Bayer Planting Intentions

Julie Wisch

9/1/2020

Executive Summary

In this document we aim to answer the following two questions: How well can we predict how many acres of corn will be planted in the United States in a given year? How quickly can we generate this prediction?

The USDA releases annual projections of national planting acreage that are usually within 3 - 4 million acres of the true value. Using the modeling approach outlined in the following pages, we can predict the national planting acreage within 1 - 2 million acres of actual planting. We can make an initial projection in December of the year prior to the start of the market year (e.g. December 2019 for a March 2021 planting) (mean average error of approximately 2.2 million acres), and then continue to refine this projection in the fall of the planting year (e.g. Sept - Nov 2020 for a March 2021 planting). In November, we can predict the spring planting within 1.3 million acres. After December sales receipts, we are able to predict spring planting within 1.15 million acres, on average. This is a substantial improvement over the USDA's projections. In order to do this, we need Bayer order data as well as the publicly available USDA projections.

USDA Projections and Customer Analytics

USDA Projection Availability

The USDA publishes corn planting projections annually. It is the Customer Analytics team's understanding that these models are developed in October, preliminarily released in November, and re-released in final form in February. When the USDA projections are published, the USDA produces projections for an 8 to 11 year range, depending on the year of publication. Projections start with a backwards looking projection to data 1 year prior to the model publication up to 7 - 10 years after the projection publication. These projections are publicly available on the USDA's webpage: https://data.ers.usda.gov/reports.aspx?ID=17821 . The first publicly available year of model projections is 1997.

To clarify the dates involved, here is an example: In October of 2019, the forecasting team at the USDA met and developed a corn acreage prediction. In November of 2019, the team released preliminary projections, with finalized projections coming in February of 2020 for the number of acres planted in March of 2020. They projected the number of acres planted in the US of corn in March 2019 (year -1), all the way up to the number of acres of corn planted in the US in March 2029 (year +9).

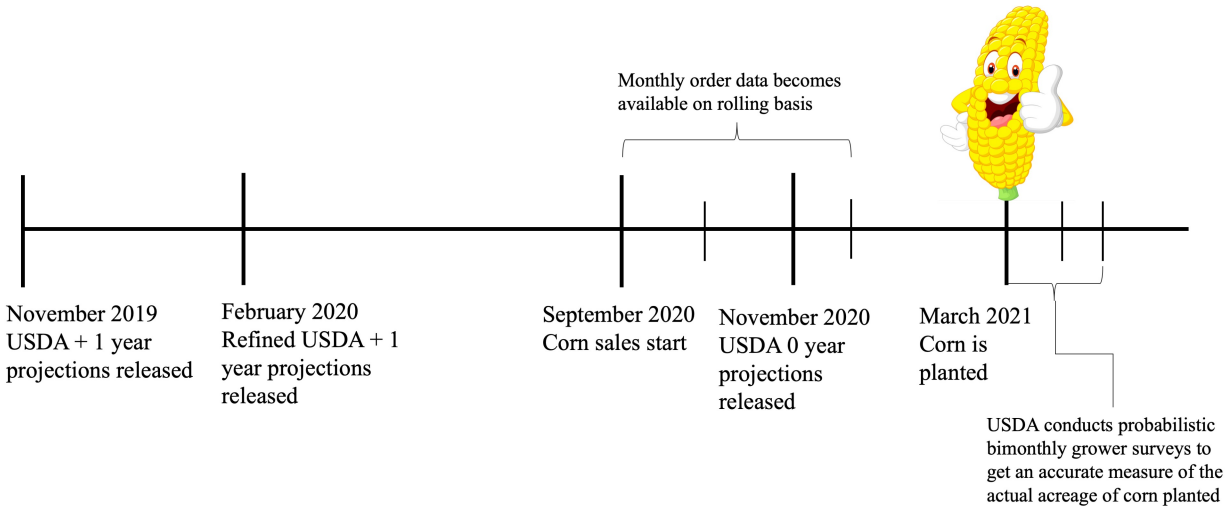
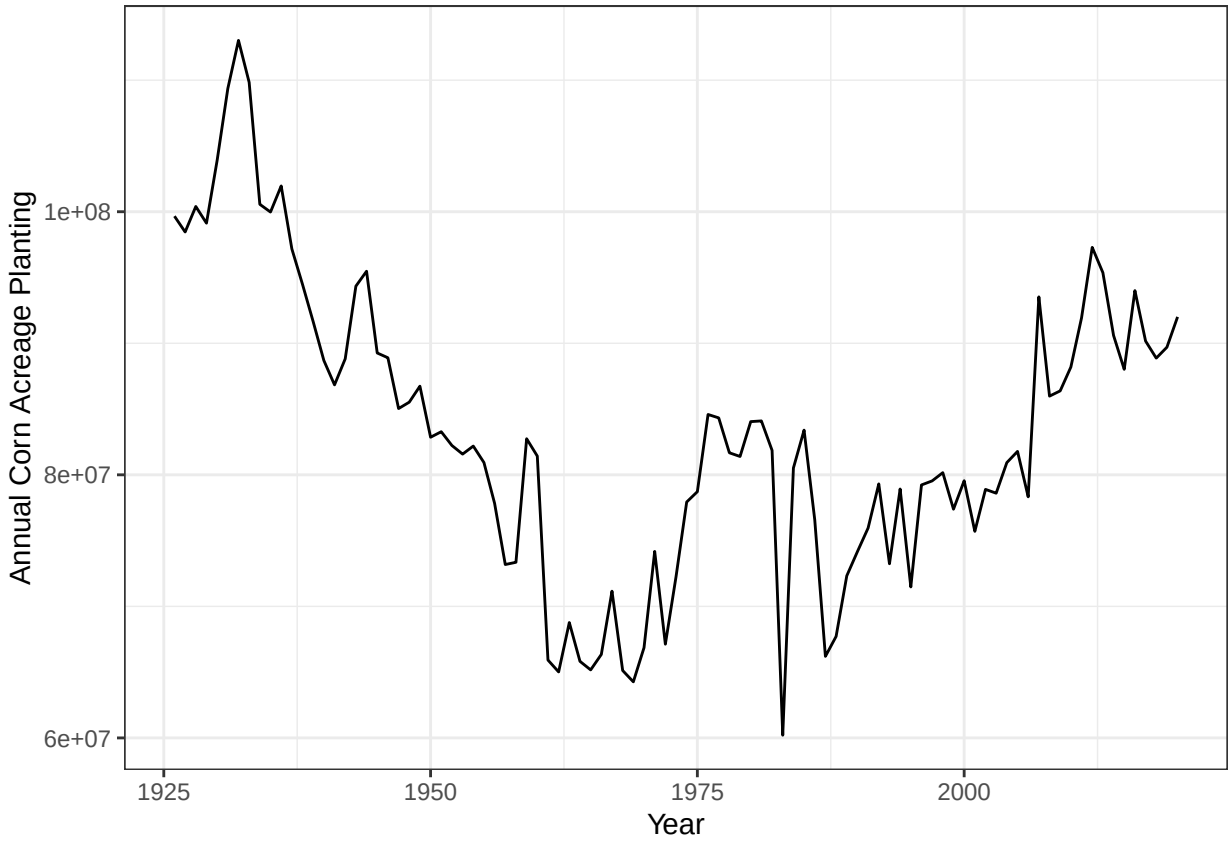


Figure 1: Timeline of USDA Projection Availability.

The USDA also conducts annual surveys in the first two weeks of March in order to project the total number of acres planted in the US for that spring. These surveys are re-conducted in June and August of the same year in order to update the planting totals. The USDA uses these probabilistic surveys to calculate the total number of acres planted annually. These totals can be found here: https://quickstats.nass.usda.gov/results/823A5054-B8F9-3341-AC31-8B03C6990BC1 . For our efforts at modeling the total planting acreage in the US, we will consider these values to be the "ground truth". The first available year of planted acres is 1926. We show all years in the figure below, but for the purposes of modeling, we will only use 2006 - present. For the case of the survey data, the year corresponds with the March where the corn was planted. If corn was planted in March of 2020, the year in this dataset is 2020.

Historical Acreage Plantings

Historical results of the USDA's annual planting survey are shown below.

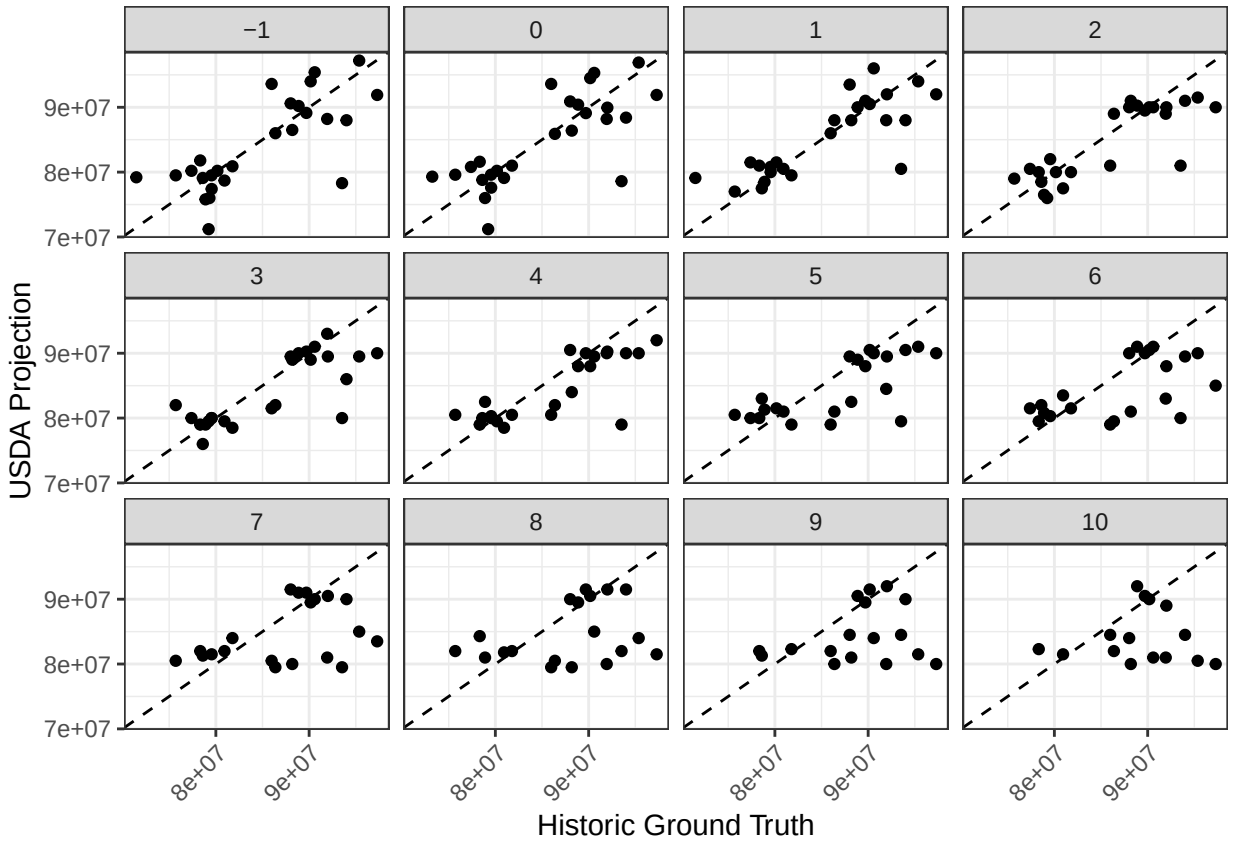


Model Projections vs Historical Reality

Model projections are only available consistently starting in 1996. We were also able to obtain projections from 1993 and 1995. Projections from 1997 on are publicly available online; years 1993, 1995 and 1996 were provided via email exchange with the USDA.

In the figures below, the heading shows the number of years between the projection and actual planting, the ground truth is on the x axis, and the USDA's projection is on the y axis. A diagonal line cuts across the plot. If a projection is absolutely perfect, the point will lie on the diagonal line. If the projection overpredicts, the point will be above the diagonal. Conversely, if the projection underpredicts, the projection will be below the diagonal.

For year -1, the projection is made with full knowledge of the year's events, and you can see that the model works extremely well. The year 0 projection refers to the year where the projection is released in February and the planting actually occurs in March of the same year. The projection also appears to be reasonably successful then. By the time the USDA is predicting 5 or more years into the future, there is a minimally discernable relationship between the ground truth and the projections.



Value - Add Opportunities for Customer Analytics

In addition to access to the USDA's publicly available models, the Customer Analytics team can access sales data for roughly one third of the US corn market going back to 2006. In theory, sales should directly correspond to acres planted, meaning we should have real time insights that the USDA does not have.

Where can the Customer Analytics team add value?

- By creating a model in the fall for the following March that performs better than the +1 year model that is available before the USDA November estimates are released
- By creating a model in the fall or winter for the following March that performs better than the 0 year model

Customer Analytics Proposed Model

The USDA will release projections for up to 9 future years. The objective of this model is to take the available projection for the upcoming year and augment it with our sales data as it becomes available.

While we considered multiple models and evaluated their performance over a variety of sliding windows, here we will present our proposed model. Complete details on the approaches considered are available in the appendix.

The model we use relies on the XGBoost algorithm and trains on 8 years of historical data. We include in-season sales, historical sales, and a variety of other features.

A model using only historical sales and the USDA's +1 projection is the baseline model. This is called the "BeforeOrders" model. As sales data becomes available, we will add in September, October, and November sales. By the end of November, the 0 year USDA projection becomes available. We advise creating two models: one model that can be deployed as early as the December prior to the in-season December (e.g. December

2019 for a March 2021 planting) and yields an average error of 2.21 million acres. By the end of November, we have a model that predicts planting acres within, on average, 1.25 million acres of actual. By the end of December, we should deploy the second model using the revised estimates and sales data. This model can be refined throughout the spring, but seems to plateau after the addition of December sales data, yielding an average error of 1.1 - 1.2 million acres.

Below we provide a detailed comparison of the error associated with the USDA projections. The +1 model can be released as early as the December prior to the December of the market year and then refined throughout the fall of the growing season. The 0 model should be released after December of the market year. Details on the +1 model and 0 model are available in the appendix. The error metrics for the USDA, as well as the Customer Analytics model, are limited to the time period for which we have sales records (2006 - 2019).

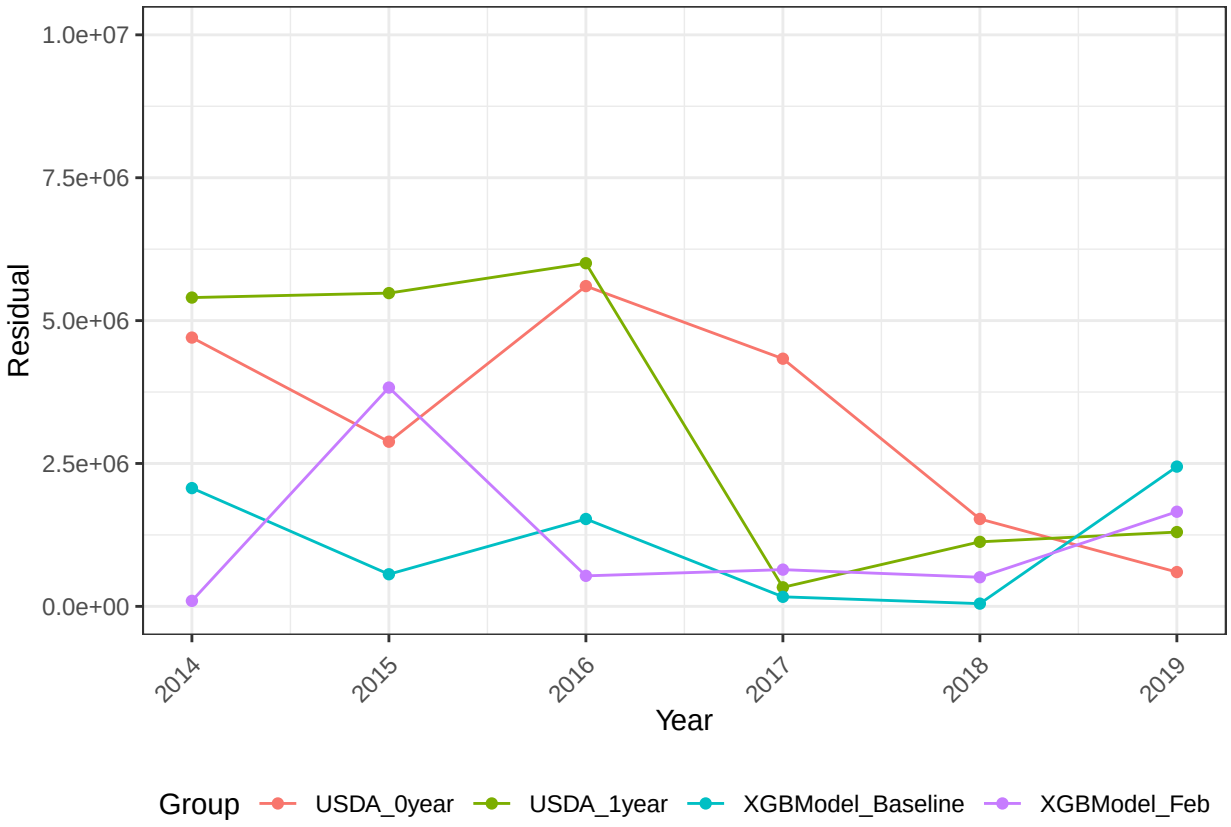
Note that the +1 projection published by the USDA outperforms, on average, the 0 year projection from the USDA. This suggests that the USDA is overly responsive to in season variables. In contrast, our model is able to produce estimates that are much better than the USDA projections (MAE = 2.21 for our Pre-Order model as compared to MAE = 3.22 for the USDA +1 projection), and then our model continues to improve as more sales records become available. About 80% of Bayer sales are completed by the end of November, and our model does not improve substantially from December - February. The model error drops to about 1 million acres after December sales records are added, and no gains are observed after December.

Table 1: Error Metrics Associated with Different Models

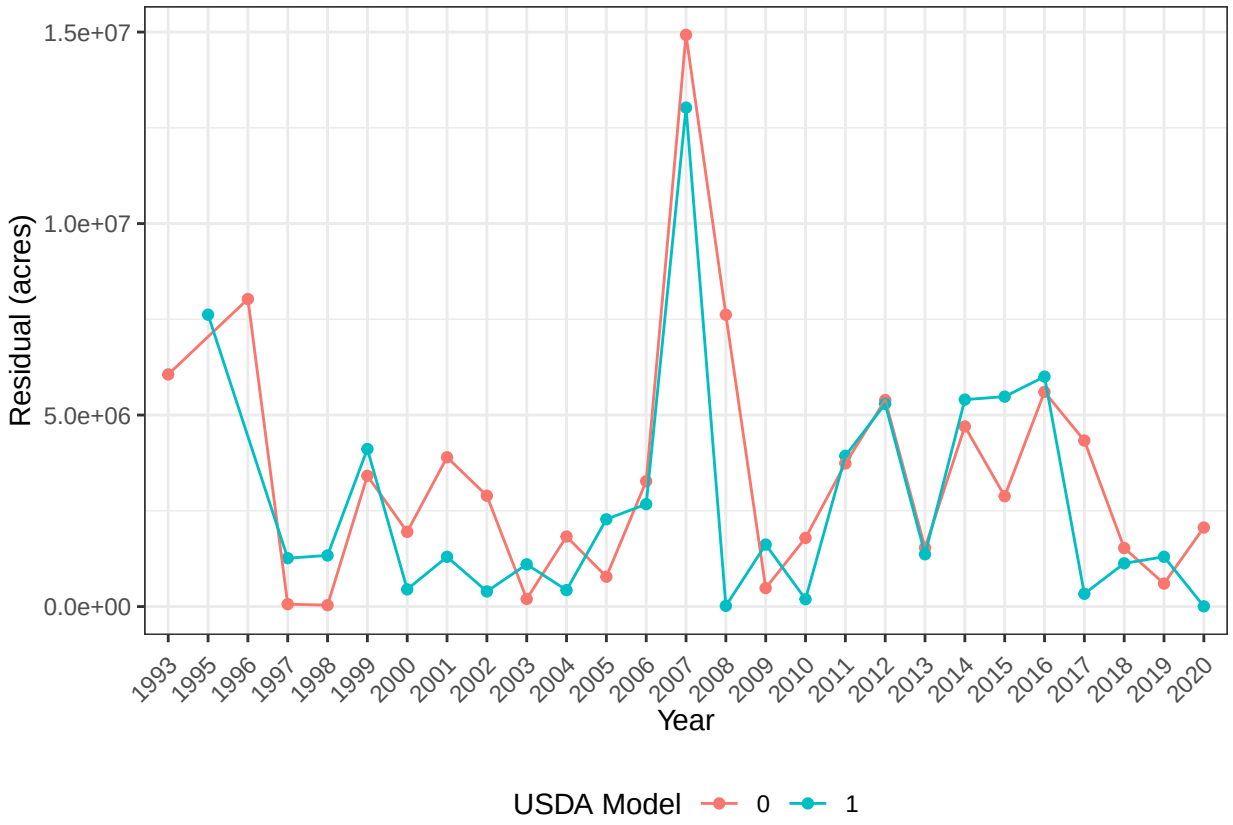
	Model	MAE	Model	MAE
3	USDA 1 Year Projection	3.22	+1 Year Baseline Model	1.22
4			Sept XGB Model	1.22
5			Sept & Oct XGB Model	1.24
6			Sept - Nov XGB Model	0.95
7	USDA 0 Year Projection	4.03	0 Year Baseline Model	1.58
8			December Model	1.23
9			January Model	1.23
10			February Model	1.21

A Word of Caution

We caution readers of this document that we are working with a limited set of data. We have sales records starting in 2006. If we rely on 8 years of training data, as in the model outlined in this main body of the document, this only leaves 6 years of data to test on. When we compare our model residuals - that is, the absolute value of the difference between our model predictions and the ground truth number - our model demonstrates a much more consistent behavior than the USDA's models, and generally outperforms the USDA.



However, use of 2014 - 2019 may mask the true variability in historical plantings. Recall the first plot of the document where acreage plantings have varied wildly in the last 100 years, even though recent history has suggested a relatively steady increase in planting acreage from year to year. We can even see this in a residual plot looking at the behavior of the USDA's in season (0) and 1 year prior to season (1) models. There are several years (e.g. 1993 - 1996, 2007 - 2008) with very large residuals outside of the 2014 - 2019 testing window we have available for our model.



Conclusions

From our work, we conclude the following:

- The USDA in-season projection is slightly worse than the USDA one-year-prior projection.
- The Customer Analytics team can produce a model that substantially outperforms the publicly available USDA model, as early as the December before the start of the market year (e.g. December 2019 for the March 2021 planting)
- The Customer Analytics team can refine this model throughout the fall, with the final useful information being added at the end of December.
- The Customer Analytics team can provide estimates that represent between a 30% and 70% reduction in error, depending on when the estimates are produced.
- Enthusiasm with respect to these results should be tempered by the knowledge that true variability may not be observed in the 6 years of available testing data. Testing of these results on future years - or with older sales data - is necessary.

Appendix

Model Features List

Our baseline model includes the following features:

- Net national brand orders in September, 1 and 2 years prior
- Net national brand orders in October, 1 and 2 years prior
- Net national brand orders in November, 1 and 2 years prior
- Net national brand orders in December, 1 and 2 years prior
- Running total of national brand orders for Sept - Oct, 1 and 2 years prior
- Running total of national brand orders for Sept - Nov, 1 and 2 years prior
- Running total of national brand orders for Sept - Dec, 1 and 2 years prior
- USDA Projections for the upcoming year

On a rolling basis, monthly sales totals are added.

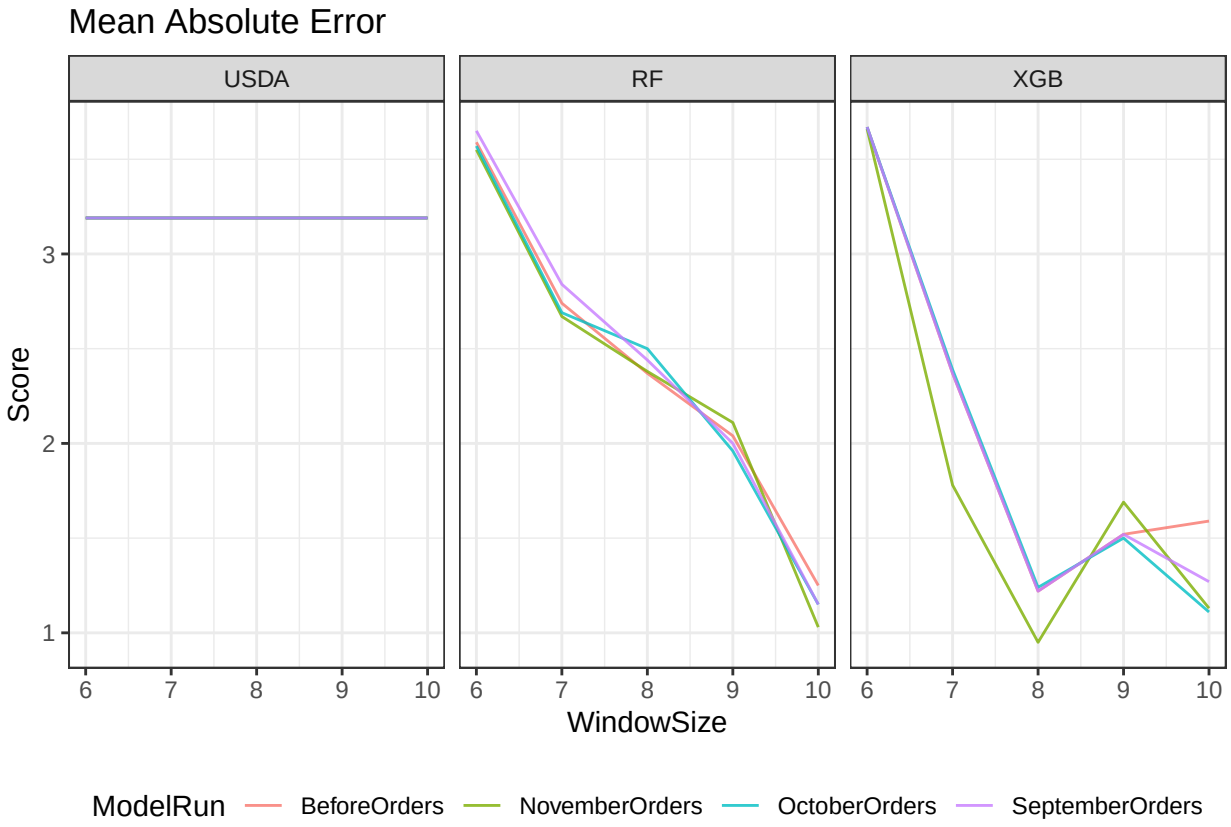
+1 Model Year Details

The USDA will release projections for up to 9 future years. The objective of this model is to take the available projection for the upcoming year and augment it with our sales data. We will consider two different models (Random Forest (RF) and XGBoost (XGB)) and compare them to the available published USDA model. With our sales-augmented models, we will include the following features:

- Net national brand orders in September, 1 and 2 years prior
- Net national brand orders in October, 1 and 2 years prior
- Net national brand orders in November, 1 and 2 years prior
- Net national brand orders in December, 1 and 2 years prior
- Running total of national brand orders for Sept - Oct, 1 and 2 years prior
- Running total of national brand orders for Sept - Nov, 1 and 2 years prior
- Running total of national brand orders for Sept - Dec, 1 and 2 years prior
- USDA Projections for the upcoming year

Further, we will simulate the progression of the fall. We will create a model with the above features. This will be the baseline model for the upcoming year, called the "BeforeOrders" model. As sales data becomes available, we will add in September, October, and November sales. By the end of November, the 0 year USDA model becomes available. At that point, it probably makes more sense to use the newer published numbers and shift modeling approaches.

For these models, we will train on a range of windows, using between 6 and 9 years of data. We will calculate error metrics based on this sliding window. In the plotted error metrics below, we see that both the random forest and XGBoost models outperform the published USDA numbers. In all cases, even using sales data from 1 and 2 years prior are sufficiently informative as to make the BeforeOrders model better than the published USDA model. The XGBoost model performs better than the random forest model for nearly all cases, so we will discuss that further.



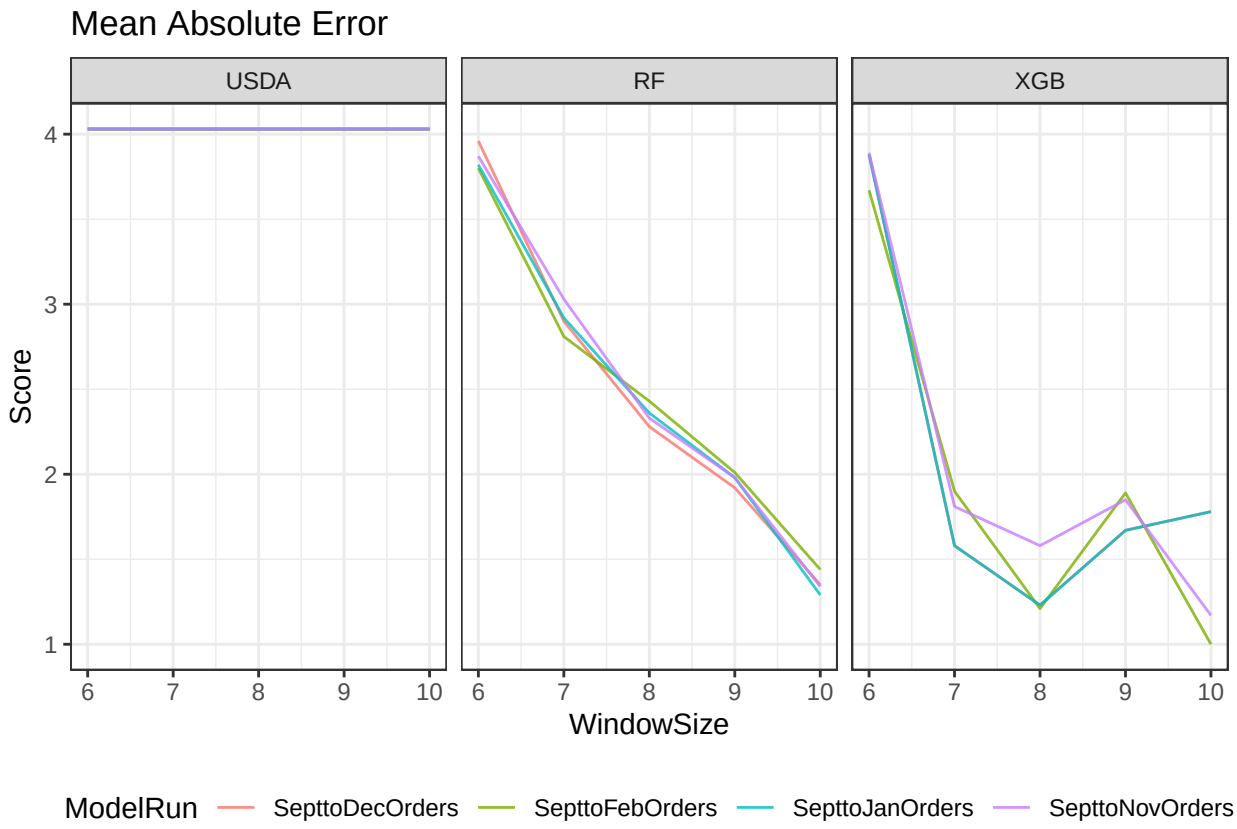
From these results, we can conclude the following:

- The Customer Analytics team can produce a model that outperforms the publicly available USDA model in the fall
- The Customer Analytics team can refine this model throughout the season

0 Year Model Details

The model described in the previous section outperforms the projections made by the USDA for planting one year out. In November, the USDA will release new projections for the following March. The objective of this model is to take the new USDA Projections, released in November, and augment them with our sales information in order to see if we can improve projections.

Similar to before, we will start with a base model, in this case a model that uses both the new USDA projections and sales records from September through November. We will subsequently add December, January, and February orders. And again we find that both models are better than the USDA numbers, and the XGBoost model is the best performing model. Much like before when the addition of one month's sales records does not yield an improvement, but two months' sales records does; we see that the model that includes sales records through December has a consistently strong performance.

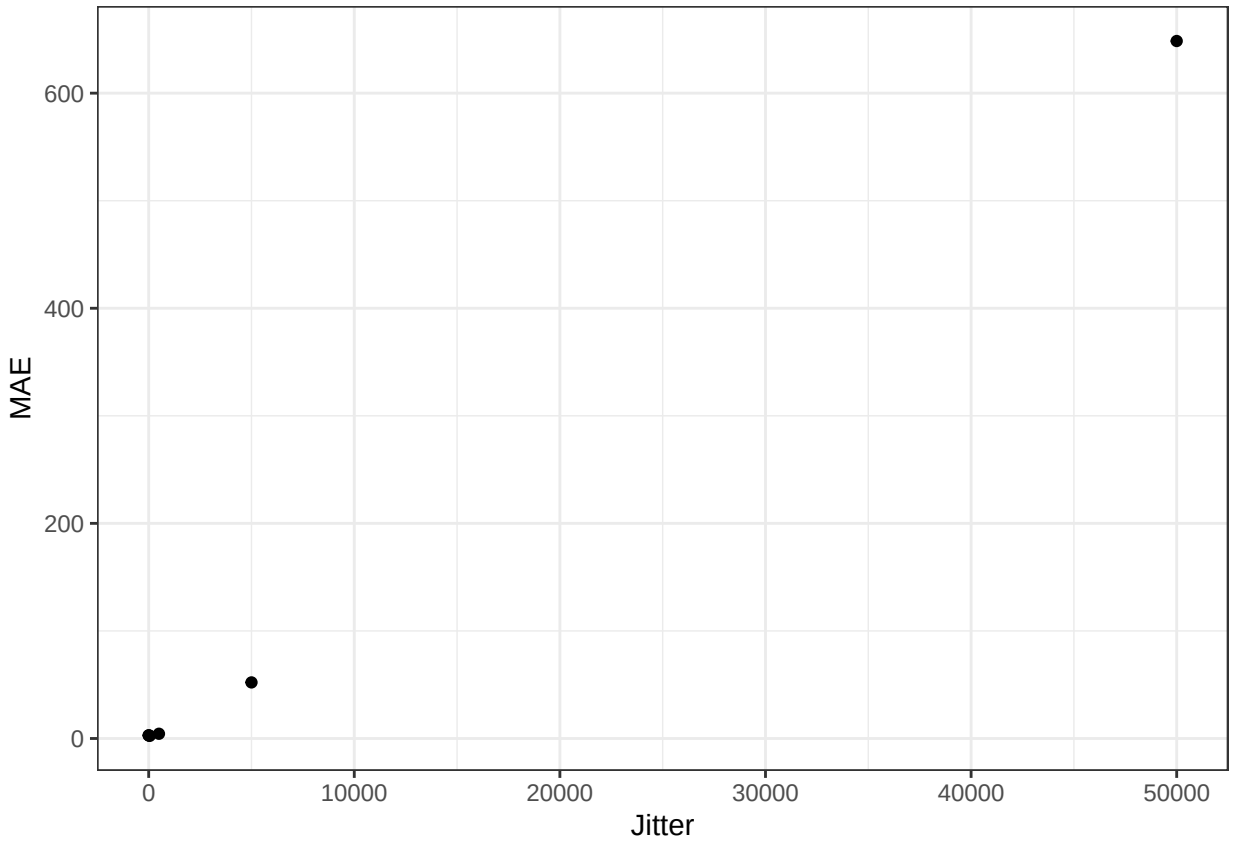


From these results, we can conclude the following:

- The Customer Analytics team can produce a model that outperforms the publicly available updated USDA model in the spring
- The Customer Analytics team can refine this model throughout the season
- The USDA performs worse in season, and this initial noise introduce by the USDA projection negatively impacts the baseline spring Customer Analytics model. After additional sales data is added, the Customer Analytics model improves beyond the best available fall numbers.

Jitter Test

In order to test model robustness, we will apply the jitter test. We will increase the amount by which we perturb the response variable and then compare the MAE from the original (pertubation free) model to the jittered models. If the model is working properly, the MAE will get progressively worse as the jitter increases. That is what we see in the figure below.



Permutation Test

Another way to assess model robustness is to randomly shuffle the years of the dataset. If we break the link between years and the response variable, the model should break. If it doesn't, the model requires further examination. Here we compare the MAE of the unshuffled model to the shuffled model.

Table 2: Results of Shuffle Test

OriginalMAE	ShuffledMAE
2.21	5.82