COVER PAGE

To bee or not to bee: Colony collapse in the US

Link to GitHub repository: https://github.com/mattlampl/bees_knees

Contributions: Each member of our team contributed equally to this project; there were no imbalances in time or energy contributed. Below, we detail each group member's primary tasks and responsibilities.

	Primary Task 1	**Primary Task 2**	**Primary Task 3**
Kai	Managed team and completed general project management work, ensuring timeliness in meeting project deadlines.	Cleaned APHIS dataset and created 10 feature variables. Identified and implemented appropriate techniques to handle data missingness, a key issue.	Led Ridge and Lasso models. Set model parameters and assessed feature importance and more.
Irfan	Created the data and file structure in Google Collab and coached team members on platform usage.	Led initial exploratory data cleaning such that the team could create baseline models to iterate on.	Led predictive modeling efforts, including decision trees and random forests.
Matt	Cleaned Honey Neonic dataset. Decided how to handle categorical values and impute values for nulls.	Identified composite primary key to merge all cleaned datasets and merged all cleaned datasets into one file for model building, training, and finetuning.	Led decision trees and random forest. Set model parameters and assessed feature importance and more.
Jack	Created modeling plan for Monte Carlo simulation and documented research process in proposal and modeling plan.	Completed initial exploratory data analysis and created visualizations and plots to understand the distribution of the data	Cleaned Air Quality and Urbanicity datasets, created previous-year and 5-year-rolling-average features for AQ data
Sara	Found and cited the majority of the data sources the team used for the project.	Completed exploratory data analysis and created plots of honey yield across time to understand trends in data.	Led team in writing proposal and final report, and creating data dictionary.

Resources: We used the following resources to complete our project. We linked to code support pages (e.g., Stack Exchange, Toward Data Science, etc.) as endnotes in the report for easy access.

Resource	**Description**
Stack Exchange	Used to resolve coding issues while cleaning, merging, and modeling data.
SK Learn, Seaborn, MatPlotLib, Numpy, Math	Used to create machine learning models, create charts, and more. Reference our GitHub repository for all libraries used to complete this project.
Instructor	Reached out to Dr. Peter Zhang with questions, attending his office hours

Problem Statement

Over the past few decades, declining numbers of honeybees have been an increasing cause for concern. While honeybees are not native to the US,[1] the American food system relies heavily on them. In fact, honeybees pollinate roughly one-third of all food eaten by Americans, ranging from pumpkins to cranberries to almonds.[2] Moreover, honey yields have declined in recent years, with US average yields declining by 23.97% from 1998 to 2017, which may suggest a concurrent decline in overall honeybee fitness (See Figure 1, where the red line is the US average).

Figure 1: Honey yields per colony: Declines across all regions in US

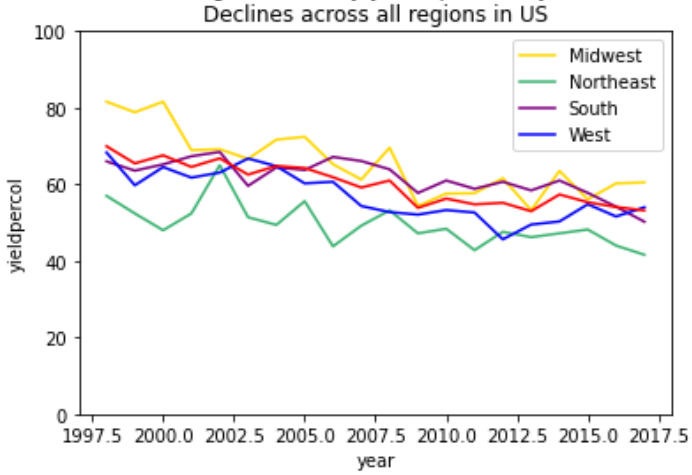
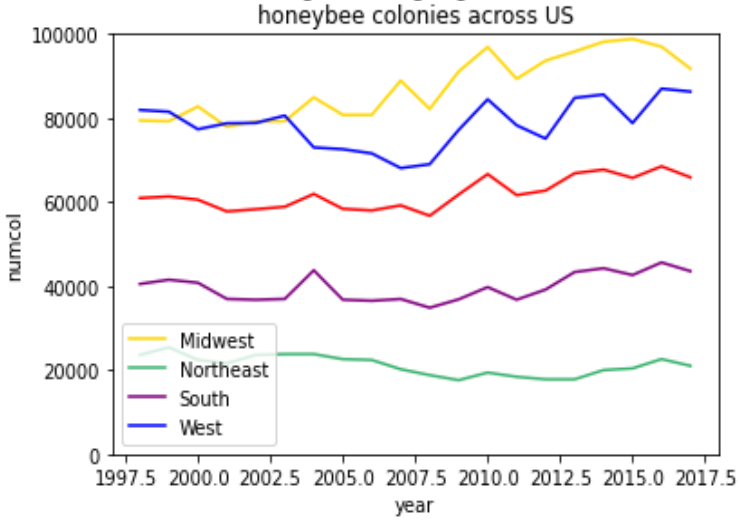


Figure 2: Slight gains in honeybee colonies across US



While the average number of honeybee colonies has not declined over the past twenty years — in fact, there are slight increases — beekeepers can offset colony losses by creating new colonies from surviving ones (See Figure 2, where the red line is the US average).[3] In turn, colony replacement can reduce the apparent loss of honeybee colonies in year-to-year comparisons, making trends appear stabler than perhaps they are. Indeed, from April 2021 to April 2022, beekeepers participating in the National Honeybee Survey reported that they lost 39.0% of their colonies, a figure which does not account for colony replacement.[4]

There are many potential causes for honeybee decline. First of note is Colony Collapse Disorder (CCD), in which the "majority of worker bees in a colony disappear and leave behind a queen, plenty of food and a few nurse bees to care for the remaining immature bees and the queen."[5] CCD surfaced as a major cause of concern in 2006-2007, when a group of beekeepers noted that 30 to 90% of their honeybees had disappeared.[6] The disappearance of worker bees spells disaster for hives: unable to support themselves, the remaining honeybees eventually die.[7]

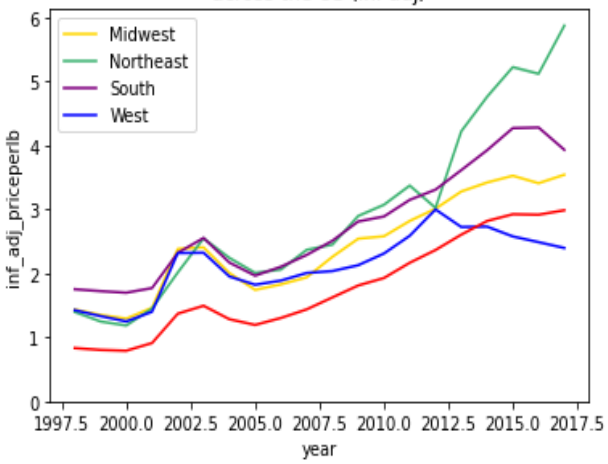
Other factors that contribute to honeybee decline include the application of certain types of pesticides, as well as the presence of parasites.[8] Neonicotinoids, common pesticides used in farming, can negatively impact honeybees' "motivation to initiate foraging, amount of nectar collected, and initiation of

subsequent foraging bouts."[9] Further, neonicotinoids are known to interfere with honeybees' memory,[10] and studies have shown that they can harm fertility rates for male bees, reducing their lifespan as well as their quantity of viable sperm.[11] In addition, the varroa mite — a honeybee parasite — is a significant threat to honeybee colonies: they "weaken bees by feeding on their fat bodies, organs vital for metabolism, and the immune system."[12] The magnitude of the mites' damage is huge: honeybee hives in North Carolina have fallen by 44% since 1990, when the mites were first observed.[13]

Another factor impacting honeybees is "poor forage," which occurs when honeybees cannot find proper sources of food and habitat due to low diversity in local flora.[14] Poor forage is of particular concern in states where most arable land is covered by monoculture crops, such as corn and soy.[15] In a field trial in Iowa, researchers found that fertility rates of queen bees tended to be lower when caged bees were fed the dietary mixes of bees that were kept in soy or coy fields, as compared to prairies.[16]

Further, climate change may contribute to honeybee decline. One study noted that climate change may create favorable habitats for invasive species that predate on honeybees.[17] Another group of researchers ran a field experiment, in which they simulated the warmer and wetter weather patterns likely to be induced by climate change in Northern Europe.[18] They found that climate change was associated with a 40% decrease in wildflowers. In turn, this could force pollinators — including honeybees[19] — to turn to less nutrient-rich flowers, thereby reducing honeybee health and fitness.[20]

Figure 3: Price per pound of honey across the US (inf-adj)



In summary, while the number of honeybee colonies has remained relatively stable over the past twenty years, honey yields per colony have decreased by nearly a quarter. The potential causes of this decline are many, among them CCD, pesticides, parasites, poor forage, and climate change. Honeybee productivity declines have significant social and economic impacts: honeybees are crucial pollinators, without which one-third of the American diet would vanish; and honey is an important agricultural product, the price of which has more than doubled in the past twenty years (See Figure 3, in which the red line is the US average). Recognizing the importance of this, we analyze the factors that shape honeybee productivity decline via predictive models and a Monte Carlo simulation. In so doing, we identify important risk factors and potential policy interventions that can ensure healthy hives — and protect the food system.

Data Summary and Limitations

We used the datasets below to complete our analyses. Please reference Table 1 in the appendix for information on the features derived from each source, data source identification process, etc. We cleaned each dataset individually and merged them, using state-year combinations — the unit of analysis for this

project — as composite primary keys. We then exported our cleaned, merged dataset to a CSV to complete our prediction and simulation tasks.

Source	Description
APHIS Dataset	Sourced from the National Honey Bee Survey, this dataset contains feature data on factors that can impact honeybee populations, including levels of varroa mites, Nosema spores, and more.
Honey Neonic Dataset	Sourced from Kaggle, this dataset includes our target variables (number of honeybee colonies and honey production levels) as well as feature variables, such as levels of exposure to various toxins that harm honeybee populations.
Urbanicity Dataset	Sourced from the US Census Bureau, this dataset contains information on the level of urbanicity per state across decennial census years to enable us to see whether states with higher levels of urbanization have impaired honey production levels.
Air Quality Dataset	Sourced from the Environmental Protection Agency, this series of datasets contains county-level information on air quality, such as the number of days in a year with good air quality, hazardous air quality, and levels of PM 2.5 and PM10.
Temperature Datasets	Sourced from NOAA, these datasets contain temperature data for states from 1895-2022.

Based on our initial research, we identified many important feature variables, including pesticide levels, climate and weather conditions, urbanicity levels, air quality, and the presence of honeybee pests. We created these features using the datasets above, sourcing our target variable (honeybee yield) from the Honey Neonic dataset from Kaggle. We also engineered additional feature variables from these datasets, such as previous-year pesticide levels, as past events or trends may inform future outcomes.

There are important limitations with regard to the data. First, the unit of analysis for our project is at the state-year level (e.g., honey yield in Illinois in 1999). However, honeybee colony viability is a more local phenomenon: the "quality and quantity of nectar and pollen"[21] often determine the success of hives, the availability of which is dictated by the local presence of flowers, grasses, etc.. While we include a state-level urbanicity variable as a proxy for local availability of nectar and pollen, we acknowledge that this is a weak substitute. That is, a honeybee colony could be located in a rural area — lush with wildflowers and swaying prairie grasses — while concurrently being located in a highly urbanized state.

Ultimately, we chose state-year level analysis because of data availability issues and time constraints. While geographic granularity is ideal, our primary hurdle lay in the fact that our target variable — honey yield — was measured at the state-year level. Thus, while we identified feature variables that offer more geographic granularity — for example, at the county level — this nuance was lost in merging to the Honey Neonic dataset. Moreover, identifying, cleaning, and merging the six datasets below was a time-intensive process. We found datasets that enabled us to engineer a rich array of feature variables, but given the time constraints of the ten-week project, were not able to account for all of the factors that can impact honeybee decline, such as crop production levels. Given more time, we would incorporate

additional datasets to engineer more features, as well as work to identify a dataset that lends to enhanced geographic granularity (e.g., at the county level), to better model honey production.

Implementation Details & Assumptions

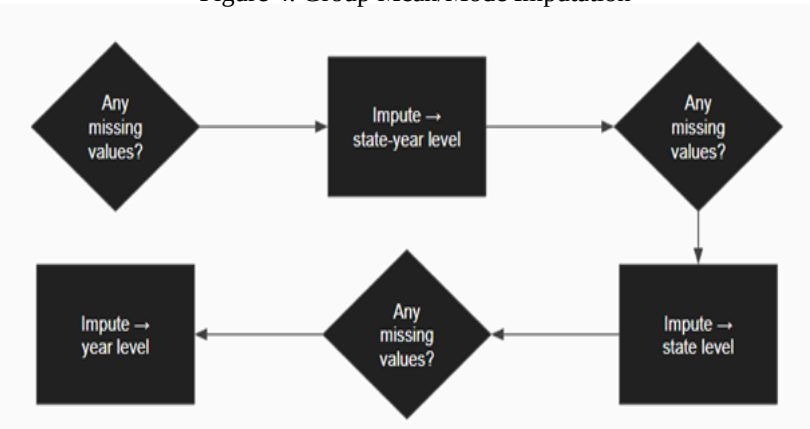
As previously stated, we decided to conduct an analysis at the state-year level. This choice has an implication on the primary assumption of this project, which is that the honeybee phenomenon may be generalized at the state level within annual basis. With this assumption, we may apply inferential statistics to this domain.

The following is a summary of the granularity adjustments for each dataset:

	Granularity Adjustment
APHIS	Value of features were summarized on its year-based (group by)
Urbanicity	The urbanicity data were originally provided at the decade level. For years up to and including 2010, state-level urbanicity data from 2000 were used. For 2011 and later years, state-level urbanicity data from 2010 were used.
Air Quality	Value of features were summarized on its state-based (group by)

After all the datasets had the same level of granularity, we combined them by using a left-join on the Honey Neonic dataset. This dataset weas chosen as the anchor in the merging process because it contains the target variable of the planned prediction model: honey yield per colony. As a result of the left join, there are some missing values in the data (also carried from the data conditions prior to the join). To handle missing values, two methods are used: (1) mean imputation for features with continuous values, and (2) mode imputation for features with binary/categorical values. Instead of imputing the missing value with the mean/mode of each feature, the mean/mode was taken from the most specific peers gradually to the most general one.[22] Figure 4 is the overall imputation strategy used in this project:

Figure 4. Group Mean/Mode Imputation



There were no missing values in the data after three phases of imputation. It should also be noted that by employing this imputation approach, we assumed that each feature in the model has a consistent average

value in each state every year. Once all missing values have been handled, we converted the feature state to dummy variables. This operation is performed with the use of a one-hot encoding, which generates a number of binary variables for each state.[23]

Other model implementation details include normalizing the data for our LASSO and Ridge models. After researching when to normalize the data,[24] we decided to first split the data into train-test sets, normalizing the sets separately. The intuition behind this is that we do not want to expose the train set to the test set prior to building our models, inadvertently biasing the normalization toward the test set.[25]

Analytical Formulations

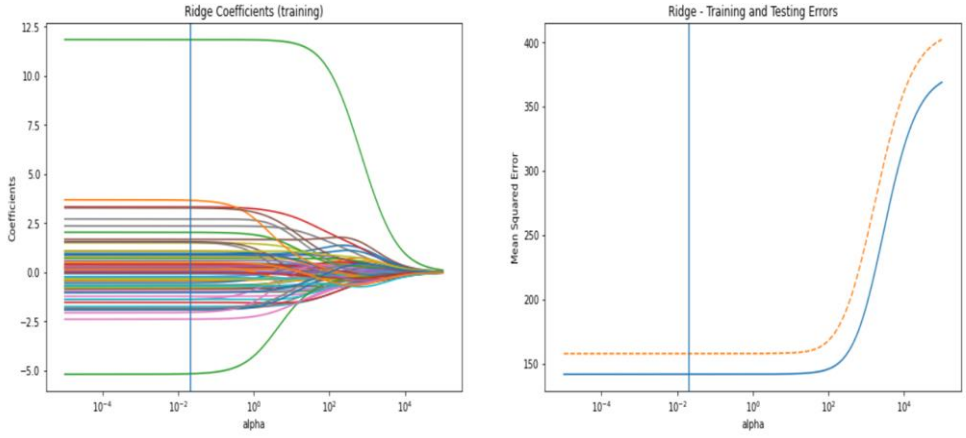
In order to create predictions on the yield per colony, two main machine learning algorithms were use: Linear Regression and Tree Based Methods. For the linear regression models, two popular regularization techniques — Ridge and LASSO — were employed.[26,27] For the tree-based methods, decision trees, random forests, and gradient boosted trees were trained. For all machine learning models, the main dataset was split into a train and test sets. Because of the temporal nature of the prediction problem, the training set was all data recorded prior to 2016, and the test set was all data recorded in 2017 and 2018. Using this method, we were assessed how well our models were able to predict the future.

Regularized Regression Methods

We first attempted to predict per colony honey yield using regularized regression methods, specifically Ridge and LASSO. To tune the parameters, we considered 200 separate alpha values spaced evenly between -5 and 5. We then fit the model to the training set using each alpha and compared the training and testing mean squared error (MSE). We picked the alpha with the lowest train MSE as the optimal parameter. We visualized the coefficients moved towards zero in Figure 5 below.

For our Ridge model, we found that the optimal alpha was 0.021 and was associated with a train MSE of 141.92 and a test MSE of 157.94. The most important features for this model were the state's previous year's yield (positive relationship) and a five-year average median of the state's Air Quality Index (AQI) (negative relationship).

Figure 5. Ridge Parameter Tuning

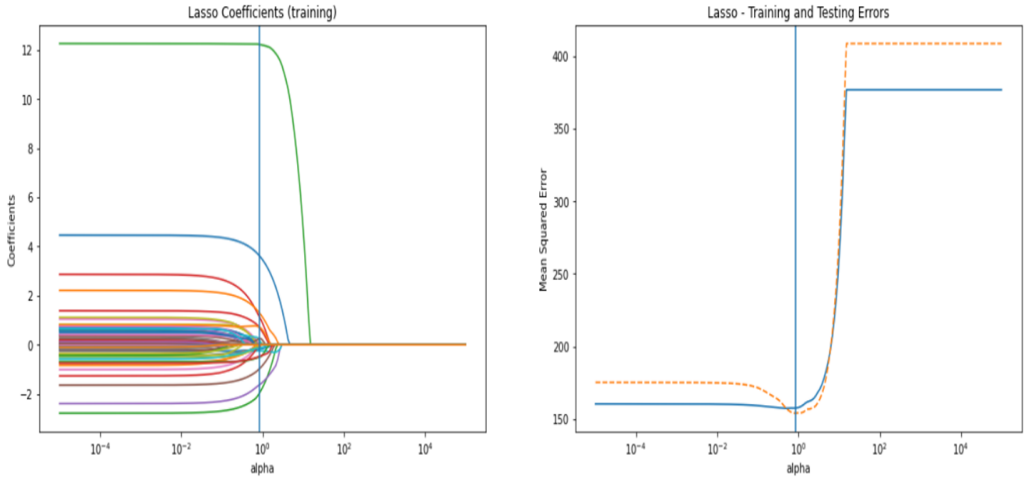


The LASSO model did best with an alpha of 0.84 and was associated with a train MSE of

157.52 and a test MSE of 153.93. The most important features for the LASSO model were the state's previous year's yield (positive), the number of colonies (positive), and the usage of two Neonicotinoids, Imidacloprid and Acetamiprid, both of which have a negative relationship with per colony honey yield.

While the Ridge model had a lower train MSE, the LASSO model had the smaller test MSE, meaning that the LASSO regularization slightly outperformed the Ridge on unseen observations

Figure 6. LASSO Parameter Tuning



Linear regression cost function:[28]

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} (y_i - \sum_{j=0}^{p} (w_j * x_{ij}))^2$$

Ridge regression cost function:[29]

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} (y_i - \sum_{j=0}^{p} (w_j * x_{ij}))^2 + \lambda \sum_{j=0}^{p} (w_j)^2$$

LASSO regression cost function:[30]

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} (y_i - \sum_{j=0}^{p} (w_j * x_{ij}))^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Tree Based Methods

We determined that tree-based methods performed the best in terms of predicting the yield per colony. A multitude of algorithms were trained including a simple decision tree regressor, random forest regressor, and gradient boosted trees. The latter performed the best in terms of minimizing the Mean Squared Error for both the training and test set. The Gradient Boosted Tree was fitted according to the method described by Jerome Friedman et. al. in the paper *Greedy Function Approximation: A Gradient Boosting Machine* (Equation 1).[31] [32] We found that a gradient boosting algorithm with 100 estimators and a learning rate of 0.1 outperformed any random forest regression, while also having a tighter fit to the distribution of the actual data, without overfitting (Figure 7).

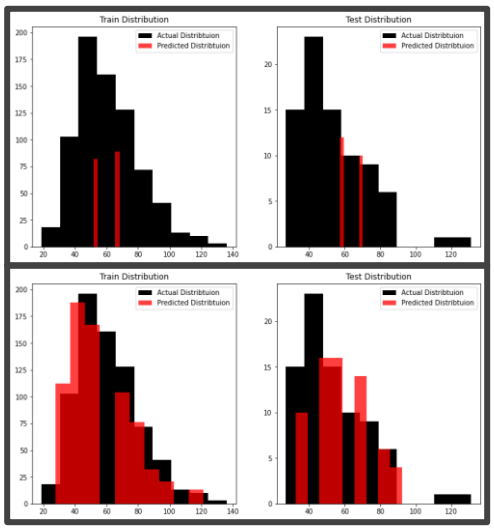


Figure 7: A random forest with 5000 trees (top) compared to a gradient boosted forest with 100 trees (bottom)

The trained Gradient Boosted Trees model's most important features were in line with those of the linear regression models. The previous year's yield was the most important feature, while the use of pesticides, 5-year rolling averages of air quality indicators, and bee colony viruses were also in the top 20 predictors.

Algebraic Formulation:

1) Initialize model with the mean value of yield per colony

$$F_0(x) = \frac{1}{n}\sum_{i=1}^{n} \quad y_i$$

2) For $m = 1$ to M ($M = 100$):

$$F_{m+1}(x_i) = F_m(x_i) + h_m = y_i$$

Therefore, each estimator h_m is fit to the residual $y_i - F_m(x_i)$

3) The final prediction is calculated by summing all trees multiplied by the learning rate (λ):

$$\hat{y}_i = F(x_i)_0 + \sum_{m=1}^{100} \quad \lambda F(x_i)_m$$

Simulation Models

To better understand the dynamics implicit in our best-performing model, we used the gradient-boosted forest described above as the basis of a Monte Carlo simulation. That is, we randomly generated a large amount of synthetic feature data and used it as the input data for our model. This analysis builds on past work by scientists at the US Environmental Protection Agency who have used simulation methods to probe sensitivities of different predictors in models of honeybee agriculture.[33]

Our synthetic data were sampled from uniform distributions over the observed range of each continuously valued feature, and we constructed these distributions by state. As a result, the simulation generates data only within the state-by-state ranges of values previously observed. Given the growing instability of the global climate, this assumption may result in an artificially narrow range of outcomes being produced.

However, we do not believe that this threatens the utility of our analysis, because we are not trying to simulate predictions. Rather, we are trying to understand feature relationships *within* the model. In other words, we are sacrificing future applicability for a better understanding of the current model's assumptions.

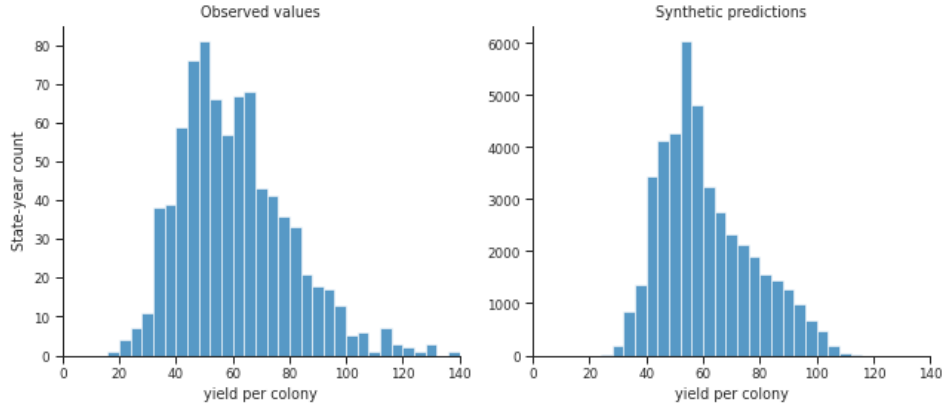


Figure 8. Annual yield per honeybee colony distributions. The left panel shows real state-year honeybee per-colony yields, and the right panel shows state-year per-colony yields predicted by our gradient-boosted tree model on synthetic data.

Figure 8 compares the observed distribution of state-by-state yield per honeybee colony to our simulated distribution. Both distributions are centered slightly below 60 with a slightly longer tail on the high end of yield. Our simulation produces a narrower range of outcomes than what is observed in real data, but our simulation produced 44,000 observations versus only 825 real observations.

We then compared two subsets of our simulated data based on predicted outcome: top 10% of state-year observations by honeybee yield, and bottom 90% by the same criteria. We ranked features by the portion of the global range of each feature that is accounted for by the difference of the means of these two groups. As expected, previous year yield is the most significant feature in absolute terms, with the 90/10 intergroup mean difference accounting for 38% of the total range. The next two most important features were location-based: the intergroup mean difference figures for Hawaii and Louisiana are 22% and 18% respectively, while the same figure is -10% for the Northeast region. This is the sort of finding that might be missed by only looking at a tree-based model's feature importance because only a small proportion of the observations, around 5%, are gathered from Hawaii and Louisiana. Thus, the features are not "important" to the model while still being strongly associated with high per-colony yield. Lastly, and perhaps surprisingly, previous-year pest levels were associated with high-yield predictions, and pesticide use was associated with low-yield predictions.

Analysis & Future Work

As our findings demonstrate, the previous year's average honey per colony is a crucial predictor variable in all of our models. It should be unsurprising that past-year yields are important predictors of current trends; the past is often indicative of the future. However, other important variables point to the impact of environmental factors in modeling honey yield per colony, in particular the median 5-year Air Quality Index and the usage of certain neonicotinoid pesticides, in particular Imidacloprid, Acetamiprid, and Thiacloprid. Overall, our predictive models suggest that honey hives tend to follow previous years' trends

and that factors like air quality and pesticide usage seem to have the largest impact on honey yields. As such, policymakers may want to consider implementing local pollution and pesticide caps in municipalities or counties with high levels of honey production.

Our simulation suggests that location is a first-order consideration for honeybee colony yield. Specifically, warmer and wetter climates like those found in Hawaii and Louisiana host colonies that our model predicts will produce the highest yield. As of now, these location effects appear to be more important for high per-colony yield than the effects of either pests or pesticides. Regarding pest levels, our simulation found that higher previous-year spore levels were associated with higher current-year yields, Our simulation cannot itself test hypotheses about why this might be, but this finding suggests that the lasting effects of a high-spore year may be attenuated by survivorship effects in future years. Our simulation also confirmed the importance of air quality found in predictive analyses.

While we unearthed important insights in this analysis, there is future work yet to be done. First, more granular data are required. As explored above, honeybee colonies are highly influenced by local phenomena. However, we studied the topic at the state-year level (e.g., honey yields in IL in 2001) because the honey yield data were aggregated to this level. If researchers had access to data with deeper geographic granularity (at the county or metropolitan area level) and with deeper temporal granularity (e.g., monthly or weekly observations instead of annual aggregations), they could pinpoint with greater precision which variables — weather, parasites, climate change, poor forage conditions, etc. — were associated with decreased yields. In turn, this would enhance the specificity of the policy implications of this report and could enable us to propose more geographically targeted interventions.

In addition, as the impact of climate change becomes more acute, future analytical work in this area should account for changing weather patterns and extreme weather events. As explored above, studies have shown that climate change may decrease the diversity of local flora[34] and create more favorable conditions for invasive species that predate on honeybees.[35] Thus, it is integral that researchers track and analyze how climate change impacts honeybees, a species that is itself integral to the US food system — and national food security.

Appendix

TABLE 1: Full Data Dictionary

Source	Link	Description	Discovery process	Variables
APHIS	Short link to data here. *Note that you must create a free online account to download the data.*	Sourced from the National Honey Bee Survey, this dataset contains feature data on factors that can impact honeybee populations, including levels of varroa mites, Nosema spores, and bee viruses. We created counts for continuous variables (number of spores, mites). For binary variables, we took the MAX as well as calculated the probability of the presence of a given bee virus for different samples from the same state-year. Variables that end in "v" are binary variables indicating the presence of a given bee virus.	Sara found this dataset in an online search for honeybee data. In addition to providing helpful feature data, the site was helpful in researching factors that impact honeybee colonies.	**Continuous variables**: Number of varroa mites, Number of spores, Previous-year mites, Previous-year spores. **Categorical, binary variables**: CBPV, DWV, IAPV, KBV, SBPV, ABPV, DWV_B, LSV2, MKV, CBPV_Prob, DWV_Prob, IAPV_Prob, KBV_Prob, SBPV_Prob, ABPV_Prob, DWVB_Prob, LSV2_Prob, MKV_Prov
Honey Neonic	Short link to data here.	Sourced from Kaggle, this dataset includes our target variables (number of honeybee colonies and honey production levels) as well as feature variables, such as levels of exposure to various toxins that harm honeybee populations.	Sara found this dataset when the team was researching potential project directions.	**Target variable:** Honey yield per colony **Continuous features:** Number of colonies, region, Clothianidin levels, Imidacloprid levels, Thiamethoxam levels, Acetamiprid levels, Thiacloprid levels, Previous-year levels of the aforementioned variables, North Dakota yield, and Previous-year honey yield

Urbanicity	Short link to data here.	Sourced from the US Census Bureau, this dataset contains information on the level of urbanicity per state across decennial census years.	Sara found this dataset online because urbanicity may be an important feature in modeling honeybee populations.	**Continuous features:** Percentage urban land per state
Air Quality	Short link to data here. *Note that you must download the CSVs separately for each year.*	Sourced from the Environmental Protection Agency, this series of datasets contains county-level information on air quality, such as the number of days in a year with good air quality, hazardous air quality, and levels of PM 2.5 and PM10.	Sara found this dataset online because air quality could impact honeybee viability.	**Continuous features:** Previous number of days: - Air quality issues, good days, moderate days, unhealthy for sensitive groups days, nhealthy days, very unhealthy days, hazardous days, Previous-year: - Max AQI, 90th percentile AQI, median AQI, days CO, days NO2, days with ozone issues, days PM2.5, days with PM10, 5 year average number of days: - With AQI, good days, moderate days, days unhealthy for sensitive groups, unhealthy days, very unhealthy days, number of hazardous days, 5 year MAX: - AQI, 90th percentile AQI, median AQI, avg. days CO, avg. days NO2, avg. days ozone, avg. days PM2.5, avg. days PM10
Statewide Time Series	Short link to data here.	Sourced from NOAA, these datasets contain temperature data for states from 1895-2022.	Irfan found this dataset in an online search for temperature-related feature variables	Continuous variables: Average temperature, Anomaly levels

COEFFICIENT FIGURES: See the feature importance and associated coefficient value for the Ridge and LASSO models below. Note that the data were normalized, so the coefficients cannot be interpreted in their original units.

Figure A1. Ridge Coefficients

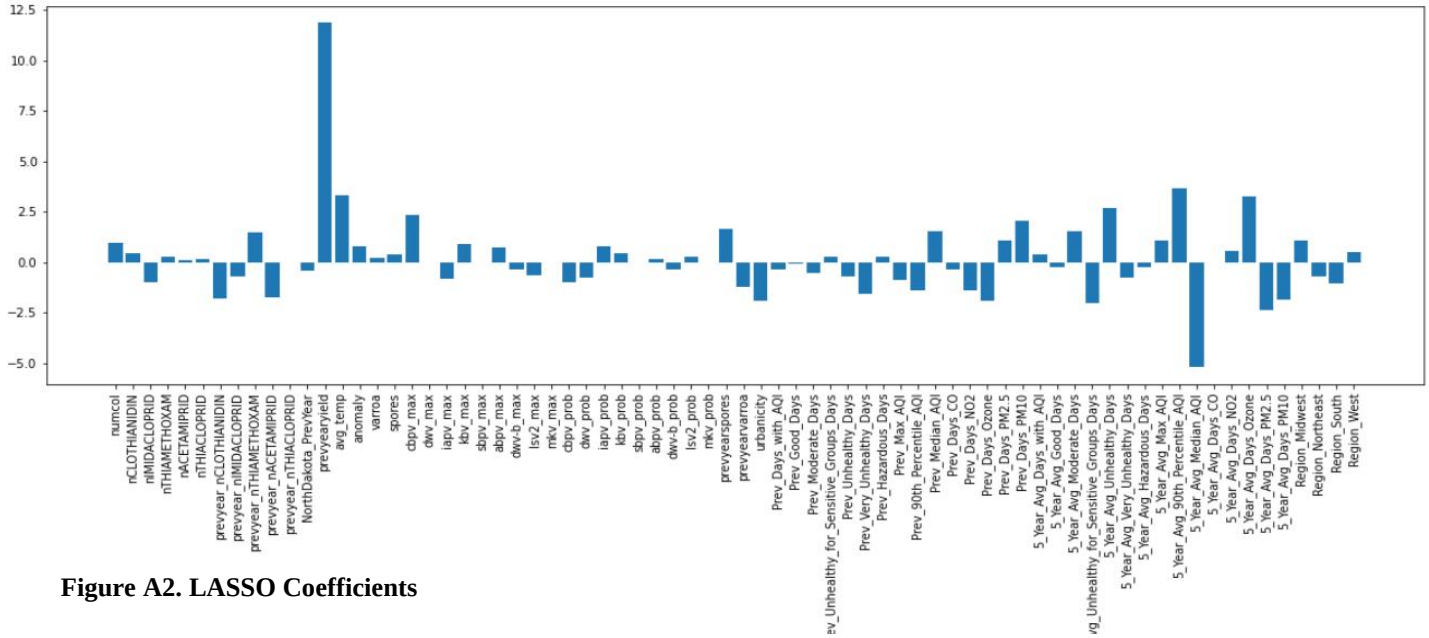
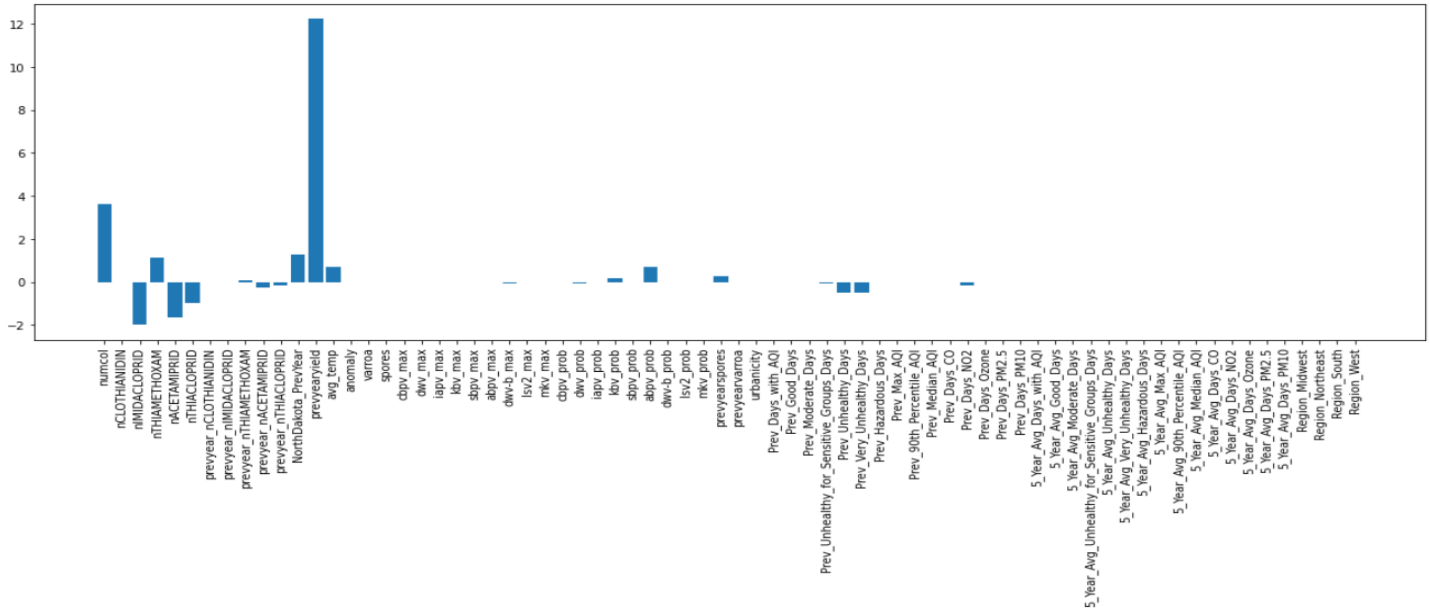


Figure A2. LASSO Coefficients



ADDITIONAL IMPLEMENTATION DETAILS: In this report, we detailed assumptions and imputations that we made. To access the code we wrote to complete our analyses, you can fork and download our GitHub repository with the relevant code here**:** https://github.com/mattlampl/bees_knees

[1] *Are Honey Bees Native to North America? | U.S. Geological Survey*. https://www.usgs.gov/faqs/are-honey-bees-native-north-america#:~:text=Honey%20bees%20are%20not%20native,and%20265%20pounds%20of%20nectar. Accessed 28 Nov. 2022.

[2]Medicine, Center for Veterinary. "Helping Agriculture's Helpful Honey Bees." FDA, Mar. 2021. www.fda.gov, https://www.fda.gov/animal-veterinary/animal-health-literacy/helping-agricultures-helpful-honey-bees.

[3] The Bee Informed Team. "United States Honey Bee Colony Losses 2021-2022: Preliminary Results from the Bee Informed Partnership." *Bee Informed Partnership*, 28 July 2022, https://beeinformed.org/2022/07/27/united-states-honey-bee-colony-losses-2021-2022-preliminary-results-from-the-bee-informed-partnership/.

[4] The Bee Informed Team. "United States Honey Bee Colony Losses 2021-2022: Preliminary Results from the Bee Informed Partnership." *Bee Informed Partnership*, 28 July 2022, https://beeinformed.org/2022/07/27/united-states-honey-bee-colony-losses-2021-2022-preliminary-results-from-the-bee-informed-partnership/.

[5]US EPA, OCSPP. Colony Collapse Disorder. 29 Aug. 2013, https://www.epa.gov/pollinator-protection/colony-collapse-disorder.

[6]US EPA, OCSPP. Colony Collapse Disorder. 29 Aug. 2013, https://www.epa.gov/pollinator-protection/colony-collapse-disorder.

[7]US EPA, OCSPP. Colony Collapse Disorder. 29 Aug. 2013, https://www.epa.gov/pollinator-protection/colony-collapse-disorder.

[8] "Honeybees Are Still on the Decline, Recent Survey Found. That Could Sting Crop Production." KCUR 89.3 - NPR in Kansas City, 18 Aug. 2022, https://www.kcur.org/news/2022-08-18/honeybees-are-still-on-the-decline-recent-survey-found-that-could-sting-crop-production.

[9] Muth, F., and A. S. Leonard. "A Neonicotinoid Pesticide Impairs Foraging, but Not Learning, in Free-Flying Bumblebees." Scientific Reports, vol. 9, no. 1, Mar. 2019, p. 4764. www.nature.com, https://doi.org/10.1038/s41598-019-39701-5.

[10] *Pesticides Can Harm Bees Twice—as Larvae and Adults*. https://www.science.org/content/article/pesticides-can-harm-bees-twice-larvae-and-adults. Accessed 28 Nov. 2022.

[11] Straub, Lars, et al. "Neonicotinoid Insecticides Can Serve as Inadvertent Insect Contraceptives." *Proceedings of the Royal Society B: Biological Sciences*, vol. 283, no. 1835, July 2016, p. 20160506. *PubMed Central*, https://doi.org/10.1098/rspb.2016.0506.

[12] Scientists Breed Honey Bees to Fight Deadly Parasite. https://www.science.org/content/article/scientists-breed-honey-bees-fight-deadly-parasite. Accessed 28 Nov. 2022.

[13] "Managing Varroa Mites in Honey Bee Colonies." Small Farm Sustainability, https://www.extension.iastate.edu/smallfarms/managing-varroa-mites-honey-bee-colonies. Accessed 28 Nov. 2022.

[14] "Honeybees Are Still on the Decline, Recent Survey Found. That Could Sting Crop Production." KCUR 89.3 - NPR in Kansas City, 18 Aug. 2022, https://www.kcur.org/news/2022-08-18/honeybees-are-still-on-the-decline-recent-survey-found-that-could-sting-crop-production.

[15] "Honeybees Are Still on the Decline, Recent Survey Found. That Could Sting Crop Production." KCUR 89.3 - NPR in Kansas City, 18 Aug. 2022, https://www.kcur.org/news/2022-08-18/honeybees-are-still-on-the-decline-recent-survey-found-that-could-sting-crop-production.

[16] St. Clair, Ashley L., et al. "Access to Prairie Pollen Affects Honey Bee Queen Fecundity in the Field and Lab." Frontiers in Sustainable Food Systems, vol. 6, 2022. Frontiers, https://www.frontiersin.org/articles/10.3389/fsufs.2022.908667.

[17] Le Conte, Y., and M. Navajas. "Climate Change: Impact on Honey Bee Populations and Diseases." Revue Scientifique Et Technique (International Office of Epizootics), vol. 27, no. 2, Aug. 2008, pp. 485–97, 499–510.

[18]Moss, Ellen D., and Darren M. Evans. "Experimental Climate Warming Reduces Floral Resources and Alters Insect Visitation and Wildflower Seed Set in a Cereal Agro-Ecosystem." Frontiers in Plant Science, vol. 13, 2022. Frontiers, https://www.frontiersin.org/articles/10.3389/fpls.2022.826205.

[19] The same species of honeybee — *Apis mellifera*, or the western honeybee — is found in both Europe and the US. In fact, European settlers introduced *Apis mellifera* to the US in the 1600s; the species is not native to the US. *See: Citation 1.*

[20] "Climate Crisis Forecasts a Fragile Future for Wildflowers and Pollinators." Mongabay Environmental News, 1 Apr. 2022, https://news.mongabay.com/2022/04/climate-crisis-forecasts-a-fragile-future-for-wildflowers-and-pollinators/.

[21]Komasilova, Olvija, et al. "Model for Finding the Number of Honey Bee Colonies Needed for the Optimal Foraging Process in a Specific Geographical Location." *PeerJ*, vol. 9, Sept. 2021, p. e12178. *PubMed Central*, https://doi.org/10.7717/peerj.12178.

[22] Viernes, Francis Adrian. "Handling 'Missing Data' like a pro - Part 2- Imputation Methods." Medium, Towards Data Science, 30 July 2021, https://towardsdatascience.com/handling-missing-data-like-a-pro-part-2-imputation-methods-eabbf10b9ce4.

[23] Pramoditha, Rukshan. "Encoding Categorical Variables: One-Hot vs Dummy Encoding." Medium, Towards Data Science, 16 Dec. 2021, https://towardsdatascience.com/encoding-categorical-variables-one-hot-vs-dummy-encoding-6d5b9c46e2db.

[24] Stack Overflow User. "Normalize Data before or after Split of Training and Testing Data?" Stack Overflow, 29 July 2020, https://stackoverflow.com/q/49444262.

[25] Ibid.

[26] Melkumova, L.E., and S.Ya. Shatskikh. "Comparing Ridge and LASSO Estimators for Data Analysis." *Procedia Engineering* 201 (2017): 746–55. https://doi.org/10.1016/j.proeng.2017.09.615.

[27] Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825-30. https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

[28] Bhattacharyya, Saptashwa. "Ridge and Lasso Regression: L1 and L2 Regularization." Medium, Towards Data Science, 28 Sept. 2020, https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

[29] Ibid.

[30] Ibid.

[31] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 2001, 1189–1232.

[32] Li, Cheng. "A Gentle Introduction to Gradient Boosting." URL: Http://Www. Ccs. Neu. Edu/Home/Vip/Teach/MLcourse/4_ Boosting/Slides/Gradient_boosting. Pdf, 2016.

[33] Kuan, A.C. et al. "Sensitivity analyses for simulating pesticide impacts on honeybee colonies." *Ecological Modelling* 376 (2018): 15-27. https://doi.org/10.1016%2Fj.ecolmodel.2018.02.010

[34] Moss, Ellen D., and Darren M. Evans. "Experimental Climate Warming Reduces Floral Resources and Alters Insect Visitation and Wildflower Seed Set in a Cereal Agro-Ecosystem." Frontiers in Plant Science, vol. 13, 2022. Frontiers, https://www.frontiersin.org/articles/10.3389/fpls.2022.826205.

[35] Le Conte, Y., and M. Navajas. "Climate Change: Impact on Honey Bee Populations and Diseases." Revue Scientifique Et Technique (International Office of Epizootics), vol. 27, no. 2, Aug. 2008, pp. 485–97, 499–510.