Apollo Agriculture Updates to EfD
April 2018

Overview

Apollo Agriculture is a technology company founded to help smallholder farmers maximize their profits. We leverage advances in machine learning, remote sensing, and mobile money to deliver input finance and agronomic advice to drive down the costs and increase the scalability of agricultural finance. With the support of Enterprise for Development, Apollo was able to hire a Data Scientist to accelerate the pace of our credit modeling R&D. Below, we provide an overview of our data science team's progress in this work. Our team has taken a multi-model approach to credit decisioning, meaning that rather than building a single model, we are using several, overlapping models using different data sources and statistical techniques. This approach prevents us from overfitting on 2017 data, which provided a relatively narrow data set. We describe this work in greater detail below.

We have also provided a summary of our efforts to understand customer yield changes (and, therefore, income changes). While our data science team is still working to build the first version of our satellite yield models from 2017 data, we have sufficient data to eventually make a first gross estimate of yield differences between Apollo and non-Apollo farmers, likely in Q3 of this year. We are currently developing a map to visualize 2017 yield data, and will plan to share this with EfD in late April. We also describe some of the challenges we've faced in data collection and how we plan to approach data collection in future seasons.

Credit Modeling Development

We have built a set of tools that enables us to extract relevant information from the customer lifecycle and convert it to features that can be processed with machine learning. This includes features processed from satellite data, questions directly asked of the farmer (e.g., how much was your yield last year?), and inferences from our communication platform (e.g. how long it takes a customer to respond to an SMS, whether or not she listened to an entire training, etc.).

For our 1,016 loans extended in our first season, we used a minimal underwriting process focused only on verifying a farmer's identity and preventing fraud. This approach allows us to train and evaluate our model based on a random sample of farmers, rather than a sample biased by previously established "intuitive" criteria. We call this approach "lend-to-learn." We are piloting our credit model for the first time in 2018 loan decisions, while also extending a subset of loans using our lend-to-learn approach. This approach allows us to test the effectiveness of our credit model against a control group while continuing to build unbiased credit model training data.

Apollo's Approach to Early Stage Credit Model Development

Rather than relying on one single aggregate credit score, we are using a few of the highest-performing models, which rely on different sources of data. Any customer that is selected by one or more of these individual models is approved. While this will not be the approach we take to credit decisions in the long-term (since the optimal credit modeling approach will become

increasingly clear over time), making lending decisions with multiple models serves as a hedge against making bad predictions with a single model in our early seasons and mitigates the risk that a single, overfit model leads us to approve an unnecessarily narrow subset of farmers. This multi-model framework helps us overcome the limitations and risks posed by the relatively small sample size of our first 1,016 loans.

Given our strong belief in the predictive power of certain data points (based on our team members' previous experience, as well as academic literature), we have augmented our credit model decisions with a small number of rule-based filters applied to **all** prospective customers. In the 2018 season, we are automatically excluding applicants who have 2 or more prior non-performing loans (or the same for their spouse) assessed through a Transunion credit report. As we gather more data, we will reduce our reliance on rule-based filters (for example, using machine learning-based approaches to identify applicants with a prior default, but who have a strong probability of repayment).

Individual Model Selection in Multi-Model Framework

As described above, our credit model in our early seasons will consist of several unique models, and a customer selected by any one of these models will be considered "approved." In selecting each unique model, we are assessing the following two characteristics:

1) Model quality and generalizability: does a candidate model have sufficient performance on 2017 data to be considered? Identifying whether a model is likely to generalize well to 2018 applicants is a challenging task. For example, in 2017, one of the two regions in which we made loans experienced significant drought conditions, and farmers in this region repaid at a measurably lower rate. A model that simply rejected all farmers in this region and accepted all farmers from the other region might appear to perform well on the 2017 evaluation data, but would be extremely unlikely to generalize well since rainfall patterns and drought conditions will certainly vary from year-to-year (and would not not be useful from an operational perspective, given the desire to make loans in both regions). Our aim is to avoid building a model that indexes against random events, and instead tries to understand the more fundamental dynamics that lead to repayment or non-repayment. We therefore used the following frameworks for assessing generalizability:

- Overall performance: trained and evaluated on both regions
- In-region performance: trained and evaluated on a single region
- Out of region performance: trained on one region and evaluated on another region
- Geographic heterogeneity of predictions (as in, a model that selects applicants across distinct geographic regions).

2) Model differentiation: Based on the set of models that passed the initial screen, which should we include for a multi-model framework (ie. evaluate as part of a group of models)? In making this determination, we prioritized models that added a differentiated perspective. Models used in the selection process are intended to be diverse both in the features they use and in model type/complexity. We describe each unique model in Table 1.

Data Modeling and Machine-Learning Techniques

We have used a variety of different approaches in our data modeling, from random forest models that let us identify which features have an impact, to neural networks that enable us to extract more subtle, less perceptible relationships between data. Each of these approaches has different advantages and risks associated with it, and using multiple techniques is another way to mitigate the risk of overfitting on a relatively small data set. Data modeling techniques we have used include:

- **Random forest:** Random forests operate by training multiple decision trees, where each tree only sees a random subset of the total dataset. The output at prediction time is the average vote of all the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.
- **Gradient boosted trees:** Gradient boosting produces a prediction model in the form of an ensemble of sequentially trained weak prediction models, each fit to the error of the previous prediction. Gradient boosting can produce a model that performs well through the aggregation of underfit models (in this case, decision trees).

In terms of machine learning ("ML") techniques, we've invested heavily in auto-encoders and transfer learning:

- **Autoencoders** are a technique for unsupervised machine learning. The fundamental limitation for traditional, "supervised" ML is the quantity of data available for training. Training a supervised ML model traditionally requires substantial amounts of labeled data (e.g., a computer could not learn whether an image is a cat or a dog without thousands of images of cats and dogs). Unsupervised learning is a technique that lets computers infer the basic structure of images or other data by teaching them about the differences between much larger sums of data. By restricting the possible categories into which satellite images can be divided, the computer begins to learn how to categorize fields from forest from roads, distinctions between types of houses, or presence of other relevant visual features within a farmer's compound without having to receive large quantities of labeled images. Since there is much more unlabeled data than labeled data, this let's most of the "hard" part of the training - the identification of predictive features in the satellite data - occur with cheap and easily available data. Labeled data is only used to train the model for the selection for the specific problem at hand - in our case, using the features to predict credit risk. In this work, we are building on recent advances in Generative Adversarial Networks ("GANs") to improve the quality of autoencoder generated features. Because the spatial characteristics of a satellite image and the location of the features in the images is central to the credit assessment we are building, we combine these approaches with convolutional neural networks which allow us to preserve spatial characteristics.
- **Transfer learning** is a technique that allows a neural network to store knowledge learned solving one problem and apply it to a different but related problem. As an example of one specific approach, we can train a neural network to map between daytime imagery and nighttime light emissions, leveraging an approach developed by Stanford scientists. Since nighttime light emissions is partially driven by income level, this allows us to first train our model to detect many features that are relevant to the prediction of assets. We can then take this model, which effectively predicts income level, and "re-train" it to determine creditworthiness, using a much smaller quantity of training data than would otherwise be required. As with unsupervised ML, being able to reduce the amount of training data

needed is critical given the expense and length of time required to collect repayment data as an agricultural credit business.

Table 1: Model Types of Multi-Model Framework

Model Number	Feature sets	Feature Set Description	Model type
1	Features extracted from satellite images using a generative adversarial network ("autoencoder"); and registration survey	Features extracted from satellite images using a generative adversarial network (see above); and registration survey data	Binary classification wrapper to random forest regression (treating the repayment as a continuous variable)
2*	Features extracted from satellite images using transfer learning	Features extracted from satellite images via transfer learning, using a convolutional network trained on nightlights imagery; see description above	Gradient boosted trees classifier ***Note:** not listed in the attached excel workbook
3	Registration survey	All customers undergo a registration survey with Apollo that collects over 200 points of socioeconomic, demographic, and behavioral data.	Random forest classifier
4	TransUnion features (both raw and processed features from credit reference bureau)	Transunion credit data provides both a singular credit score as well as "raw" features like number of open and closed loans, number of non performing loans, etc. We are using both credit score and raw components for this model.	Gradient boosted trees classifier
5	Features standardized by region (eg. standardized separately in Njoro and Bahati): • Registration survey • Remote sensing-based features (eg. rainfall and soil)	This feature set includes both registration data and remote-sensing based features, including rainfall and soil data from publicly available data sources. Standardizing by region addresses the challenge poised in our first year, in which one region experienced drought and one did not.	Random forest classifier
6	Registration survey using a continuous variable	This feature set also includes registration survey data, but treats repayment as a continuous variable. Since someone who repaid 50% is likely a different risk than someone who repaid 2%, and success	Binary classification wrapper to random forest regression (treating repayment as a continuous

		or failure may be driven by extrinsic factors (like an unexpected illness or good weather), this model treats different levels of default differently. We trained this model by dividing repayment into quantiles. We optimized the score for the middle three quantiles (where the top quantile was set to 1 and the bottom was set to 0). This allows the data to determine how strongly we should weight someone that repaid 50% compared to someone that repaid 2%	variable) .
7	Subset of features from TransUnion and registration survey feature sets selected via recursive feature elimination	This feature set include features from TransUnion as well as registration survey data selected via recursive feature elimination, wherein features with low signal are selectively and automatically eliminated to make the model more effective.	Gradient boosted trees classification
8	Ongoing work: Features extracted from satellite data via supervised learning techniques, for example yield model	We have built a data pipeline for satellite yield modeling that will enable more automation of yield comparisons via satellite in future seasons. Once we have an effective yield model, (expected in Q2/Q3 of this year) this feature set can be used in credit decisions.	TBD

The question of which model "performs best" is not simple and, at this stage, not yet possible. Model performance is not only a function of predictive skill and default rate. We also consider, for example, acceptance rate and cost to Apollo. For example, if a model accepted only 1% of farmers but had a 99% repayment rate, that model would not be considered high-performing. However, in terms of expected reductions in default rates, we expect the regression models based on registration survey data to achieve the greatest reductions on 2018 customer selections. With the caveat that these particular models are at a higher risk of overfitting, we currently expect ~20% default rates among customers selected with these models.

Harvest Data and Satellite Yield Modeling Update

In our early seasons, we must physically collect harvest boxes to build our satellite-yield model. Once we have collected sufficient harvest measurements, we can correlate actual yields with pixels from satellite imagery of customer farms, which allows us to accurately predict yields via satellite alone. Unfortunately, harvest data collection was a significant challenge for us in our first year. We were not able to collect harvest data for as many of our customers and our control group as planned, due to a mix of operational and timing challenges related to harvest and the late and varied onset of rains. As a result, we have less harvest data for yield model training this year than

we anticipated and an inconclusive understanding of how customers' yields compare against our control group.

Despite the limitations of data collected for the 2017 season, we gained substantial insight into how to better manage this process in future seasons, and we have successfully built a data pipeline for satellite yield modeling that will enable more automation of yield comparisons via satellite in 2018 and beyond. As noted in the table above, our satellite yield modeling work is an ongoing part of our data science efforts and credit decision-making R&D, and we expect to have a first version yield-model by Q2/Q3 2018. This will allow us to make an informed estimate of average yield increase for Apollo farmers relative to others, though this estimate will be influenced by certain challenges and potentially confounding factors.

Challenges:
We found that individuals in our control group (farmers who had gone through our registration process but were randomly denied) were -- perhaps unsurprisingly -- disinclined to allow us to collect harvest measurements. We also recognize that there is significant potential for selection bias in harvest measurement collections this fall (ie. non-Apollo farmers that have good yields may be more likely than those with poor yields to allow our field agents to collect data, especially if they want to become customers in the future). Another potential opportunity for selection bias, though likely less important, is that agents dispatched for data collection may have been more likely skip over non-Apollo farmers vs. Apollo farmers whose harvests were negatively affected by pests. It's plausible that such selection bias would mask improvement in yield from being an Apollo farmer. Many control group farmers also appear to have harvested earlier than our average customer, perhaps because Apollo farmers planted in the same time window.

Implication for 2018
Once we have an effective yield model, we can compare yields between control farmers and Apollo farmers using predicted yield from a satellite yield model. The first version of this yield model capability will likely be ready in Q2/Q3 of 2018. In the meantime, we are proactively addressing challenges experienced in 2017 through several different approaches:

- We will set clearer up-front expectations (eg. in loan contract and reiterated at other points) that Apollo has the right to take yield measurements for all customers. The goal is to minimize selection bias and sample regardless of repayment, yield, etc. While we recognize that getting yield measurements from non-repayers may be harder in practice, clear communication from the outset may help.
- We are evaluating best practices for establishment and engagement of a control group, to mitigate unwillingness to participate in harvest data collections. To that end, we are reviewing lessons from the One Acre Fund describing their practical experience with different yield comparison frameworks. Options include randomly denied farmers (in an RCT), interested neighbors (which may minimize selection bias), or newly enrolled farmers.

We are also considering providing financial incentives for control group farmers to ensure participation and generate goodwill.

- To the extent possible, we will ensure that the temporal distribution of harvest measurements is similar between Apollo and non-Apollo farmers.