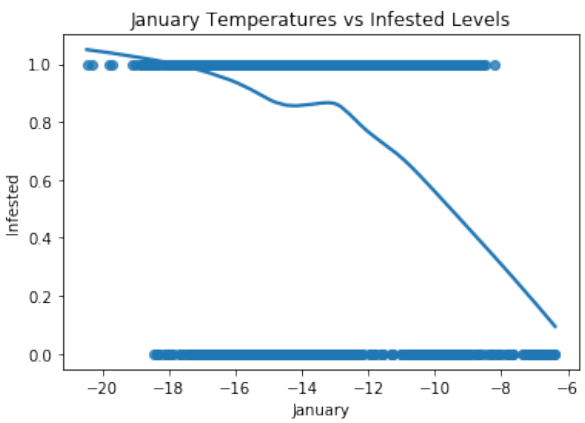
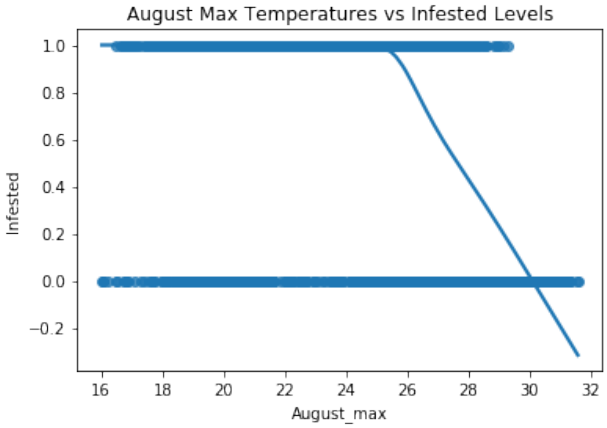
Michael Pixton

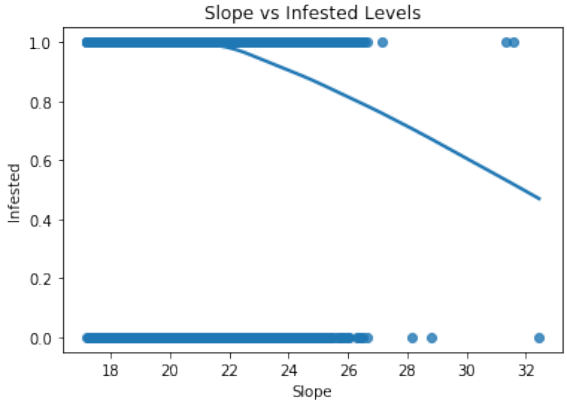
Stat 330 Final – Pine Beetle Infestation

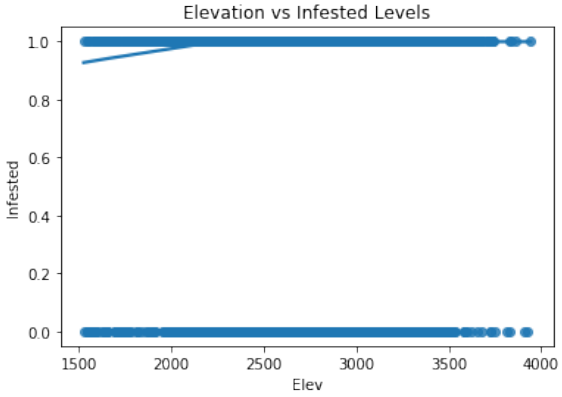
1. The mountain pine beetle lives natively in the Norwest forest of the United States. Its presence helps the decomposition of older and dead pine trees to make room for new pine trees. However, due to changing temperatures and precipitation, today in some locations, the mountain pine beetle has started to drastically decrease pine tree life spans. We would like to look into this recent change and determine why it has happened. A large dataset taken from the Colorado State Forest Service with information regarding the mountain pine beetle will be analyzed using statistical modeling. This analysis will look into what conditions are leading to areas becoming infested by the mountain pine beetle. With better understanding of what variables are causing the epidemics, precautions may be taken to reduce these epidemics. We also will be able to use our modeling to predict where potential new epidemics may start; allowing the Forest Service to react and strengthen those areas in whatever measures are possible.

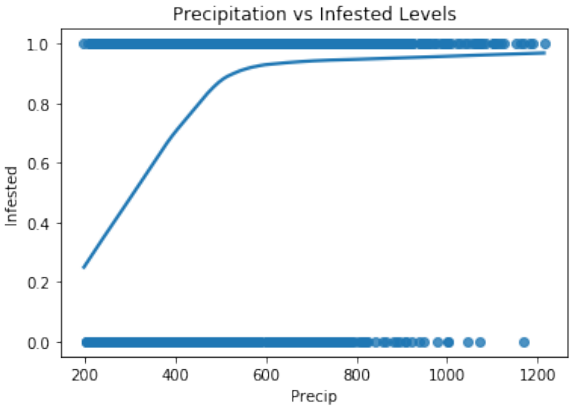
   The dataset is first explored, looking at what characteristics appear to have the greatest effect on whether a location is infested or not. The following series of scatterplots with smooth curves look at the quantitative variables within the dataset and compare them to whether or not the area was infested. We are looking for strong upward or downward trends.





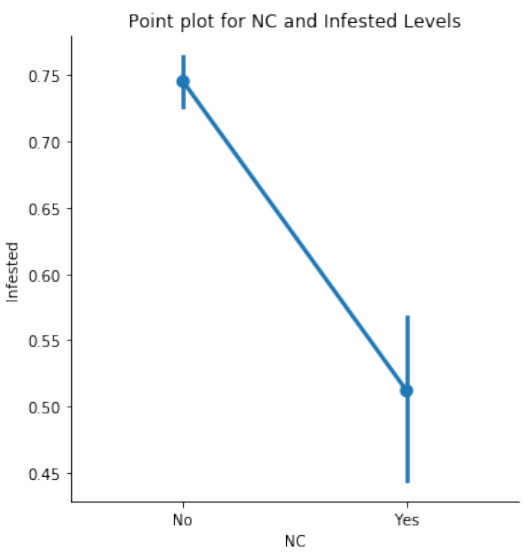


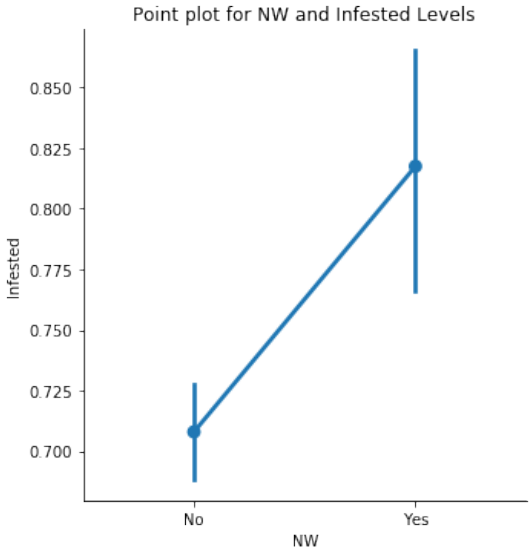


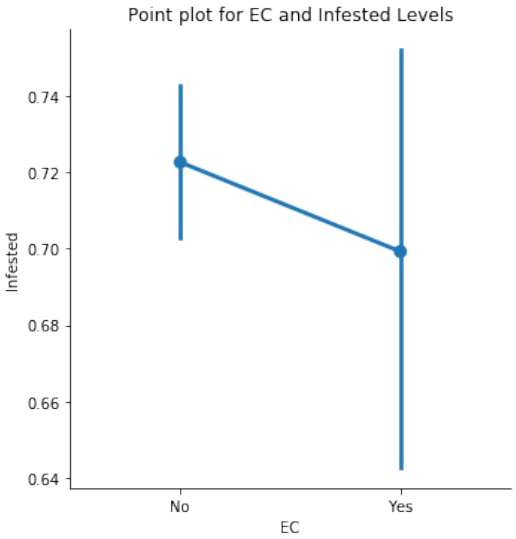


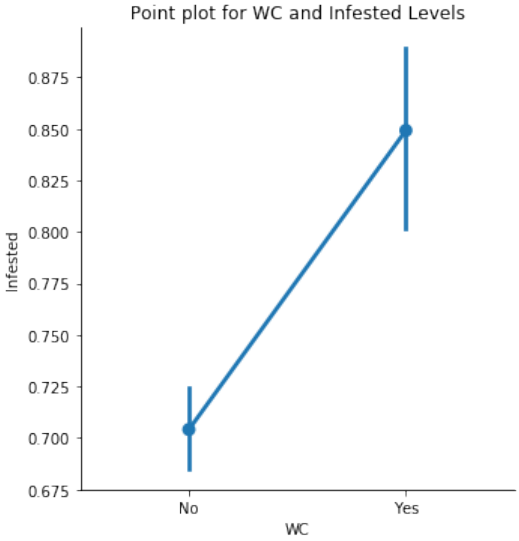
These scatterplots show a what appears to be a strong trend from both temperature-gauge variables. August maximum and January minimum temperatures appear to influence whether or not an area is infested fairly heavily, both showing that higher temperatures seem to be a factor for lower infestation rates. The slope and elevation variables appear to have less effect on infestation, very high slopes seem to have lower infestation rates and elevation seems to have little to no effect. Precipitation levels appear to be fairly significant based upon the trends seen, with higher precipitation leading to higher infestation rates. The next series of plots show our categorical variables and, as all the categorical variables are locations for: North central (NC), Northwest (NW), East central (EC), West central (WC), Southeast (SE), South central (SC), Southwest (SW). They are each compared to the rest of the datapoints to show relative higher or lower infestation rates. Again, we are looking for strong trends.

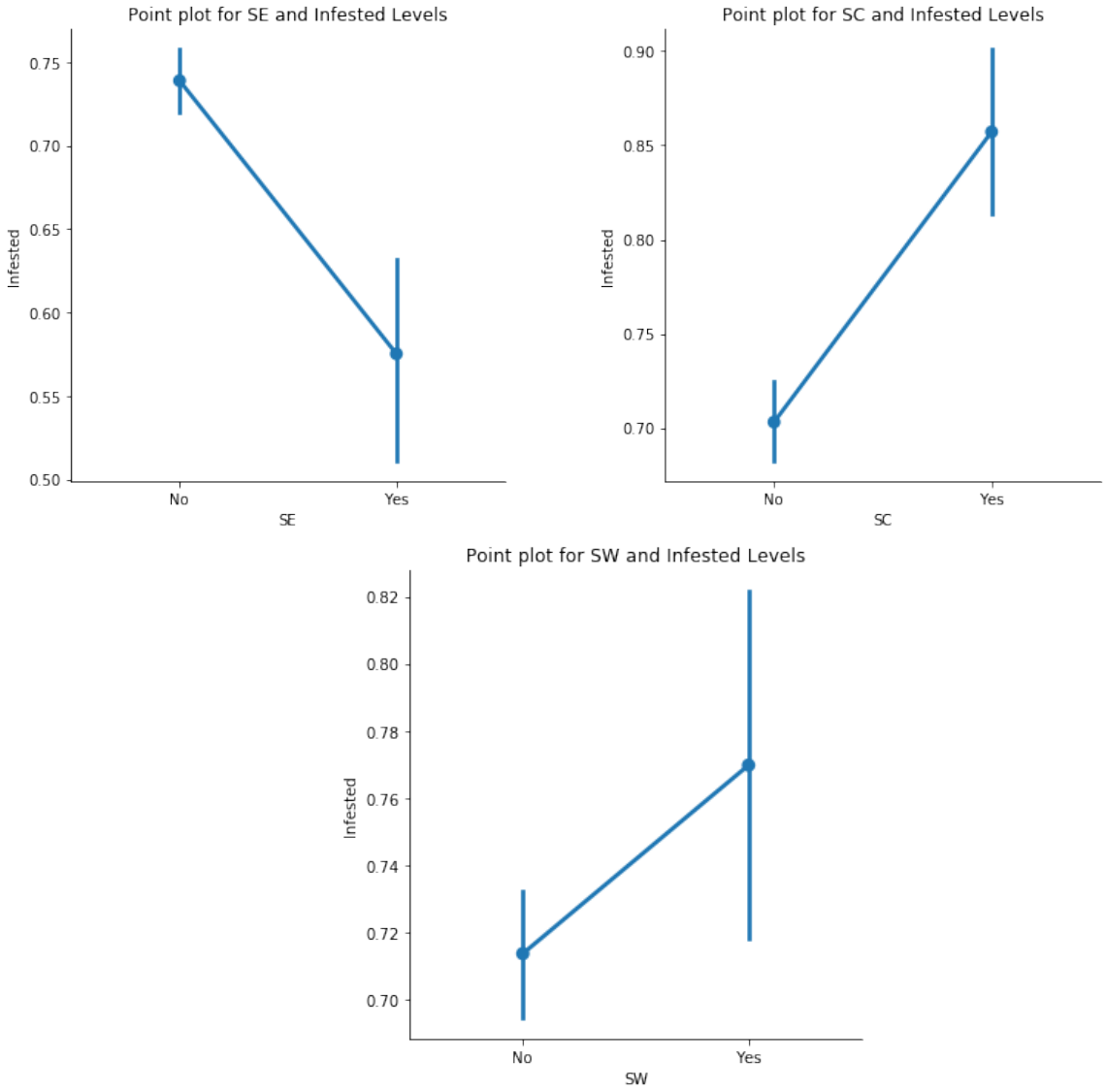
Michael Pixton







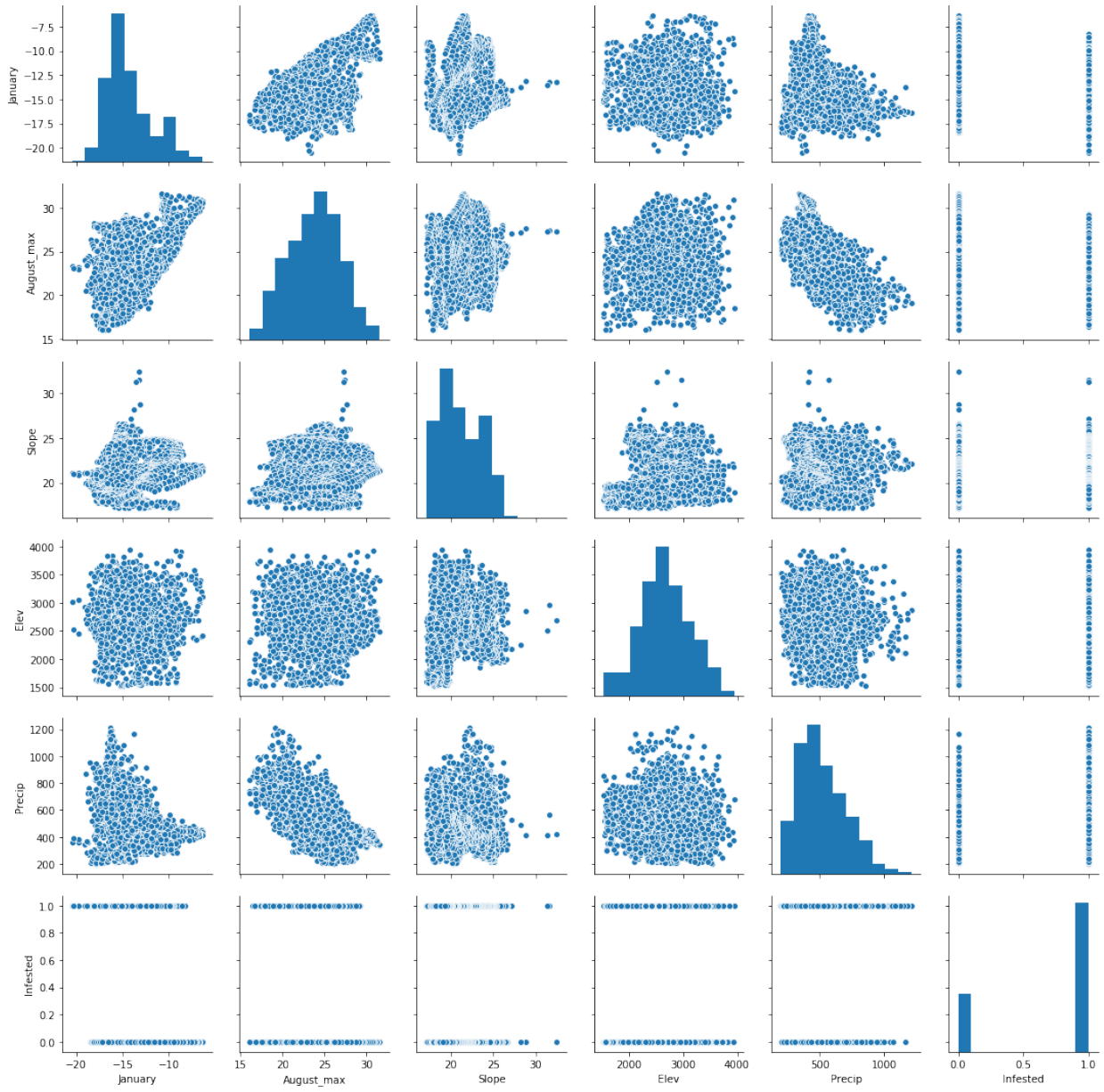




These plots show us trends, but it is important to look at the scale and see how much the infested rate actually goes up by. The areas of NW, WC, SC all show minor increases in infested levels. The areas of SW and EC appear to have minimal effect on infested levels and the areas of NC and SE appear to have moderate decreases in infested levels. Another metric for exploring the data is to use a summary analysis. From the summary, we learn that there are 2310 observations and that the mean of infested variable is .72, so this means we have more data points representing infested areas than non-infested areas. Finally, for our exploratory data analysis, we have a create scatterplot matrix.

This will let us look at all the quantitative variables and see if there might be some collinearity; collinearity is where one dependent variable affects another dependent variable.

Scatterplot Matrix of Quantitative Variables



The scatterplot matrix shows that most variables appear fine, but the two temperature variables might have some collinearity issues as plots appear somewhat similar. This is

logical as a location with higher August temperatures is likely to have higher January temperatures. We will look into this later and see if there is a significant problem with collinearity.

We now must decide how to fit a model to our dataset. Immediately, we know that our response variable is categorical, it is either yes, an area is infested, or no, an area is not infested. This means that the conventional multiple linear regression (MLR) model would not work for our dataset. We can however, fit this data with a logistic regression model. This model takes advantage of the Bernoulli variable, that is either a yes or no. The dependent variable, if an area is infested or not, is modeled by a Bernoulli distribution and then we fit that distribution to a model. This is known as logistic regression and will be the statistical method we use to approach this dataset.

2. When creating a model, we wish to only include those variables that have a significant effect on our dependent variable. Too many independent variables can lead to inflated errors and less accurate predictions from our model. Two metrics for variable selection: AIC and BIC exist. AIC is typically used for models that will predict while BIC is used for models that will look at variable inference. We are more interested in what variables are significant and how much they are significant; therefore, the BIC method was used when creating this model. There are various algorithms for choosing what variables will fit the response variable; we selected the ideal possible variable group with the best variable selection method. Best variable selection is only possible when the number of independent variables is less than 40. Since our data set only has 12 independent variables, best variable selection was used. Using BIC and best variable selection, we determined that the best model has the following independent variables: January's minimum temperature (January), August's maximum temperature (August_max), mean annual precipitation in inches (Precip), the area is North Central, and the area is South East.

Now that we have the independent variables, we can write out a model that we will fit to our data. Our model will have the following form:

$$y_i = \sim ind\ Bern(p_i)$$

$$\log[p_i / (1 - p_i)] = B_0 + B_1*NC + B_2*SE + B_3*January + B_4*August\_max + B_5*Precip$$

We can interpret this model using the Beta (B) values. For example, as a quantitative variable such as Precip increases by 1 and all other variables are held constant, the location is $100 * (\exp\{B_5\} - 1)$ percent more likely to be infested. This is a bit hard to interpret with all the transformations and without any numbers but it is simply a small change in the likelihood that an area will be infested or not. Another example, as a categorical variable such as NC changes from, “no” to “yes” and all other variables are held constant, the location is $100 * (\exp\{B_1\} - 1)$ percent more likely to be infested.

In order to use a logistic regression model, we must make some assumptions for this type of model. Linearity and independence are the two assumptions that we must justify for our model to be considered valid. Linearity can be checked by looking at the first series of plots in our exploratory analysis. We are looking to make sure that general trend lines are linear. And we can confirm that for our quantitative variables, the trends do appear to be mostly linear. They both curve a bit, but nothing that we believe violates our linearity assumption so linearity is assumed. Independence is our next assumption that we need to consider. Independence assumes that each data point has no influence on any other data point from our dataset. We believe each spot is distinctive so each point is independent. We also note, some data points may border one another, which could violate our independence assumption. However, we believe each spot is unique enough, so independence is assumed. This fulfills the assumptions necessary for a logistic regression model. Outside of assumptions, the variance inflation factors (VIFs) are calculated to check for potential problems with multicollinearity among our variables. All the values are less than 5 and typically, only values greater than 10 are considered potential multicollinearity issues so we assume the independent variables are not collinear.

3. With a base model chosen, variables selected, and all assumptions justified, we can now fit our data to our model. The model is fit and we get the table below of Beta values each with a 95% confidence interval.

Coefficients (Beta) Values with 95% Confidence Interval			
variable	**coef (Beta)**	**[0.025**	**0.975]**
Intercept	-0.1574	-1.884	1.569
NC[T.Yes]	-1.2183	-1.512	-0.924
SE[T.Yes]	-0.9191	-1.218	-0.62
January	-0.1466	-0.193	-0.1
August_max	-0.0851	-0.132	-0.038
Precip	0.0029	0.002	0.004

Our model, now with the Beta values, has the following form:

$$\log[p / (1 - p)] = -0.1574 + -1.2183*NC + -0.9191*SE + -0.1466*January + -0.0851*August\_max + 0.0029*Precip$$

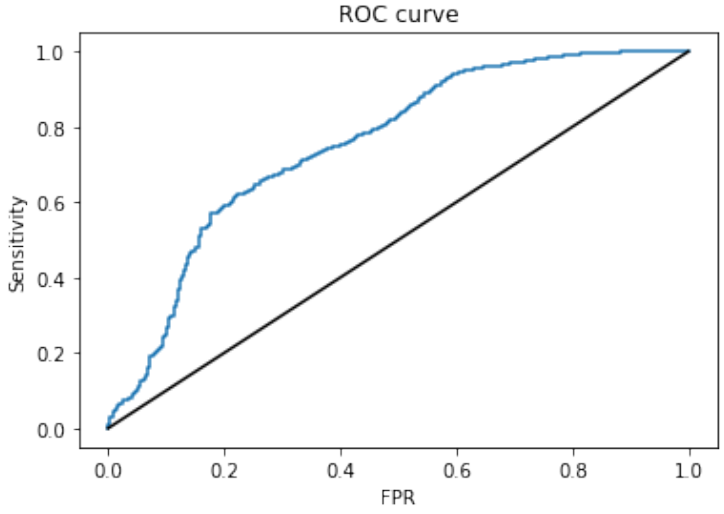
This table and formula illustrate the effect of each of our independent variables on whether an area is infested or not. We see that most of the values are negative, meaning that they decrease the likelihood of an area being infested. We can also interpret the effect inversely where each variable now increases likelihood of infestation. These values are best interpreted within a confidence interval to give awareness to the range of their possible effect. We can untransform the intervals into an easier to understand percentage range by following the example below. Assume we are looking at the effect of Precip.

$$100 * (\exp\{(0.002, 0.004)\} - 1) = (.2\%, .4\%)$$

We interpret this interval as; we are 95% confident that as Precip increases by 1 the likelihood of an area being infested increases by between .2% and .4%.

We would like to now analyze how effective our model is at fitting the data. First, we report a pseudo R-squared of .141. This is a measure between 0 and 1 that shows how

effective our model is at perfectly modeling pine beetle infestations. The small number of .141 would seem to indicate our model is not strong, but generally pseudo R-squared values are smaller and we believe that our model is still much better than guessing based upon the average. Another method for testing how good our model fits is by finding a threshold to minimize our misclassification rate so we do not error as much when classifying infested areas. The Receiving Operator Characteristic (ROC) Curve in the graph below was constructed by plotting the sensitivity against the false positive rate for each possible threshold. At .495, we are at moment where the curve stops going up as much and starts to flatten out more; which is where the ideal threshold rate is, high sensitivity with a low false positive rate. Sensitivity is a measure of how many locations that are infected were correctly identified as infected areas by our model. False positive rate is a measure of how many locations that are not infected were incorrectly identified as infected areas by our model. We can also use this curve to report the (area under the curve) AUC of this model of 0.755. An AUC of .5 means that the model is doing the same as just flipping a coin to test if an area is infested or not, so we should be higher than .5; and an AUC of 1 means we have a perfect model that always correctly predicts. The value of .755 is right in the middle, showing that we are predicting with our model over all thresholds correctly 75.5% of the time.



Another measure of the effectiveness of our model is to now take our ideal threshold rate of .495 and predict our dataset with the model we created. This will create a confusion matrix as seen below:

	Predicated No	**Predicated Yes**
True No	245	402
True Yes	82	1581

We can now report the sensitivity of our model with our current dataset as 0.951; this indicates we are correctly identifying infected locations as being infected 95.1% of the time. This is an excellent sensitivity and means that our model excels at finding true positive datapoints. Another metric to put up next to our sensitivity is our specificity. Specificity is the number of times our model correctly identifies a location as not being infected. The specificity of our model is .379, or only 37.9% of the time are we correctly identifying a un-infected location as not being infected. This would seem to be bad but generally, if we have high sensitivity, we will have a lower specificity; and for our model we are more concerned about determining if an area is infected rather than identifying areas that are not infected. A couple of other numbers we learn from our confusion matrix are the positive predictive values and negative predicative values. Each of these is the percent of either correctly predicted yes's or no's. The reported positive predictive rate of our model with the ideal threshold is .797 or, 79.7% of the time when our model predicts yes, the area is actually an infected area. And the reported negative predictive rate of our model with the ideal threshold is .749 or, 74.9% of the time when our mode predicts no, the area is actually a non-infected area. In summary, our model is doing an effective job with the chosen threshold; the model is generally picking correctly, about 75% of the time. We are erroring more on the side of accidently predicting an area is infected when it is not rather than erroring on predicting an area is not infected when it actual is.

We now would like to see how well our model will do predicting new locations. We can create a cross validation test where a few data points are pulled out of our dataset and the rest of the data is used to create a new model. Then the new model will try and predict the pulled-out data points. This test is run 1000 times and the average of the sensitivity, specificity, positive predictive rate, negative predictive rate, and AUC are calculated. Our predictive capabilities for other locations can be judged upon these numbers. Sensitivity is .948, so the model is correctly identifying a new infected location as being infected 94.8% of the time. Specificity is .374, so the model is correctly identifying a new un-infected location as being un-infected 37.4% of the time. Positive predictive rate is .795, so the model is correctly identifying an infected location 79.5% of the time. Negative predictive rate is .753, so the model is correctly identifying a non-infected location 75.3% of the time. Finally, the AUC for all our tests averaged to be .754 or 75.4% of the time our model is correctly choosing from all possible threshold values. The predictive qualities of our model seem to closely match those values from our dataset so we can conclude that our model is being as effective as possible when predicting values outside the dataset and has good predictive capabilities.

Ecologists wish to use our model to predict pine beetle infestations in an area using forecasts for future temperatures and precipitation. Using our model, we predicted the likelihood that the area will become infested. The following table shows each year and the likelihood it will become infested.

Table of Model Predictions	
Year	Prediction
2018	0.862524
2019	0.9056
2020	0.87846
2021	0.673542
2022	0.860631
2023	0.750776
2024	0.89227
2025	0.874927
2026	0.89594
2027	0.832852

The model predicts that in most years, the area will become infested with over 80% likelihood. The model was built from data in a single time frame so and it can be difficult access how our model predicts over a series of time frames like the above. For example, and in particular, we are not as confident if the 86.2% likelihood is true for the first year. However, based on the overall predictions for the next 10 years and with average of the next 10 years being 84.3% likelihood of being infested; we believe it is highly likely that the area in question will become infested within the next 10 years. We also believe it would therefore be beneficial for the Forest Service to concentrate their efforts on this area to prevent infestation.

4. Based upon the analysis done, we now believe that we can accurately predict pine beetle infestations based upon the variables: January minimum temperature, August maximum temperature, annual precipitation, and if the area in question is in the Northcentral or Southeast. We also can conclude that pine beetle infestations increase when precipitation increases, decrease when temperatures increase, and decrease when located in the Northcentral or Southeast areas. Using this model, the forest service can know where to better protect forests that are more likely to become

infested and the forest service will continue to measure the key variables, looking for changes in an area that could indicate an increase in the likelihood of infestation.

A few more steps could help better understand key factors and how they affect our response variable, pine beetle infestations. One step would to look into our independence assumption. Another analysis would be performed where independence in not assumed among the data point areas, as some are not spatial independent. Another step could be to monitor a new set of variables such as forest fires, pollution, human activity, and more that might be considered key to pine beetle infestations. Other influential variables could help our model be more accurate when making predictions and help the forest service better protect and respond to pine beetle infestations. Finally, this data set was all taken from the same area, within Northern Colorado. Another analysis of the same variables could be done in another area with current pine beetle infestations to check if the same results are produced, increasing the validity of our analysis.