Predicting Blueberry Yield Using Machine Learning

Brian Mann

Introduction

Blueberries are an essential part of the breakfast table for many Americans. In terms of total fruit sales, blueberries account for over 5% of the market share[4]. I am from Georgia, the nation's top grower of blueberries. In 2021, Georgia growers accounted for 4.15 million pounds of cultivated blueberries valued at over $130 million[5]. For the agricultural companies responsible for growing and harvesting them, it is critical to be able to predict how external factors influence crop yield.

The overall goal of this project was to generate a regression model that can accurately predict the yield of blueberries in an area given weather conditions and bee pollination density. The data used in this project was derived from the Wild Blueberry Pollination Simulation Model[1][2][3], which is an open-source program that has been validated by field data collected in Maine and Canada. Although the data is simulated, it has generated samples based on real-world data used as input.

It will be helpful to first elaborate on some of the variables that were used:

- **Clone size (m^2)**: represents the average area covered by the clones that the blueberries grow on, which can grow to hundreds of feet long.
- **Bee density (m^2 / min)**: four variables represent the average density of some different types of bees - honey, bumble, andrena, and osmia.
- **Temperature** (°F): there are six temperature variables, which include the maximum, average, and minimum temperature recorded for all of the highs and lows during the growing season.
- **Rain (days)**: two variables describe the number of days of rainfall and the average days of rainfall.
- **Blueberry characteristics**: there are three variables related to the characteristics of the average blueberry - fruit set (g), fruit mass (g), and number of seeds. Fruit set is the first stage of fruit development where a fruit's potential size is determined[6].
- **Yield (kg)**: total mass of blueberries in the field, which is what we would like to predict.

This project was conducted in three phases. The first phase included gathering the data, converting it into a dataframe, and performing some exploratory data analysis. The next phase involved cleaning the data, assessing features, transforming features, and then selecting only the most relevant features for model building. The final phase was to split the data into a training and test set, fit various regression models to the data, then select the model that best minimizes error between the expected and predicted results.

Milestone 1

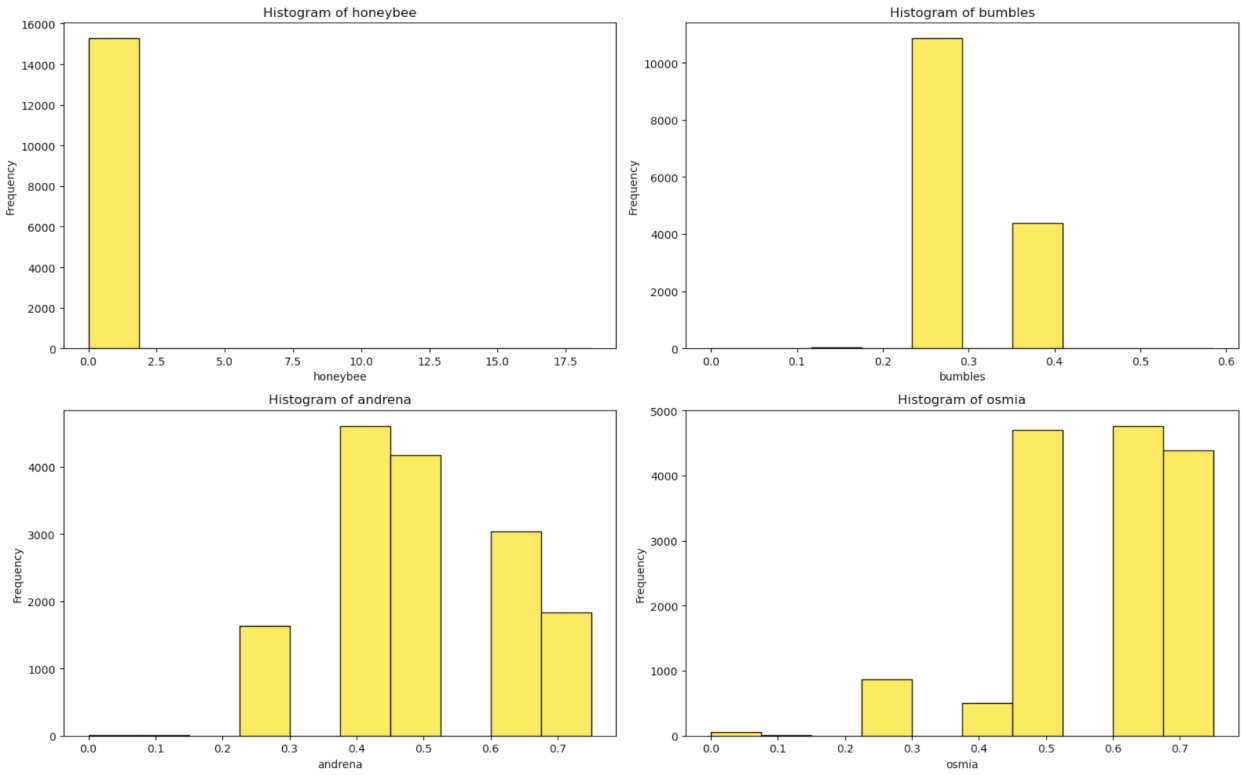

Figure 1: Histogram of Bee Densities

The first set of graphs are histograms (Fig. 1) for each of the three primary categories of variables - bees, weather, and fruit characteristics. Aside from the honeybees, each bee type seems to have densities lying on only a few discrete bins between 0.2-0.7. As for the honeybees, there seem to be a couple of outliers far above the norm, as evidenced by the boxplot. These will be considered in the subsequent phase. A similar story can be found with the number and average of rainy days. There appear to be many repeat values in only a handful of bins. Additionally, the distribution of rainy days seems to be roughly equal among those bins. On the other hand, fruit mass, fruit set, seeds and yield all seem to be following a mostly normal distribution, all with a slightly left/negative skew.

The next set of graphs (Fig. are of correlation values among the various variables. The first correlation matrix is one comparing all 18 values to one another. Since this is quite difficult to read, the values were separated into the same three categories as before. None of the bee densities has much of a correlation with yield, with osmia having the highest with only +0.2. Temperature had close to zero correlation with yield, but the number of raining days had a moderate negative correlation (-0.48), indicating that the blueberries may have benefitted from more days in the sun. Unsurprisingly, each of fruit set, fruit mass, seeds and yield all had a very high positive correlation (+0.83-0.94) with each other. This shows that bigger, riper berries with more seeds are a very indicative of having a high yield.

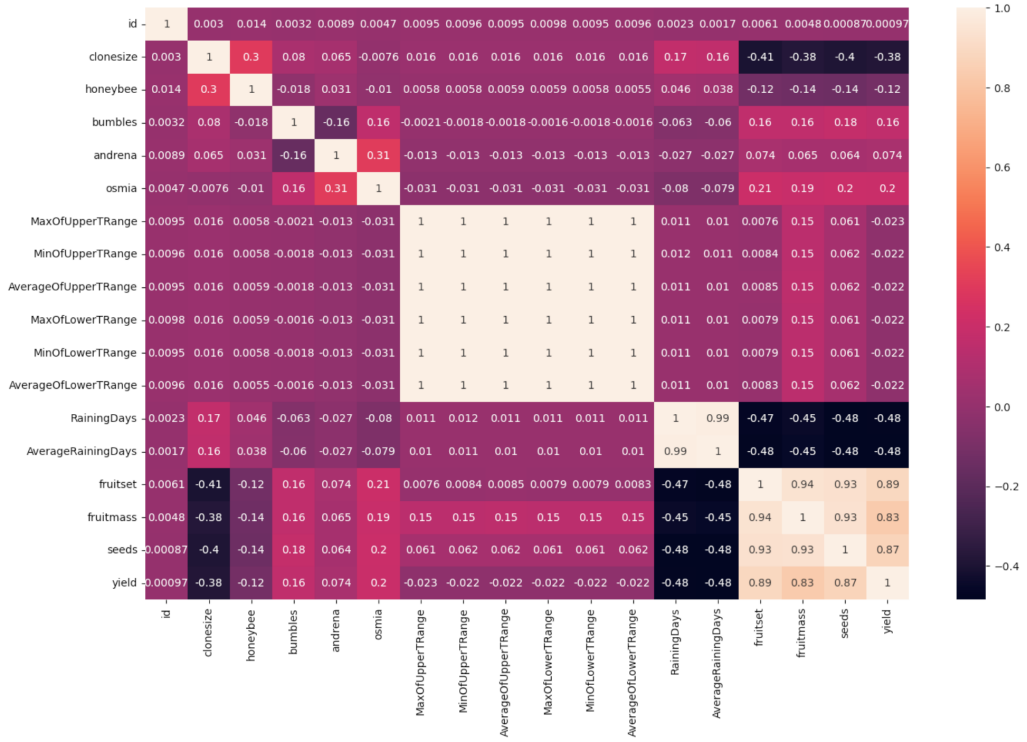

Figure 2: Correlation Among All Variables

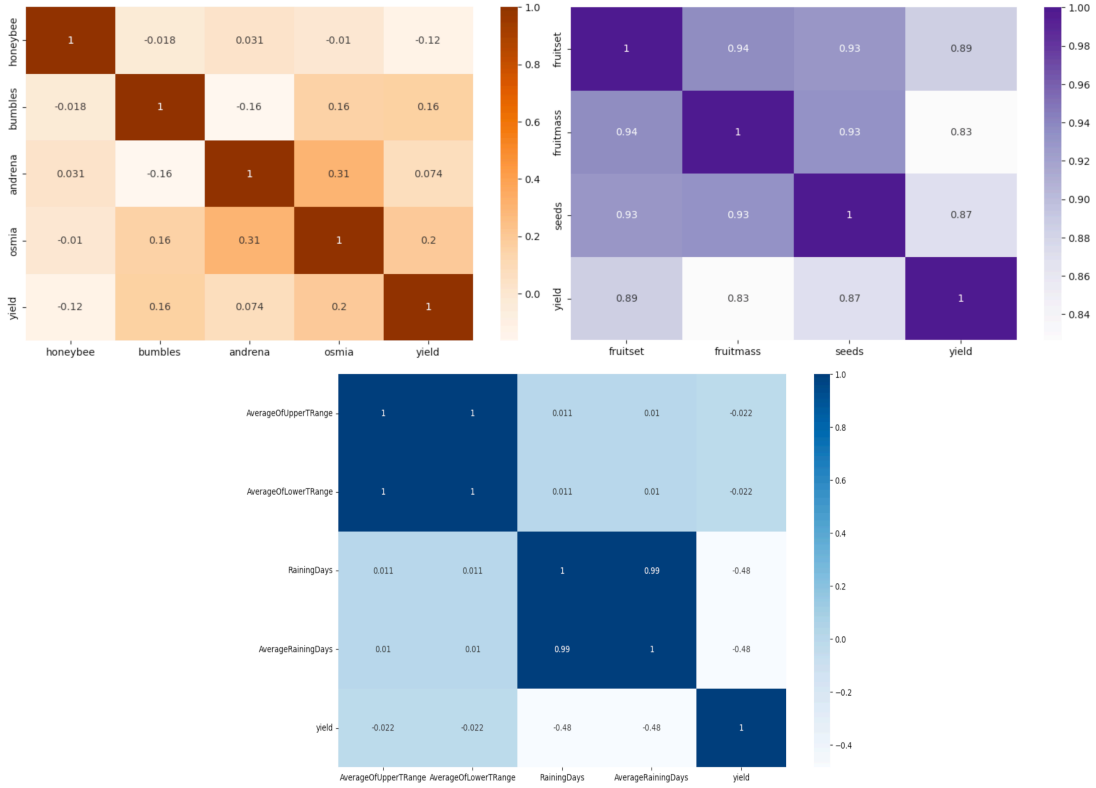

Figure 3: Snapshot of Correlations with Yield

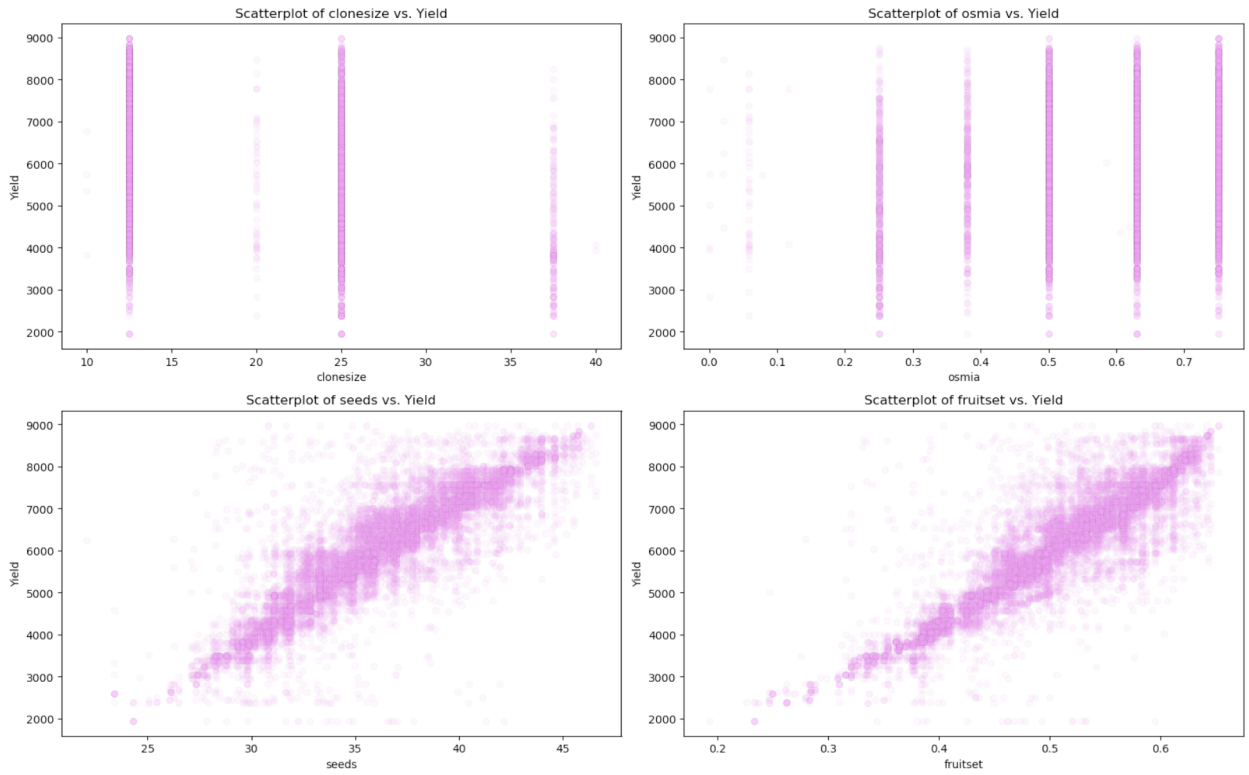

Figure 4: Scatter Plots Containing Yield

Lastly, a few different variables (clonesize, osmia, seeds, and fruitset) were plotted in scatterplots against yield. Similarly to how the histograms had shown, there seem to be only a few discrete values that the vast majority of datapoints fall into for the osmia bees and clonesize. This is most likely also true for the rest of the bees and the number of rainy days. On the other hand, fruitset and seeds had a much broader and continuous range of values, with a clear positive linear relationship. This supports the data given from the correlation plots.

Milestone 2

	id	clonesize	honeybee	bumbles	andrena	osmia	MaxOfUpperTRange	MinOfUpperTRange	AverageOfUpperTRange	MaxOfLowerTRange
0	0	25.0	0.50	0.25	0.75	0.50	69.7	42.1	58.2	50.2
1	1	25.0	0.50	0.25	0.50	0.50	69.7	42.1	58.2	50.2
2	2	12.5	0.25	0.25	0.63	0.63	86.0	52.0	71.9	62.0

MinOfLowerTRange	AverageOfLowerTRange	RainingDays	AverageRainingDays	fruitset	fruitmass	seeds	yield
24.3	41.2	24.0	0.39	0.425011	0.417545	32.460887	4476.81146
24.3	41.2	24.0	0.39	0.444908	0.422051	33.858317	5548.12201
30.0	50.8	24.0	0.39	0.552927	0.470853	38.341781	6869.77760

Figure 5: First Few Rows of Initial Dataset

The id column serves little to no purpose in creating a model, so we excluded it. Because the rain_days column and the rain_avg column are essentially two ways of describing the same quantity, we decided to only keep the rain_avg column. As we saw from the correlation plot in milestone 1, each type of temperature data was highly correlated with all the other temperature data. Thus, in order to mitigate the potential for multicolinearity in our regression analysis, we opted to create a separate variable that tracks the difference between the highest and lowest temperatures (temp_diff), while simultaneously dropping all of the other temperature variables.

Despite all the variables in the dataframe being numeric, there were a significant number of repeated values for clone size, temperature, rain, and all the bee-related columns. To better reflect the pseudo-categorical nature of the data, we created two separate sets of data - one for the original features, and one of one-hot encoded features. A PCA and a variance threshold were then set up to be used for feature extraction/selection during the modeling phase.

Milestone 3

DUMMY	TRANSFORMATION	FEATURES	MODEL	ALPHA	R2	RMSE
False	NONE	10.0	RIDGE	3.727594	0.818420	575.110717
True	NONE	62.0	RIDGE	24.420531	0.819837	572.861472
False	PCA	6.0	RIDGE	0.000450	0.793287	613.622348
True	PCA	36.0	RIDGE	2.559548	0.808218	591.045318
False	VTHRESH	2.0	RIDGE	0.828643	0.226024	1187.355701
True	VTHRESH	20.0	RIDGE	5.428675	0.388887	1055.061571

Figure 6: Results of Regression Testing

The target variable that we would like to predict is blueberry yield, which is a real-valued numeric quantity. In this case, linear regression was a clear choice for a starting point, as it is simple and easily interpretable. Ridge regression was added to further improve model building due to its improved feature selection and reduction of overfitting. Additionally, lasso regression was also added, but due to issues with convergence - possibly due to multicolinearity - it was dropped.

Grid search cross-validations were conducted across two sets of training/test features (dummy-transformed and original) and three types of transformations (PCA, variance threshold, and no selection/extraciton). Ridge regression was a clear winner in every test, unsurprisingly, as linear regression was simply used as a baseline. In both the R^2 and RMSE metrics, the dummy-transformed features out-performed the original features in each test. Despite this, the improvements were relatively slim. The PCA extraction did not produce an improvement in R^2 or RMSE, nor did the variance threshold, although the variance threshold produced significantly worse results than all other models.

In the end, the dummy-transformed features with no selection/extraction produced the lowest RMSE (572.8) and the highest R^2 (0.819). However, only using the original features produced nearly the same result (RMSE 575.1 and R^2 0.818). Considering that the dummy-transformed data included 52 more features than the original data, with larger volumes of data, it may be more efficient to forgo this type of transformation. Therefore, it is recommended that the first ridge regression model be used with the original non-transformed features.

Conclusion

This project was focused on developing a model that could accurately predict the level of blueberry yield that would be produced in a given field based on factors such as bee density, rainfall, temperature, and characteristics of the individual blueberries. After producing six sets of data for testing, it was determined that the one-hot encoded data produced the ideal ridge regression model, with the vanilla (unprocessed) dataset coming in a close second. PCA extraction and variance threshold selection had no positive effect on making a better model. With a R^2 value of ~82% and RMSE of ~570, one of these ridge regression models would be ready to be deployed now. RMSE is less than half a standard deviation (1337) from the average value of yield (6025) in the dataset. Some future steps will be to try to find other characteristics involved in growing blueberries that might have an additional affect at predicting yield. Additionally, some of the variables such as clone size and bee density should be calculated to a more specific degree, so as to avoid the overly generalized estimations used in this dataset.

Sources

- [1] https://www.kaggle.com/competitions/playground-series-s3e14/
- [2] https://www.kaggle.com/datasets/shashwatwork/wild-blueberry-yield-prediction-dataset
- [3] https://data.mendeley.com/datasets/p5hvjzsvn8/1
- [4] https://www.ers.usda.gov/publications/pub-details/?pubid=107357
- [5] https://www.agmrc.org/commodities-products/fruits/blueberries
- [6] https://lodigrowers.com/improving-fruit-set/