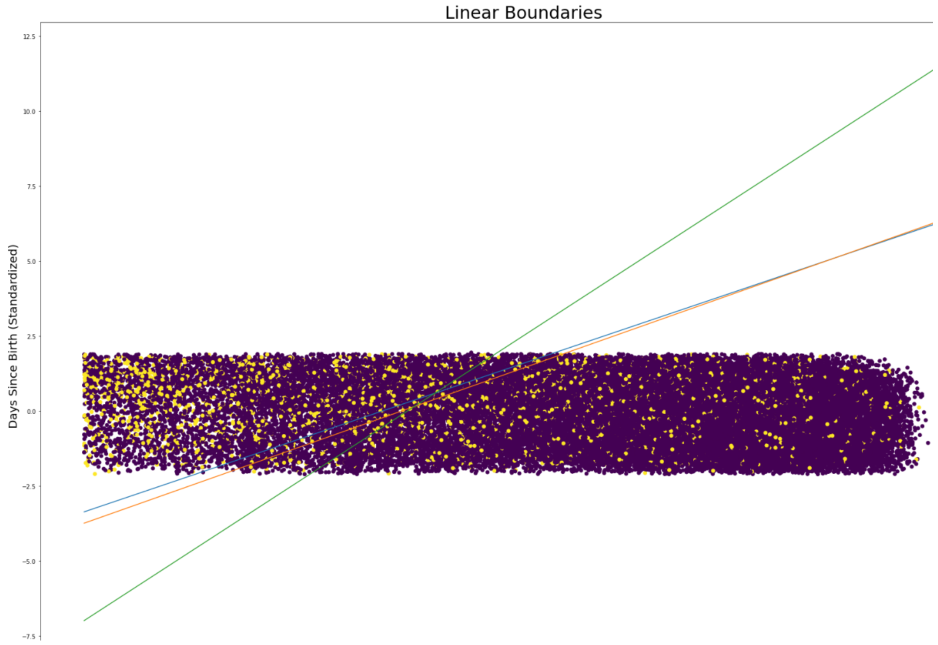
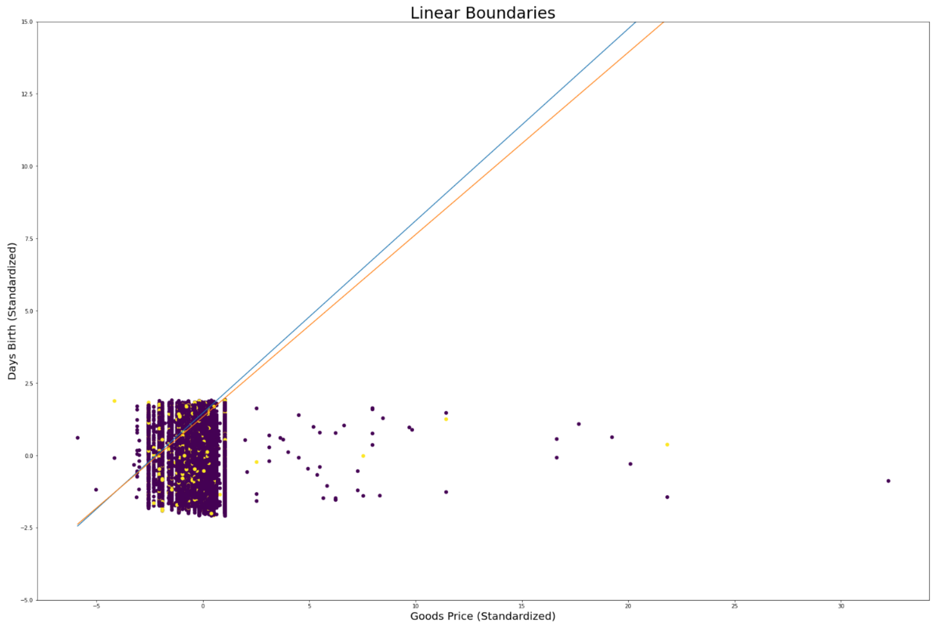
Model Comparison

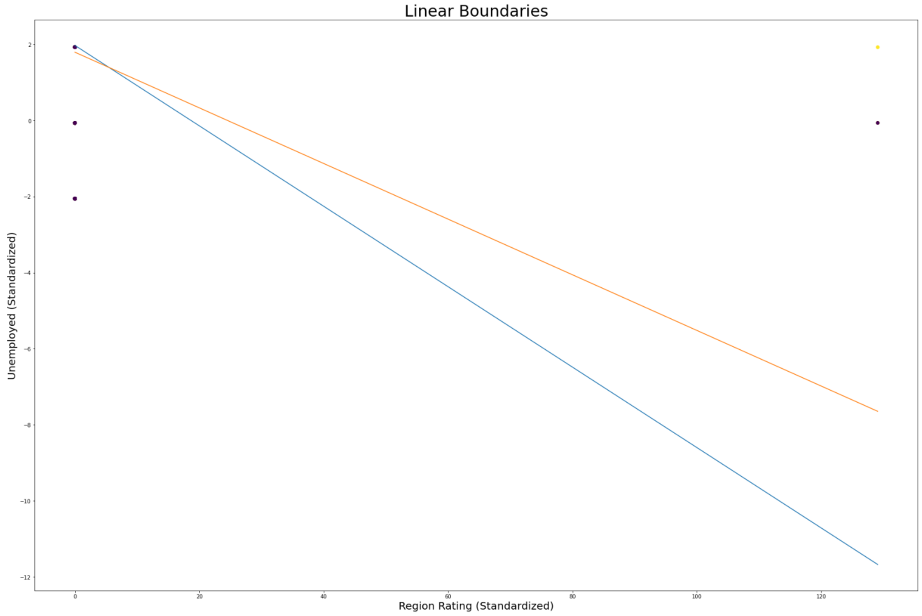
For this project we implemented classification models using logistic regression, support vector machines and a perceptron (with a ridge penalty). For logistic regression, we give it an input of data and get an output between values of 0 to 1. The value outputted from the model gives us the probability - the closer that probability is to 0 or 1, the more confidence the model has that it is classified in that respective group. SVMs mainly find the best margin between two groups of data - the output we get from our linear function and depending on its value between 1 and -1, we assign it to the respective class. Based on these class assignments, the model learns the best way to "separate" this data into two distinct groups. For perceptrons, we aggregate a sum of the product of our data and some determined weights to get an output. This output then is compared to a threshold value and classified as 1 or 0 based on such.

We chose some combinations of two variables so we could visualize the linear separator for each model. We can see from the plots below that Perceptron and Logistic Regression will tend to give similar predictions. Their goal seems to be similar, in addition to putting points on the right side of the boundary, they want to maximize the "distance" as well. However, SVM also wants to maximize the margins, which leads it to have a different line.

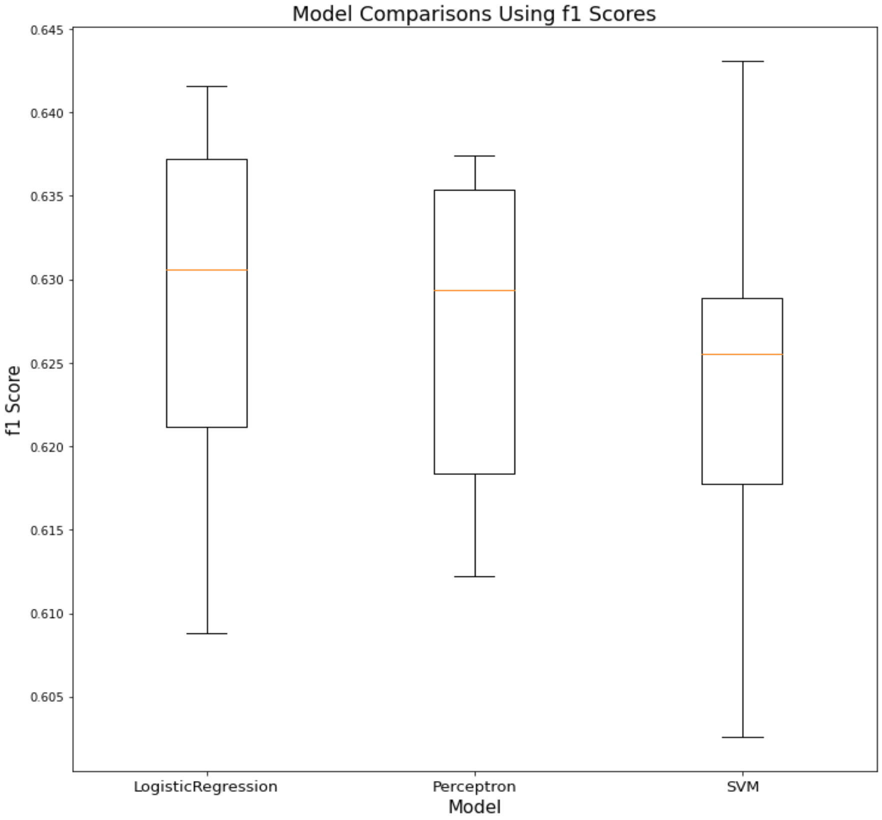




In fact SVM failed many times despite class weights because of how messy the data was. Its goal of maximizing boundaries made it pick a line way out there off the graph that put all points on one side.



This plot below reports the distributions of f1 scores coming from each fold for k-fold stratified cross-validation scores for each model. It appears, while SVM has the greatest range of scores, in general, logistic regression may have the best overall f1 score. Although, there is quite a bit of overlap in the distributions for all three models, so it is tough to say whether any one has the advantage over the others.



Below are the different F1 scores and Accuracy for the final three models each for class weights and for our own implementations using balanced sub-sampling.

F1 Scores	Logistic Regression	SVM	Perceptron
Sklearn	0.221953755	0.223485988	0.22860202
Our Implementation	0.397007	0.393098	0.435143

Accuracy	Logistic Regression	SVM	Perceptron
Sklearn	0.797	N/A	0.788
Our Implementation	0.755	N/A	0.790

Based on both f1 scores and accuracy for our own implementations the Perceptron produced the best predictions. In contrast, according to sklearn implementations the Perceptron still had the best f1 score, although much worse then what was according to our implementation, but the Logistic Regression had the best accuracy score.