Machine Learning Engineer Nanodegree

Capstone Proposal:

September 27, 2018

Michael Radvansky

Proposal

Domain Background

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. [i]

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). [ii] 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. [i]

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. [ii] More recently, it has been theorized that neonic pesticides are at least partially responsible for CCD. In fact, just this year, the European Union has banned the three main neonicotinoids: clothianidin, imidacloprid, and thiamethoxam. Several states of the United States have also restricted use of neonicotinoids out of concern for pollinators and bees. [iii]

As I have an agricultural background, I am aware of the importance of bees to the population of the world and am deeply concerned as to the impact of the loss of these critical pollinators on the world's food supply.

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 decline in number of honey bee colonies and honey yield per bee colony basis the amount of specific neonic pesticides used over time for this regression problem.

Datasets and Inputs

Data for this study were put together by Kevin Smith and are sourced from Kaggle. [iv]

The dataset has 1132 rows 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:

- USGS data for pesticides 1992-2016, same as prior version
- [from Honey Production in the USA] : The National Agricultural Statistics Service (NASS) 1998-2017, same as prior version
- [New!] by OCR conversion (Messy, Ugly, Brute-force uploading) of NASS data from scanned PDF files. 1991-1997

The data contained in the dataset are as below:

From USDA 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*)

Solution Statement

The problem of predicting the decline in number of honey bee colonies and honey yield per bee colony basis the amount of specific neonic pesticides used over time is a classic regression problem. To solve this problem the data first need to be inspected and cleaned, formatted, and restructured if necessary. Then the data will need to be segregated into training and testing sets. Once the data sets are prepared, different regression models will be evaluated in terms of the honey production vs neonic pesticide prediction that is being sought.

Benchmark Model

Several data scientists have inspected these data and have concluded that neonic pesticides are negatively impacting bee colonies in the United States; however, neither data scientist has applied a regression model to predict the decline in number of honey bee colonies or honey yield per bee colony basis the amount of specific neonic pesticides used over time. [v] [vi] These researchers have determined, that while there is a general inverse correlation of number of bee colonies and the use of neonic pesticides across the United States, there are regional/state differences, as well as differences involving the use of specific neonic pesticides.

As there is no benchmark model to compare the optimized model from this study, the K-Nearest Neighbors algorithm will be used on the same dataset to provide a benchmark.

The intent of this study, is to expand upon the investigation performed by these two researchers and come up with a predictive model for the number of honey bee colonies and honey yield per bee colony basis the amount of specific neonic pesticides used over time. Predictions will be performed across the entire data set as will as regions/states after inspection of the data.

Evaluation Metrics

The primary metric that will be used to evaluate the predictions of the models tested will be the R Squared metric . Other metrics that will be observed are Mean Absolute Error and Mean Squared Error.

R Squared (R^2 or r^2) is the proportion of the variance in the dependent variable that is predictable from the independent variables. [vii]

Mean Absolute Error is a measure of the difference between two continuous variables.~~ [viii]~~

Mean Squared Error of an estimator measure the average of the squares of the errors—that is, the average pf the squared difference between the estimated values and what is estimated. [ix]

Project Design

The steps required to complete this study are below:

1. The honey production/pesticide usage data need to be loaded into a dataset.
    - In this step the, the data will be loaded into a pandas dataset.
2. The data need to be inspected to determine if any cleaning, formatting, or restructuring is required (preprocessing); if so, the data need to be preprocessed.
    - In this step. The data need to be reviewed, checking for missing data. If data are missing, and the proportion of missing data are very low (< 10%), the missing data will be removed from the dataset. Otherwise, the missing data will be imputed by a yet to be determined imputing approach.
3. Charts and visualizations will be produced to determine which features are the best candidates for modelling.
    - In this step, basis the findings of Chen and Zmith, the data may need to be subdivided into regional/state datasets to understand if there are specific correlations in data due

to geographic area. On the same note, data may need to be subdivided by year, to understand if there are specific correlations in certain years.

4. The data need to be inspected to determine if any features are skewed; if so the data need to be normalized.
    - Again after further inspection, if the pesticide features (*nCLOTHIANIDIN*, *nIMIDACLOPRID, nTHIAMETHOXAM* , *nACETAMIPRID* , *nTHIACLOPRID*)are found to be skewed, those features will need to be normalized through a normalizing process such as logarithmic transformation [x]
5. The data need to be split into training and testing sets.
6. The Supervised Learning models (Decision Trees, LightGBM, and Support Vector Machines) will be evaluated against the data.
7. The results of the model analysis will be summarized and the predictability of decline in number of honey bee colonies and honey yield per bee colony basis the amount of specific neonic pesticides used over time will be presented along with final conclusions.

[i] Greenpeace. The Role of the Bee, 2014
https://sos-bees.org/situation

[ii] Wikipedia. Colony Collapse Disorder, 2018
https://en.wikipedia.org/wiki/Colony_collapse_disorder

[iii] Wikipedia. Neonicotinoid, 2018
https://en.wikipedia.org/wiki/Neonicotinoid

[iv] Zmith, Kevin. Honeybees and Neonic Pesticides. vHoneyNeonic_v03.csv 2018
https://www.kaggle.com/kevinzmith/honey-with-neonic-pesticide

[v] Chen, Mike Honey Bee Colonies and the Use of Neonicotinoids 2018
https://www.kaggle.com/xiangtic/honey-bee-colonies-and-the-use-of-neonicotinoids

[vi] Smith, Linda. U.S. neonic usage and honey production insights 2018
https://www.kaggle.com/skiventist/u-s-neonic-usage-and-honey-production-insights

[vii] Wikipedia. Coefficient of Determination, 2018
https://en.wikipedia.org/wiki/Coefficient_of_determination

[viii] Wikipedia. Mean Absolute Error. 2018
https://en.wikipedia.org/wiki/Mean_absolute_error

[ix] Wikipedia. Mean squared error, 2018.
https://en.wikipedia.org/wiki/Mean_squared_error

[x] Wikipedia. Data transformation (statistics) 2018
https://en.wikipedia.org/wiki/Data_transformation_(statistics)