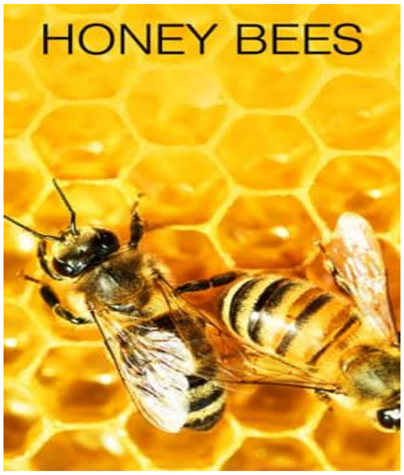
Machine Learning Engineer Nanodegree

Capstone Project



Michael L. Radvansky

November 14, 2018

I. Definition

Project Overview

Bees have a critical role in providing food to people across the world. One third of all the food produced in the world is dependent upon bees for pollination. Without bees, worldwide food production would be devastated. From a purely economic standpoint, it is estimated that the value of bee's pollination is around 265 billion € worldwide. (Greenpeace. The Role of the Bee, 2014, 2014)

Mysteriously, bee populations in North America and around the world are declining and jeopardize agricultural production. This phenomenon is known as Colony Collapse Disorder (CCD). In the United States, 40% of commercial honeybees have been lost since 2006. In Europe, 25% of commercial honeybee populations have been lost since 1985. (Wikipedia - Colony Collapse Disorder, 2018)

It is unclear what is causing CCD. Suggested causes include parasitic infections, malnutrition, pathogens, genetic factors, immunodeficiencies, loss of habitat, or a combination of all of these. (Wikipedia - Colony Collapse Disorder, 2018) More recently, it has been theorized that neonic (also known as neonicotinoid) pesticides are at least partially responsible for CCD. In fact, just this year, the European Union has banned the three main neonic pesticides: clothianidin, imidacloprid, and thiamethoxam. Several states of the United States have also restricted use of neonic pesticides out of concern for pollinators and bees. (Wikipedia - Neonicotinoids, 2018)

This study intends to predict the decline in number of honey bee colonies and honey yield per bee colony basis the amount of specific neonic pesticides used over time as single regression problems. Several regression models will be evaluated to make the predictions.

Data for this study were put together by Kevin Zmith and are sourced from Kaggle. (Zmith, n.d.) The dataset has 1132 rows and 17 features of yearly United States honey production metrics combined with neonic pesticide usage by state from 1991 through 2017.

This dataset is version 3 and joins data from three sources:

1) USGS data for pesticides 1992-2016, same as prior version 2) [from Honey Production in the USA] : The National Agricultural Statistics Service (NASS) 19982017, same as prior version 3) [New!] by OCR conversion (Messy, Ugly, Brute-force uploading) of NASS data from scanned PDF files. 1991-1997

Problem Statement

Decline in bee populations threaten the world's food supply; and, it is theorized that the decline is related to the use of neonic pesticides. From United States honey production data collected by the National Agricultural Statistics Service (NASS) combined with United States neonic pesticide usage collected by the United States Geological Service (USGS), this study intends to predict the number of honey bee colonies and honey yield per bee colony basis the amount of specific neonic pesticides used over time using several regression models. The steps that will be performed to make the predictions and evaluate the models are:

1. The honey production/pesticide usage data will be loaded into a dataset.
2. The data will be inspected to determine if any cleaning, formatting, or restructuring is required (preprocessing); if so, the data will be preprocessed.
3. Charts and visualizations will be produced to determine trends as well as which features are the best candidates for modelling.
4. The data will be inspected to determine if any features are skewed; if so the data will be normalized.
5. The data will be split into training and testing sets.
6. The Supervised Learning models (DecisionTreeRegressor (SciKitLearn - DecisionTreeRegressor, n.d.), LightGBM (Microsoft - Lightgbm, n.d.), and SupportVectorRegressor (SckiKitLearn - sklearn.svm.SVR, n.d.)) will be evaluated against the modelled features. KNeighborsRegressor (SciKitLearn - SVR, n.d.) will be used to establish a baseline.

7. The r2 scores of the feature predictions will be compared for each supervised learning model evaluated against the baseline KNeighborsRegressor r2 scores to determine the best model to use to evaluate these data.

Metrics

The primary metric that will be used to evaluate the predictions of the models tested will be the R Squared metric. (Wikipedia - Coefficient of Determination, n.d.) R Squared (r2) is the proportion of the variance in the dependent variable that is predictable from the independent variables. Vii

R-squared = Explained variation / Total variation
R-squared is always between -100 and 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits the data.

I chose to use the r2 metric to evaluate these predictions because it is probably the most popular evaluation metric used for regression Further study of r2 metric shows that it is able to deliver more robust results due to its “squared” nature which prevents cancelling the positive and negative error values when summed. However, again because of its “squared” nature, r2 gives a higher weighting to large outliers than small outliers which are treated equally in other metrics such as Mean Absolute Error (MAE). Since r2 is highly impacted by many large outlier values, it should be evaluated as a metric only when all the outliers are removed from the dataset. (Srivastava, 2016) In retrospect, after completing the regressions, MAE may have been the more appropriate metric to use to evaluate the predictions as I did not remove the outliers (which were many) from these data. MAE is more robust to outliers as it does not make use of square. (Aydore, 2015)

II. Analysis

Data Exploration

The data contained in the dataset are as below:

• state: The state code (in the United States) where the data were collected
• year: The year in which the data were collected
• statename: The name of the state (in the United States) where the data were collected
• Region: The region (of the United States) where the state is located
• fips: The United States Federal Information Processing Standard State Code

From NASS data

• numcol: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year

• yieldpercol: Honey yield per colony. Unit is pounds
• totalprod: Total production (numcol x yieldpercol). Unit is pounds
• stocks: Refers to stocks held by producers. Unit is pounds
• priceperlb: Refers to average price per pound based on expanded sales. Unit is dollars.
• prodvalue: Value of production (totalprod x priceperlb). Unit is dollars.

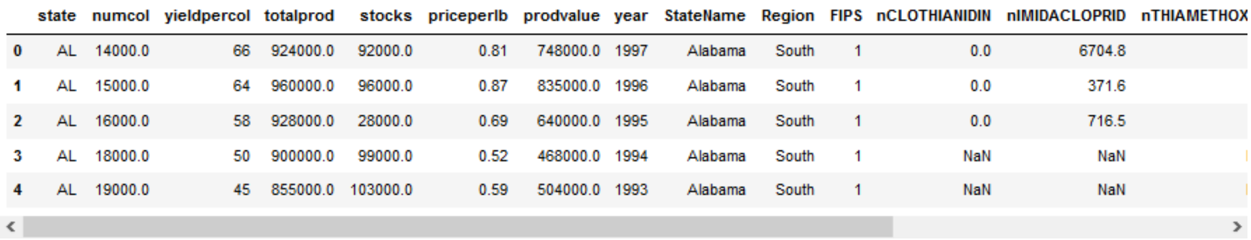
From USGS Data

• nCLOTHIANIDIN: The amount in kg of CLOTHIANIDIN applied
• nIMIDACLOPRID: The amount in kg of IMIDACLOPRID applied
• nTHIAMETHOXAM: The amount in kg of THIAMETHOXAM applied
• nACETAMIPRID: The amount in kg of ACETAMIPRID applied
• nTHIACLOPRID: The amount in kg of THIACLOPRID applied

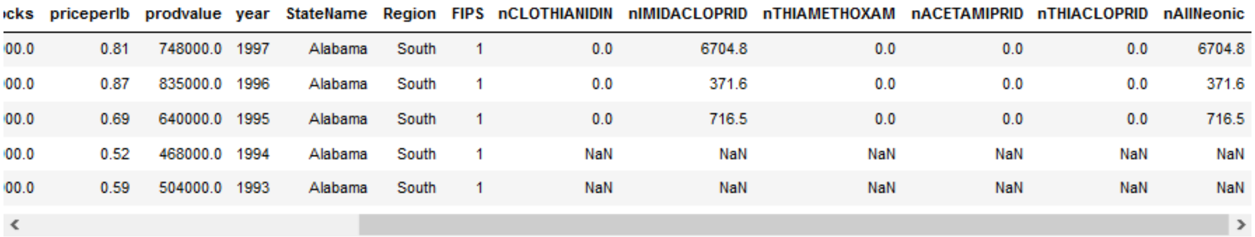
- nAllNeonic: The amount in kg of all Neonics applied = (nCLOTHIANIDIN + nIMIDACLOPRID + nTHIAMETHOXAM + nACETAMIPRID + nTHIACLOPRID)

The data are in a time series by year. A sample of data from the initial combined NASS/USGS dataset is presented in Table 1. below:

Table 1. Sample Data from the Combined NASS/USGS Dataset



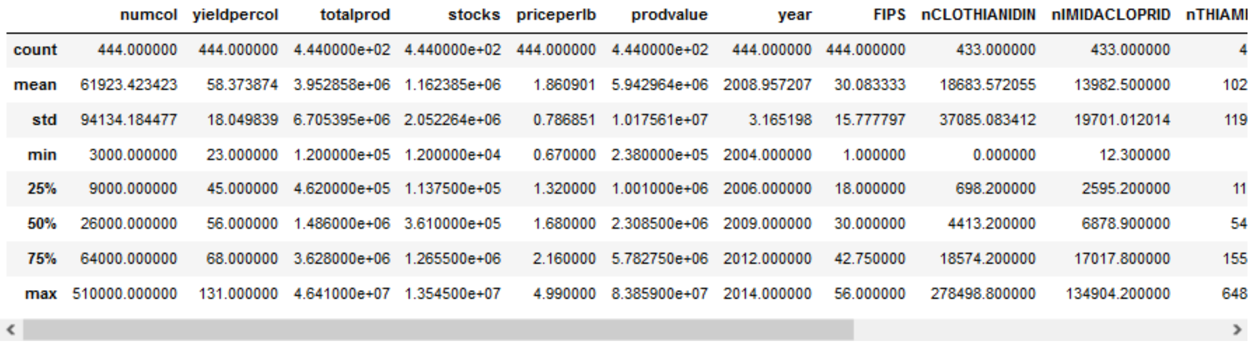
	state	numcol	yieldpercol	totalprod	stocks	priceperlb	prodvalue	year	StateName	Region	FIPS	nCLOTHIANIDIN	nIMIDACLOPRID	nTHIAMETHOX
0	AL	14000.0	66	924000.0	92000.0	0.81	748000.0	1997	Alabama	South	1	0.0	6704.8	
1	AL	15000.0	64	960000.0	96000.0	0.87	835000.0	1996	Alabama	South	1	0.0	371.6	
2	AL	16000.0	58	928000.0	28000.0	0.69	640000.0	1995	Alabama	South	1	0.0	716.5	
3	AL	18000.0	50	900000.0	99000.0	0.52	468000.0	1994	Alabama	South	1	NaN	NaN	
4	AL	19000.0	45	855000.0	103000.0	0.59	504000.0	1993	Alabama	South	1	NaN	NaN	



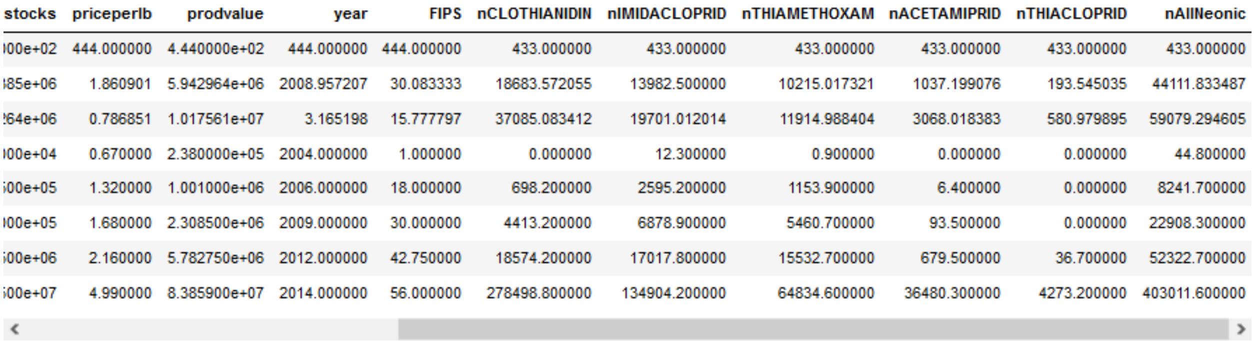
cks	priceperlb	prodvalue	year	StateName	Region	FIPS	nCLOTHIANIDIN	nIMIDACLOPRID	nTHIAMETHOXAM	nACETAMIPRID	nTHIACLOPRID	nAllNeonic
00.0	0.81	748000.0	1997	Alabama	South	1	0.0	6704.8	0.0	0.0	0.0	6704.8
00.0	0.87	835000.0	1996	Alabama	South	1	0.0	371.6	0.0	0.0	0.0	371.6
00.0	0.69	640000.0	1995	Alabama	South	1	0.0	716.5	0.0	0.0	0.0	716.5
00.0	0.52	468000.0	1994	Alabama	South	1	NaN	NaN	NaN	NaN	NaN	NaN
00.0	0.59	504000.0	1993	Alabama	South	1	NaN	NaN	NaN	NaN	NaN	NaN

Descriptive Statistics of the initial combined NASS/USGS dataset are presented in Table 2. below:

Table 2. Descriptive Statistics of the Combined NASS/USGS Dataset



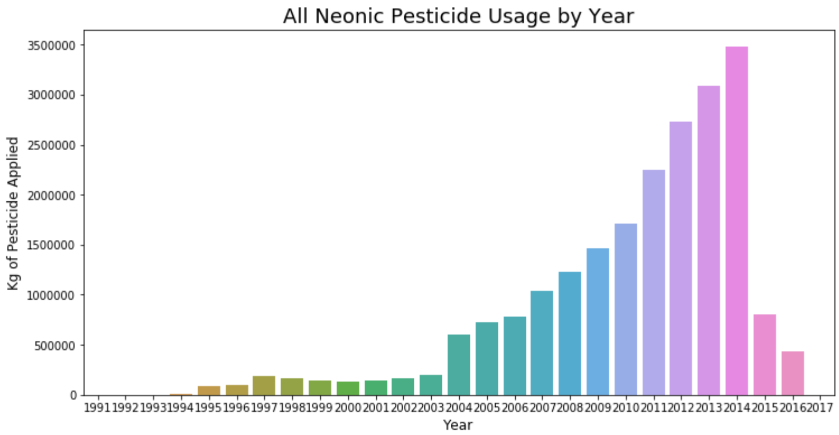
	numcol	yieldpercol	totalprod	stocks	priceperlb	prodvalue	year	FIPS	nCLOTHIANIDIN	nIMIDACLOPRID	nTHIAMI
count	444.000000	444.000000	4.440000e+02	4.440000e+02	444.000000	4.440000e+02	444.000000	444.000000	433.000000	433.000000	4
mean	61923.423423	58.373874	3.952858e+06	1.162385e+06	1.860901	5.942964e+06	2008.957207	30.083333	18683.572055	13982.500000	102
std	94134.184477	18.049839	6.705395e+06	2.052264e+06	0.786851	1.017561e+07	3.165198	15.777797	37085.083412	19701.012014	119
min	3000.000000	23.000000	1.200000e+05	1.200000e+04	0.670000	2.380000e+05	2004.000000	1.000000	0.000000	12.300000	
25%	9000.000000	45.000000	4.620000e+05	1.137500e+05	1.320000	1.001000e+06	2006.000000	18.000000	698.200000	2595.200000	11
50%	26000.000000	56.000000	1.486000e+06	3.610000e+05	1.680000	2.308500e+06	2009.000000	30.000000	4413.200000	6878.900000	54
75%	64000.000000	68.000000	3.628000e+06	1.265500e+06	2.160000	5.782750e+06	2012.000000	42.750000	18574.200000	17017.800000	155
max	510000.000000	131.000000	4.641000e+07	1.354500e+07	4.990000	8.385900e+07	2014.000000	56.000000	278498.800000	134904.200000	648



stocks	priceperlb	prodvalue	year	FIPS	nCLOTHIANIDIN	nIMIDACLOPRID	nTHIAMETHOXAM	nACETAMIPRID	nTHIACLOPRID	nAllNeonic
00e+02	444.000000	4.440000e+02	444.000000	444.000000	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
85e+06	1.860901	5.942964e+06	2008.957207	30.083333	18683.572055	13982.500000	10215.017321	1037.199076	193.545035	44111.833487
64e+06	0.786851	1.017561e+07	3.165198	15.777797	37085.083412	19701.012014	11914.988404	3068.018383	580.979895	59079.294605
00e+04	0.670000	2.380000e+05	2004.000000	1.000000	0.000000	12.300000	0.900000	0.000000	0.000000	44.800000
00e+05	1.320000	1.001000e+06	2006.000000	18.000000	698.200000	2595.200000	1153.900000	6.400000	0.000000	8241.700000
00e+05	1.680000	2.308500e+06	2009.000000	30.000000	4413.200000	6878.900000	5460.700000	93.500000	0.000000	22908.300000
00e+06	2.160000	5.782750e+06	2012.000000	42.750000	18574.200000	17017.800000	15532.700000	679.500000	36.700000	52322.700000
00e+07	4.990000	8.385900e+07	2014.000000	56.000000	278498.800000	134904.200000	64834.600000	36480.300000	4273.200000	403011.600000

The data set does contain abnormalities as there are many data elements that are missing values. Specifically, certain states are missing data for one or many of the neonics over the years. Hawaii has no data for neonic usage in any year. South Carolina has neonic data usage for only one year. Alabama, Colorado, South Dakota, North Dakota, Nebraska, Kansas, and Wyoming are missing most values for acetamiprid usage. Alabama, Wisconsin, Utah, Texas, South Dakota, South Carolina, Nevada, New Mexico, Nebraska, North Dakota, Mississippi, Minnesota, Maine, Louisiana, Montana, Wyoming, Arkansas, Arizona, Iowa, Colorado, Hawaii, Florida, California, Illinois, Kentucky, and Missouri are missing most values for thiacloprid usage. In addition, as shown in Figure 1 below, neonic usage in the United States did not become prevalent until 2004. After 2014, neonic usage was severely curtailed in the United States as many states banned the use of the pesticides.

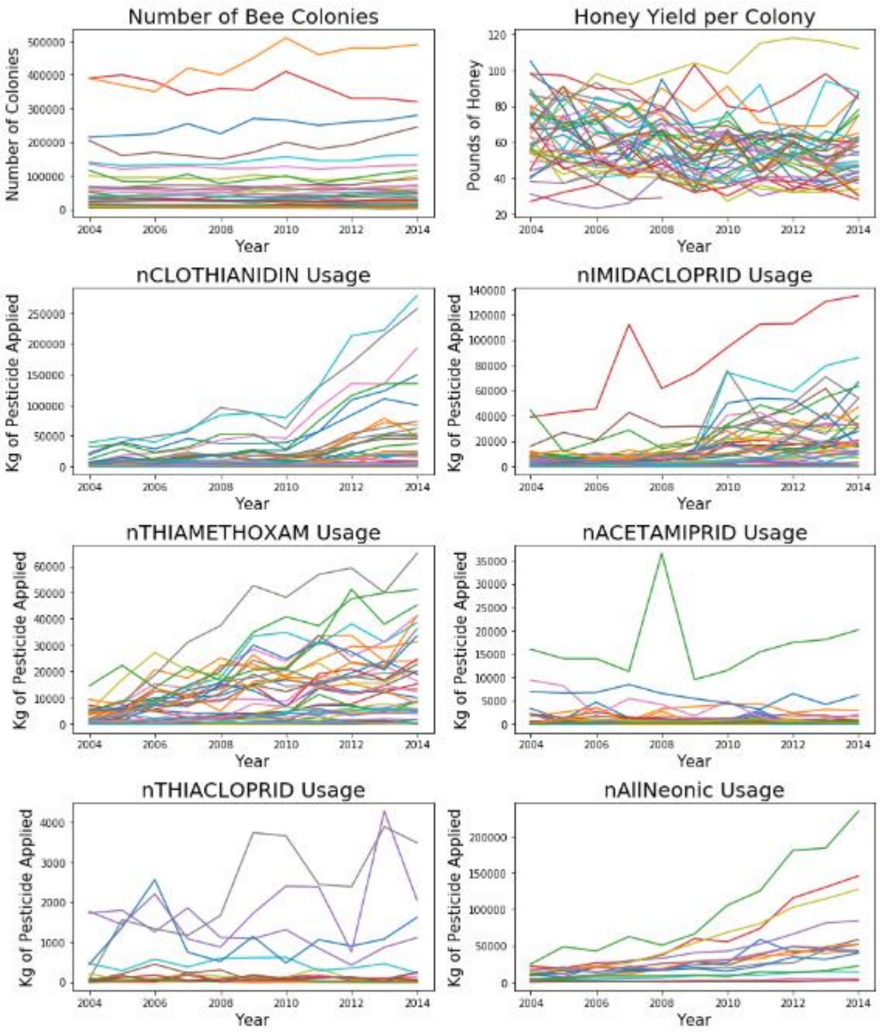
Figure 1. All Neonic Pesticide Usage by Year (Across All States)



Exploratory Visualization

In Figure 2 below, graphs or each feature over the years are presented.

Figure 2. Graphs of Features Across Years for Each State From 2004 Through 2014



I chose to graph each feature by year to determine if there are any trends that are present in these time series data. Visually, it appears the trend for clothiandin, imacloprid, thiamethoxam, and allNeonic pesticide usage increased from 2004 through 2014. In addition, while not as evident, it appears that there may be a trend for Honey Yield Per Colony to decrease over the same period.

In Figures 3a and 3b below, scatter plots of each pesticide application per number of bee colonies and honey yield per colony across all years and states are presented:

Figure 3a. Plots of Pesticide Usage by Number of Bee Colonies Across All Years and States from 2004 through 2014.

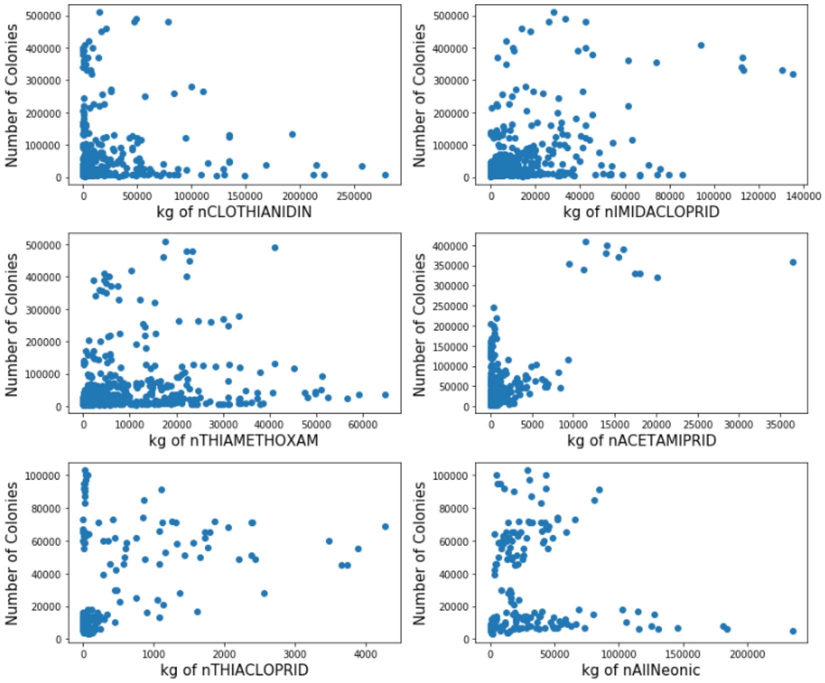
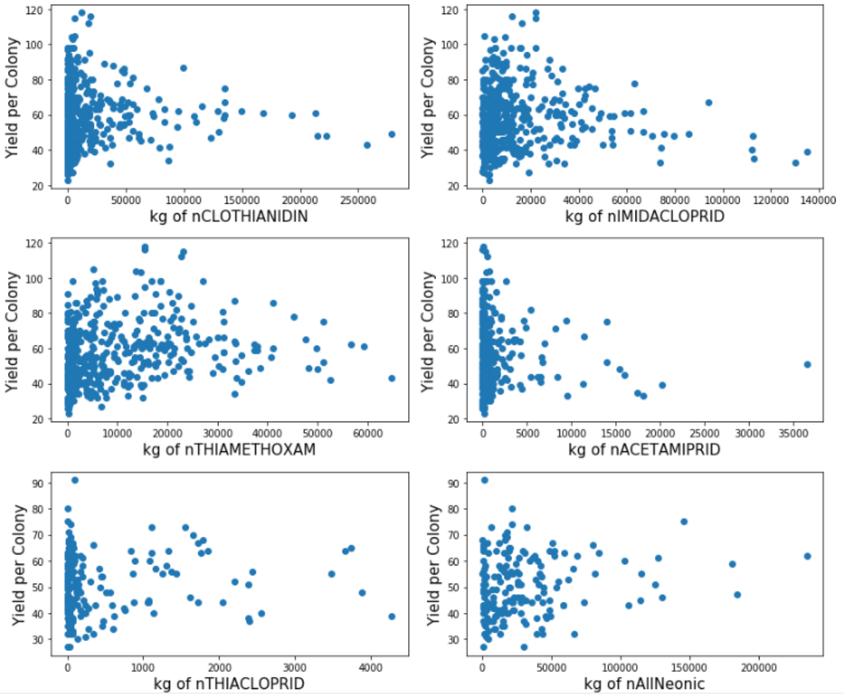
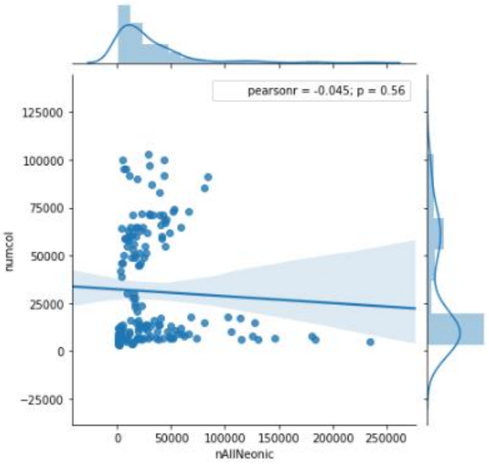


Figure 3b. Plots of Pesticide Usage by Pounds of Honey Per Colony Across All Years and States from 2004 through 2014.

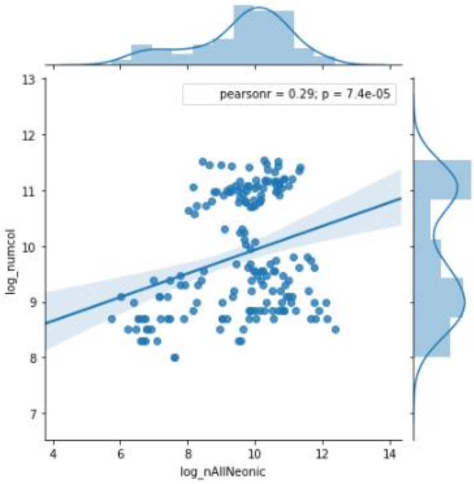


I chose to use scatterplots to visually detect the best features to model. From the scatterplots above, there are no clear correlations of any specific pesticide application in terms number of bee colonies or honey yield per colony. What is clear is the data are heavily skewed and will need to be normalized before analysis. To get a clearer picture of the strength of any correlation in these data and the extent the data are skewed, I ran a simple linear regression for one of the pesticide applications (allNeonic) for unnormalized and normalized data as presented in Figures 4a and 4b below:

Figure 4a. Simple Linear Regression of AllNeonic Pesticide Application Per Number of Colonies (Unnormalized Data)



. Figure 4b. Simple Linear Regression of AllNeonic Pesticide Application Per Number of Colonies (Log Normalized Data)



The unnormalized and normalized plots above show very weak correlations as confirmed by the Pearson correlation coefficients and confidence intervals. Even after performing logarithmic normalization of the data (which did reduce the skew) the correlation only marginally improved.

In Figures 5a and 5b below, scatter plots of each pesticide application per number of bee colonies and honey yield per colony across all years and states are presented:

Figure 5a. Plots of Pesticide Usage by Number of Bee Colonies Across All Years and States from 2004 through 2014 (Log Normalized Data)

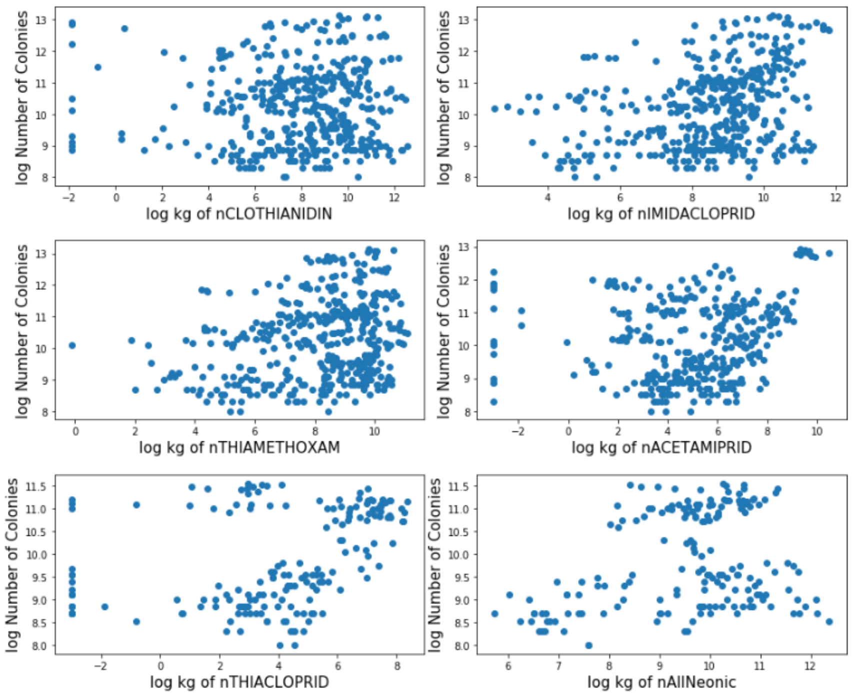
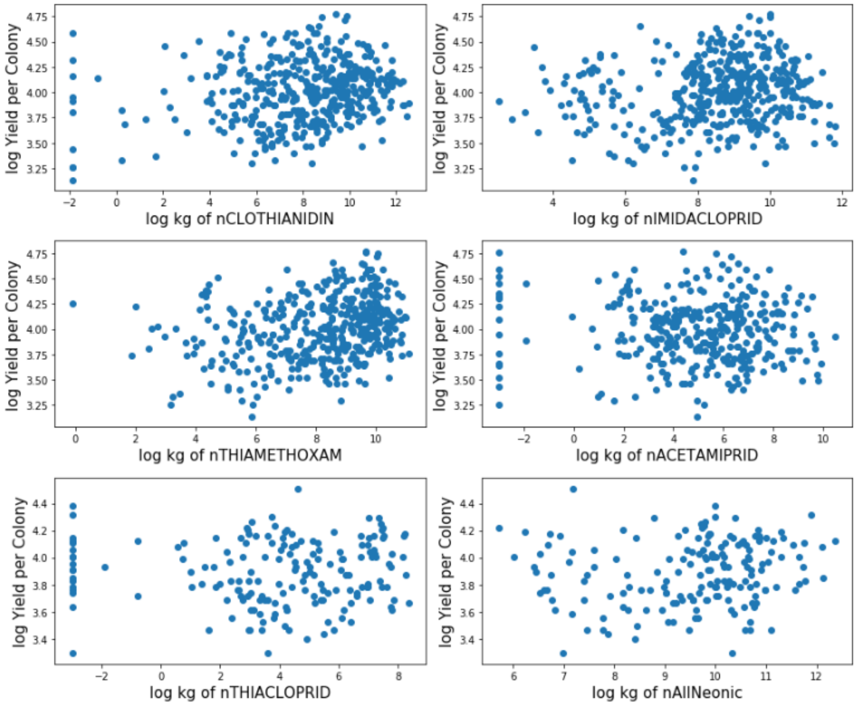


Figure 5b. Plots of Pesticide Usage by Honey Yield Per Colony Across All Years and States from 2004 through 2014 (Log Normalized Data)



As with the unnormalized scatterplots, there does not appear to be any clear correlation of any specific pesticide application in terms number of bee colonies or honey yield per colony.

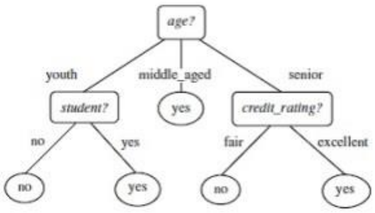
Algorithms and Techniques

Again, this is a typical supervised learning regression problem that deals with continuous data collected over a series of years.

As such, I have chosen the following three algorithms to perform the predictions:

1) DecisionTreeRegressor (DTR) – This is a very intuitive model where decisions are made to traverse down branches of a tree based upon the best performing decision at each node of the tree.
Pros: Easy to interpret and understand, great at learning complex relationships
Cons: Can be prone to overfit. May not generalize well. May be slower and require more machine resources.

DTR builds regression models in the form of a tree structure. When training the model, by using a frequency table it splits the dataset into smaller and smaller subsets that contain data of similar value while at the same time and associated decision tree is incrementally developed; the result is a tree with decision and leaf nodes. (Sayad, n.d.) The core algorithm for building a decision tree is called ID3. (Wikipedia, 2018) which can be used to construct a decision tree for regression by replacing information gain (as in a classification problem) with standard deviation reduction (reducing the error). When descending decision tree for a regression problem, the goal is to find the next node that returns the largest standard deviation reduction. After a split if the similar data are pure, (homogenous) the splitting stops; otherwise the split keeps going using some other attribute. Other termination criteria can be set such as when the standard deviation of the branch becomes smaller than a certain fraction of the standard deviation for the full dataset. (Sadawi, 2014) The process is then run recursively on the non-leaf branches until all data are processed. Finally, when the number of instances is more than one at a leaf node, the average is calculated as the final value for the target.

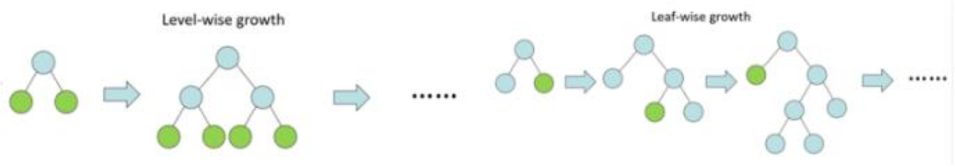


A simple decision tree for predicting whether a person will buy a computer. In this case, the model would predict that a young student would buy a computer, whereas a senior without an excellent credit rating would not.

IMAGE

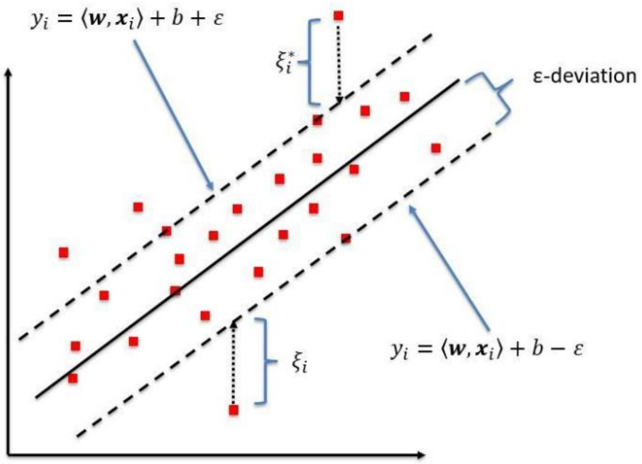
2) Lightgbm (LGBM) – LightGBM is Microsoft's latest entry in the machine learning space. This new algorithm is a boosting algorithm based on Decision tree algorithms.
Pros: Very Fast! May result in much better accuracy.
Cons: There are many parameters to tune.

As mentioned above, LGBM is based on decision tree models; however, with LGBM is in a category of models known as Gradient Boosting Decision Trees (GBDT). GBDT's combine the predictions of multiple decision trees by adding them together. In doing so, GBT's make predictions that generalizes well when compared to a simple Decision Tree Regression. GBDT's are trained iteratively one tree at a time. (Kurita, 2018) In training the data to start, a GBDT would first train from a simple weak decision tree of the data. The next tree is then trained to minimize the loss function when its outputs are added to the first tree. LGBM distinguishes itself from the general class of GBDT's, in the way it optimizes the predictions. LGBM uses the leaf-wise growth strategy when growing a decision tree rather than the level-wise strategy used in other GBDT's. The level-wide strategy maintains a balanced tree while the lead-wise strategy splits the only the leaf that reduces loss the most. Leaf-wide wise training is more prone to overfitting but is more flexible. (Kurita, 2018)



3) SupportVectorRegression (SVR) – A very popular algorithm used for both classification and regression that finds the optimal hyperplane to find the optimal solution.
Pros: Robust. Works well on small datasets. Provides multiple kernel implementations.
Cons: Does not do well with noisy data. Can require high machine resources.

SVR tries to fit a line to data by minimizing a cost function; however, with SVR non-linear kernels can be implemented to provide for non-linear regressions (fitting a curve to the data). As opposed to a s simple regression where the focus is minimizing the error rate, SVR tries to fit the error within certain thresholds. Visually these thresholds appear as boundary lines which are some distance from the hyperplane (let's say 'e' or epsilon). So, the boundary lines are drawn at "+epsilon" and "-epsilon" distance from the hyperplane. (Bhattacharyya, 2018)


$y_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b - \varepsilon$

Solid line is the hyperplane. Dashed lines are the upper and lower boundaries.

Assuming the hyperplane is a straight line going through the Y axis, the equation of the hyperplane is *Wx + b = 0* and the equations of the boundary lines are *Wx + b=+e* and *Wx +b=-e* respectively. When fitting the model, SVR takes into account only those points that fit within the boundary lines meaning it considers only those points that have the least error rate thus providing a better fitting model.

Benchmark

The benchmark model I selected to use for this project is the KNeighborsRegressor algorithm as it is very easy to understand and should be able to be improved upon by more sophisticated regression algorithms. Benchmark KNeighborsRegressor r2 scores of the log normalized data are presented in Table 1. below:

Table 1. KNeighborsRegressor r2 Scores of Pesticide Usage by Number of Bee Colonies Across All Years and States from 2004 through 2014 (Log Normalized Data)

	CLOTHIANDIN	IMIDACLOPRID	THIAMETHOXAM	ACETAMIPRID	THIACLOPRID	AllNeonic
Number of Bee Colonies	-0.49921006	-0.310901863	-0.528572407	-0.04809638	0.382243083	0.617039342
Honey Yield Per Colony	-0.477376076	-0.541283035	-0.382270723	-0.339114404	0.431792068	0.574686196

III. Methodology

Data Preprocessing

As mentioned above, the data set used in this analysis does contain abnormalities as there are many pesticide usage data that are missing. The strategy I used to deal with these abnormalities is to drop the specific pesticide usage data for an entire state from the dataset if 90% or more of that state's data are missing for that specific pesticide usage. (John Paul Mueller, n.d.) For the remaining data elements that are missing for a specific pesticide, I used the strategy of adding one half of the least specific pesticide usage value found in any state to all the data points for that specific pesticide. (Stahel, 2008)

So, as described in Figure 1 above, neonic usage in the general United States did not become prevalent until 2004. After 2014, neonic usage was severely curtailed in the United States as many states banned the use of the pesticides. Because much of these data were missing prior to 2004 and after 2014 I decided to only analyze those data between 2004 and 2014. Data before and after this range were removed from the dataset.

Observation of the data also showed that there were no pesticide data captured for any year in the state of Hawaii; therefore, the Hawaii data were dropped from the dataset.

Further observation of the data showed that acetimaprid application data were missing from the dataset for Alabama, Colorado, South Dakota, North Dakota, Nebraska, Kansas, and Wyoming; therefore, acetimaprid data were dropped from the dataset for these states.

In addition, it was observed that thiacloprid application data were missing from the dataset for Alabama, Wisconsin, Utah, Texas, South Dakota, South Carolina, Nevada, New Mexico, Nebraska, North Dakota, Mississippi, Minnesota, Maine, Louisiana, Montana, Wyoming, Arkansas, Arizona, Iowa, Colorado, Hawaii, Florida, California, Illinois, Kentucky, and Missouri; therefore, thiacloprid data were dropped from the dataset for these states.

Also, as South Carolina only had data for one year of pesticide usage, I decided to remove South Carolina from the dataset as well. Removing South Carolina became mandatory when I tried to split the data and stratify by state (more below).

As mentioned above, because the data are heavily skewed, I normalized the data by performing a log transformation of the pesticide usage, number of bee colonies, and honey yield per colony data.

Finally, I decided to remove those columns from the dataset that would not be used in this investigation. Total Production, stocks, priceperlb, prod, StateName, region, and FIPS were dropped from the dataset.

The result of preprocessing these data resulted in the creation of three distinct data sets from which prediction models will be evaluated:

1) excludeHI_data - has 432 data points with 12 variables each. The log transformed values for Clothiandin, imidacloprid, and thiamethoxam will be evaluated to predict the number of honey bee colonies and honey yield per bee colony .
2) nacetimaprid_data - has 355 data points with 8 variables each . The log transformed data for nacetimaprid will be evaluated to predict the number of honey bee colonies and honey yield per bee colony .
3) nathiacloprid_data - has 176 data points with 10 variables each. The log transformed data for thiacloprid and all neonic will be evaluated to predict the number of honey bee colonies and honey yield per bee colony .

Implementation

Determining correlations/skewness of data

To determine trends and amount of skewness in the data, I used matplotlib.pyplot to produce scatterplots of each pesticide application per number of bee colonies and honey yield per colony across all years. Visually then, I determined that the data showed no clear correlation of any specific pesticide application in terms number of bee colonies or honey yield per colony. I quickly determined the data were heavily skewed. After log transforming the data, I found that while skewness was removed from the data, visually there still were no clear correlations of any specific pesticide application in terms number of bee colonies or honey yield per colony.

To provide a more quantitative estimation of any correlations in the data, I took a the allneonic usage data and used seaborn.jointplot (Wascom, n.d.) to perform a simple linear regression of allneonic versus number of bee colonies (both on the non-log transformed and transformed data). The Pearsonf (Wikipedia - Pearson Correlation Coefficient, 2108) and probability values returned from the seaborn jointplots showed no strong correlation in either the non-log transformed data or log transformed data.

Normalizing the data

As the data are heavily skewed, it was not possible to perform the predictions without normalizing the data. To normalize the data, I chose to use numpy.log to calculate the natural logarithm of the pesticide usages as well as number of bee colonies and honey yield per colony. An example of my log normalizing statement for clothiandin is below:

```
excludeHI_data['log_nCLOTHIANIDIN'] = np.log(excludeHI_data['nCLOTHIANIDIN']+(0.5*minimum_nCOTHIANDIN_value_for_any_state))
```

Initially, I ran into a problem when calculating the logs for those remaining data points that had zero values for one of the pesticide usages. For those zero data points, the log function ran into a division by zero situation and errored out. So, for these zero data points, as I described above, I added one half of the least non-zero pesticide usage value found in any state to all the data points for that specific pesticide which allowed for the logarithmic transformations to complete successfully. (Stahel, 2008)

Splitting the data

The data to be evaluated are unique by state and year. There is only one data point for each feature for each state and year. When splitting, so as not to introduce bias for any state, I chose to stratify the splits by state so that training and testing sets contained data uniformly split across each state. I used sklearn.train_test_split to split the data An example of my split statement is below:

excludeHI_data_train, excludeHI_data_test = train_test_split(excludeHI_data, test_size=0.25, random_state=59, **stratify=excludeHI_data[['state']**])

When first running this split statement, I ran into issues with the South Carolina data for which there was only one data point. The split failed because it could not uniformly split the data into training and testing sets from only one data point; therefore, I decided to drop the South Carolina data point from the data set.

Running the regression algorithms

For each regression algorithm (KNeighborsRegressor, DecisionTreeRegressor, LightGBM, and SupportVectorRegressor) regressions were performed for each specific pesticide usage to predict the number of bee colonies and the honey yield per colony, I fit the model with the training data, predicted the regressors from the testing data, and then determined the r2 score of the predictions.

Summarizing the r2 scores

The r2scores found in the executions for each specific pesticide usage to predict number of bee colonies and honey yield per colony were collected in a table where the best performing r2 was identified for each of the features analyzed.

Refinement

In performing the analysis, using the clothiandin data to predict number of colonies from the excludeHI_data, I took the steps below to refine the models:

KNeighborsRegressor (baseline algorithm): In an attempt to refine this model, I tried different n_neighbors from one through 10 to use for the queries (KNeighborsRegressor(n_neighbors=x). All values of n_neighbors tried resulted in the same r2 score for the model.

DecisionTreeRegressor: To refine this model, I tried different values for splitter (‘best’,’random’) and criterion (‘mse’, ‘friedman_mse’, and ‘mae’) which all resulted in the same r2 score for the model.

Lightgbm: To refine this model, I tried different values for parameters (boosting_type, num_leaves, learning_rate) when training the model. The model performed best with boosting_type=’gbdt’, num_leaves=31, and learning_rate=.01). Changing the learning_rate had the biggest impact on the r2score: r2 = -0.002234109800022255 for a learning rate of 0.05 vs -0.000825042194555480 for a learning rate of 0.01.

SupportVectorRegressor: To refine this model, I tried to use various kernel types (‘linear’,’rbf’, ’sigmoid’,and ’poly’) when building the model. The best performing kernel was ‘linear’. R2 scores found were -0.14253744220411013, -0.14343232043819887, -0.14342992160542112, and never completed for ‘linear’,’rbf’, ’sigmoid’,and ’poly’ respectively.

IV. Results

Model Evaluation and Validation

In Tables 2a and 2b below, the r2 scores for each regression algorithm (KNeighborsRegressor, DecisionTreeRegressor, LightGBM, and SupportVectorRegressor) for each specific pesticide usage of the number of bee colonies and the honey yield per colony are presented. The best performing scores are highlighted in YELLOW.

Table 2a. r2 Scores for Each Regression Algorithm for Each Specific Pesticide Usage of the Number of Bee Colonies.

	Number of Bee Colonies					
	nCLOTHIANDIN	nIMIDACLOPRID	nTHIAMETHOXAM	nACETAMIPRID	nTHIACLOPRID	nAllNeonic
KNeighborsRegressor	-0.49921006	-0.310901863	-0.528572407	-0.04809638	0.382243083	0.617039342
DecisionTreeRegressor	-0.920527491	-0.780779305	-1.119874801	-0.542409798	-0.211912231	-0.49312502
LightGBM	0.002192813	0.079377781	0.026118548	0.158465136	0.294692253	0.142536545
SupportVectorRegressor	-0.001790117	0.064641303	0.043099155	-0.049791466	0.048911942	-0.046990546

Table 2b. r2 Scores for Each Regression Algorithm for Each Specific Pesticide Usage of the Yield of Honey Per Bee Colony.

	Honey Yield Per Colony					
	nCLOTHIANDIN	nIMIDACLOPRID	nTHIAMETHOXAM	nACETAMIPRID	nTHIACLOPRID	nAllNeonic
KNeighborsRegressor	-0.477376076	-0.541283035	-0.382270723	-0.339114404	0.431792068	0.574686196
DecisionTreeRegressor	-0.71783473	-1.121540727	-0.862464003	-1.014427081	-1.385148602	-0.460354643
LightGBM	-0.004258604	0.026831258	0.015146934	0.008409943	-0.041808539	0.091334261
SupportVectorRegressor	-0.02820708	-0.01331021	0.002947266	-0.007053852	-0.028891219	-0.015254388

It is important to note that none of the models predicted very well. The best r2 score of any regression performed was less than 0.62. Most r2 scores were below 0.1 and many were negative. Overall, the best performing model over all 12 regressions (6 each pesticide usage x number of colonies and yield per colony) was the lightGBM algorithm; however, there were specific differences in the best performing model by pesticide usage analyzed and number of colonies and yield per colony. Lightgbm had the best r2 scores for clothiandin, imidacloprid, and actamiprid usage to predict Number of Bee Colonies as well as the best r2 scores for clothiandin, imidacloprid, thiamethoxam, and actamiprid usage to predict Honey Yield Per Colony. SupportVectorRegressor proved to provide the best r2 score for predicting Number of Bee Colonies from Thiamethoxam usage; however, the r2 score was not very good. The r2 scores for predicting number of bee colonies and honey yield per colony were highest for the baseline KNNRegressor model than any other algorithm; the dataset used in this analysis contained the least number of data points of all the datasets used as so many states were missing these specific pesticide usage data.

To determine the robustness of one of the models, I chose to perform a sensitivity analysis using the lightGBM model of the predicted number of colonies based on acetamiprid usage for the state of Illinois. Table 3 below shows those actual versus predicted values:

Table 3. Actual Versus Predicted Values of Number of Colonies from Acetamiprid Usage for the State of Illinois Using the LightGBM model

		Number of Colonies	
Year	ACETAMIPRID (kg)	Actual	Predicted
2004	456.9	7000	19310
2005	74.2	8000	18988
2006	74.9	10000	18988
2007	73.1	9000	18988
2008	24.9	8000	22015
2009	123.2	8000	22502
2010	24.5	9000	22015
2011	98.2	7000	22342
2012	152.7	7000	20668
2013	178.9	7000	20103
2014	92.1	8000	22342

The sensitivity analysis above demonstrates how poorly the model is predicting the number of colonies basis acetamiprid usage. Again, none of the r2 scores for any model used to predict across the entire United States over all the years was very high. With that said, from the data as they are, the lightGBM model appeared to have the overall highest r2 scores. The limiting factor in the suitability of the model for these data is not the model but the data itself. There is just no strong relationship between pesticide usage and number of colonies or honey yield per colony when regressing the data in aggregate across all years and states. Due to the low r2 scores, none of these models can be trusted to predict either the number of colonies or the honey yield per colony from any of the pesticide usages.

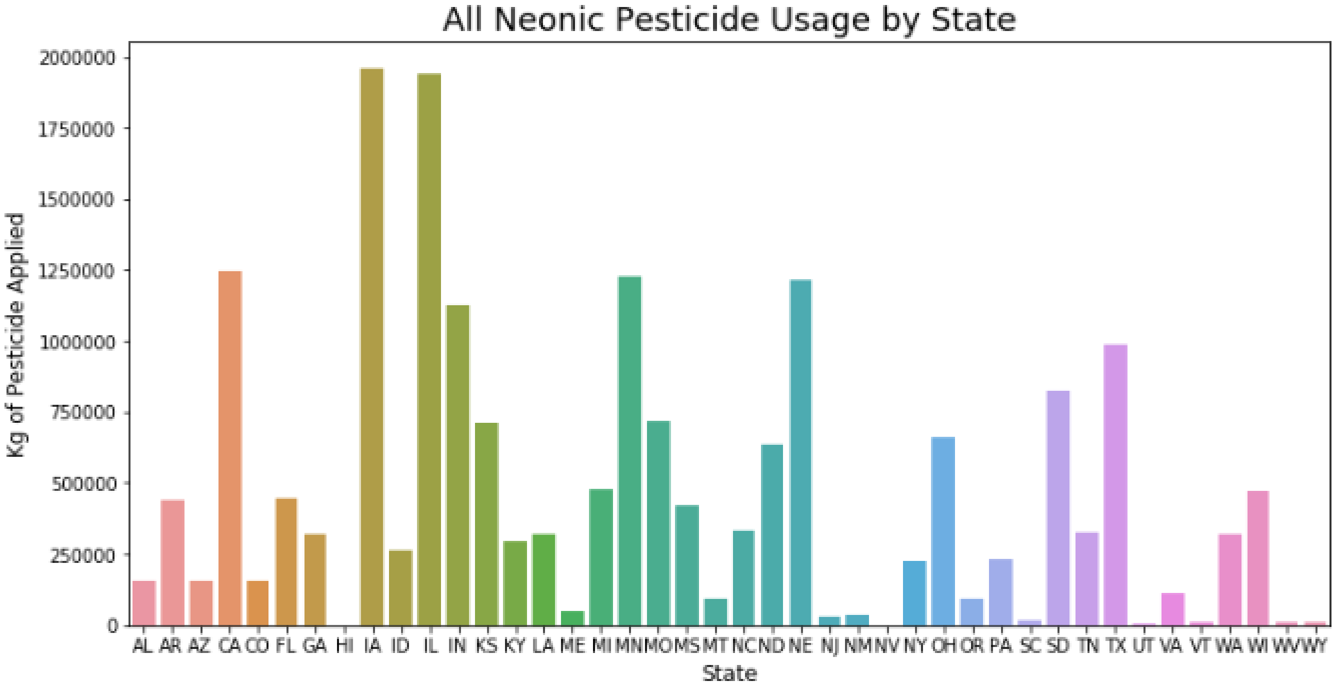
Justification

As mentioned above, none of the models predicted very well. When compared against the benchmark, not counting the thiacloprid data set which had 85% of its data dropped from the original dataset, in general the lightGBM model resulted in the highest r2 scores for the regressions. For only one pesticide usage (thiamethoxam) SupportVectorRegressor provided the best model when predicting number of bee colonies. Again, the limiting factor here are data. There is just too much variability in the data to make any significant predictions that can be trusted.

V. Conclusion

Free-Form Visualization

Figure 6. Graph of All Neonic Pesticide Usage across all years from 2004 through 2014 by State.



I provided the graph above to show how much variability there is in the pesticide usage across states. The midwestern states appear to have the highest usage of pesticide while the southwest states appear to have the lowest usage. From this graph, along with the graphs provided in figure 2 above, there is tremendous variability in pesticide usage as well as number of bee colonies and honey yield per colony per year. This is important to consider when evaluating any prediction model for these data as none of the coefficients of determinations were very high. The data collected for this investigation are purely empirical meaning that no other factors such as weather, specific agricultural practices by state, etc. are not available to consider in the analysis.

Reflection

To recap, in this investigation, the NASS/USGS data were loaded into a dataset. The data were then inspected to determine the presence of missing data. Next the data were preprocessed to handle the missing data that were found. Scatterplots of the data were completed to determine if there were visible trends in the data or if the data were skewed. As the data were skewed, the data were log transformed to normalize the data and scatterplots were created for these normalized data. Next the data were split into training and testing sets. The Supervised Learning models (DecisionTreeRegressor, LightGBM, and SupportVectorRegressor) were evaluated against the modelled features. Finally, the r2 scores of the feature predictions were compared for each supervised learning model evaluated against the baseline KNeighborsRegressor r2 scores to determine the best prediction model for these data.

The exploratory investigation of the data used in this project was very interesting. It was interesting to see the great variability of data by state and year for the various pesticide usages as well as the bee colony/honey production metrics. It was also interesting to observe that these data were presented as a time series.

As for the challenging aspects of this project, I found the time series nature of the data to be new to me in terms of how to deal with year as a factor in the analysis. So how should the data be analyzed? By year, by state, by region, in aggregate? For this effort, I chose to look at the data in aggregate; however, it may make sense to later look at the data by the other variables. Also, logically, can these specific pesticide usages be analyzed independently of the others? Are they truly independent or is there some interaction between them in terms of the predictions that are trying to be made. Multiple regression may be more appropriate that the single regressions that were performed. As the data are empirical only, is there confounding of the data by other variable that are not present in the dataset (for example rainfall, temperature, crops planted, etc.). Ultimately, the data itself presented the biggest challenge in this investigation.

Handling missing data was I bit of a challenge as I need to find some accepted practice for handling the missing data which I found through research on the web. Also, I did spend a lot of time thinking about an appropriate splitting strategy. It did not seem correct to just split the data randomly across the entire dataset. I was concerned that using just a random split could introduce bias by state or year. As the data are unique by date and year, I could not stratify the splits using both elements. I chose to stratify the splits by state to remove any state bias. Initially, this caused a problem in that my initial split stratified by state failed because there was only one data point for South Carolina. After dropping that data point the stratified splits by state were successful.

As for the final result of this study, none of the simple regressions performed were very good at predicting the number of bee colonies present or the honey yield per colony. For me this was a disappointment; however, it does seem logical because of the high variability in the data and the fact that these data are empirical in nature. More study of this dataset as well as additional datasets through a multiple regression approach
Are needed before any useful predictions can be made. The single regressions using the lightGBM or any other model used in this study of these data in aggregate are not useful to make predictions.

Improvement

To improve this study the one part of my implementation that I would improve is to analyze the data by specific region of the United States rather than analyze the data in aggregate across the entire United States. In evaluating by region, I also would have experimented with a custom ensemble model combining those that I have tried. While I have never created a custom ensemble model, I plan to experiment with one later at some point in time.

Again, in looking at these data in aggregate, the slight improvements of the other models gained from the benchmark model do not provide for any strong predictor. More data and/or different ways to group these data for analysis need to me made to determine stronger trends.

VI. References

Aydore, S. (2015, February 15). *Mathematics and Machine Learning - What is the difference between squared error and absolute error?* Retrieved from Quora : https://www.quora.com/What-is-the-difference-between-squared-error-and-absolute-error

Bhattacharyya, I. (2018, June 29). *Support Vector Regression or SVR*. Retrieved from Coinmonks - Support Vector Regression: https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf60ff

Greenpeace. The Role of the Bee, 2014. (2014). *The Role of the Bee*. Retrieved from Greenpeace: https://sos-bees.org/situation

John Paul Mueller, L. M. (n.d.). *Dummies - Identifying Missing Data for Machine Learning*. Retrieved from Dummies: https://www.dummies.com/programming/big-data/data-science/identifying-missing-data-machine-learning/

Kurita, K. (2018). *LightGBM and XGBoost Explained*. Retrieved from Machine Learning Explained: mlexplained.com/2018/01/05/lightgbm-abd-xgboost-explained/

Microsoft - Lightgbm. (n.d.). *Welcome to LightGBM's Documentation!* Retrieved from LigntGBM: https://lightgbm.readthedocs.io/en/latest/

Sadawi, N. (2014). *Regression with Decision Trees*. Retrieved from YouTube - Regression with Decision Trees: https://www.youtube.com/watch?v=nSaOuPCNvlk

Sayad, S. (n.d.). *Decision Tree Regression*. Retrieved from An Introduction to Data Science: www.saedsayad.com/decision_tree_reg.htm

SciKitLearn - DecisionTreeRegressor. (n.d.). *sklearn.tree.DecisionTreeRegressor*. Retrieved from SciKitLearn: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

SciKitLearn - SVR. (n.d.). *sklearn.neighbors.KNeighborsRegressor*. Retrieved from SciKitLearn: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

SckiKitLearn - sklearn.svm.SVR. (n.d.). *sklearn.svm.SVR*. Retrieved from SciKitLearn: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

Srivastava, T. (2016, February 16). *7 Important Model Evaluation Metrics Everyone Should Know*. Retrieved from Analytics Vidhya: https://analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/

Stahel, W. A. (2008). Statistical data analysis. An introduction to natural scientists 5th Ed. In W. A. Stahel, *Statistical data analysis. An introduction to natural scientists 5th Ed.* ISBN 978-3-8348-0410-5. Retrieved from https://translate.google.com/translate?sl=auto&tl=en&js=y&prev=_t&hl=en&ie=UTF-8&u=https%3A%2F%2Fstat.ethz.ch%2F%7Estahel%2Fstat-dat-ana%2F&edit-text=

Wascom, M. (n.d.). *seaborn.jointplot*. Retrieved from seaborn: statistical data visualization: https://seaborn.pydata.org/generated/seaborn.jointplot.html

Wikipedia - Coefficient of Determination. (n.d.). *Wikipedia - Coefficient of Determination*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Coefficient_of_determination

Wikipedia - Colony Collapse Disorder. (2018). *Colony Collapse Disorder*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Colony_collapse_disorder

Wikipedia - Neonicotinoids. (2018). *Neonicotinoids*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Neonicotinoid

Wikipedia - Pearson Correlation Coefficient. (2108). *Pearson correlation coefficient*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Wikipedia. (2018). *Wikipedia - ID3 Algorithm*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/ID3_algorithm

Zmith, K. (n.d.). *vHoneyNeonic_v03.csv, Honeybees and Neonic Pesticides.* . Retrieved from Kaggle: https://www.kaggle.com/kevinzmith/honey-with-neonic-pesticide