IN SEARCH OF A PREDICTIVE MODEL FOR AFLATOXIN INSURANCE CLAIMS BASED ON TEMPERATURE DATA

Fausto J. German-Jimenez, UNC Charlotte
Dr. Gabriel Terejanu, College of Computing & Informatics

INTRODUCTION

Aflatoxin is a carcinogenic product of mold that affects thousands of corn-producing farms in the United States alone every year. Being able to predict when a county will have compromising aflatoxin levels in cornfields would be beneficial for insurance companies and farm owners alike.

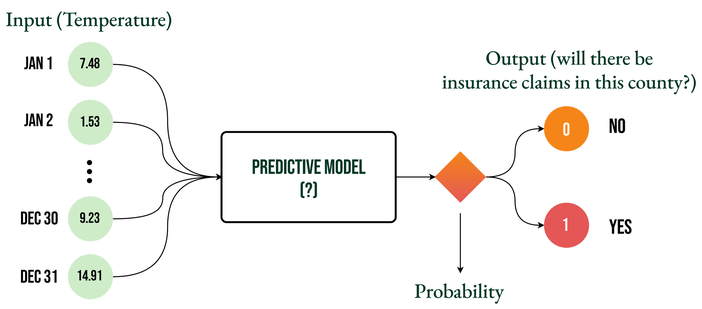
https://unsplash.com/photos/v9r31Dxg0X0

https://www.ckwri.tamuk.edu/news-events/aflatoxin-whats-quail-manager-do

https://webapp.agron.ksu.edu/agr_social/m_eu_article.throck?article_id=1516

OBJECTIVES

With this research, our objective is to create a machine learning model capable of predicting whether or not a U.S. county will have aflatoxin insurance claims based on the daily average temperature for that county in a particular year.



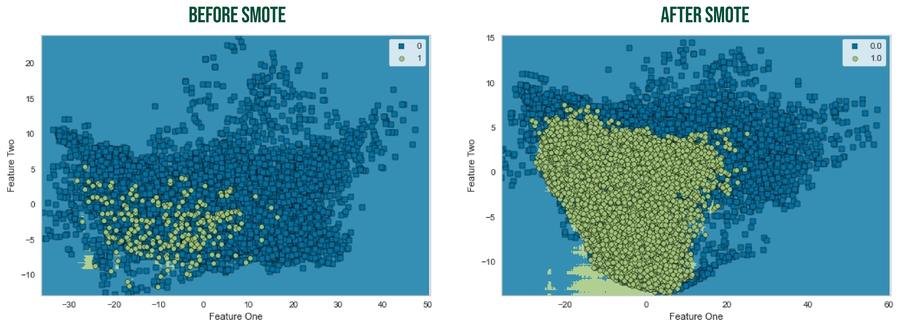
MATERIALS & METHODS – BUILDING THE MODEL

MATERIALS:

- The daily, average temperature data from all U.S. counties from 2010 through 2020, as well as reported insurance claims due to aflatoxin for each of the counties.
- The Python library "PyCaret" to easily find the best performant machine learning model for this type of problem.

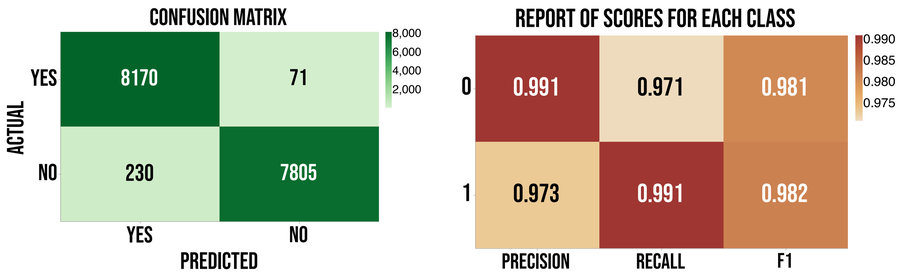
SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE):

- While performing exploratory analysis of the data, we found that there was a severe imbalance in the number of counties that reported losses vs. the ones that did not.
- We solved this problem by applying a Synthetic Minority Oversampling Technique (SMOTE) to the data from 2010 through 2019, and left the year 2020 for validation.



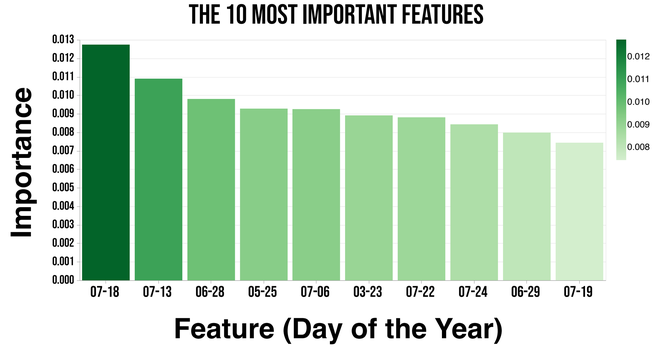
THE MODEL:

We trained an extra-trees, binary classification model using PyCaret and Scikit-Learn. The inputs for this model were the average temperature for each day of the year in a particular county from 2010 through 2019, and the model would classify the county as either having aflatoxin insurance claims or not having aflatoxin insurance claims. This model had an average accuracy of 98.15% on the testing dataset.



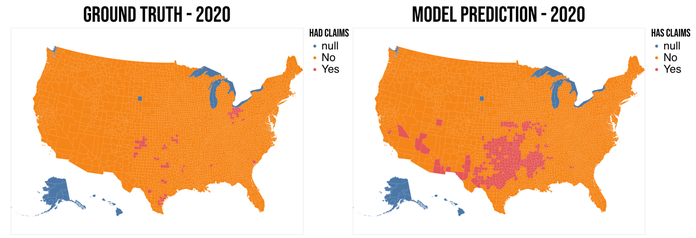
FEATURE IMPORTANCE:

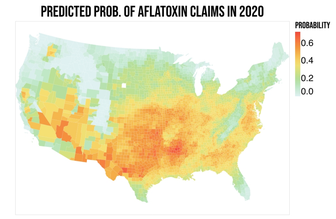
By analyzing the importance of the features learned by the model, we found that July 18th was the most important day of the year when determining if there would be any aflatoxin claims for the county, followed by July 13th and June 28th respectively.



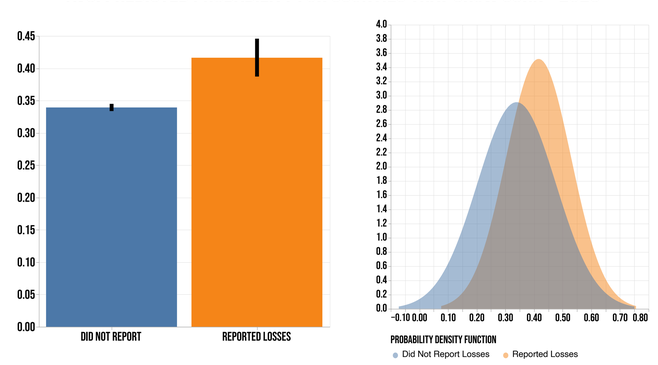
FINDINGS/PRELIMINARY RESULTS

After making predictions on the 2020 data alone, the model seemed to produce many false positives.





However, by performing Welch's t-test on the model's predictions for counties that grow corn and reported aflatoxin losses vs. counties that grow corn and did not report aflatoxin losses in 2020, we found that there was a 99.99% statistical difference between the two groups of predictions, with a p-value of ~1.18x10-6.



CONCLUSION/FUTURE WORK

As demonstrated by this experiment, training an extra-trees, binary classification model using temperature data alone can be an effective way of making predictions about aflatoxin insurance claims in U.S. counties.

However, since aflatoxin production is also affected by humidity and water content in the farms, including humidity data could potentially improve the performance of future models.

Additionally, training a model that can accept variable input sizes would allow for predictions to be made before the end of the year.