Predicting 2020 Crop Loss Indemnities

Elizabeth Hoepfinger elh33168@uga.edu
Dillon Loubser dcl11282@uga.edu
Jake Hobbs jgh21456@uga.edu

Abstract

There are many issues that can arise in the world of farming which can cause a profit loss. In this paper, we looked at the data from the Risk Management agency for the Summary of Business for Crops to determine if a model can be created to predict different crop indemnities across America, and if there is more than one model, which one can give the best prediction? Previous studies have been done on similar crop data, but we could not find any that strongly resemble the results of this paper. The data set we used was large, and preprocessing was necessary. This included adding column names in excel followed by further preprocessing within Python to prepare the dataset for prediction modeling. The models chosen for predicting Indemnity Amount include Multiple Linear Regression (MLR), Bagging Regressor, and a Neural Network. Overall, the Bagging Regressor performed substantially better than our other two prediction models, with a substantially lower mean squared error and significantly higher R2 when compared to the other two models, with over 98% of the total variability in the dataset being explained by the model. With the work done in this paper, we believe that it opens the door for discussion about regression techniques which can be used in future studies on the prediction of crop indemnities.

1. Introduction

In the world of farming, there are many different issues that can arise and cause a major loss in profit. In these cases, there is insurance that can save the farmers from losing too much money, known as indemnities. There are a lot of factors that go into determining indemnities in the real world: where the person lives, what the weather is like there, what are the chances of issues, etc. In these cases, one can suspect that it would be difficult to determine how much indemnities should be given to who, what kind of policy they need, etc. The goal of this paper is to determine if, given the data from the Risk Management Agency for the Summary of Business for Crops, is there a model that can be created to predict the different indemnities across America? Another question this paper seeks to answer is, if there is more than one model which predicts the indemnities well, which model can be used to give the best prediction?

What we found in our experiments, when creating three regression models, is that regression models like the ones we have used have good predictive power. Of the three we tried, multiple linear regression, neural networks, and bagging, which is an ensemble method, we found that the order from best to worst in terms of both mean squared error and root mean squared error is bagging, then neural networks, and finally multiple linear regression, with

all of these methods resulting in at least decent predictive power for crop indemnities. In this paper, we will break down the previous literature related to the topic of crop indemnities or tied to the methods we were using, discuss the data set as a whole and give a breakdown on how we preprocessed it in order to create our models, then we will describe how we set up the models that we tested, give the results of said models through mean squared error and r-squared, and finally discuss how this research can be used in future work in relation to crop indemnities.

2. Related Work

When discussing the idea of existing literature related to the given topic of crop indemnities, there are many cases of predicting crop indemnities, but they either tackle the creation of models in a different way than what we are intending or the dataset used does not correlate with ours. In a paper from 2013 by Robert Finger entitled "Investigating the performance of different estimation techniques for crop yield data analysis in crop insurance applications," which tackles the subject of creating models for different crop insurance ideas, one of which being crop indemnities [3]. The author of the paper used different methods than we are using, that being OLS-, M-, MM-, and TS-estimators, to try to see if they can find one to help reduce the risk of crop loss. He found that all of the estimators were essentially equally strong in estimating crop indemnities. This paper is close to what we are trying to do, but is built on papers and data from around the world, whereas we are focusing on a dataset from the United States.

Another similar example comes from a paper written in 2015 by Bruce A. Babcock, titled "Using Cumulative Prospect Theory to Explain Anomalous Crop Insurance Coverage Choice," which also discusses the idea of predicting the indemnity payout [1]. Similar to the previous paper discussed, Babcock gives validity to the idea of being able to predict crop indemnities, but applies a different method of prediction unrelated to regression. He also does not use the dataset we are using; this is more centered around the question of proper coverage of plants in order to optimize loss aversion as an outcome, rather than just find the predictive powers of different regression models for this task.

Other papers we found in our research are worse in relation to what we are going for, as they are disconnected from the question we aim to solve at large. For example, The paper "Spatio-Temporal Modeling of Agricultural Yield Data with an Application to Pricing Crop Insurance Contracts", written in 2008 by Vitor A. Ozaki, Sujit K. Ghosh, Barry K. Goodwin, and Ricardo Shirota, explore similar topics of creating models for crop insurance [9]. Ozaki, et al is more focused on the pricing of the contracts themselves than the indemnities paid out, and is centered on Brazil instead of the United States. The paper "An application of ranked set sampling for mean and median estimation using USDA crop production data", written in 2005 by Chad E. Husby, Elizabeth A. Stasny, and Douglas A. Wolfe, uses the dataset from the same source as we do, and while it uses a different sample dataset, it still has the same set up [5]. The only difference is that it attempts to use ranked set sampling to test its general predictive power, rather than focus on crop indemnities. The paper "Quantifying uncertainty and variable sensitivity within the US billion-dollar

weather and climate disaster cost estimates", written in 2015 by Adam B. Smith and Jessica L. Matthews discusses similar loss payouts and predicting them, but focuses more on case studies rather than year-round datasets, and uses Monte-Carlo simulations rather than regression models [10].

The paper "When does USDA information have the most impact on crop and livestock markets?", published in 2020 by Olga Isengildina-Massa, Xiang Cao, Berna Karali, Scott H. Irwin, Michael Adjemian, and Robert C. Johansson, focuses on impacts of different elements on the market, rather than looking at indemnities at all [6]. The paper "Cropland Area and Net Primary Production Computed from 30 Years of USDA Agricultural Harvest Data", written in 2004 by Jeffrey A. Hicke, David B. Lobell, and Gregory P. Asner use the same dataset as us, but was used to look at net primary production rather than crop indemnities [8]. Finally, the papers "Iowa's Agricultural Situation: USDA's year-end summaries show livestock contraction and unexpected increases in crop production", "Crop producer risk management survey: A preliminary sum-mary of selected data; a report from the understanding farmer risk managementdecision making and Educational Needs Research Project", and "Survey of U.S. Multiple Peril Crop Insurance Literature Since 1980", all focus less on discussing results related to the dataset and rather focus on providing looks at properties of related topics which could help build models for the dataset [4][2][7].

Overall, while there are a few related papers which can guide expectations for the results of this given paper, there is surprisingly little in terms of strongly related papers out there in the world.

3. Data Summary

The dataset we are using is a dataset we built from 3 different State/County/Crop/Coverage Level datasets found on the USDA Risk Management Agency website under their State/County/Crop Summary of Business[11]. The purpose of these datasets was to work in tandem with the other two datasets in the Summary of Business to give anyone looking for information regarding crop insurance a dataset for a specific year, with the tools necessary to work with the datasets given. Specifically, we are looking at the datasets which correlate with the consecutive years of 2017, 2018, and 2019. We chose not to work with 2020, as upon further inspection the numbers for that year would create larger errors, since the pandemic caused differing outcomes. The datasets are delimited by the — symbol, such that it would be easy to create .csv files from the datasets given. These datasets are large, each with approximately 133 thousand entries.

Each column of the dataset correlates to a different item, with some items in the dataset correlating with one another. For example, the first column is the year of the policies recorded, and does not correlate to any other column in any meaningful way. The next two columns are for the Location State Code and Location State Abbreviation. Both of these elements deal with the same value, the Location State of the specific entry, but one is a numeric value and the other a string value. This is the same layout for Local County, Commodity, and Insurance Plan, which all have a numeric code and either a string name or abbreviation. There is also a Coverage Category (a string for identifying the type of coverage

between Buyup, CAT, Existing Coverage Policy, and Limited Coverage), Delivery Type (which is a string that identifies the delivery between Reinsured CAT coverage, Reinsured Buyup coverage, Federal CAT coverage, and Federal Buyup coverage), Coverage Level (which is a decimal number that has the complement of the deductible, variables for how many policies were sold, earned premium, or were indemnified, variables for number of units which earned premium or were indemnified, and a string variable for the type of the quantity discussed in the specific entry). Other elements include a variable for the number of whichever quantity was reported as being planted adjusted by the insured's share in the commodity, the number of whichever quantity was insured under an endorsement and/or companion option, the amount of liability the entry was allowed to have, the total premium for the entry, the subsidy amount for the premium, the subsidy from the state, any additional subsidy, the premium discount, the indemnity amount, which is what we are trying to predict, and the loss ratio between indemnity and total premium. The full list of the columns and their meanings are given as a pdf alongside the datasets, which can be used with the datasets to denote the columns.

It would also be pertinent to note that, in some instances of the columns regarding county and commodity, they use the number 999 or 9999 to denote All Other Counties or All Other Crops, a detail which would put into question the validity of any model, dependent on how far outside of the numbers given beside them they are.

4. Data Preprocessing

The first bit of preprocessing we did was add the column titles onto the dataset in order to use them correctly. We only added the column names to the first dataset, the one regarding 2017, for reasons discussed below. The first bit of preprocessing we did was to organize the three datasets, 2017, 2018, and 2019, into one large dataset. We did this by feeding them through Microsoft Excel, which allowed us to use the delimiter — to create the columns and set up the dataset. We organized the data in a way which allowed us to create column names out of the first row of the 2017 file, as discussed before we made the column names. Once the preprocessing within excel was completed, the Pandas toolkit within Python was used to further preprocess the data. First, to combat problems caused by unit disparities, we removed all instances in the dataset that did not have their 'Quantity Type' variable measured in acres. Secondly, all string columns that had a corresponding numeric code were removed, as well as variables that were useless or harmful to prediction, such as 'Year' and 'Quantity Type'. Furthermore, string columns without corresponding numeric code columns were factored using one-hot encoding. In this process, a column is created for each level of each variable needing to be factored, and is given a value of 1 or 0, depending on which value of the original variable is held for each instance. Once this process was completed, the original columns were removed from the dataset. Lastly, the prediction variables (i.e., all variables in the dataset other than the target variable, 'Indemnity Amount') were scaled using the standard scaler.

5. Model Development Process

To develop our models, we split the dataset into separate training and testing sets (with a 75/25 ratio) and trained Multiple Linear Regression (MLR), Bagging Regressor, and Neural Network models using the training set. For MLR, we simply fit our linear models to the training dataset. Next, we created our Bagging Regressor model. This was slightly more complicated than MLR, which required us to first make a decision tree regressor to create our base estimator. Using the base estimator, we made a bagging variable which held the BaggingRegressor function and we included the following variables: the base estimator formerly mentioned, an estimator set equal to 25, and both max features and max samples variables set to 1.0. We then used the variable just created with the bagging function to fit it to the training dataset. After that, we were able to make predictions using the testing dataset. Finally, we created our neural network. Using tensorflow and keras, we first made an optimizer for the network. We chose Adam as our optimizer for a stochastic gradient descent approach to building our network. We tried several learning rates, but eventually settled on 0.005, as it gave us the best results. Next, we initialized Sequential from keras.models and Dense from keras.layers to initialize and add layers to the model. Our hidden layer used the "relu" activation function, which takes element-wise maximums of the input. Lastly, we used mean squared error as our loss measure, and began training the model to the dataset. After many attempts at predicting the data using various values for our epochs and bach size hyperparameters, we settled on a batch size of 500 with 250 epochs.

To evaluate our models, we measured the Mean Squared Error and R^2 on each of the three model development techniques. These measures of loss and explanation of dataset variability should give us a decent idea of how well our models perform at predicting future data.

6. Analysis

The first model we will be analyzing is the Multiple Linear Regression model. This model showed a mean squared error of 48,886,632,941.024445 and a $R^2 = 0.6968$. Second, our Bagging Regressor had a mean squared error of 2,909,418,450.81651 and a $R^2 = 0.9819$. Finally, our neural network had a mean squared error of 32,597,150,634.80516 and a $R^2 = 0.7978$.

Overall, the Bagging Regressor performed substantially better than our other two prediction models. With a tenfold reduction in mean squared error and over 98% of the variability of the data explained by the model, there is no question as to which model we should choose for our future predictions.

Despite the high predictive power of our final model, the results of these models could potentially be pushed further to an even higher level of predictive power, given more time and resources (particularly the neural network).

7. Future Work

Our work can be used in any future work which involves the need for prediction of crop indemnities. Using the results in this paper, as well as following our steps to preprocess the dataset and create the models, one could easily find a way to extend this work. By either experimenting further to find even better models for the use of prediction, or perhaps using the models we have given to understand the dataset more, our work can be extended to create new ways for insurance companies to estimate the needed values. Our work also suggests that there are ways in which one can use data mining techniques for the same types of tasks as the ones presented in the papers we discussed in our literature review. Future work could also use this paper and other related works to determine the most viable solution for whichever problem is at hand regarding the dataset. Overall, our paper opens the door for discussion about regression techniques which can be used in future studies on prediction regarding the dataset.

References

[1] Bruce A. Babcock. *Using Cumulative Prospect Theory to Explain Anomalous Crop Insurance Coverage Choice*. June 2015. URL: https://onlinelibrary.wiley.com/doi/full/10.1093/ajae/aav032.

[2] Keith H. Coble et al. *Crop producer risk management survey: A preliminary summary of selected data; a report from the understanding farmer risk management decision making and Educational Needs Research Project*. Jan. 1999. URL: https://ageconsearch.umn.edu/record/15805.

[3] Robert Finger. *Investigating the performance of different estimation techniques for crop yield data analysis in crop insurance applications*. Jan. 2013. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/agec.12005.

[4] Chad Hart. *"Iowa's Agricultural Situation: USDA's year-end summaries show livestock contraction and unexpected increases in crop production"*. Aug. 2015. URL: https://lib.dr.iastate.edu/iowaagreview/vol9/iss1/3/.

[5] Chad E. Husby, Elizabeth A. Stasny, and Douglas A. Wolfe. *An application of ranked set sampling for mean and median estimation using USDA crop production data - journal of agricultural, Biological and environmental statistics*. URL: https://link.springer.com/article/10.1198%2F108571105X58234.

[6] Olga Isengildina-Massa et al. *When does USDA information have the most impact on crop and livestock markets?* Apr. 2020. URL: https://www.sciencedirect.com/science/article/abs/pii/S2405851320300143.

[7] Thomas O. Knight and Keith H. Coble. *Survey of U.S. multiple peril crop insurance literature since 1980*. Mar. 1997. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/1349683.

[8] David B. Lobell and Gregory P. Asner. *Cropland area and net primary production computed from 30 years of USDA Agricultural Harvest Data*. July 2004. URL: https://journals.ametsoc.org/view/journals/eint/8/10/1087-3562_2004_008_0001_caanpp_2.0.co_2.xml?tab_body=fulltext-display.

[9] Vitor A. Ozaki et al. *Spatio-temporal modeling of agricultural yield data with an application to pricing crop insurance contracts*. Nov. 2008. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8276.2008.01153.x.

[10] Adam B. Smith and Jessica L. Matthews. *Quantifying uncertainty and variable sensitivity within the US billion-dollar weather and climate disaster cost estimates - natural hazards*. Mar. 2015. URL: https://link.springer.com/article/10.1007/s11069-015-1678-x.

[11] *State/County/Crop Summary of Business*. URL: https://www.rma.usda.gov/en/Information-Tools/Summary-of-Business/State-County-Crop-Summary-of-Business.