ADVANCING CROP INSURANCE RISK MANAGEMENT THROUGH CUSTOMIZED MACHINE LEARNING APPLICATIONS

by

Jake J. Zhu

A Technical Report Submitted to the Faculty of the

DEPARTMENT OF MATHEMATICS

In Partial Fulfillment of the Requirements
For the Degree of

M.S. STATISTICS & DATA SCIENCE

In the Graduate College at

THE UNIVERSITY OF HOUSTON

2023

ACKNOWLEDGEMENTS

I stand on the shoulders of many as I present this work, and it is with profound gratitude that I acknowledge the contributions, encouragement, and support that have made this journey possible.

First and foremost, I am extremely fortunate to have been a part of an incredible full-time team at Tokio Marine HCC. My heartfelt thanks to Jing Yu, Jacob Gogan, and Jim Dole, who have been exceptional companions on this journey. Our brainstorming sessions, debates, and shared moments of learning will always be fondly remembered. These exchanges have enriched my understanding and honed my skills immensely.

To my fellow actuarial interns, who have shared the summer days filled with laughter, I thank you. You all have made this time an unforgettable experience and the workplace a truly vibrant place. Our shared experiences, discussions, and the camaraderie have indeed brightened my summer.

In this pursuit of knowledge, my heartfelt appreciation goes to the faculty of the data science program at the University of Houston and University of Houston - Downtown. The knowledge imparted by you all laid the very foundation of my understanding of the subject. Without this guidance, I would not be in this place in my life and career. Your teachings have been invaluable.

And finally, to my family, who has been my backbone throughout this journey. My parents and my sister, your unflinching support and faith in me have been my stronghold. You have given me the strength to pursue my dreams, standing by me, cheering, and often comforting, through the highs and lows of this endeavor. Your love and belief in me have been my constant source of inspiration.

Contents

List of Figures

List of Tables

List of Abbreviations

AIP	**A**pproved **I**nsurance **P**rovider
FCIC	**F**ederal **C**rop **I**nsurance **C**orporation
GLM	**G**eneralized **L**inear **M**odel
LR	**L**ost **R**atio
MPCI	**M**ulti **P**eril **C**rop **I**nsurance
RMA	**R**isk **M**anagement **A**gency
RMSE	**R**oot **M**ean **S**quared **E**rror
RY	**R**einsurance **Y**ear
SCD	**S**ales **C**losing **D**ate
TMHCC	**T**okio **M**arine **H**ouston **C**asualty **C**ompany
USDA	**U**nited **S**tates **D**epartment **o**f **A**griculture
XGBoost	e**X**treme **G**radient **Boost**ing

ABSTRACT

In this paper, we present an innovative approach to risk management in crop insurance, focusing on employing Machine Learning (ML) techniques to improve prediction accuracy for low-yield crop instances, a fundamental factor in determining insurance payouts. Specifically, we explore the application of the XGBoost algorithm, customized with an objective function tailored to our unique insurance assessment requirements.

We leverage the XGBoost algorithm, renowned for its robust predictive capabilities and computational efficiency, as the backbone of our model. In a departure from standard use cases, we tailor the XGBoost model with a customized objective function specifically designed to enhance the focus on low yield crop instances. This approach enables us to significantly improve the accuracy of predicting insurance payouts under such conditions, thereby refining our risk management capabilities and financial planning procedures.

Our research establishes the significant potential of ML in crop insurance risk assessment. Through our application of the XGBoost algorithm and the integration of a custom objective function, we are able to enhance our prediction model, resulting in more accurate payout predictions. This leads to improved risk sharing, lower financial volatility, and enhanced service delivery to our policyholders.

Chapter 1 Introduction, Background, & Motivation

1.1 History of Tokio Marine HCC (TMHCC)

Tokio Marine HCC is a prominent global insurance company known for its diversified specialty insurance products and services. As part of the Tokio Marine Group, one of Japan's largest insurance entities, Tokio Marine HCC operates across multiple continents, offering a wide array of insurance solutions. Tokio Marine HCC was established through the integration of Houston Casualty Company (HCC) and Tokio Marine Holdings Inc. in 2015. Tokio Marine Holdings, with its roots dating back to 1879, is one of the oldest and largest insurance groups in Japan. Houston Casualty Company, on the other hand, was a leading specialty insurer in the United States since its inception in 1974. The combination of these two entities culminated in the creation of Tokio Marine HCC, a global specialty insurer (HCC, 2023).

FIGURE 1.1: TMHCC Logo

1.2 Introduction of Crop Insurance

Crop insurance is a subset of multiple insurance products that is offered by TMHCC. It offers an extensive suite of crop insurance products designed to meet the diverse needs of America's farmers and ranchers. These include multi-peril crop insurance (MPCI), which provides comprehensive protection against weather-related losses and market fluctuations, and crop-hail insurance for localized storm damage.

Additionally, TMHCC offers a range of named peril policies, covering specific risks such as excessive rainfall, freeze, or drought, as well as livestock insurance and specialty crop insurance for high-value crops. TMHCC is also a leading provider of federal crop insurance programs, working closely with the United States Department of Agriculture (USDA) to help farmers manage risk and stabilize their operations.

1.2.1 Overview of Crop Insurance

The Risk Management Agency (RMA) of the USDA plays a pivotal role in overseeing and directing the Federal Crop Insurance Corporation (FCIC). The process of establishing rates for each policy is an intricate and data-driven procedure managed by the RMA. To determine these rates, the RMA undertakes comprehensive evaluations of various factors including historical yield data, pricing trends, climatic conditions, and emerging agricultural risks. This analytical approach ensures that the rates are reflective of the inherent risks associated with each policy, making it both fair to the farmer and financially sustainable for the FCIC (Tsiboe and Turner, 2023).

Furthermore, the RMA constantly reviews and revises these rates, recognizing that agricultural risks aren't static. Changes in global warming patterns, pest

migrations, or even evolving farming practices can shift the risk landscape. Thus, periodic reassessments ensure that policy rates remain updated and relevant to current conditions.

1.2.2 Fund Designation: Commercial vs. Assigned Risk

Once rates are determined, the RMA's oversight continues. They entrust Approved Insurance Providers (AIPs), such as TMHCC, with the responsibility of selling and servicing federal crop insurance policies. A distinctive feature of this system is the flexibility it offers to AIPs in classifying their underwritten policies.

AIPs have the right to allocate each individual policy they write into one of two distinct funds (RMA and USDA, 2009):

- Commercial Fund: Policies categorized under this fund are typically those where risks are more predictable and manageable. The AIP assumes a larger portion of the risk and, consequently, stands to gain more from the potential profit. However, they are also more exposed to potential losses.

- Assigned Risk Fund: Contrarily, policies that exhibit higher unpredictability or have historically registered frequent losses are designated into the Assigned Risk Fund. In this arrangement, the risk and rewards are shared more equally between the AIPs and the FCIC. This approach ensures that even high-risk agricultural entities can access crucial insurance, thereby fostering inclusivity.

1.2.3 Insurance Pricing and Reserving: Overview

In the realm of insurance, understanding the intricacies of pricing and reserving is important. These two components play a pivotal role in ensuring the financial stability of insurance companies while also safeguarding the interests of policyholders.

- Pricing in insurance denotes the mechanism by which premium rates are determined. These rates are meticulously calculated to encompass anticipated claims, operational costs, and a margin for profit.
- Reserving in insurance is the practice of allocating funds to cater to future claim-related obligations. This fiduciary responsibility ensures that insurance firms remain solvent and honor their commitments to policyholders.

1.3 Motivation of Technical Research

At the time of deciding fund allocations, there are many uncertainties to the loss ratios (LR) of policies due to unknown yield, price, and acreage. The loss ratio can be defined as the following:

$$\text{LossRatio} = \frac{\text{Losses Incurred in Claims + Adjustment Expenses}}{\text{Premiums Earned for Period of Time}} \tag{1.1}$$

For this report, we will focus on the pricing side; specifically, the prediction of yield estimates where the low yields occur and is more likely to trigger an insurance claim. The pricing team in the actuarial department of TMHCC have been working to improve the yield and price prediction; and the reserving team have been working on updating the acreage estimation. One important task in

MPCI is the fund designation of policies, i.e., putting low risk policies to the Commercial Fund (C Fund) and high-risk ones to Assigned Risk Fund (A Fund). Ideally, the net profit can be maximized by ceding the high-risk policies to the government.

Each reinsurance year (RY), the crop harvesting season usually occurs anywhere from June to December, dependent on the harvested crop itself. The reinsurance process typically starts in March, where farmers may sign up for varying insurance policies dependent on what they prefer. There exists a sales closing date (SCD), usually one month after the process starting, where AIPs must communicate FCIC and decide whether each policyholder goes to A or C fund. In order to help the underwriting process during this time, a model predicting and helping the calculations of LRs may be used to forecast profitability in decision making.

1.4 Previous Work Done by TMHCC

Previously, the crop yield models were done using a Generalized Linear Model (GLM) with five feature variables up until recently, where a new model using the popular supervised learning library, XGBoost (eXtreme Gradient Boosting), was implemented. Over time, more specific feature variables were added such as precipitation, soil moisture, drought conditions, etc., in order to bring more accuracy to the model (there are 158 predictor variables total). An iterative process of training using data gathered from 2006-2021 was used to narrow down relevant predictor variables to twenty variables.

Chapter 2 Data and Methods

2.1 Data Overview

Given the data from insurance forms filled out by policyholders, TMHCC also used the following data sources in the creation of the XGBoost model:

- RMA county yields
- NWS CPC Climate Outlook
- Soil Survey (SSURGO)
- NOAA Palmer Drought Severity Index
- NMME Temperature and Precipitation Forecasts
- Elevation from World Soils Harmonized World Soil Database
- NOAA Snowfall & Snowpack
- Copernicus and NOAA Soil Moisture

Previous testing was done on the following crops:

- corn
- cotton
- grain sorghum
- rice
- soybeans
- wheat

2.1.1 Feature Selection

The feature selection process implemented previously by the full-time team ensured the model's accuracy, applying a feature selection process to identify the most impactful features using Shapley values, which represents the average marginal

contribution of a feature considering all possible combinations. Leveraging the SHAP Python package to perform this analysis. The selected variables were then used to build regression county yield models for all crops.

Shapley values explain the output of the model by attributing a value to each feature, showing how much each feature in the data set contributed, either positively or negatively, to a particular prediction. With the subset of features best selected, we can refit the XGBoost model to see if performance improves (or at least remains stable with fewer features) (Lundberg and Lee, 2017).

2.2 XGBoost's Loss Function

At the heart of many machine learning algorithms is the concept of a loss function. It quantifies how well the model's predictions align with the actual outcomes, guiding the learning process by highlighting areas for improvement. For our regression task, the default objective implemented by XGBoost is mean squared error (MSE). It measures the average of the squares of the differences between actual and estimated values (Chen and Guestrin, 2016).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

Where:

- n is the total number of observations.
- y_i represents the observed value for the i-th observation.
- $\hat{y}_i$ represents the estimated value for the i-th observation.

The MSE captures the average squared discrepancy between observed and predicted values. A value of zero indicates perfect predictions, but such a scenario is rare in real-world applications. MSE also forms as a great loss function for fitting the center of data sets. But, our overarching goal is to more accurately predict low yield (e.g. bottom 20% of all yield records), pick up high risk records, and then make better decisions on fund designations. Therefore, we can focus our approach on certain parts of the data, in this case, some arbitrary percentile defined by previous experience (can be 20th percentile, 15th percentile, etc.). Hence, we will define our own loss function to capture such needs.

2.2.1 Customized Loss Function: Asymmetric Pseudo-Huber

The Huber loss is a piece-wise loss function used in robust regression that attempts to combine the best of both worlds of squared loss (MSE), $L(a) = (y_i - \hat{y}_i)^2$, and absolute loss (MAE), $L(a) = |(y_i - \hat{y}_i)|$ (Huber, 1964). In order to blend the benefits of both, the pseudo-Huber loss function may be implemented as a smooth approximation of Huber loss (Hartley and Zisserman, 2004).

$$L_\delta(a) = \delta^2 \left(\sqrt{1 + \left(\frac{a}{\delta}\right)^2} - 1 \right) \tag{2.2}$$

Where:

- δ is a user-defined parameter. When δ is large, the function resembles MAE, and when δ is small, it approaches MSE.
- a is the difference between the predicted and actual values.

The characteristics of the pseudo-Huber loss:

- Continuous and Differentiable: Unlike MAE, it's differentiable everywhere, even at zero. This makes it suitable for gradient-based optimization[1].
- Robustness: Like MAE, it is robust to outliers, especially when the error value is large.
- Smooth Transition: It transitions smoothly from MAE to MSE properties depending on the magnitude of the error, making it flexible.

After selecting the appropriate loss function, we then apply our own needs to the loss function; to apply a weight to the loss function according to the lower percentile of yield values predicted. After consultation with the full-time team, 20% was selected to be the threshold for current testing.

The customized function has the following characteristics:

- Asymmetric: For all values under the 20% threshold, we will apply a weight that is twice as large as prediction values above the threshold.
- Gradient and Hessian: We apply these weights to the gradient and hessian in order for the optimization process of XGBoost.
- The variables 'scale' and 'scale_sqrt' are the first and second order derivatives respectively with respect to the residual, α.
- Delta is predefined as 1.345, but can be changed at a later time (Fox and Weisberg, 2011).

[1]The gradients in a loss function are used to update the model parameters during the training process. If the gradients are not continuous, it can lead to unstable updates and make it difficult for the model to converge to an optimal solution. This is because the gradient descent algorithm relies on the smoothness of the loss function to make small adjustments to the model parameters in the direction of the steepest descent. If the gradients are discontinuous, the algorithm may overshoot or undershoot the optimal solution, leading to poor performance. Therefore, it is important for the gradients to be continuous to ensure stable and efficient training of the model.

```python
def rmse_function(y_true, y_pred):#, delta):# , underprediction_penalty, bottom_20_penalty):
    """
    https://github.com/froukje/articles/blob/main/02_custom_loss_xgboost.ipynb
    https://github.com/dmlc/xgboost/blob/03cd087da180b7dff21bd8ef34997bf747016025/src/objective/regression_loss.h

    This function is defining a custom objective function for a gradient boosting algorithm.
    We use an asymmetric pseudo-Huber because of its allowing of continuous distribution through zeros.
    Attempt to penalize more for bottom 20% quantile for true values.

    """
    # y_true = pd.Series.to_numpy(y_true)
    # global delta, underprediction_penalty, bottom_20_penalty
    delta = 1.345

    # calculate error and absolute error
    residual = y_pred - y_true # y_true - y_pred could be used but keeping consistent with XGB backend
    # abs_residual = np.abs(residual)

    # calculate the 20th percentile threshold we are interested in
    # and apply appropriate weights according to threshold
    threshold = np.percentile(y_true, 20)
    weights = np.where(y_true <= threshold, 2, 1)

    # from pseudo-huber math
    scale = 1 + (residual / delta)**2
    scale_sqrt = np.sqrt(scale)

    # Calculate the weighted gradient and hessian of pseudo-Huber
    grad = weights * (residual / scale_sqrt)
    hess = weights * ((1 / scale) / scale_sqrt)
    # print(grad, hess)
    return grad, hess
```

FIGURE 2.1: Customized Asymmetric Pseudo-Huber Loss Function

Chapter 3 Data Analysis/Simulation Results

3.1 Crop Testing: Corn

A total of 73,440 county records for the crop 'corn' were used to train with XGBoost using the customized loss function. The cross-validation process was implemented by splitting different years into different groups of reinsurance years (RY), making sure each year was balanced in number of records. Thus, all data from 2006-2020 was used in the training process. The data from year 2021 had 4814 rows with the following points of interest:

- Bottom 5% = 240 records
- Bottom 10% = 481 records
- Bottom 15% = 722 records
- Bottom 20% = 962 records

An outline for the process is as follows:

1. Finding hyper-parameters using the Optuna package in Python, considering the top variables as previously tested by the full-time team; training data used was 2006-2020, with 2021 used as a holdout year.
2. Use said hyper-parameters to cross-validate 2021 data as validation set. Find predictions on 2021 data.
3. Compare truth values (sorted by 'Predicted' values only[2]) to the predicted values for 2021 for both original default loss function (RMSE) and new loss

[2]We only compare with predicted values because during implementation of the loss function, there may not be actual values to use as we are predicting into the future.

function.

3.1.1 Results & Discussion

Included are the plots (Figure 3.1) and table (Table 3.1) of varying quantiles of interest showing yield prediction RMSEs measured with differing thresholds in their respective observations. At the 30%, 25%, and 20% threshold, we can observe the RMSE to be similar for our RY of 2021. This is to be expected as our loss function should behave similarly to the default function above the arbitrary threshold that was predefined in our customized function. When we go below that threshold however, we start to observe a widening gap between our default function and the customized function. As the threshold quantile we observe decreases, our custom function's RMSE decreases, which results in an improvement in comparison to the default function as shown in the table below. This can be interpreted as an refinement on accuracy when pertaining to our insurance use case. The more accurate we are at the bottom end of the predicted yield, the more accurate we can calculate the loss ratio (LR) and maximize profit by bettering our decision making for the fund designation process.

TABLE 3.1: Comparison of loss functions using RMSE for the holdout year of 2021. (RMSE Units are in bushels per acre)

Threshold Quantile	Default RMSE	Customized RMSE	Percent Improvement
20%	30.528	30.059	1.53%
15%	31.752	30.311	4.53%
10%	32.349	29.124	9.96%
5%	31.711	26.681	15.86%

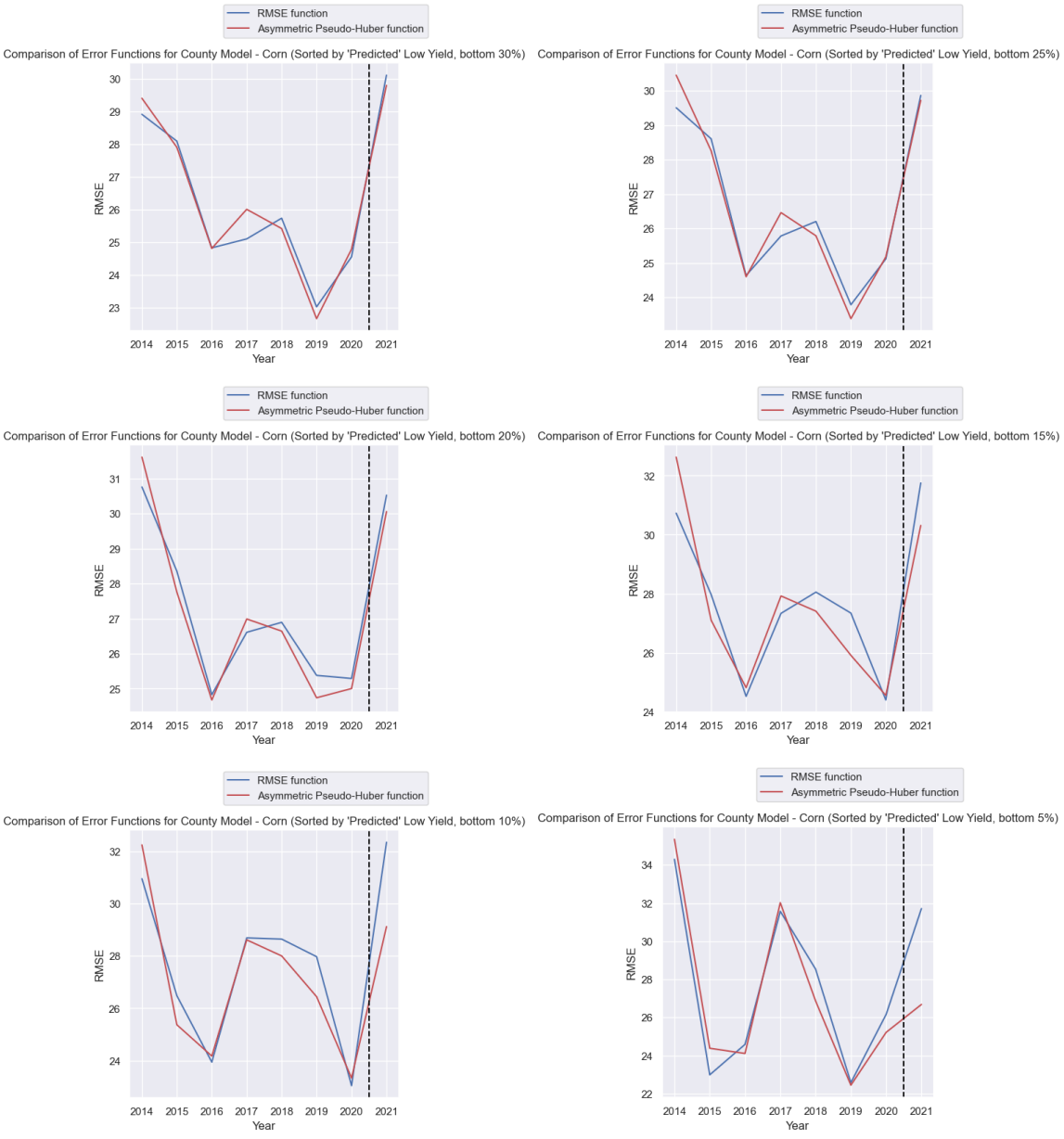

FIGURE 3.1: Results with varying thresholds of interest

Chapter 4 Conclusion and Future Work

4.1 Potential of Customized Loss Function

Our research has shown significant improvement over the existing default method using RMSE, in regards to prediction accuracy in extreme cases where farmers may have very low or zero yield due to catastrophic events; thereby, enhancing TMHCC's risk sharing and lowering our financial volatility.

4.1.1 Next Steps

There has been discussion on potential items for future improvement, as listed in the following:

- Instead of building a county model of overall counties with insurance claims, we can further divide the model into state-based models, increasing the overall accuracy of certain states.
- There has been ideas of other customized loss functions to explore, such ideas include:
  - Incorporation of a loss function where we can focus on certain parts of the data (what was done in this research paper) in conjunction with focus on point estimation accuracy (overestimation vs. underestimation).

  - Using a weighted RMSE, where we inverse transform the weight to incorporate the yield as the weight and normalize according to min-max scalar.
  - With our current customized loss function, we can further improve on the parameters of δ and our threshold; ideally, there could be a threshold for each crop and/or state.
  - A customized version of Root Mean Squared Log Error (RMSLE) where we can assign different weights according to overestimation/underestimation.
- We can add feature engineered predictor variables that are potentially more useful and may be included in the model, such as sentiment analysis using Natural Language Processing (NLP) for more weather data for different states.

Bibliography

Chen, Tianqi and Carlos Guestrin (Aug. 2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785.

Fox, John and Sanford Weisberg (2011). *An R Companion to Applied Regression*. Thousand Oaks: Sage.

Hartley, R. I. and A. Zisserman (2004). *Multiple View Geometry in Computer Vision*. Second. Cambridge University Press, ISBN: 0521540518, p. 619.

HCC, Tokio Marine (2023). *TMHCC: Our Story*. URL: https://www.tmhcc.com/en-us/about-us/our-story. (accessed: 07.09.2023).

Huber, Peter J. (1964). "Robust Estimation of a Location Parameter". In: *Annals of Mathematical Statistics*.

Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: URL: https://arxiv.org/abs/1705.07874.

RMA and USDA (2009). *Fund Designation Guidelines*. URL: https://legacy.rma.usda.gov/FTP/Publications/M13_Handbook/2009/approved/REC09EXH.PDF. (accessed: 05.27.2023).

Tsiboe, Francis and Dylan Turner (2023). *Crop Insurance at a Glance*. URL: https://www.ers.usda.gov/topics/farm-practices-management/risk-management/crop-insurance-at-a-glance/. (accessed: 07.11.2023).