Machine Learning Driven Optimizations to Soybean Yield

Noah Kasmanoff (nsk367)
Peter Simone (ps4021)
Rachana Swamy (rms816)
Rahul Zalkikar (rz1567)

December 9, 2019

Abstract

Soybean is an essential crop for American farmers. With greater production comes tremendous benefit to the American economy at large. In this report, we examine the relationship between exogenous and endogenous variables to soybean yield. By doing so, we identify the optimal permutations of these factors which portend themselves to a successful harvest season. After wrangling the necessary data, we created forecasting models that allow us to predict the total yearly yield of a given state within an 11% margin of error, sufficient to validate this need for a model such as this, and warrant further research into making an AI-based crop yield system.

1 Business Understanding

Soybean is one of the most popular crops worldwide, especially so in the United States. Until 2017, United States was the leading producer and exported of Soybeans. In 2016 alone, this crop's industry was valued over $40 billion, and continues to grow. [1] It is a staple to American farmers, and is intimately tied to issues such as climate change, politics, and international trade. [4] [2] [5]

This relationship is in jeopardy. With significant challenges such as climate change, a growing population, unsustainable production and consumption patterns, and increasing competition from other countries like Brazil and Argentina, the agricultural industry is in need of solutions. Similar to other industries facing challenges today, it is possible to search for these solutions in the context of data science. For this reason, we seek to provide data-driven insights to address the largest obstacles facing the agricultural industry today; primarily, we seek to optimize the crop yield of soybean for American farmers and, in doing so, provide insight into how to handle their soybean due to changes in climate and substance applications. There is also utility from the perspective of supply chain, as using this predictive model could help farmers forecast the need for additional resources come harvest.

A data mining solution is a feasible approach. Specifically, we can wrangle from an array of agricultural related databases, and extract the relevant features to generate a model which informs us of the optimal strategies for farmers to pursue for optimizing yield.

2 Data Understanding

To properly characterize soybean yield, the features are separated into three distinct domains:

1. **Crop Data:** Data related to the historical planting and harvesting cycles of soybean across the dataset.

2. **Environmental Data:** Historical information on temperature, and moisture levels in the various regions of interest.

3. **Soil Data:** Proxy variables for the quality of the soil growing soybean, such as fertilizer, herbicide, insecticide, and fungicide concentrations and usages.

2.1 Crop Data

We collected crop-specific data via. the U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service online portal [7] and gathered the historical information of soybean yield from all available US States informative to modern crop cycles. From the USDA, we obtained our target of Soybean yield, expressed as bushels per acre ($\frac{bu}{ac}$) and additionally collected data on the total area planted, condition of soybean in previous harvesting cycles, and more.

2.2 Environmental Data

Previous studies have indicated that climate conditions are essential to soybean growth. In fact, using climate data alone proves to be an effective indicator of soybean yield in our experiments. Incorporating climate data into our model allows us to truly extract the influence of the actionable variables discussed below. In particular, we seek to characterize the climate at which soybean is grown over a particular cycle with the information associated with temperature, precipitation and soil moisture. This data contains 11 variables that analyse monthly temperature, precipitation and drought conditions aggregated at State-level and was collected from the U.S. Climate Divisional Database [3].

2.3 Soil Data

The profile of the soil in which soybean is grown is characterized with proxy variables, collected from USDA surveys [7]. In addition to the data provided, these surveys also validate our choice for selected features in this section, as these substances were curated by domain experts as important to soybean growth, further motivating this task of maximizing soybean yield. In particular, we select fertilizer, insecticide, fungicide, and herbicide concentrations in a given crop cycle. Perhaps most importantly, the concentrations of these substances is a process which is indeed actionable, presenting opportunity to apply machine learning and data science to the agricultural industry.

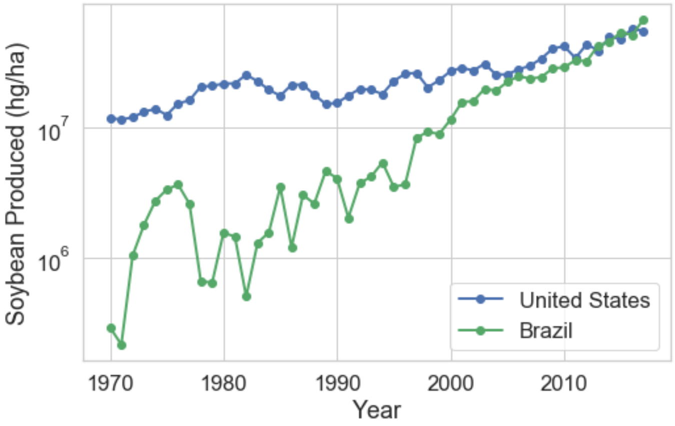

Figure 1: Yearly trends of soybean production between the United States and Brazil.

Our specified time domain is from 1996 to 2019. Until the late 1990s, the United States soybean industry stood alone as the main producer and exporter, but now faces tremendous competition from Brazil. Using this time window, we find an ideal balance between collecting a sufficient number of data points while still containing information relevant to today. [7] This particular period provided an optimal trade-off between total data instances (equal to $n_{states} * t_{lookback}$)

After collecting and preparing these features, we forecast the yield of soybean and extract insights related to the crop, climate, and soil data summarized above [1].

Since not all states are equally represented in the data, we faced challenges with selection bias. Surveys are not sent out consistently, and some states do not have information for certain years. This poses an issue because some features in our dataset that are inherent to a given state, such as climate and area, might be over or under represented. To circumvent this issue, we develop a model agnostic to state and year directly. This is to help reduce the bias, and also to make the model more robust and usable.

3 Data Preparation

We combined seven unique USDA datasets on farm operations, crop condition, crop yield, fertilizer, herbicide, insecticide, and fungicide applications. Indexing on state and year, this agricultural data was merged with Environmental data obtained from NOAA. The final dataset had 722 instances.[2].

3.1 Sparsity Challenges

After merging, we had to handle missing values. Some USDA data contained NA values, or values with random strings where there should've been numeric values. To handle this, we interpolated based off the mean of a year and state grouping, and extrapolated where possible as well.

[1] For a more thorough description of the features used, please refer to Appendix A.
[2] Roughly speaking, 31 states were covered, for approximately 23 years worth of data.

3.2 Data Leakage

We sought to eliminate any form of data leakage. Features such as crop condition, crop moisture stress index and area harvested are collected almost simultaneously with yield calculations, and present no actionable insight (for instance, we cannot simply tell farmers to harvest more or control drought conditions). Accordingly, they were transitioned into lagged features, since we anticipated they may still be related to the following year's harvest. Similarly, if merging on year, another form of leakage that occurs is when the climate information for months after harvest are included in that year's index. Since this information is still relevant as well, we applied the same lagging effect so that climate information for a given year-state was used in the prediction of yield for the following year, where it still maintained relevance. Additionally, we one hot encoded categorical variables to allow for binary features.

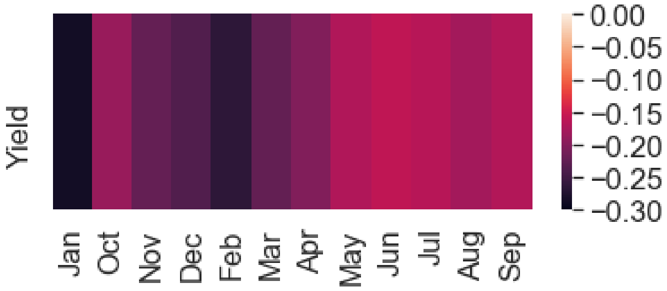

Figure 2: Monthly temperature information has a non-trivial correlation with yield.

An encouraging result of our exploratory data analysis was the discovered distribution of our target variable. We find that yield follows approximately normal behavior. This implies that a parametric and non-parametric methods will be applicable. With this in mind, we can begin to plan which models will be tested in the process.

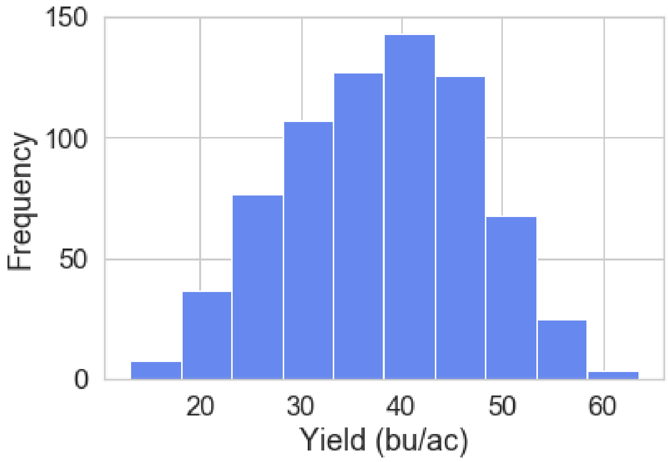

Figure 3: Distribution of yearly soybean yields for US states.

3.3 Feature Selection

To deal with feature selection in an attempt to reduce the dimensionality of our dataset, we first assessed instances of collinearity in the data. We have domains of features based on the source of data that they originate from. These include: Crop Condition Data, Soil Data, and Environmental Condition Data. We generated a correlation matrix and pinpointed features highly correlated with others outside their domain (i.e. if a "Crop Condition" feature was highly correlated to a "Soil Data" feature). We also examined variables that were highly correlated to the target variable and removed those, since we are trying to produce actionable insight instead of modeling obvious relationships (ex.Lagged Area Harvested in Acres with Yield in BU per Acre).

Next, after standardizing our feature sets, we performed Principal Component Analysis (PCA) to better understand the explained variance in our data. The obvious trend we found was that after approximately 100 principal components, nearly all of our explained variance in the data was accounted for. This prompted us to seek ways which greatly cut down on the total number of features, and we assess those results.

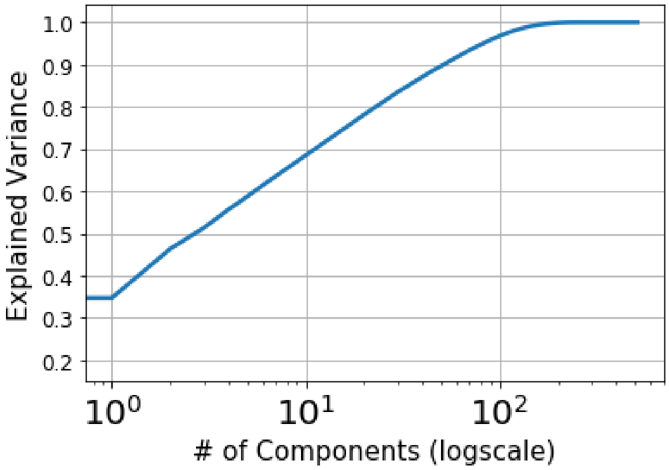

Figure 4: Cumulative variance in the total dataset.

Next, we adopted regularization on the final dataset [8] to reduce the dimensionality of this dataset (which contains more columns than rows, we first employed lasso regression. Lasso regression puts constraints on the size of the coefficients associated to each variable, it is all the more important that we standardized since variables of different scale would not contribute equally to the analysis and can create bias. Using LassoCV, we selected the best model through 5-fold cross-validation. The training/validation set was randomly split (80-20) for cross-validation. We overlay this with a meta-transformer (SelectFromModel) [3] that performs feature selection based on the feature importance from Lasso regression by setting a threshold to select features corresponding to the L1 norm penalty. After adjusting the convergence tolerance for each iteration to higher value, at the cost of losing accuracy by compounding error, we obtained a subset of 82 features with non-zero coefficients from LassoCV. This reduction from 792 to 82 could be due to feature collinearity.

4 Modeling & Evaluation

4.1 Baseline Model

First, we established the baseline model to improve from, using lasso linear regression as mentioned previously. LassoCV served as both a method of feature selection moving forward, and a baseline model. Due to the challenges of having more features than data points, traditional linear regression would fail due to simple geometry, as it is

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html

impossible to draw a line in space with the number of points being fewer than the total number of dimensions. To overcome this challenge, we employed shrinkage to find a sparser representation of our feature space.

Non-parametric algorithms such as decision trees (and the ensemble random forests) also naturally perform such a feature selection, but for purposes of taking a simplest model first approach, lasso regression permits us to develop this baseline model and simultaneously identify the most important features, which will be fed through more sophisticated algorithms.

To evaluate our yield forecast capabilities, mean absolute percentage error (MAPE), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 all provide well defined scores for us to use. These metrics are generally well accepted in the field and provide insight into a variety of relationships.

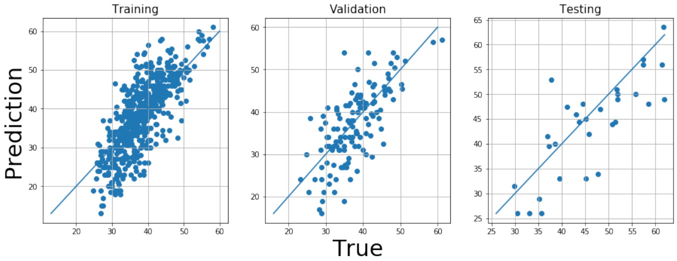

Figure 5: Lasso regression predicted yield vs. expected yield for training,validation, and testing sets.

	R^2	**MAPE**	**MAE**	**MSE**	**RMSE**
Training	0.56	14.34%	4.79	37.80	6.15
Validation	0.51	15.05%	4.84	37.53	6.13
Testing	0.52	13.30%	5.19	44.33	6.66

Table 1: Evaluation scores of lasso regression show a strong baseline to start with.

Feature Name	Abs(Coefficient)
AREA_PLANTED_IN_ACRES	2.65
MAX_TEMP_Month8_lagged	1.52
APPLICATIONS_IN_NUMBER_AVG_HERBICIDE_24-D 2-EHE	0.84
CONDITION_5_YEAR_AVG_IN_PCT_EXCELLENT	0.77
Precipitation_Month12_lagged	0.73
Crop Moisture Stress Index (CMSI)_lagged	0.68
MAX_TEMP_Month1	0.22
MAX_TEMP_Month7_lagged	0.21
TREATED_IN_PCT_OF_AREA_PLANTED_AVG_HERBICIDE_BENTAZON	0.17
TREATED_IN_PCT_OF_AREA_PLANTED_AVG_HERBICIDE_24-D DIMETH. SALT	0.15
APPLICATIONS_IN_LB_/_ACRE_/_APPLICATION_AVG_HERBICIDE_24-D DIMETH. SALT	0.15
TREATED_IN_PCT_OF_AREA_PLANTED_AVG_HERBICIDE_SAFLUFENACIL	0.12

Table 2: Most important variables found in lasso regression baseline.

The results of lasso regression demonstrate a powerful baseline, and prove the need for a study such as this. Based on the established knowledge that climate has a known relationship with crop harvests, it would have been reasonable to see only climate and previous harvest features as the most important variables to crop yield. Instead, we discover that herbicides occupy many spots in the list of most important features found after shrinkage. This surprising result is further backed by already relatively successful evaluation scores, where the R^2 score already

	R^2	**MAPE**	**MAE**	**MSE**	**RMSE**
Training	0.92	5.63%	1.94	6.79	2.61
Validation	0.53	15.14%	4.76	35.48	5.96
Testing	0.62	10.96%	4.79	35.35	5.95

Table 3: Random Forest Regressor with default parameters on 82 features selected from LassoCV results.

demonstrate that this model is capable of a significant portion of the variance of this data, with better models looming. The baseline lasso regression demonstrates that actionable features based on substances farmers place in the soil such as herbicides play an important role in determining soybean yield.

4.2 Random Forest

In our next experiment, we consider the random forest. Random forests select at each candidate split in the decision tree learning process, a random subset of the features (feature bagging). Without feature bagging, if one or a few features are strong predictors for yield, it is likely they will be present in many of the decision trees created, increasing the correlation between estimators and causing learners to be biased in favor of features that appear highly predictive in the training set, but fail to be as predictive for unseen data points. This is one reason why we prefer to use Random Forest for our data. The training algorithm of a Random Forest bags decision tree learners by selecting a random sample of the training data, with replacement. By training on different parts of the same training set and averaging decision tree regressors, Random Forest reduces variance, possibly at the cost of increased bias, while limiting overfitting. Using an out of the box random forest algorithm, we already achieve improved results displayed in table 3.

4.3 Tuned Random Forest

Capitalizing on this success, we build via grid and random searches a tuned random forest algorithm with even more impressive results. The details of this search are detailed below:

We leverage randomized search with 5-fold cross validation to optimizing our random forest, so all parameters that influence the learning process are searched simultaneously. This allowed us to develop a model that generalized well and weigh features accurately, and include the maximum number of features to consider at each node split (every node of a decision tree is associate with a set of data-points), the minimum number of samples to split a node, the maximum levels in a tree, and the minimum samples in each leaf node. In addition, our algorithm samples with replacement via bootstrap.

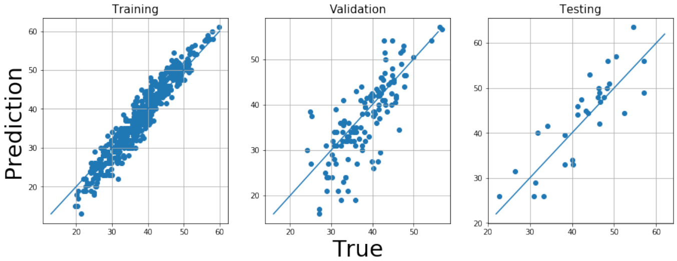

Figure 6: Tuned Random Forest Model Predictions vs. True values of Yield

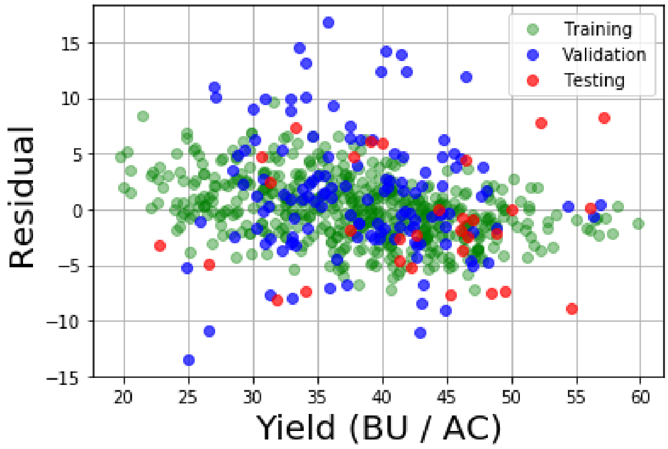

Figure 7: Tuned Random Forest Model Residuals Plot

	R^2	**MAPE**	**MAE**	**MSE**	**RMSE**
Training	0.98	2.62%	0.88	1.44	1.20
Validation	0.58	13.99%	4.30	31.86	5.64
Testing	0.72	10.58%	4.23	25.50	5.04

Table 4: Tuned Random Forest Regressor for 82 features selected from LassoCV results.

The randomness in the residuals plot tells us that the error of the model is not bias towards low or high ends of the target value. This is good, and shows that the tuned random forest model is robust and capable of predicting a wide range of values.

We also observe an increase in R-squared value and a decrease in MAPE, MAE, MSE, and RMSE across corresponding Training, Validation, and Testing sets, with particularly impressive MAPE (10.58%) and RMSE (5.04) for 2018. After establishing what appears to be a generalized and therefore optimal model, we also confirm the validity of its use in practice by the nature of its features of highest gini importance.

Feature Name	**Importance**
Crop Moisture Stress Index (CMSI) - lagged	0.082599
Min-Temp-Month12 - lagged	0.071355
Hexazinone	0.066945
MCPA Dimethyl Salt	0.053902
Halauxifen-Methyl	0.049862
Area-Planted-in-Acres	0.047258
Bicyclopyrone	0.044854
Condition-5-Year-Avg-in-PCT_Excellent	0.040763
Pyroxsulam	0.036613
Metsulfuron-Methyl	0.031348
Application-in-Number-Avg-Herbicide-24-D2-EHE	0.029802
Diuron	0.026242

Table 5: Features of Highest Gini Importance in Tuned RF model.

4.4 Post-Modeling Feature Discussion

The three most important features to our most powerful model present an opportunity to interpret our results.

The variable with the highest feature importance is the lagged is Crop Moisture Stress Index (CMSI), which measures effects of drought and intense wetness on national soybean yield. It makes intuitive sense that either a lack or an excessive amount of soil moisture during critical phases of the soybean growth cycle (especially in high growth areas) impacts average crop yield.

Next is the minimum temperature in December, which a proxy for the intensity of the winter prior to the soybean planting and harvest. Upon research this appears to be a legitimate result, as colder winters allow for the soil to better retain the nitrogen accumulated the previous Fall, meaning a richer soil for soybean to grown in, and as a result, grow more.

The third most important feature is the presence of hexazinone, a popular herbicide. [6] Although the purposes of this study are not to study the effects of specific herbicide types, it is certainly encouraging to find that one particular substance's presence holds an impact on yield prediction more than others.

Model	R^2	**MAPE**
Lasso Baseline	0.52	13.30%
Default Random Forest (on Lasso Features)	0.62	10.96%
Tuned Random Forest (on Lasso Features)	0.72	10.58%

Table 6: Comparing performance metrics of baseline model with out of the bag Random Forest, and tuned Random Forest

5 Deployment

This model will not be deployed with a purpose of attaining real-time results, but rather, it will be used to assess the impact that the selected features have on future soybean yield. The model can serve two purposes if deployed in a real setting. First will be to explain to farmers what aspects positively and negatively affect yield based on previous data. The second is that once crops are planted, and the farmer is surveyed for usage of fertilizer, insecticide, herbicide, and fungicide usage, the model will use this data along with exogenous climate data as input to predict their soybean yield. This has immediate impact on the farmer because it will better help the farmer get a sense of how to handle their supply chain, since higher yield (bushels per acre) could mean that they need more resources to accommodate.

As our separation into test data suggests, we intend to deploy this model dynamically to predict yearly soybean yield. Ideally, we see the results of this model serving a tool to farmers, providing additional guidance in their efforts to maximize yield. The deployment will be dynamic in the sense that the model can be updated on new incoming yearly data, to better reflect what is to be expected in the upcoming years.

It is worth noting again that the model is agnostic to state. This is important from a deployment perspective because it performs robustly regardless of which state a farmer using our model is harvesting soybean from. A risk that might arise with the proposed plan is if new types of fertilizers, insecticides, herbicides, or fungicides develop in the market or there are changes in the pesticide regulations, that the current model does not take into account. In this case, the model might need to be retrained entirely to select the most important variables, since the model currently will not know how these new variables behave.

6 Conclusion and Future Work

There is an abundant amount of information available online regarding the agricultural industry. In this report we identified key variables, some actionable, that play essential roles in determining the yield of soybean for a given state in a given year. The biggest challenges we faced with these datasets was that USDA surveys on soil applications are not sent out annually, and thus create a sparsity issue for characterizing yield. In spite of this, interpolation techniques allowed us to fit basic machine learning models to this data to recover promising results. Consequentially, we anticipate that similar techniques could prove useful for other high value crops. Furthermore, improved granularity, such as examining yield at a county level would produce even more insightful results that could be used by industries dependant on soybeans to manage their import and exports plans.

Speaking explicitly on our own data structure, one potential improvement may come from turning this into a classification task. Since the target variable falls within a discrete and normal distribution, it is feasible to consider binning such results into different classes, and predicting yield values for threshold values to serve as the various classes. Sparsity challenges in the feature values would still remain, but by re-forming this task it allows for different evaluation metrics to be tested, which may be more suitable to certain use cases.

Further exploration into a data-driven approach to farming could prove a useful measure to tackle important challenges such as famine and climate change, and this study demonstrates the potential for more advanced approaches.

7 Appendix A

Feature Category	**Source**	**Description**
Fertilizers	USDA	Total applications in lbs, $\frac{lbs}{acre}$, count, $\frac{\frac{lbs}{acre}}{yr}$ and % of area planted treated for fertilizers used.
Insecticides	USDA	Total applications in lbs, $\frac{lbs}{acre}$, count, $\frac{\frac{lbs}{acre}}{yr}$ and % of area planted treated for insecticides used.
Fungicides	USDA	Total applications in lbs, $\frac{lbs}{acre}$, count, $\frac{\frac{lbs}{acre}}{yr}$ and % of area planted treated for fungicides used.
Herbicides	USDA	Total applications in lbs, $\frac{lbs}{acre}$, count, $\frac{\frac{lbs}{acre}}{yr}$ and % of area planted treated for herbicides used.
Temperature	NOAA	Maximum, average, and minimum monthly temperatures.
Precipitation	NOAA	Total monthly inches of rainfall.
Palmer Z-Index	NOAA	Measures short-term drought on a monthly scale.
Palmer Drought Severity Index (PDSI)	NOAA	Measures the duration and intensity of droughts and dryness based on precipitation and temperature data.
Palmer Modified Drought Index (PMDI)	NOAA	Operational version of the PDSI.
Palmer Hydro-logical Drought Index (PHDI)	NOAA	Measures hydro-logical impacts of drought which take longer to develop and longer to recover from.
Crop Moisture Stress Index (CMSI)	NOAA	Measures the impact of both lack and abundance of soil moisture on the national crop yield of Soybeans.

8 Appendix B

Contributions

Rachana Swamy:

Scraped NOAA and USDA sites for necessary datasets.

Peter Simone:

Cleaned USDA and NOAA data, merged and structured all datasets, PCA, feature selection, developed baseline model.

Rahul Zalkikar:

Data interpolation and extrapolation, developed Random Forest model, grid search.

Noah Kasmanoff

Debugging, provided support to other teammates. Wrote and proofread majority of report.

9 Appendix C: Additional Modeling

We also attempted an XGBoost regression by tuning a tree booster with a parameter grid scoring based on mean squared error. Our initial grid was for the tree model and consisted of the same parameters as the Random Forest Regressor. We also re-optimized on a range of L1 regularization terms for weights and finally tuned on a lower learning rate. Our tuned XGBooster Regressor performed extremely well on the training set and slightly worse across all metrics on the validation and testing sets in comparison to the Tuned Random Forest Regressor. We were suspicious of overfitting, and noticed a clear instance in the residual plot of the model error.

References

[1] *2017 Soystats: a reference guide to important soybean facts and figures.* SoyStats.com, 2007.

[2] CNN. 'we've gone this far': Farmers stick with trump over trade. https://www.cnn.com/2019/12/07/politics/trump-soybean-farmers-trade-aid-payments/index.html, December 2019. Accessed on 2019-12-07.

[3] NOAA National Centers for Environmental Information. Climate at a glance. https://www.ncdc.noaa.gov/cag/, Nov 2019. Accessed on 2019-11-11.

[4] Harvest Public Media. Study: Climate change will affect soybeans in 2 ways that cancel each other out. https://www.harvestpublicmedia.org/post/study-climate-change-will-affect-soybeans-2-ways-cancel-each-other-out, January 2019. Accessed on 2019-12-07.

[5] PBS. Why soybeans have out-sized importance in u.s. trade talks. https://www.pbs.org/newshour/economy/why-soybeans-have-outsized-importance-in-u-s-trade-talks, August 2019. Accessed on 2019-12-07.

[6] PubChem. Hexazinone. https://pubchem.ncbi.nlm.nih.gov/compound/Hexazinone, Dec 2019. Accessed on 2019-12-09.

[7] USDA National Agricultural Statistics Service. Quick stats. https://www.nass.usda.gov/Quick_Stats/Lite/, May 2018. Accessed on 2019-11-11.

[8] Jennifer Zhao. More features than data points in linear regression? https://medium.com/@jennifer.zzz/more-features-than-data-points-in-linear-regression-5bcabba6883e, Nov 2017. Accessed on 2019-12-07.