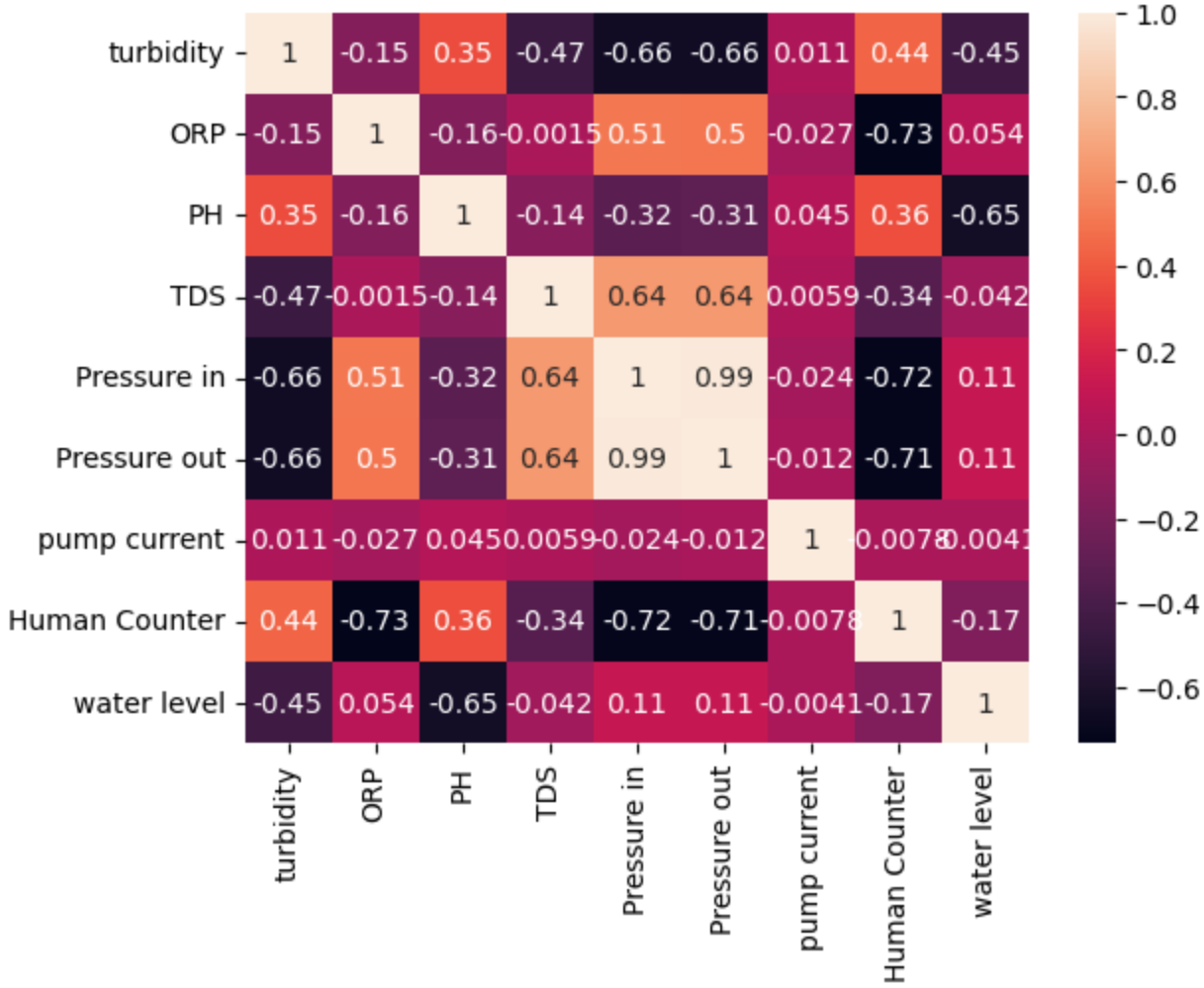


Please note, that the results of these models are for their ability to predict pH given a set of sensor features and relies heavily on properly functioning sensors. Please refer to the following two images for a comparison of model performance.





At this time, our ability to forecast pH into the future is not possible. Please refer to the heatmap below of the non-constant features to understand the correlation between features.



We propose that ANTLER focuses on an ensemble based model which can predict pH based on other sensor data. The next step is to explore the possibility of creating a time series regression model which uses the currently existing data features to predict pH trends. Something that may be helpful is having timestamps as to when chemicals may be added to the pool at this time which can cause some of the changes in things like TDS and turbidity. This way we can connect any future modeling with real-world actions.