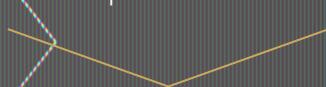


COVERED IN BEES!!

Exploring causes of colony
loss and the predictive
power of data



COLONY COLLAPSE

Documented back to 1869

Disappearing disease

Spring dwindle

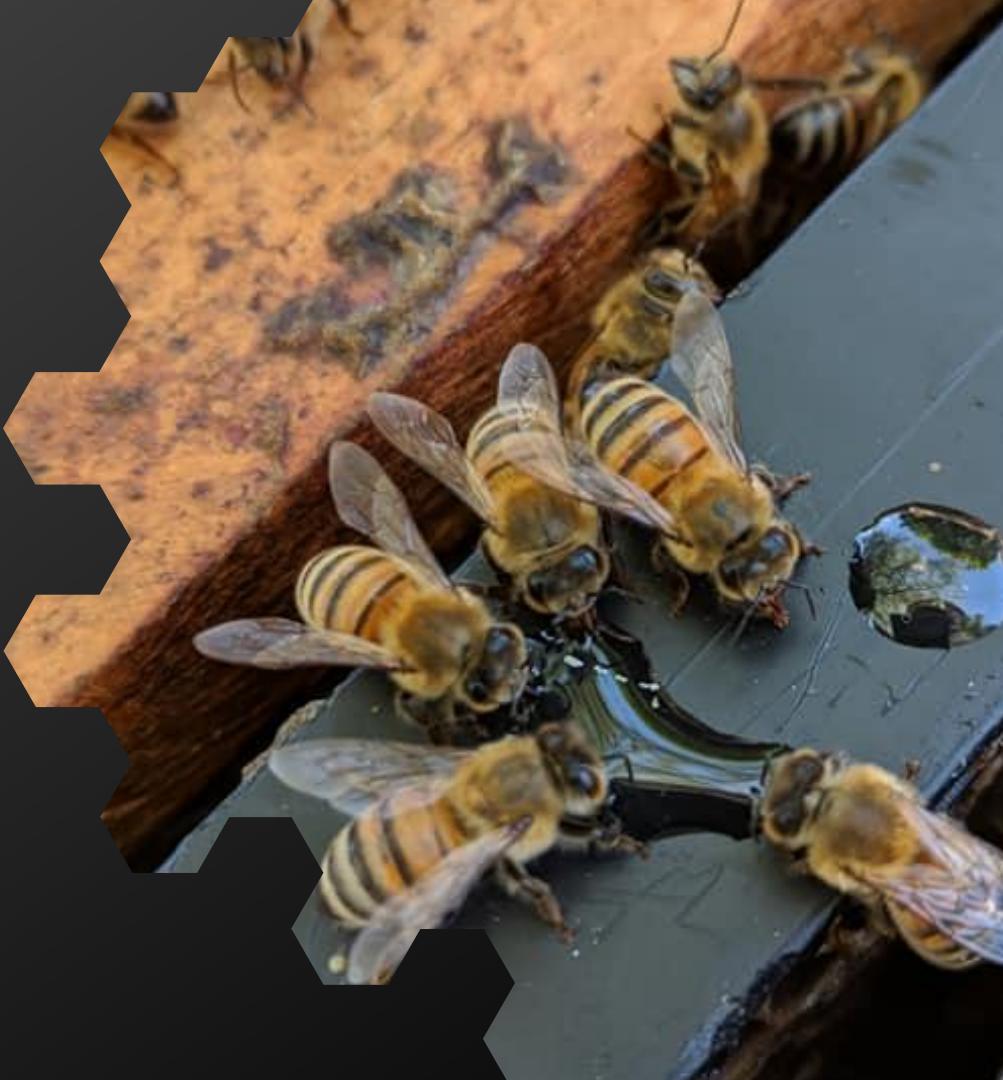
May disease

Autumn collapse

Fall dwindle...

Named colony collapse disorder
as of 2006, when attrition climbed
dramatically

The USDA has been keeping data
on bees since 2015.



REAL ECONOMIC IMPACT

FOOD 

Almonds & Avocados

 INDUSTRY

Cotton & Timber

SANITY

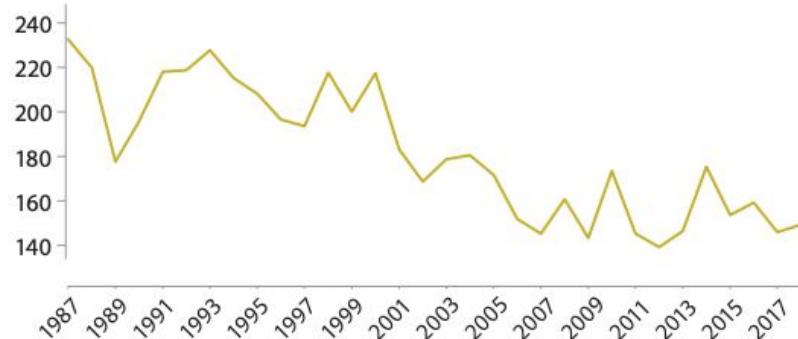


Coffee & Tea & Wine

 MEDICINE

Aspirin & Morphine

Fig. 3. U.S. Honey Production, 1987 – 2018 (million pounds)



Source: USDA NASS, *Honey Report* (various years).

WHAT DATA DO WE HAVE?

SEASONAL COUNTS

Starting colony and lost colony numbers

01

CAUSES OF DEATH

Varroa mites, pesticides, other diseases, pests, and “unknown”

02

WHERE AND WHEN

State, year, quarter.
Plus, extrapolated
region and subregion

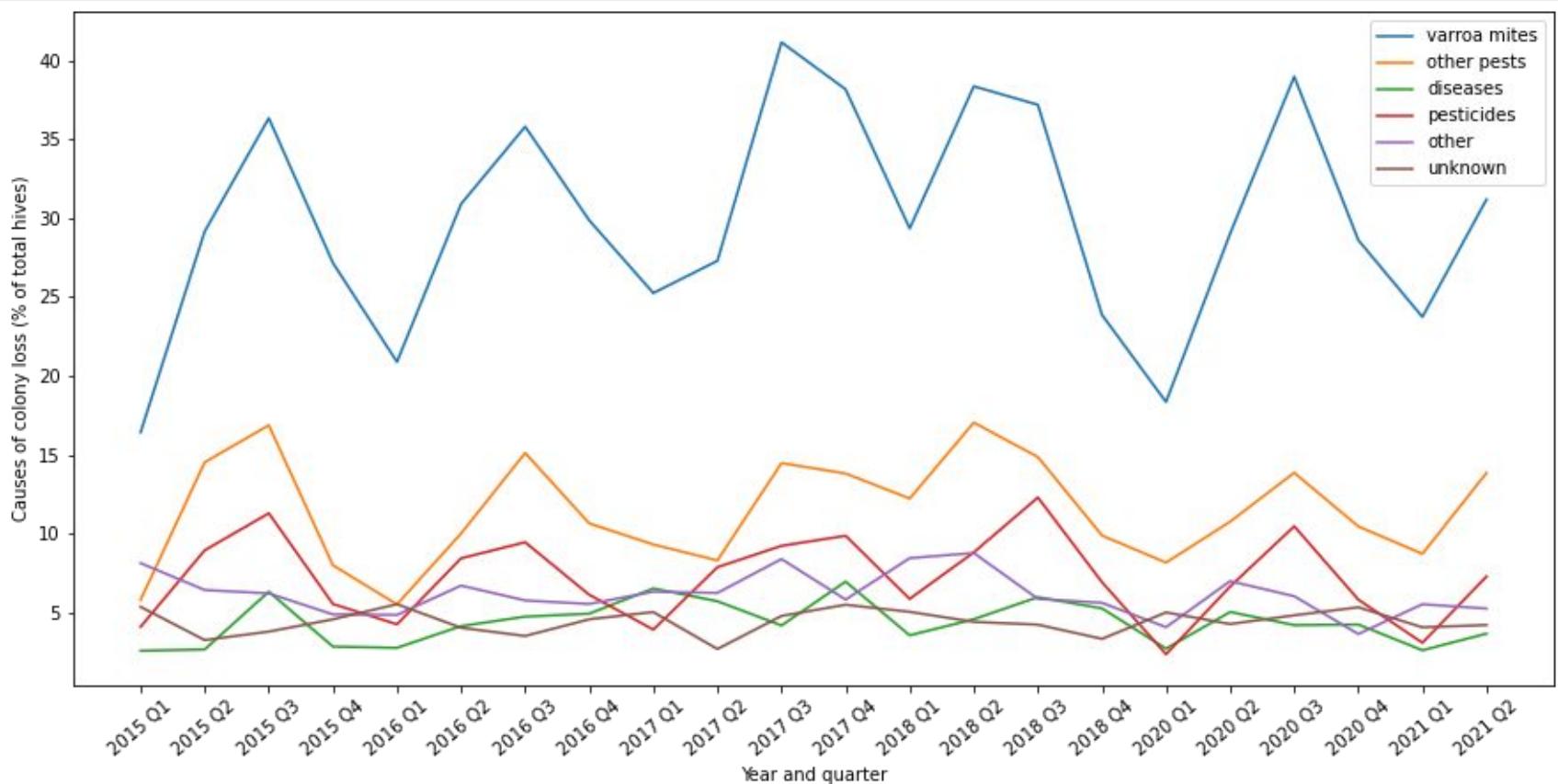
03

ABOUT THE PROJECT

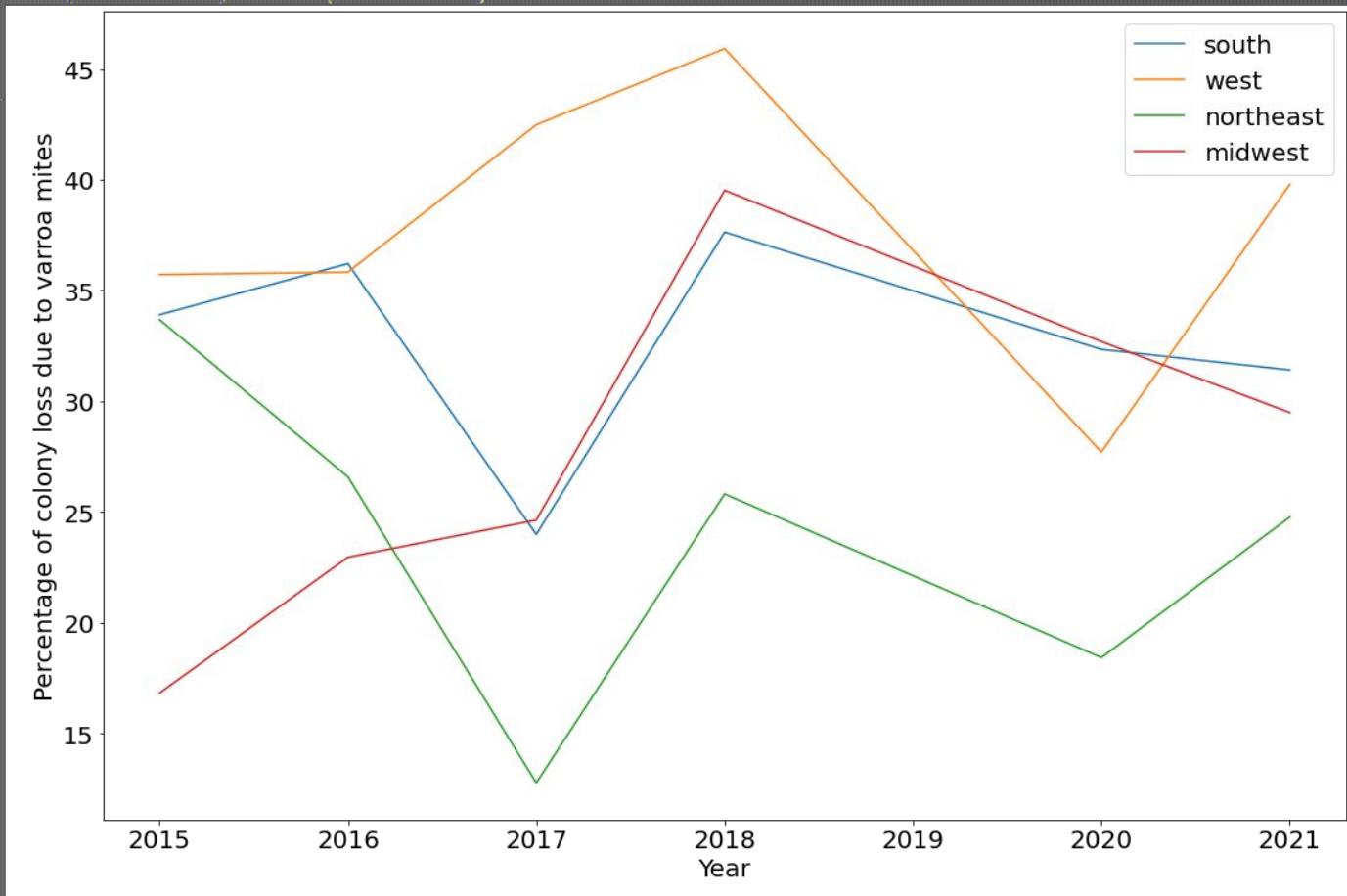


Is it possible to predict the largest
contributors to colony loss with the data
available and, if so, what are those
predictors?

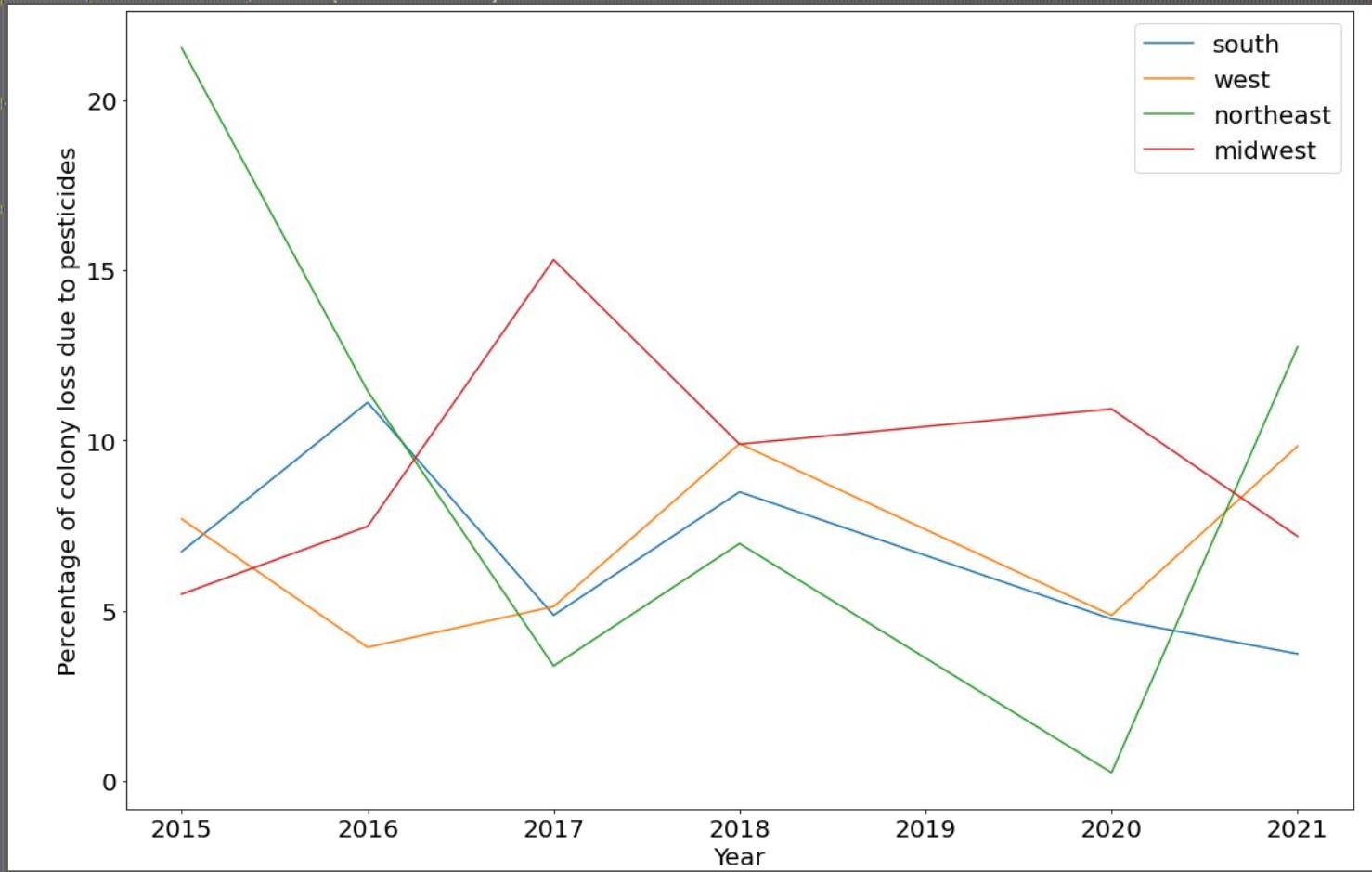
STRONGLY SEASONAL DATA



NO OBVIOUS PATTERNS



NO OBVIOUS PATTERNS



PREDICTION ENGINE



ABOUT THE ANALYSIS

RIDGE
REGRESSION

A sub-type of linear
regression good for
sparse data

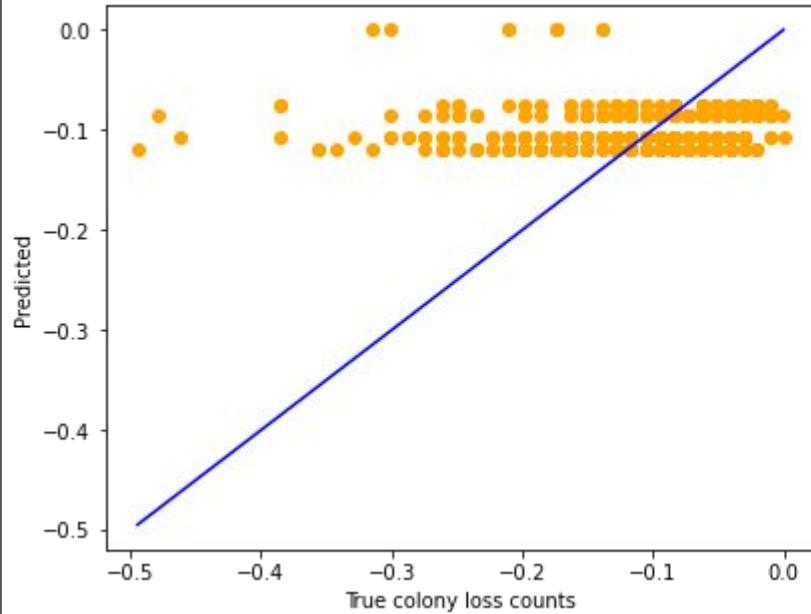
RANDOM
FOREST

A lot of decision trees

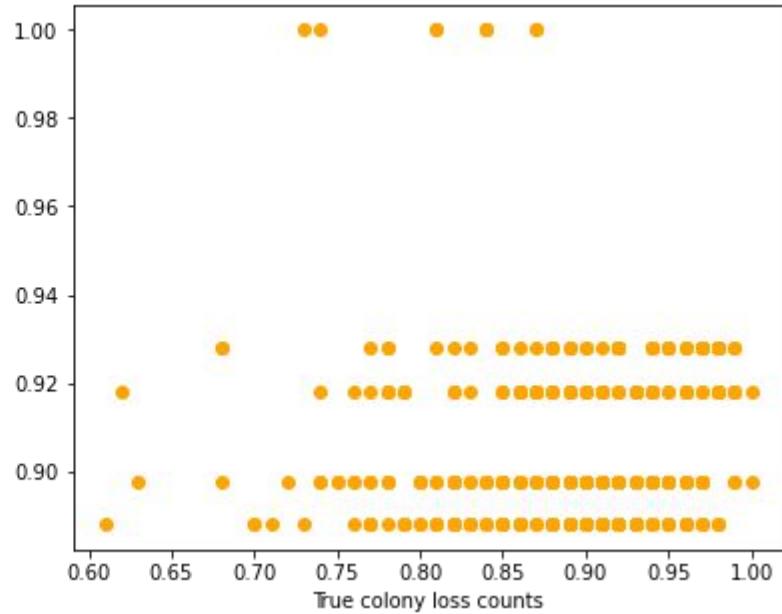
SUPPORT
VECTOR

Seldom used but well
suited here due to time
constraints

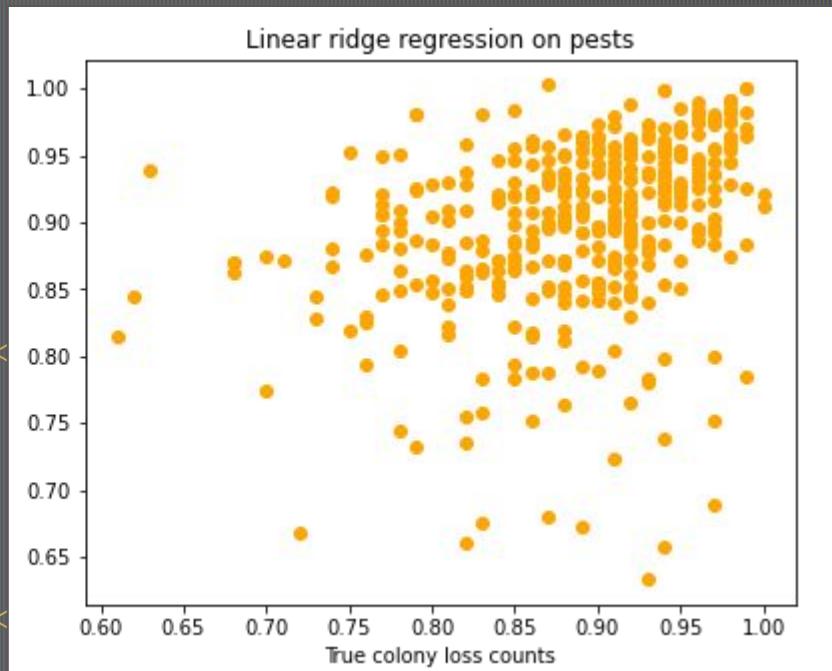
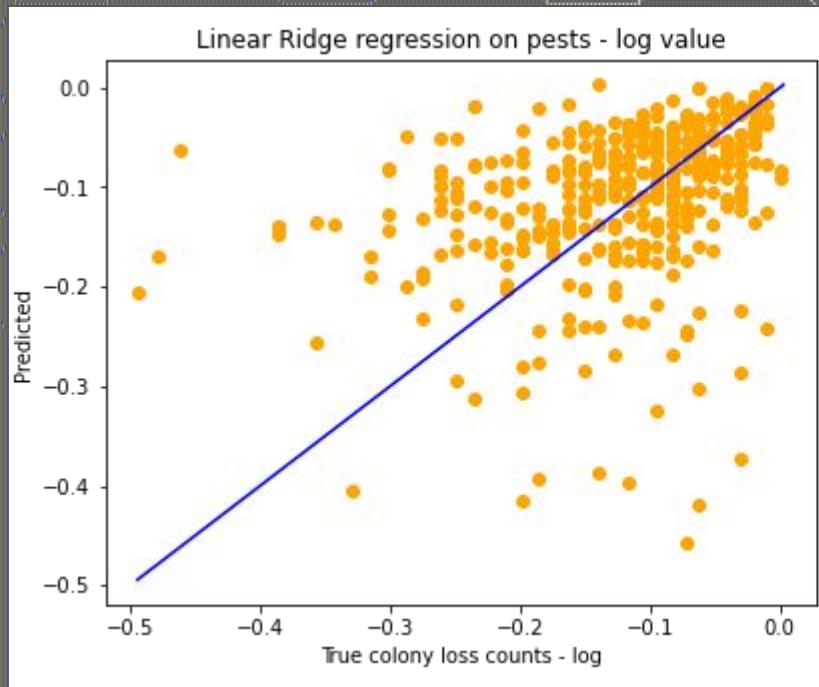
Linear Ridge regression on regions - log value



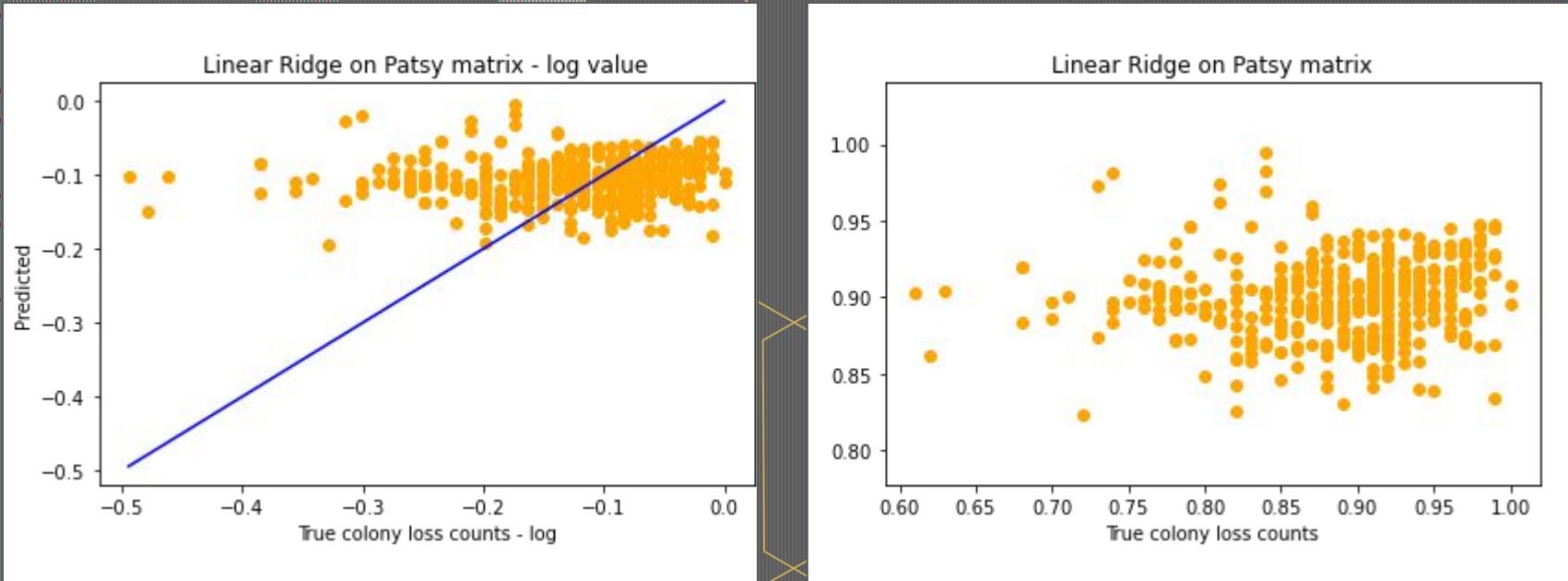
Linear Ridge regression on regions



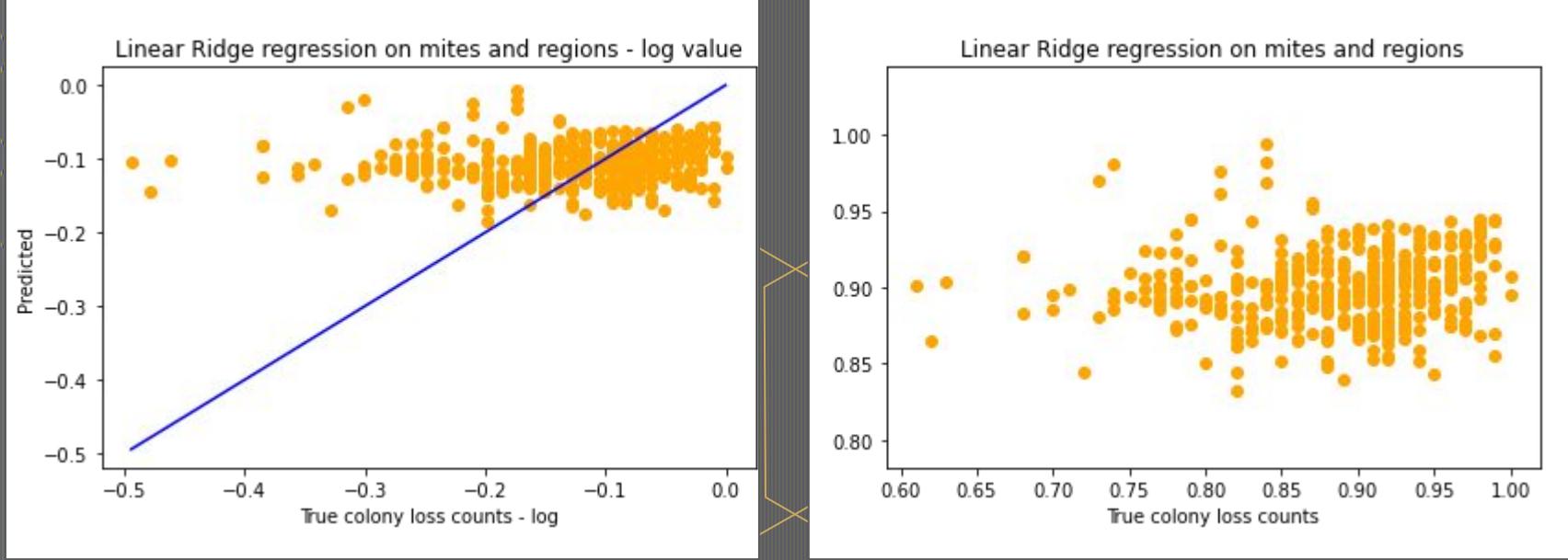
ONE OF THE TOP PERFORMING MODELS:
PREDICTIONS BASED ON REGION - LINEAR RIDGE



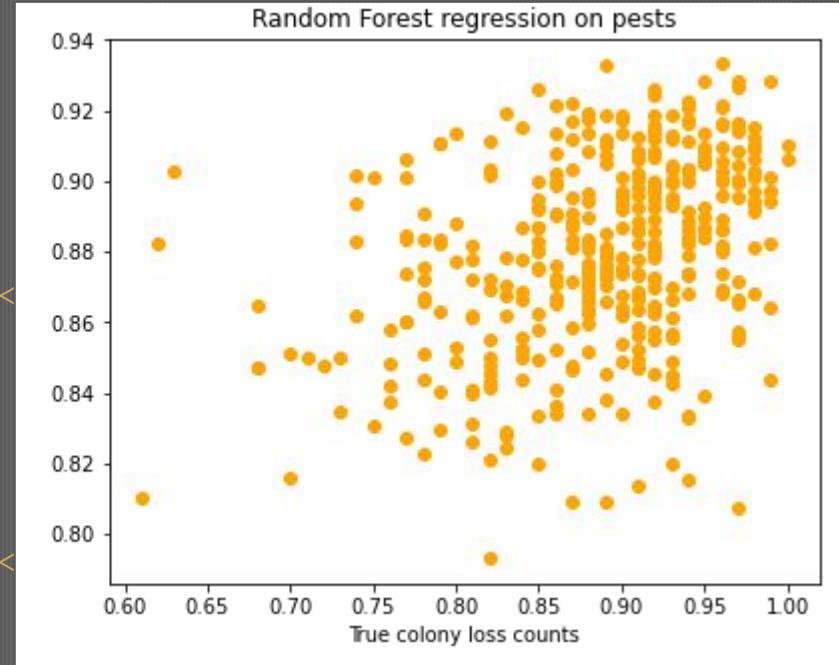
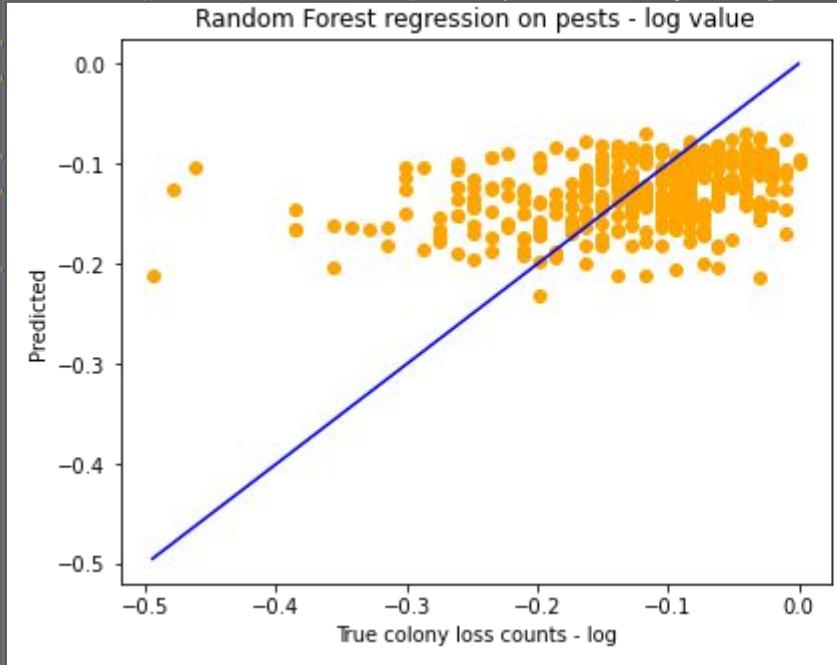
ONE OF THE TOP PERFORMING MODELS:
PREDICTIONS BASED ON ALL PEST DATA - LINEAR RIDGE



ONE OF THE TOP PERFORMING MODELS:
PREDICTIONS BASED ON PATSY MATRIX - LINEAR RIDGE



ONE OF THE TOP PERFORMING MODELS:
PREDICTIONS BASED ON VARROA MITES AND REGIONS - LINEAR RIDGE



ONE OF THE TOP PERFORMING MODELS:
PREDICTIONS BASED ON PESTS - RANDOM FOREST

BY THE NUMBERS

THE BEST MODEL

0.047

Mean absolute
error

0.077

RMSE

0.132

R-squared

Random
forest - all
pest data

ABOUT ME



I've been a beekeeper since spring of 2019 and hosting hives since 2017.

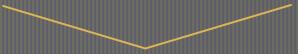
Springboard Data Science student
Oct 2021 - Apr 2022

www.github.com/squareleaf

www.linkedin.com/in/tiffanydgreen/



THANKS



Many thanks to my mentor, Julian Jenkins III, and his endless patience

USDA bee data obtained thru:

www.nass.usda.gov

<https://usda.library.cornell.edu/concern/publications/rn301137d?locale=en>

<https://www.kaggle.com/elljes15/bee-colony-data-cleaning-usda-data>

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik