PPG Paints Student Analysis

Machine Learning Study of Paint Qualities and Popularity

Exploratory Data Analysis

Intro

For my final project in my undergraduate Introduction to Machine Learning class at the University of Pittsburgh, I was given the opportunity to stretch my newly acquired legs in machine learning concepts with real data from PPG.

In this presentation, I'll go over the general process of exploring and analyzing the data I was given, the regression and classification models I fit and trained, and some of the conclusions I came to about how the inputs related to the predicted output.

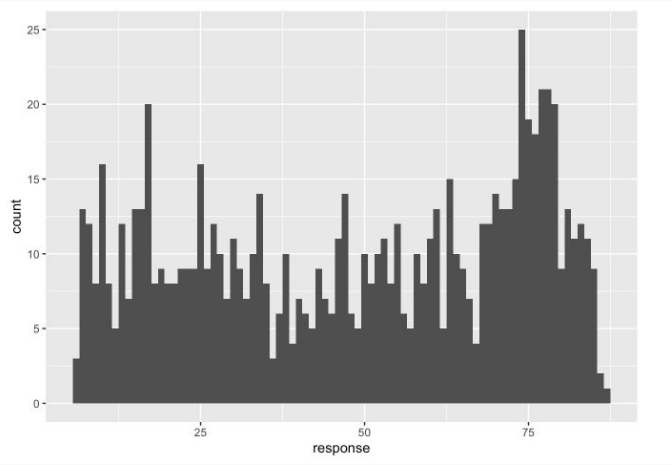
Data Overview

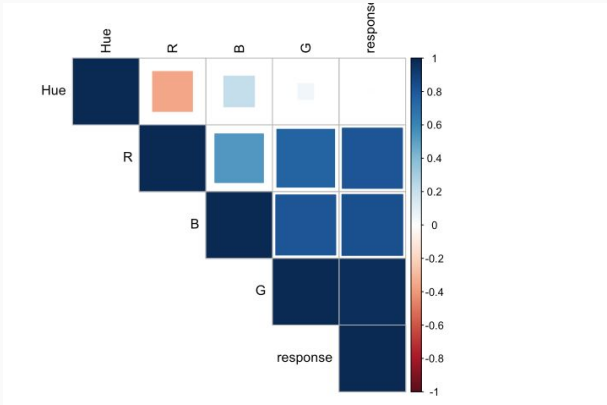
The project specifications provided us with a training set where each entry described two models of color (RGB and HSL formats) representing a paint color, as well as two outputs associated with that color. One for popularity (called `outcome` in the data), and another for some property of the paint that we were intentionally kept blind to (called `response`).

It was our job to fit models which would accurately predict these values. `response` was bounded from 0 - 100, and `outcome` was a binary value, either 0, or 1, meaning popular or unpopular.

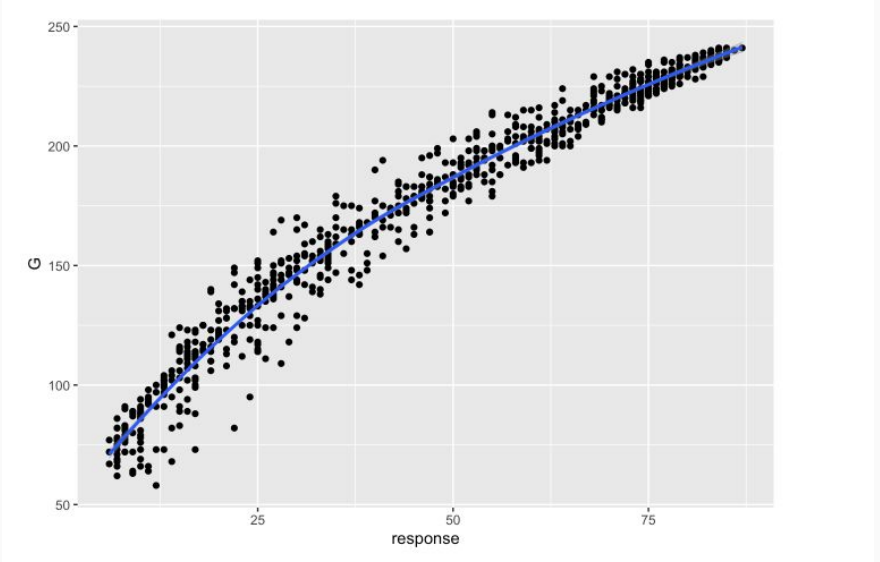
To do this, I trained a regression model on `response` and a classification model on `outcome`. But first, I took a look at all of the data we had access to by modeling it in different visual plots.

Exploratory Data Analysis: Visualizations





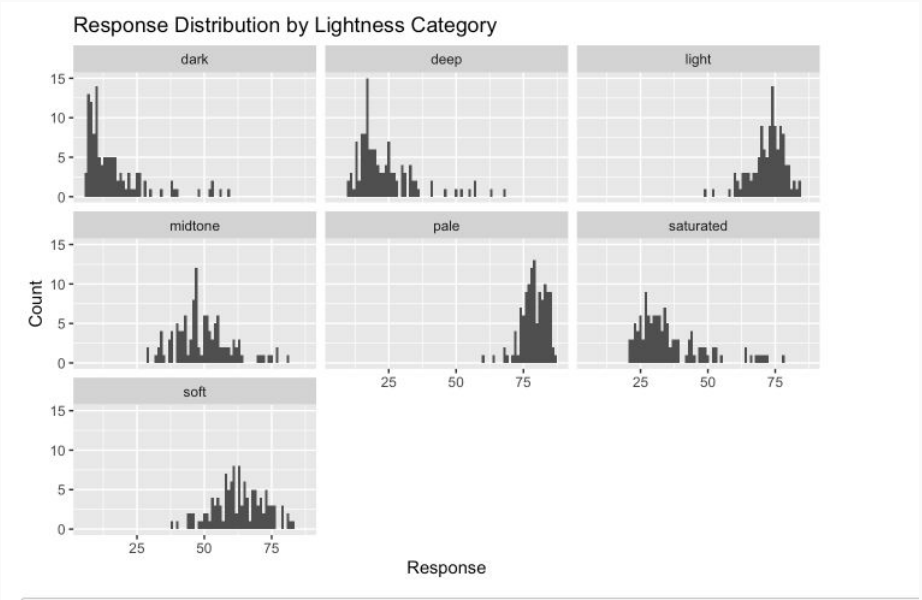
I took a look at the distribution of the response (left), all of the color variables, and several of the correlations between the variables. On the initial exploration of the data, I found a correlation between the value of `G` (for green) and the `response`. You can see that best in the correlation plot in the lower left, which measures correlation, and in the trend line just below.



Exploratory Data Analysis: Lightness

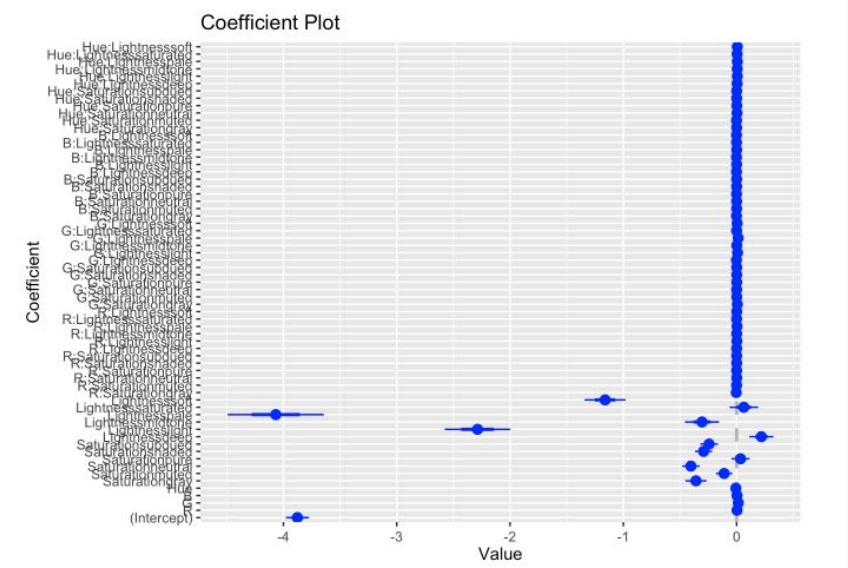
What I ultimately became more interested in was, however, the correlation between certain responses and certain shades of **light**. There were two variables that were categorical, `Saturation` and `Lightness`, which could be used to facet responses into different categories.

I noticed early on that Lightness seemed to play some role in predicting what kind of value `response` returned.



Regression

Regression: Models to Consider



For linear regression, we fit models by taking the coefficients and plug them into a number of different multiplicative or additive combinations and see how effective those combinations are. We can measure how effective they are through a variety of different metrics.

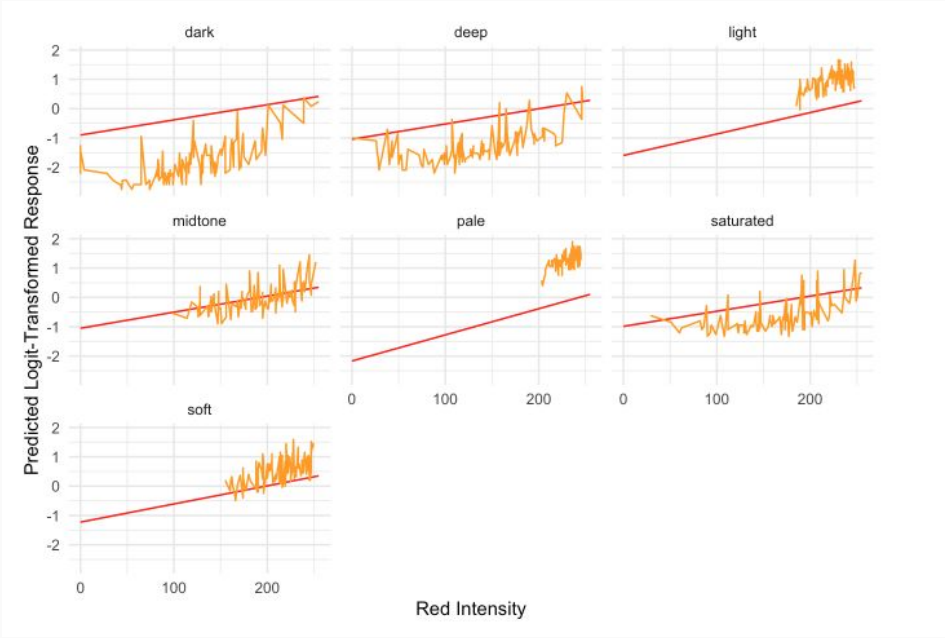
For my regression problem, I used RMSE, a typical metric for a problem like this, but others do exist, and there are pros and cons to each.

To the left, you can see a model that has an extremely high number of factors involved in the calculation, all manipulations on the existing data, and then you can see which variables played the biggest role in impacting the response by seeing how they break away from the origin. These are ‘statistically significant’.

Regression: Trends and Fit

I fit various models throughout the Regression section of my project. The visualization to the right shows one instance where, in each shade of `Lightness`, I tried to check and see if there was any kind of trend with the predicted response along `R` (Red) and the averaged out values of the others.

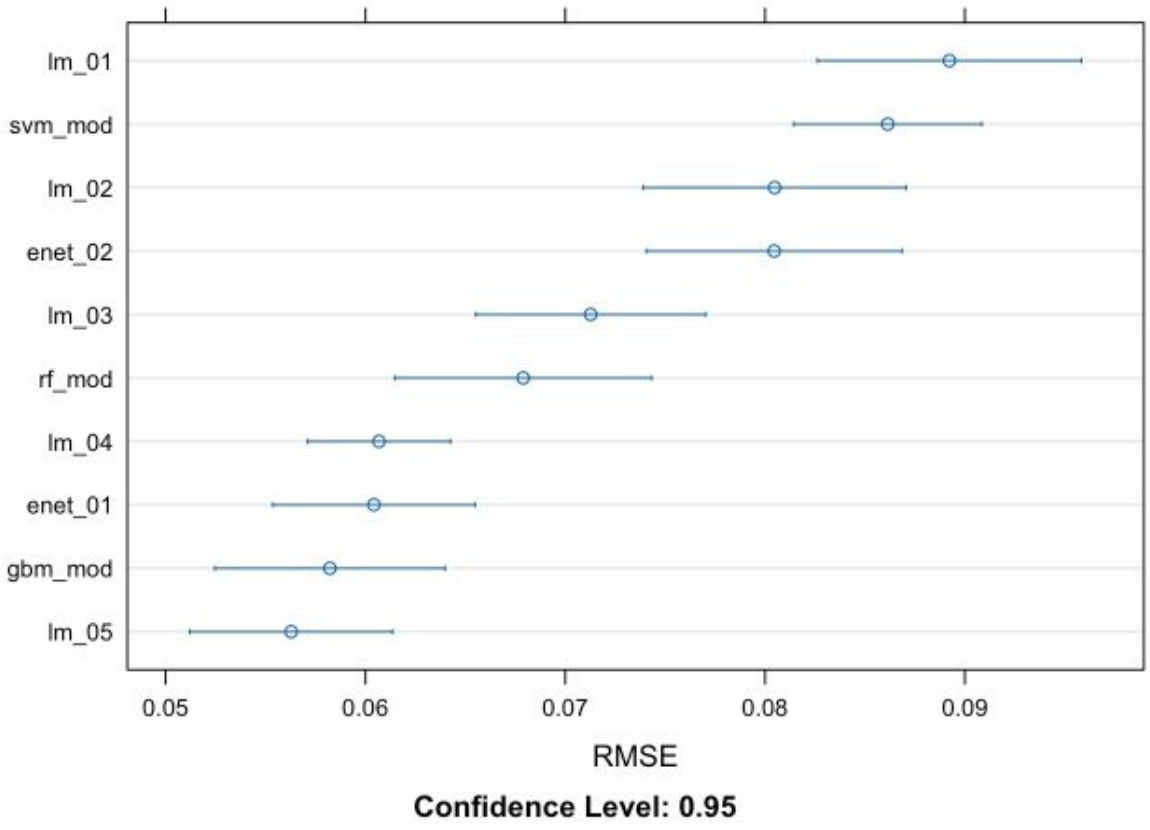
This was a promising start. It shows that the model I tested here was fitting the “trend” of the actual response (in orange).



Regression: Model Selection

By the end of the regression selection, I had trained several different kinds of models (listed on the left) and evaluated their performance against one another. This visualization shows the general process of doing that.

The model I selected was a linear model.



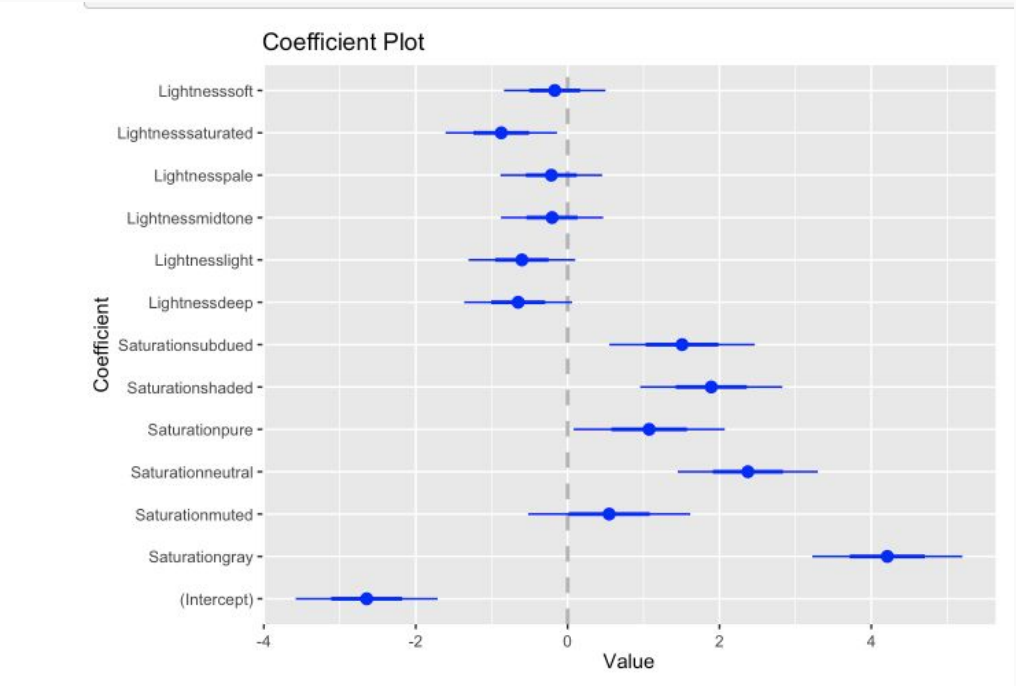
Binary Classification

Classification: Lightness

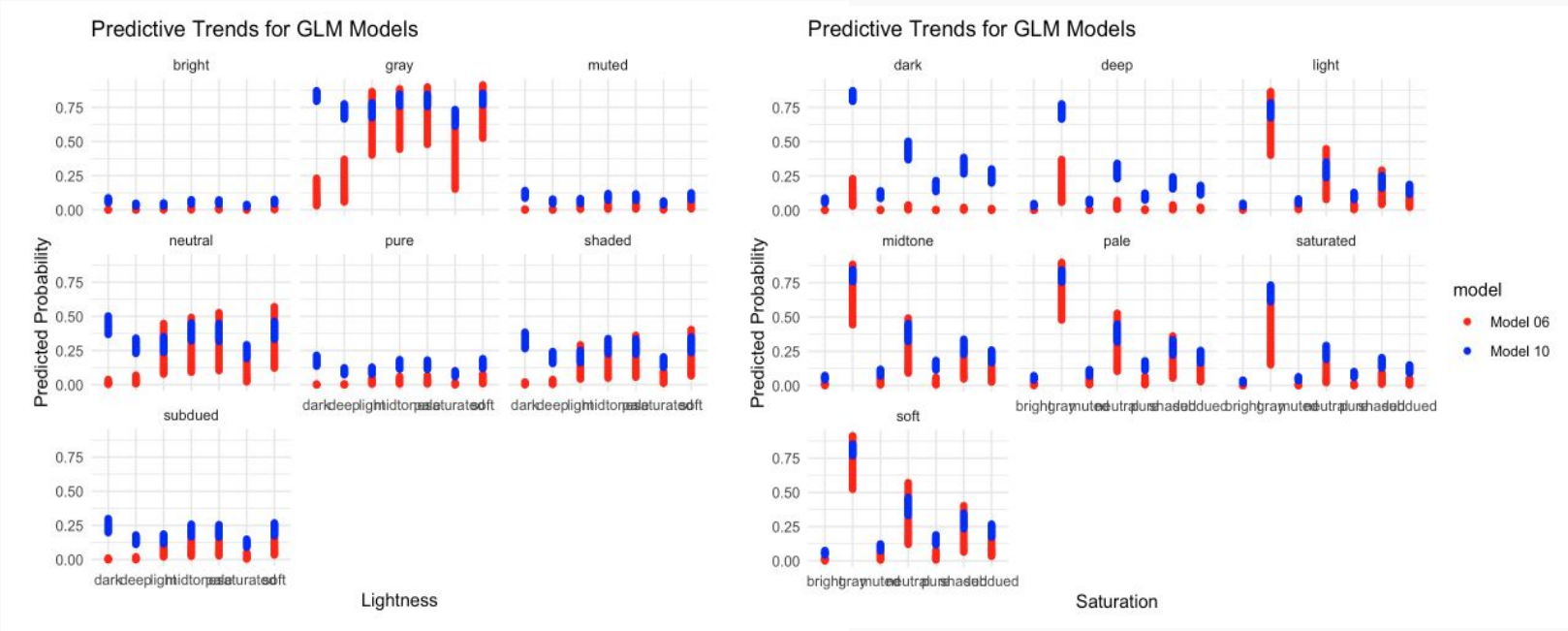
For classification, I had to predict *probabilities* on a binary value, the probability representing the odds of `outcome` being a 0 or a 1 – an event or non-event, popular or unpopular.

What I noticed quite early on was that the best performing models in both the Regression and Classification section were primarily altering and manipulating the Lightness and Saturation values.

This was true for Regression, as seen in the coefficient plot back a few slides, and it remained true here (see the figure to the right), which lined up nicely with what I noticed earlier on in the data exploration.



Classification: Lightness (cont.)



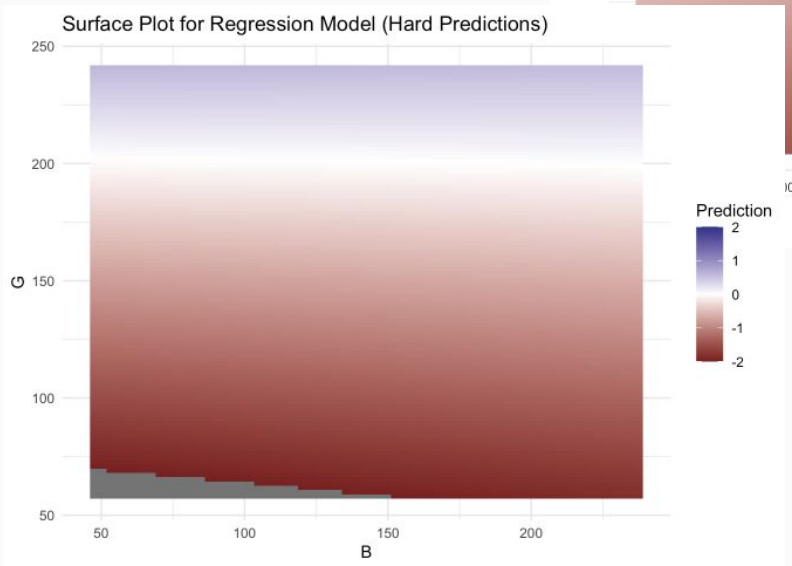
One of the more interesting visuals that I found regarding the categorical values was here. When it came time to predict the probabilities of `outcome` being popular or unpopular, I noticed that when we faceted the values with each other twice, you could see the most impactful values of `Lightness` “jumping” out of the saturation plots. It’s clear that across any value of `Saturation`, the `gray` value of `Lightness` is more heavily correlated with an `outcome`.

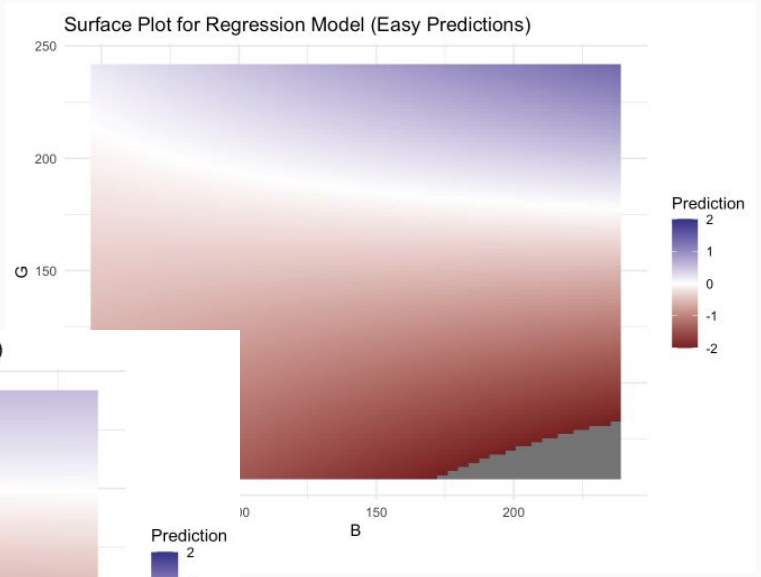
Interpretation

Interpreting

In the final section, we were tasked to observe the hardest-to-predict values of `Lightness` and `Saturation` and see what models had to say about the `response` when faceted across some of the continuous inputs.

I admitted struggled to implement and understand the visualizations here, but what stood out to me for the `regression` plot was that the model seemed to have more opinions about the easy predictions, while it didn't seem to know what to do with the G and B values in the harder-to-predict values of `Saturation` and `Lightness`.



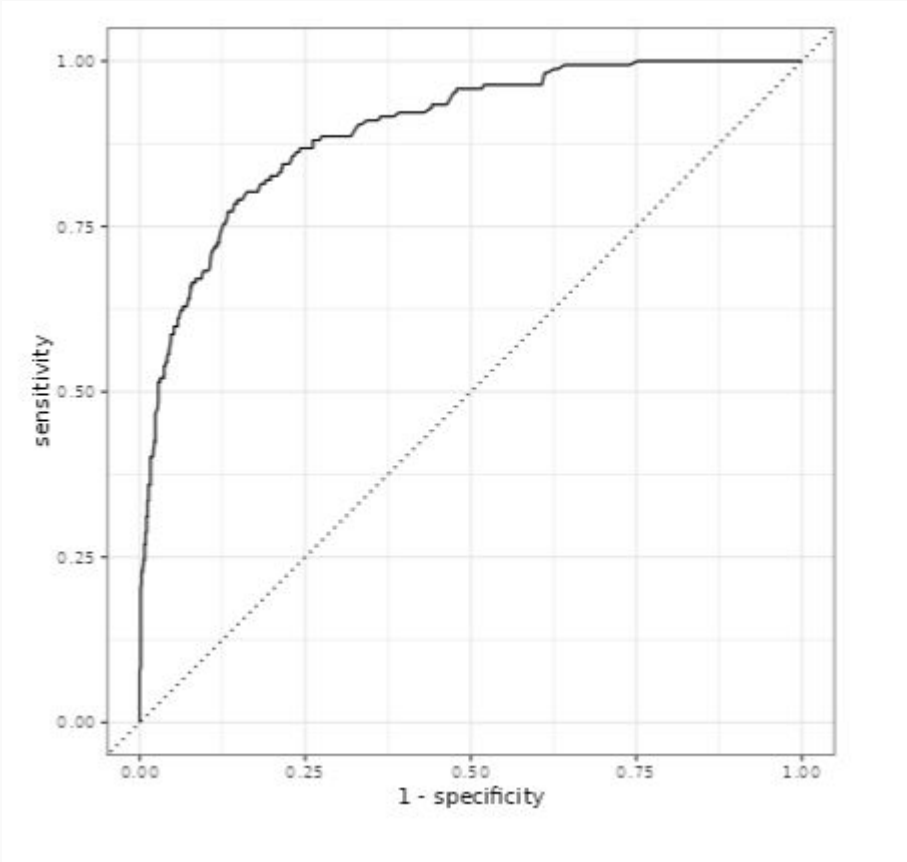


Scoring

As a part of our project, we had to take the models that we fit and test them against some more data provided by PPG and upload them to PPG's website to test out the values and get back a list of metrics, telling us how well our models performed.

In terms of regression, the models I plugged into the website did very well.

Something strange happened with my score in terms of the classification metrics that I was never able to figure out before the deadline – however, I did get to see an effective ROC AUC curve for my model, which did at least demonstrate that it was working properly.



Conclusion

All in all, it was an extremely interesting opportunity to use R and some basic principles of machine learning to create, fit, and test models on real data with an actual purpose.

It would be my unlikely hope that some of these observations are useful to PPG. Mainly, the observations around `Lightness`, and how that affected *both* the `outcome` (popularity) and `response` more than seemingly any other variable.

Thanks!

Brendan Szewczyk

hey@brendan.ms