Modelling Food Insecurity in Ethiopia

Towards a machine learning model that predicts the transitions in food security using scalable features.

Joris Westerveld[1]
J.j.l.westerveld@students.uu.nl
and
Sjoerd Stuit[2]
S.m.stuit@uu.nl

Marc van den Homberg[3]
Mvandenhomberg@redcross.nl
and
Dennis van den Berg[4]
Dlvandenberg@rodekruis.nl

Stijn Heemskerk[4]
Sheemskerk@rodekruis.nl
and
Aklilu Teklesadik[5]
ATeklesadik@redcross.nl

[1]MSc. Artificial Intelligence Graduate (3922693), Utrecht University
[2]Department of Experimental Psychology, Utrecht University
[3]Scientific Lead and Applied Researcher, The Netherlands Red Cross
[4]Volunteer, The Netherlands Red Cross
[5]Technical Project Manager

May 2, 2019

1 Abstract

Insecurity concerning food resources has become a more and more serious problem. Therefore an initiative of the Netherlands Cross, called the 510, wants to use data to positively impact faster and more (cost) effective humanitarian aid. The goal of this study is to create a machine learning model (in our case a Xgboost model) that uses scalable features (like satellite imagery) to predict the transitions in food insecurity with the 510. After optimization through resampling, feature engineering and hyperparameter tuning we validated the Xgboost model by comparing it against several baselines. The Xgboost model performance (f1 macro score of 0.526), on average, got close to the benchmark (predictions of *Famine Early Warning Systems Network* (2019), which had a f1 macro score of 0.637). Nevertheless our model did identify improvements in food security (f1 score of 0.506) of livelihood zones better than the benchmark (f1 score of 0.498). Other results of this study is that the features that the Xgboost model identified as most relevant, corresponds with the study of Misselhorn (2004), like climate and land drivers. Furthermore the performance of the Xgboost model is also better for varying prediction intervals (4 or 12 months ahead) compared to the baselines. Lastly the Xgboost model also revealed that there is a spatial dependency between livelihood zones, since similar predicted livelihood zones seem to be clustered. All in all, this study showed that our Xgboost model has predictive value, which gave new insights but also opens new doors. There is potential to improve it further, by adding more features and taking the spatial dependency into account. This can, in the future, hopefully get us closer to optimizing the decision-making of the humanitarian assistance and give more insight about the complex phenomena of food security.

Contents

2 Introduction

Even in this modern age, with a rapidly growing population and tesla's being launched into space, 124 million people across 51 countries faced crisis levels of acute food insecurity or worse in 2017 (Food Insecurity Information Network, 2018). Food security exists when all people, at all times, have physical and economic access to sufficient safe and nutritious food to meet their dietary needs and food preferences for a healthy and active life (FAO, 2003). It is troublesome however that there seems to be an upwarding trend in food insecurity since in 2017 an estimated of 124 million people faced acute food insecurity, compared to the 108 million people in 2016 (Food Insecurity Information Network, 2018). More research is needed how agencies active in the humanitarian aid can support and bring relief to these food insecure regions more efficiently and what tools these agencies can use to optimize their decision making.

2.1 Context description

The amount of data that has been collected over these past years has grown exponentially. Usually the term Big Data has been used to define these large volumes of data, which can be structured, unstructured or aggregated data sets (Kitchin, 2013). Moreover, using these heaps of data to our advantage cannot only lead to just understanding patterns better, but also to forecasting these patterns. Subsequently, when we forecast these patterns this could potentially also lead to taking action to prevent these identified patterns from happening. This is relevant, because there is a growing understanding that timely finance prior to a disaster can be more cost-effective than post-disaster expenditures (Guimarães Nobre et al., 2018). Similarly 510, which is an initiative of the Netherlands Red Cross, wants to use data, to positively impact faster and more (cost) effective humanitarian aid. Moreover it wants to shape the future by converting data into understanding, and put it in the hands of humanitarian relief workers, decision makers and people affected, so that they can better prepare for and cope with disasters and crises (510, 2018). To give a more concrete example, 510 predicts impending disasters on vulnerable people that live in areas prone to natural disasters. 510 calls this Impact Based Forecasting which consists out of three steps (the Red Cross Red Crescent Climate Centre, GRC and 510, 2018). The first step is to understand the risk and to determine which areas are most vulnerable. The second step is identifying the impact which should help

to identify trigger level, so it can give an identification what level of risks need to be reached to start the overall process of the Forecast Based Financing. The last step is called Forecast Triggered Action, which means that when a certain threshold has been reached funds will be released to allow people in the impending disaster areas to get the means in order to protect themselves and take action. Our study is closely related to this project (IBF), however in our case the focus is not on natural disaster prediction, but on predicting food security.

Moreover finding a variable that could summarize food security also poses a challenge. Fortunately, *Famine Early Warning Systems Network* (2019) has created the Integrated Phase Classification (IPC) class, which consists out of 5 different ordinal classes (see figure 17 in appendix A.1). The IPC classes have been created by *Famine Early Warning Systems Network* (2019) in order to make the conceptualization about food security easier. At first, different humanitarian agencies used different variables to define food security. Having one more general definition made it easier to share their conclusions with each other with regard to food security and of course take action if need be. By creating this harmonized approach *Famine Early Warning Systems Network* (2019) made sure that that this framework can be used across countries and regions, and over time.

2.2 Problem Statement

Supporting areas that lack food security can be a big challenge. Organizations like the Interchurch Organisations for Development Cooperation (ICCO) and the Netherlands Red Cross (NLRC) try to increase and stabilize food security where this is needed by intervening, creating development programs and supporting different areas during a food crisis. Most of the humanitarian aid organizations use survey data to measure the status of food security (Barrett, 2010b). The Household Food Insecurity Access Scale (HFIAS) would be an example of a questionnaire used in these situations (Swindsdale & Bilinsky, 2006). However the constraint with this survey measurement is that it can take a lot of time, is not always available in every region and mostly for small areas, and usually is not regularly updated (Barrett, 2010b). Secondly even though survey data can give an indication for food security as a snapshot at that moment, it cannot predict food security.

2.3 The Study

2.3.1 Goal of this study

The main goal of this study is to develop a model that can determine whether the IPC improves or deteriorates in a livelihood zone in the future. A livelihood zone is defined as a geographical area within which people share basically the same patterns of access to food and income (that is, they grow the same crops, or keep the same types of livestock), and have the same access to markets. So it basically divides the country into homogeneous zones within which people share broadly the same pattern of livelihood (HEA, 2018;Grillo, 2009) (see figure 1). We decided to call the transition from one IPC state to another: a change event. By emphasizing on these transitions, we want to determine which variables determine an improvement or deterioration in food security. By predicting these transitions in the IPC, we can get closer to the goal of optimizing the humanitarian assistance and contribute to the effectiveness of the impact based forecasting.

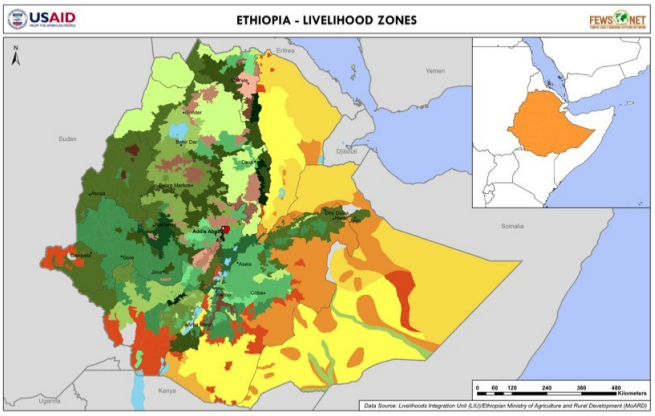


Figure 1: A figure depicting the different livelihood zones, which is retrieved from (*Famine Early Warning Systems Network*, 2009). In each of these homogeneous zones people share broadly the same pattern of livelihood.

Moreover one of the sub goals is that the model should be scalable, in such a way that it would be easy to make predictions for different countries. More specifically, the features of the model have to be available for several countries and be detailed enough so it can be aggregated on livelihood zone level. In order to overcome these challenges we opted to use satellite imagery. Satellite imagery should be detailed enough so we can aggregate the data on livelihood zone level and

also be scalable. Another sub goal is to validate the model by testing it against several baselines. These baselines do not use machine learning modelling, but are for example based on heuristics. The baselines are important to determine whether our model performs more effectively than chance and simple heuristics. With all these goals in mind, it is important to note that this model will most likely not solve the food insecurity problem. However it might at least give more insight, a better situational awareness of food security and optimize the process to send humanitarian assistance so intervention can be set up.

2.3.2 Scope of this study

The countries in Africa are highly relevant case studies because, these have the highest prevalence (25 %) of people living in households where at least one adult has been found to be severely food insecure, as a percentage of the total population (FAO, IFAD, UNICEF, WFP and WHO, 2017). However due to time limitations, we can't test every country. Thus we opted to limit the geographical scope to Ethiopia. To elaborate, Ethiopia is a good case study because it is prone to food insecurity. More specifically a study from Geda and Stoecker (2011) showed that in the Sidama Zone (which is one of the most populous zones in southern Ethiopia), a majority of households were suffering from food insecurity. We are aware that in order to validate the the model properly testing it on other countries would be a good option. However due to time, we are forced to focus on one country.

Moreover, since our goal is to forecast change events a supervised machine learning method will be used. A supervised machine learning method allows us to be more specific about the definition of the labels compared to an unsupervised technique. Fortunately, the ordinal multi classed IPC categories output label can be used in this case to identify when transitions between states happen. To elaborate, the IPC labels represent different classes of food security [1]. The exact supervised learning technique that will be used to forecast these change events, will be determined during the process of data collecting and preprocessing.

[1] for the exact meaning of these classes please refer to figure 17 in section A.3 in the appendix

2.4 Research Question

With the problem statements in mind the research question that we want to answer in this study is:

Can we create a valid model, that uses scalable features, to predict change events in food security per livelihood zone in Ethiopia?

3 Literature review

3.1 Food Security

There are several dimensions to food security. According to Holloway (2003) these dimensions can be categorized in four pillars. The first pillar is called availability and refers to the availability of sufficient quantities of appropriate quality food (Bora, Ceccacci, Delgado, & Townsend, 2011). To give an example a study from Hesselberg and Yaro (2006) showed that a dry season can affect the availability of food. The second pillar, called access, is whether people have the resources to access sufficient, safe and nutritious food but also whether they have the resources (like infrastructure or money) to either produce or buy food (Food Security Cluster, 2016;Godfray et al., 2010;Rooyen, 2000;FFSSA, 2004). The third pillar, stability, is that people should not experience a decrease in food security due to conflict, economic crisis, neutral disaster or global climate change other shocks (Sassi, 2017; Cheeseman, 2016). The last pillar is called utilization and refers to the ability to physically use the available food, such as proper food preparation and feeding practices. Next to this this pillar also refers to the biological ability to utilize the food that is consumed (so absence of diarrhea or other diseases that could hinder the use of food) (Denny et al., 2018; Connolly-Boutin & Smit, 2016). Interference with one or more of these pillars can lead to food insecurity (Denny et al., 2018; Bora et al., 2011). All in all the literature shows that food security itself is a complex phenomenon which consists of several dimensions and complex interactions. First a framework for the drivers of food security will be discussed, afterwards we will look into existing machine learning models of food insecurity.

3.1.1 Drivers of food security

Since food security is such a complex phenomenon with multiple dimensions, it is to be expected that there are a wide range of drivers that influence food insecurity. As a consequence, summarizing all these drivers is challenging. However a study from Misselhorn (2004) tried to identify the most cited drivers of food insecurity in Africa (see figure 2). Misselhorn (2004) made a distinction between direct and indirect divers. Specifically, indirect divers could potentially initiate other drivers. From these indirect drivers climate and environmental stressors but also social and political unrest due to war have the highest citation rate. Moreover Misselhorn (2004) shows that these drivers can be indicated as either a shock or chronic. Nevertheless these direct and indirect drivers, found by Misselhorn (2004), account for around 80 to 81 percent of food security. Even though the study from Misselhorn (2004) is slightly outdated it still seems to correspond with drivers identified by more recent studies like that of Barron, Tharme, and Herrero (2013). An interesting note from the study from Barron et al. (2013), is that there are complex interactions between drivers which in turn can reinforce the impact of these drivers on food security. All in all, these studies show that the drivers of food security have a complex relationship (either direct or indirect) with each other that can be either chronic or as a shock.

3.1.2 Models of food insecurity

There are already several models that predict food security. However most of these work on a different scale like on household level. An example of such a model is the study of Okori and Obua (2011) which showed that it is possible to classify famine on household level (in this case in Ugunda). Another example is that some models use more localized information. To be more specific, several models, like that of Mbukwa (2013) and (Gubert et al., 2010), mainly use data from surveys, which can be time intensive to collect, process and conduct (Barrett, 2010b). In our case we want to use more scalable information that can be easy to collect, process and can be used for several countries. Like we have discussed in the introduction, in order to achieve this we chose to use satellite imagery. There are however other studies that also tried to incorporate satellite imagery when predicting food security or the closely related phenomenon famine.

First off, the study from A Quinn, Okori, and Gidudu (2010) created classifica-

Figure 2: Adapted from the study from Misselhorn (2004). This table shows the direct and indirect most cited drivers according to Misselhorn (2004) study. Moreover the table also has information about how these identified drivers related to the access and availability pillars and whether these drivers are indicated as shocks or chronic.

		related to		indicated as		indirect drivers in %
	Cited	Access	Availability	Shock	Chronic	Cited
Climate and environmental stressors	12	33	67	43	57	17
Poverty	7	72	28	15	85	21
Increase in food prices	5	100	-	70	30	-
Absence ofproperty rights and land access	5	15	85	7	93	-
Unavailability of employment	5	93	7	29	71	-
Lack ofeducation	5	92	8	8	92	3
Poor market access	4	100	-	-	100	-
Pests and diseases of crops and livestock	4	44	56	39	61	-
Poor human health	4	77	23	9	91	4
Low regional cereal availability	4	100	-	95	5	-
Poor distribution networks and Infrastructure	4	91	9	10	90	-
in- and out-migration	4	50	50	15	85	4
Inflation	4	82	18	25	75	-
Social and political unrest or war	3	59	41	32	68	12
Sale of assets	3	82	18	83	17	4
Insufficient agricultural inputs	3	-	100	39	61	-
Formal and informal government policies	3	76	24	41	59	5
Low regional cereal availability	-	-	-	-	-	4
Prevalence of HIV/AIDS	-	-	-	-	-	5
Population pressure	-	-	-	-	-	3
total	80	65	35	33	67	81

tions of food security on household level by defining food insecurity as a calorific intake of less than 1800 kcals/day. Subsequently they added household data (like the size of the household or ownership of livestock) to satellite observation which in return resulted in a better accuracy in making famine predictions at a household level. The problem however with household level data is that it is usually not available for every region and not consistently updated. In our study we want to create a model that is updated timely but also efficient with regard to data collection. As a result our model won't be using household data for now. Secondly, in the study from (A Quinn et al., 2010) they only used the Normalized Difference Vegetation Index and the Rainfall Estimation and they argue that adding more features to satellite imagery could be beneficial for the model. Thus, it might be usefull to also look for features outside satellite imagery. IPC Global Partners (2012) also make a prediction for their IPC class quarterly. However, they don't make predictions of change events and would not be as flexible as our model that should be able to also predict on a monthly basis instead of just on quarters. Moreover building a custom model has the benefit of having more insight in what the model actually does which results in having more confidence in decision making.

3.2 Knowledge Gap

Here we first showed that food security as a phenomenon is complex and multidimensional. Because of this we have to take into account that there are a wide range of drivers like the study from Misselhorn (2004) showed. Secondly, even though there are already existing models that either try to predict food security or the closely related famine, these models are either based on non scalable features (survey data) or a non scalable target variable (calorific intake through surveys). Moreover these studies usually operate on household level which in turn can be dependent from the more time consuming data collection through data surveys. Next to all of this the existing models don't focus on the change events of the food security, or in other words they don't focus on the transitions. Lastly the model from IPC Global Partners (2012) is not as flexible, scalable, doesn't predict change events and can't predict on a monthly basis. All in all our goal is to create a model with scalable features that can predict change events for larger areas in stead of household level. Thus we want to be able to get a general impression of the change events per livelihood zone with our model. It seems from the literature review that this is not available yet.

4 Design Approach

4.1 Design Approach

Figure 3 gives an overview of the different steps that will be taken during the project with regard to the model. First we will have determine which data is available and relevant to predict food security. After we have decided which sort of data we want to collect we can process it in the correct format that we need it to be.

After this step we have to determine which metric we will use too optimize and validate the model. Next to the metric, we have to also determine which algorithm we are going to use, what kind of cross validation and how we are going to set up baselines and a benchmark.

When we have determined all the concepts that we need, we will create a model. We have to both optimize and validate the performance of the model. For optimizing the model we will pay attention to three strategies; resampling, feature engineering and hyperparameter tuning. During the research it will get more clear which strategy (or combination of strategies) will be most useful. After optimization we will validate the model by testing it against the baselines that we have determined and the benchmark that we have created.

The last step is to get insight from the model. Which features are identified as most important, is there any spatial dependency or spatial relationships and does the model perform well for each prediction interval (can it predict one month ahead but also 10 months ahead). Of course the overall performance of the model will also give insight for further research and answer the research question.

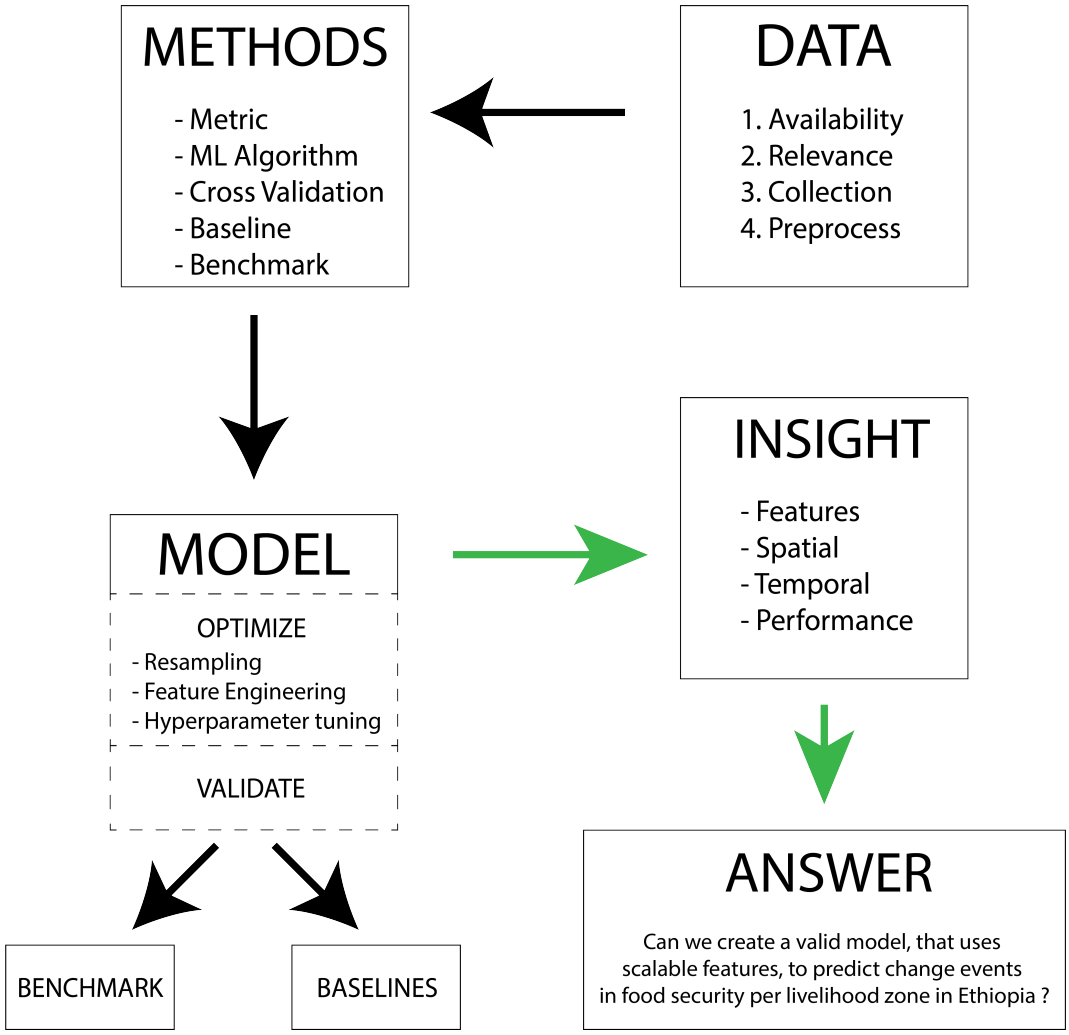


Figure 3: A research flow diagram which shows the different steps that this study will go through.

4.2 Sub-questions

In order to answer the research question outlined in section 2.4, it is now clear that several sub-questions have to be addressed

1. **Feature related sub-questions:**
   (a) Which features can we use to predict change events in food security?
   (b) Are the found features scalable?
   (c) Can we create usable features through feature engineering?
2. **Modelling related sub-questions:**
   (a) Which modelling technique(s) is/are the most appropriate for this scenario?
   (b) Which metric should we use to validate the model?
   (c) Which cross validation should we use?
   (d) How can we optimize our model properly?
3. **Validating and insight related sub-questions:**
   (a) Does the model perform better than the baselines and a benchmark?
   (b) For which livelihood zones does the model predict the food security correctly (spatial aspect)?
   (c) Does changing the prediction interval influence the performance of the model (temporal aspect)?

5 Methods

5.1 Data Availability

Many efforts have been made to make a good measure of food security and its four dimensions. However, like Upton, Cissé, and Barrett (2016) explain, due to the unobservability and multidimensionality of the food security status and the challenges that arise when establishing measurements that can be aggregated from individuals to regions (and visa versa) it can be hard to maintain fidelity to these pillars (Barrett, 2002; Barrett, 2010a; Headey & Barrett, 2015). In other words, there are limitations and challenges with regard to data collection and aggregation for food security, which in turn make it more difficult to make a reliable measurement. These challenges in combination with the fact that we want to find scalable information further limits the the data collection. However in order to find, collect and process the data into the correct format, we have to decide on what aggregation level the data should be, but also which kind of data is scalable and (openly) available.

First off the most relevant aggregation level, from a theoretical perspective, would be the livelihood-zones. This is because the livelihood zones divide areas into zones, in which people share the same pattern of access to food, income and the same access to markets. In other words it makes it possible to classify geographical regions based on the pattern of livelihood that these people share in these zones.

Secondly, since our goal is to create a model that is also easily scalable in the future, data that is available on global level is preferred. However, this requirement also limits the data types and data sets that we can use in this study. Thus, the data should be detailed enough so we can aggregate it on livelihood zones yet also be scalable (to other countries) for in the future. Next to the spatial aspect, the data should also be detailed enough with regard to the temporal aspect. In other words, the data should be updated regularly and also be timely. Another requirement is that the data should be easily accessible using code, preferably through an API. As this makes it easier to update the model when the data has been updated.

Lastly, since we want to use data that is easily accessible and available, this limits our data to open data. This is because getting access to data sources that are not open, cost either money or time (to get all the certificates of the companies in order to use this data). As a result, this study will use open data.

In order to fulfill our three requirements, we have decided to mainly use (open)

satellite imagery. More specifically, the Google Earth Engine (GEE) python package was used to collect different sorts of satellite data. We made sure that the data that we extracted from the Google Earth Engine for each variable had the same format, so we can merge it easily. By using the GEE we have fulfilled the requirements of scalable, yet detailed open data. Moreover using the GEE it is also possible that we can aggregate the data on livelihood zones. Lastly, using the GEE we also solved a more practical issue. Namely, we have used functions that run on the GEE cloud engine, instead of your local machine. As so, this made it possible to reduce the run time to process the satellite imagery in the correct format and download it. Other data sources, than the GEE, that we might use should also fulfill the three requirements.

5.2 Data Relevance

We first had to identify which satellite imageries we could use and were relevant for predicting food insecurity. Using the results from the meta-analysis from Misselhorn (2004) and the study from (Barron et al., 2013) we have chosen several satellite imageries from the GEE database (see table 1 for a quick overview). As you can see from the table, we did not succeed in identifying every single driver. This is due to several reasons. First off, as we have told in the previous section, we want to identify features that are scalable yet detailed enough to be aggregated on livelihood zones. As a result this limited the types of data that we could use. Secondly, since we are only working with open data, we are also limited by the data that is not sensitive. Or in other words, finding open data about political unrest is challenging and up to now we did not find that sort of data on the internet. Moreover finding data about poverty also posed a challenge. For example, we were not able to identify a data set on the gross domestic product that was scalable, timely updated and more detailed than on country level. Nevertheless the framework from Misselhorn (2004) and table 1 shows that we did succeed in identifying some of the most cited drivers that should be relevant for predicting food security. For example, the climate and land drivers that we have identified should give an indication of the climate and environmental stressors which was cited the most by Misselhorn (2004) as a direct driver.

Table 1: An overview of the different features used for our model and the different identified drivers found by the study of Misselhorn (2004) of food security. Even though we did not succeed in finding every single driver as a scalable yet detailed feature, we did identify most of the drivers that were cited the most according to Misselhorn (2004).

	Direct Drivers Cited in %	Indirect Drivers Cited in %	IPC	Climate and Land Drivers	Market Prices	Infrastructure	Demographical	Livelihood Zone Characteristics	Conflict
Climate and environmental stressors	12	17	X	X					X
Increase in food prices	5	-			X				
Poor market access	4	-	X						
Poor distribution networks and Infrastructure	4	-				X			
Social and political unrest or war	3	12							X
Insufficient agricultural inputs	3	-	X					X	
Population pressure	-	3					X		
Poverty	7	21							
Absence of property rights and land access	5	-							
Unavailability of employment	5	-							
Lack of education	5	3							
Pests and diseases of crops and livestock	4	-							
Poor human health	4	4							
Low regional cereal availability	4	-							
in- and out-migration	4	4							
Inflation	4	-							
Sale of assets	3	4							
Formal and informal government policies	3	5							
Low regional cereal availability	-	4							
Prevalence of HIV/AIDS	-	5							

5.3 Data Collection and Data Processing

The first processing step took place before we could use and collect the data from the GEE. To be more specific, we had to import the shapefile, containing the polygons of the different livelihood zones, into the GEE so we can aggregate the data on this level. After data has been collected, it still needs to be processed in the correct format for the model, for which we used a technique called Spatial Reduction, which is when we would aggregate each pixel of satellite imagery using different statistical methods (which methods we used per variable are listed in the table) for each livelihood zone using the polygons that we have created before. Before we continue with explaining the processing steps that we have conducted for each feature, it is important to note that table 2 gives an overview of all the different satellite imagery that we have collected and processed for our model. This table also contains some other variables that we retrieved from other sources than the GEE. These sources, like the data set from Uppsala Conflict Data Program (2018), still fit the requirements of being scalable, open and detailed enough.

5.4 Features

5.4.1 Target variable

The target variable that this model will be using is called the change event, which is a derivative of the IPC class from the *Famine Early Warning Systems Network* (2019). The IPC is an ordinal multi-class scale, containing information about the state of food security (see table 17 in the appendix for the exact meanings of each state). The change event is a variable containing information on whether the food security of a livelihood zone improved, deteriorated, or stayed the same as before. This change event makes it so that we emphasis on predicting when transitions happens, instead of the absolute IPC value.

In order to create the change event, several steps had to be taken. First, each IPC shapefile from the *Famine Early Warning Systems Network* (2019) 2010 until 2018 had to be downloaded. Afterwards we had to aggregate these shape files in the correct format for the model. In order to achieve this, we had to change the IPC shapefile to a raster so we could use spatial reduction. To be more specific, with spatial reduction we could use the mode to determine which IPC class is the most represented per livelihood zone. After the IPC class was aggregated per livelihood

Table 2: Summary of the different features used in this study and where we retrieved them from.

	Retrieved from	**URL**
IPC		
Current Situation	FEWSNET	http://fews.net/fews-data/333
HA0	FEWSNET	http://fews.net/fews-data/333
Climate and Land Drivers		
NDVI	GEE	https://code.earthengine.google.com
Rain Precipitation	GEE	https://code.earthengine.google.com
Soil Moisture	GEE	https://code.earthengine.google.com
Elevation	GEE	https://code.earthengine.google.com
Market Prices		
Food Market Prices (absolute)	HDX	https://data.humdata.org/dataset/wfp-food-prices-for-ethiopia
Infrastructure		
Accesibility	GEE	https://code.earthengine.google.com
Friction	GEE	https://code.earthengine.google.com
Demographical		
Population Density	GEE	https://code.earthengine.google.com
Population Count	GEE	https://code.earthengine.google.com
Livelihood Zone Characteristics	FEWSNET	http://fews.net/livelihoods
Main Crops	FEWSNET	http://fews.net/livelihoods
Main Stocks	FEWSNET	http://fews.net/livelihoods
livelihood zone Type	FEWSNET	http://fews.net/livelihoods
Conflict		
UCDP Fatalaties	UCDP	https://ucdp.uu.se/

zone, we could then transform these variables to change events which we express through the following formula:

$$CE_t = -(IPC_{t+n} - IPC_t)$$

In this formula t is the current month and n is the number of months ahead the CE should express. With this formula in mind, whenever the IPC class got higher in the next month (thus a deteriation of the food security) the class got a *Deterioration* which we will from now one be called *deterioration*. Whenever the IPC class got lower in the next month in a livelihood zone (a improvement of foodsecurity) it got a 1, which we will refer to from now on as *Improvement*. And lastly, whenever the IPC class stayed the same in the next month it got a 0 which will be called *No Change* (see table 3).

Table 3: An overview of the change events and their corresponding value

Abbreviation	Value
Deterioration	-1
No Change	0
Improvement	1

It is important to note that the the periods of when the IPC class were released were shifted since 2016. From the years 2010 until the end of 2015, the IPC classes were released by *Famine Early Warning Systems Network* (2019) on January, April, July and October. Unfortunately the period in which the IPC class was released was changed starting from 2016 to February, June and October. This means that our target variable differs, with regard to the temporal aspect, beginning with 2016. As a result, we decided to interpolate the IPC class. Since the IPC classes are integers, we had to round off the number when interpolating the IPC class. We've decided to this for two reasons. One being that we can now use more information, since we can collect the data from our features on a monthly basis instead of periods. The other reason is that we now can merge the IPC classes from the periods 2010 until 2015 and 2016 until 2018.

We have also included the IPC with several time lags and the binary HA0 variable from each shapefile of the *Famine Early Warning Systems Network* (2019). If HA0 is 1 then the IPC would most likely be at least one phase worse without current

or programmed humanitarian assistance. Whereas, if HA0 is 0 this means that humanitarian assistance is likely not significant enough to change the phase level (*Famine Early Warning Systems Network*, 2019).

5.4.2 Climate and land drivers

Imagery from NASA's Moderate Resolution Imaging Spectrometers (MODIS) satellite surface reflectance composites is used for determining the Normalized Difference Vegetation Index (NDVI). The NDVI might be relevant for food security since it is estimated that yield or crop production to food availability is estimated at 60% (Frelat et al., 2015). This index is generated from the Near-IR and red bans of each scene with the following formula (NIR - Red) / (NIR + Red). After loading in the satellite data our goal was to get the mean and median NDVI per region, per month for the years 2010 until Decemeber 2018. In other words we wanted to do a spatial reduction of the MODIS imagery. The end result was a dataframe containing the mean and median NDVI for each month per livelihood zone. We decided to use both metrics since usually the median is less sensitive for outliers due to for example clouds. Since clouds can disrupt certain imagery from the satellite and influence their quality (Alvera-Azcárate, Sirjacobs, Barth, & Beckers, 2012; Holben, 1986; Champion, Le-Hir, Massera, & Bellaiche, 2017).

The rain precipitation was also extracted from the GEE using the same method as for the NDVI. The satellite imagery from the TRMM was used (Tropical Rainfall Measuring Mission (TRMM), 2011). The TRMM uses an algorithm to produce a single best-estimate precipitation rate and also a precipitation error estimate. Just like the NDVI we opted to use the mean and median for each month.

Soil moisture's mean and median per livelihood zone per month was also extracted through the GEE database. The NASA-USDA SMAP global soil moisture data set provides soil moisture information across the globe. This data set was developed by the Hydrological Science Laboratory at NASA's Goddard Space Flight Center in cooperation with with USDA Foreign Agricultural Services and USDA Hydrology and Remote Sensing Lab. Next to surface soil moisture this satellite imagery also includes subsurface soil moisture, soil moisture profile ad surface and subsurface soil moisture anomalies (E. Mladenova et al., 2017; Bolten & Crow, 2012 ;Bolten, Crow, Zhan, Jackson, & Reynolds, 2010; Entekhabi et al., 2010 ;O'Neill, Chan, Njoku, Jackson, & Bindlish, 2016).

Another driver that we have extracted is the elevation. The Global Multi-

resolution Terrain Elevation Data from 2010 is used and contains elevation information on global scale (Danielson & Gesch, 2011). This data is not regularly updated, since the elevation doesn't differ that much between a couple of years. As so we have used the satellite imagery data from 2010 for each month for each year up to June 2018. In this case we used the mean and median metric per livelihood zone for this driver.

5.4.3 Market Drivers

Another important driver in which several prepossessing steps had to be done is the food market prices from the World Food Programme (2019) (WFP). According to Godfray et al. (2010) the global food prices are indicators for whether food is available and whether people can afford and have access to world markets. In order to preprocess this data in the correct format we first had to account for missing data. More specifically, data was missing in certain months for some livelihood zones. In order to overcome this problem we decided to linearly interpolate the missing values instead of throwing them away. We also decided to only include the three most traded products, namely wheat, maize and sorghum. The other products were not well maintained, with regard to data collection by the World Food Programme (2019) and contained a lot of missing values. Next to this we also included prices for both retail and wholesale as separate features. We aggregated these six features on admin level 1 with the livelihood zones. It is important to note that even though this database is quite comprehensive, some smaller markets might not be included.

5.4.4 Conflict

It has been argued that one of the main drivers of the rise in food insecurity is the increasing conflict and insecurity around the world (FAO, IFAD, UNICEF, WFP and WHO, 2017; Food Insecurity Information Network, 2018). Thus, it would only be right to also try to include a feature that could represent conflict. In our case we can use the data set from Uppsala Conflict Data Program (2018). Uppsala Conflict Data Program (2018) is a large database with a history of around 40 years. Moreover, this open database has global coverage and has a long time series which get updated annually (Uppsala Conflict Data Program, 2018). The Uppsala Conflict Data Program (2018) differentiate between two types of conflict. First, minor conflicts are those that pass the 25 battle-related deaths thresholds but

have less than 1000 deaths in a year. Secondly, major conflicts pass the 1000 annual deaths threshold. Literature reveals that there are a lot of different predictors that relate to conflict (Uppsala Conflict Data Program, 2018). Nevertheless, there are some drawbacks with using this database. First, this data set is mostly based on publicly accessible news reports which contain information about individuals killed or injured. Like Uppsala Conflict Data Program (2018) already notes, that due to the lack of (credible) information sources in many conflict zones, these numbers are low and rough estimates and could be biased. Lastly Uppsala Conflict Data Program (2018) also discusses that their estimations are partly dependent on estimates of other sources. All in all there might be some reliability issues with regard to this data set. Nevertheless, at this moment it seems that this data set would be our best option to represent conflict as a feature.

In order to preprocess this data in the correct format we had to overcome several issues. First, since the shapefile from the (Uppsala Conflict Data Program, 2018) is worldwide data, we had to filter the data to Ethiopia. The next step was to assume that conflict events, spread out over multiple years or months, had approximately a constant number of fatalities per month within that time span. Afterwards we could merge the data from Uppsala Conflict Data Program (2018) with the rest of the data on our aggregated livelihood zone level.

5.4.5 Infrastructure

The GEE was also used to extract the accessibility to cities. The GEE database contains satellite imagery in which the land-based travel time to the nearest densely-populated area are calculated for the year 2015. These highly dense populated areas are defined as continuous cells with a density of at least 1500 inhabitants per km2 or a majority of built-up land cover types coincident with a population centre of at least 50000 inhabitants (J. Weiss et al., 2018). J. Weiss et al. (2018) created this data set using a combination of data about roads, railways, rivers, water bodies, land cover types, topographical conditions and national borders (J. Weiss et al., 2018). Combining these maps they created a friction surface map which included information about the travel based speed within pixels. Afterwards they used a least-cost-path algorithms in combination with this friction map to calculate the travel time from all locations to the nearest densely-populated area (J. Weiss et al., 2018). In our model we have included the accessibility data set but also the friction data set. One is used to determine how long the travel time would take to reach the

nearest densely populated area, the other map to find out how fast the travel based speeds are within certain pixels. We chose the friction map mostly with the reason that if the travel based speed is fast, it probably means that within a livelihood zone the infrastructure is better than in areas in which the travel based speed is slower. It is important to note however that there might be a bias in this. If a livelihood contains a big city in which presumably the infrastructure is better, and thus the land based travel speed is faster, but also contains pixels in which the infrastructure is worse this could potentially result in biased numbers. As a result this area might get a number that seems to illustrates that the infrastructure is good for the whole area, whereas the infrastructure is only good in a certain part within that area. Nevertheless, finding detailed (open) data for infrastructure posed a challenge and for now this is the best that we could find. Like the climate drivers, we used the mean and median metric for the accessibility map and also the friction map.

5.4.6 Demographical variables

Next to NDVI, rainfall, land-cover and accessibility, the GEE was also used to determine population density. The data set from (CIESIN, 2016) contains approximate population density per grid cell. This data set from (CIESIN, 2016) has been adjusted to match the 2015 Revision of UN World Population Prospects country totals. It is important to note that this data is available for the years 2005, 2010, 2015 and a forecast for 2020. The same applies for the population count data set that we have used from (Center for International Earth Science Information Network (CIESIN), 2016). Subsequently we decided to linearly interpolate the missing values for the years in between for each month per livelihood zone. The same functions were used to preprocess the data in the same format as the NDVI data (but obviously with a different satellite imagery; the population density and count). However we opted to use the sum and mean of the population as features for the model.

5.4.7 Livelihood zone Characteristics

The livelihood zone shapefile from the (*Famine Early Warning Systems Network*, 2009) also contains extra information per livelihood zone. However, it is important to note that this data is from 2009. There is however no other data available with regard to these livelihood zones. Nevertheless, we have decided to include several features from this shapefile into our model. These features that we have included contain

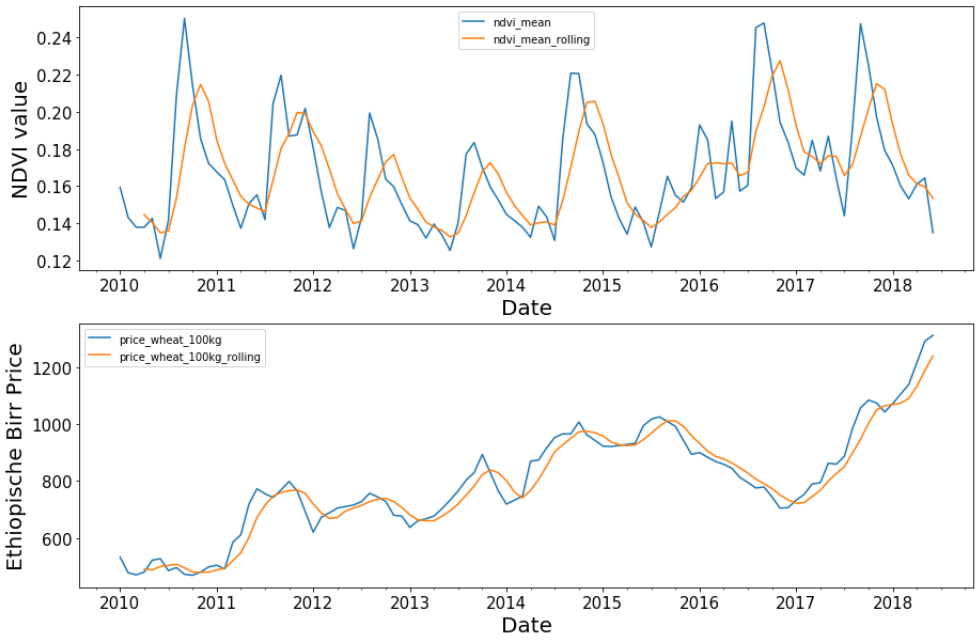
information about the main stock, crops and also what kind of livelihood zone types these areas are (for example urban or pastoral). For each of these features we had to conduct several steps in order to process the data in the correct format. First, for the main crops and stocks we had to parse the sentences containing information about their crops and stocks. After we identified each unique crop and stock we assigned a column to each of these and used binary values to indicate if a livelihood zone either used the crop or stock (with a 1) or not (with a 0). Lastly we also included a feature from the livelihood zone shapefile with information about what kind of livelihood zone type this zone is. This contained several different type of zones ranging from urban areas to pastoral. It also contained some missing information. In other words, there were some livelihood zones that had the type 'unknown'. We decided to keep this as a separate livelihood zone type and not fill these unknown values with the median or mean, since this would probably not be representative for these unidentified zones.

5.5 Feature Engineering

In order to make sure that each feature can be used to its full potential we have also spent some time on feature engineering. First off we have created features with regard to the different seasons of Ethiopia: the Belg, Kiremt and Bega seasons of Ethiopia (as dummy variables). Next to this, we have also included time lags for the IPC and HA0. In order to create this we have shifted each feature X with a time period i. In our case i ranged from 1 till 6 months. We did not include the time lags of the NDVI, rain precipitation, soil moisture and food market prices. This is because we decided to smooth out these variables using the rolling mean metric. In this case we use the mean value per feature for the four preceding months. Like said before using this metric we can smooth out the rough edges of the data and hopefully lead to more understandable data for the model. In order to visualize the effect of the rolling mean, please refer to figure 4, in which the orange line is the rolling mean and the blue line the normal mean value for the NDVI and food market prices (100 kg). We have also created three different binary variables from the change event. These three variables are a cumulative sum of the times a livelihood zone has deteriorated, improved or changed (so both improved and deteriorated) with regard to food security. Using this, we can give our model some more historical context. We also decided to include the cumulative sum of the fatalities per livelihood zone from the data set from Uppsala Conflict Data Program (2018). This should again

also give an indication of the historical context with regard to conflict and stability per livelihood zone.

Figure 4: These figures showcase the effect of a rolling window (orange lines) for one livelihood zone (ET02) in the Afar region. This figure illustrates how the values of the NDVI and food market prices get smoothed out because of the rolling mean. The blue lines represent the original mean values.



Lastly we have also included a binary feature whether market prices went up or down compared to the previous month. We did this for both retail and wholesale. Table 4 shows an overview of all the different statistical metrics to express the features that we have collected, processed and created.

5.6 Class imbalance

In order to overcome the class imbalance between the *No Change* compared to the *Deterioration* or *Improvement* of the food security we have to use either weights or a resampling technique to overcome this so the model can train better on the (train) data. We opted to choose for a resampling technique called Adaptive Synthetic

Table 4: Summary of the different metrics used to express our chosen features

	Binary	**Mean**	**Median**	**Min**	**Max**	**Sum**	**Cumulative Sum**	**Rolling Mean**	**Time Lags**
IPC									
Current Situation	X								
Change Event	X						X		X
HA0	X								X
Climate and Land Drivers									
NDVI		X	X	X	X			X	
Rain Precipitation		X	X	X	X			X	
Soil Moisture		X	X	X	X			X	
Elevation		X	X						
Market Prices									
Food Market Prices (absolute)		X		X	X			X	
Food Market Prices (binary)	X								
Infrastructure									
Accesibility		X	X						
Friction		X	X						
Demographical									
Population Density		X				X			
Population Count		X				X			
Livelihood Zone Characteristics									
Main Crops	X								
Main Stocks	X								
livelihood zone Type	X								
Conflict									
UCDP Fatalaties		X					X		

sampling approach (ADASYN) which is created by He, Bai, Garcia, and Li (2008). This technique uses a weighted distribution for the minority class examples according to their degree of difficulty in learning. Or in other words classes that are harder to learn gets more synthetic data generated compared to classes that are easier to learn (He et al., 2008). Like He et al. (2008) already notes, by using ADASYN we can not only reduce the bias which can be introduced by the class imbalance but also shift the decision boundary to the more difficult examples.

5.7 Metrics

Since it is important to validate the model properly we had to choose metrics that are fitting to the problem itself and which can validate the model's performance reliably. As a result we decided to use the f1 and accuracy score to identify how the model performs. However, to understand what the f1 metric is and what this means for our validation process it's important to shortly describe the components of the f1 score, namely precision and recall. Precision is the proportion of positive identifications that were actually correct (Google, 2019b). Thus it identifies proportion true positives compared to all positives (true and negative). Recall is the proportion of actual positives that was identified correctly (Google, 2019b). Thus, it identifies the proportion true positives compared to samples it should have found (true positives and false negatives). By combining these two metrics we get the f1 score which is the harmonic mean of recall and precision. In order to validate the model properly while having imbalanced classes we have decided to report the f1 values for each class and also the f1 macro average score from the package Pedregosa et al. (2011). Using the macro average score we give more weight to the less occurring classes (*Deterioration* and *Improvement*) than with the non-macro alternative.

We will also use the accuracy score to get an overview of how well the model can identify each livelihood zone. Google (2019a) defines accuracy as the fraction of predictions the model got right. As, the number of accuracy basically shows the ratio of which input samples it could predict correctly (true positives and true negatives) compared to all input samples. The accuracy score in this case will be used to map each region on the chart of Ethiopia to give an indication of the performance of the model to correctly identify each livelihood zone's change event. It is however important to note that we won't use this metric to either optimize our model nor to validate it is performance over all regions. This is because our target variable is imbalanced, in which the class *No Change* is more common compared to the other

Table 5: Summary of the different Metrics used to validate the model. We will use the f1 macro score as the primary optimization metric.

	Definition	**Formula**
Accuracy	The ratio of which input samples it could predict correctly compared to all input samples.	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	The proportion true positive compared to all positives found by the model.	$\frac{TP}{TP+FN}$
Precision	The proportion true positives compared to the samples it should have found.	$\frac{TP}{TP+FP}$
F1	The harmonic mean of recall and precision.	$2 \times \frac{Precision \times Recall}{Precision+Recall}$
F1 macro	Calculate metrics for each label, and find their unweighted mean.	$\frac{\sum_{i=1}^{n}(F1_i)}{n}$

two classes. This is the reason why we will use the f1 macro and f1 individual scores to optimize an validate the model (see 5 for an overview of all the metrics that we have discussed). All in all, using the f1 and accuracy scores should give us a relatively good image of the performance of the model.

5.8 Machine Learning Algorithm

For our research we have decided to use an extreme gradient boosting algorithm (Xgboost). Xgboost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. (*Python API Reference*, n.d.;Chen & Guestrin, 2016). In order to briefly explain the Xgboost algorithm we have to understand two techniques, Gradient Descent and Boosting, and the combination of the two namely, Gradient Boosted Models.

The Gradient Descent optimization will update and correct weights learned by the model in order to minimize the cost, which is given by the cost function. A cost function measures how close the actual vales are to the predicted values. Subsequently the model weights are adjusted based on the cost function's outputs.

With Boosting an ensemble of weak learners is created, in which the missclassifications are boosted in order to predict them correctly in later models.

These two techniques are used in Gradient Boosted Models, which are trees that are built sequentially and of which the weighted sum is taken of multiple models. The difference with normal Boosting is that in this case the weights assigned to the models are not derived from missclassifications of the previous models but from minimizing the cost function by using the Gradient Descent.

Xgboost can be seen as a implementation of Gradient Boosted Models but with some improvements. As Chen and Guestrin (2016) explains, the most important factor behind the success of Xgboost is its scalability and fast runtime compared to other algorithms. One reason for this is that Gradient Boost Models are built sequentially, while Xgboost is parallalized. Moreover using the Xgboost algorithm also gives some more head space with regard to optimizing. This combined with the success stories with different challenges hosted by machine learning competition site Kaggel as noted by Chen and Guestrin (2016), makes this a solid machine learning choice for our multi-classification problem.

5.9 Cross Validation and Optimizing

The first step in optimizing the model, but also validating the model is setting up a proper pipeline for cross validation. Our first step will be to split the data into a train set and a holdout set. The train data will range from 2010-01-01 until 2016-05-31, the hold out set from 2016-05-31 until 2018-06-01. So basically our holdout-set is two years. We will then use our train set to find the best parameters using a self made grid search for the Xgboost classifier and optimize on the f1 scores (ideally both the macro and f1 scores of each class individually).

Within this grid search we will use a time series cross validation, as a way to do repeated cross validation while taking the temporal aspect into account. Firstly, repeated cross validation is important because situations out of 2017 could potentially differ greatly compared to situations in 2015. Thus in order to find the best parameters and thus a robust model it is important to train on different periods and test it on different periods as well. And secondly, taking the temporal aspect into account is of vital importance for preventing information leakage from future data points. As a result, we chose to do time series cross validation. To elaborate, the train set will get larger with each k number of splits that we would choose beforehand. The size of the test size will remain the same size but will however also slide

over until the last k split. For each k step split we will resample the imbalanced training set using the ADASYN algorithm which was explained briefly before. We will then average the performance over these k number of splits and choose the best parameters for optimizing the chosen metrics (see image 5 for an example of a time series split).

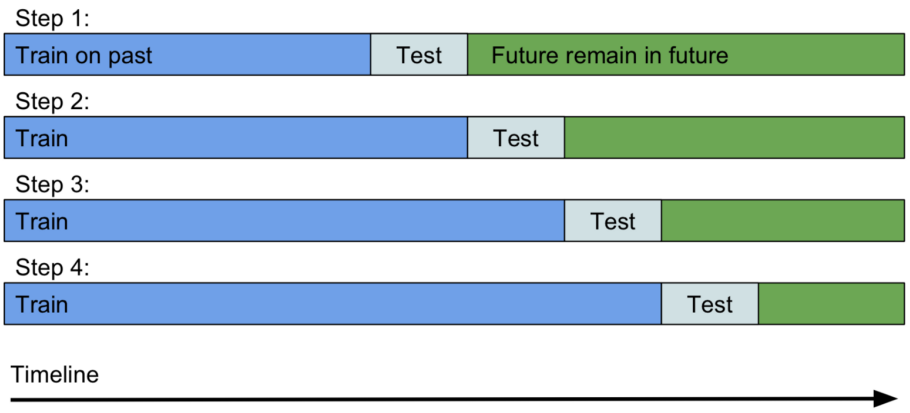


Figure 5: An example of a time series split retrieved from Osipenko (2018). As you can see each k split moves along the time line in which the train set expands while the test set slides until all data has been used.

We will conduct this grid search stepwise for the Xgboost on our train data. In the first step we will determine the *n_estimators* that we will use for determining the other parameters. After we have determined the *n_estimators* we will start with a relative large learning rate. After this we will optimize some tree based parameters namely the *max_depth*, *min_child_weight*, *gamma*, *subsample* and *colsample_bytree* (for a short description of each of these parameters please refer to documentation of the Xgboost from (*Python API Reference*, n.d.)). After we have determined which parameters of each of these work the best for our metric, we will run a grid search to optimize the regularization parameters to control the complexity of the model. The last step in this optimizing process is to increase the *n_estimators* and try different learning rates in order to reduce overfititng and further control the complexity of the model.

Using these found hyper parameters we will then retrain on our train set and test on our hold out set, how well our model generalizes and performs. In order to validate the model we will also look at the performance of the model with regard

to the temporal and spatial aspect. For the temporal aspect we will use different prediction intervals (for example predicting one month ahead and 8 months ahead) and measure the performance of the model. For the spatial aspect we want to determine which livelihood zones are predicted the most accurate and the least.

5.10 Baseline Models

A good way to compare our model's performance on the holdout set is to compare it to more simple heuristic models that can function as a baseline. Subsequently, we decided to create several baseline models that uses simple heuristics to determine how the performance of our Xgboost is.

The first heuristic model (called DCS) is based on the dummy classifier from sklearn which uses the stratified option. This model generates predictions by respecting the training set's class distribution (Pedregosa et al., 2011). The second heuristic model (HN) uses the historical norm. In other words it tests the assumption that the mode of the change event (CE) over the train period per livelihood zone is sufficient to predict the CE in the next period. More specifically, it uses the historically most occurring situation. The third heuristic model (called FP) assumes that the future equals the present. Or in other words the most recent history of a livelihood zone can predict the livelihood zone in the future. It is important to note that there is a chance that the second and third model give the same result. This is due to the fact that usually the Historical Norm for each livelihood zone is that there is no transition or change event. If there is no change event this means that the future equals the present. The fourth model called (called HNT) uses the same assumption as model 1 (the historical norm) however in this case it takes the temporal aspect into account. In other words, it is the historically most occurring situation in a specific month. The last model (called RO) uses the assumption that the most recent observation for a livelihood zone with the months taken into account is a good prediction for the future. The different baselines models are summarized in table 6.

5.11 Benchmark

In order to get an indication how well our model performs compared to other models we have also decided to create a benchmark. This benchmark is based on the performance between the actual change events from *Famine Early Warning Systems*

Table 6: Summary of the different baseline models that we are going to use to validate the model.

	Abbreviation	Description
Model 0	*DCS*	A dummy classifier that generates predictions by respecting the the training set's class distribution.
Model 1	*HN*	This model test the assumption that the mode of the CE over the train period per livelihood zone is sufficient to predict the CE in the next period. Or in other words the historical norm.
Model 2	*FeP*	Future equals the present.
Model 3	*HNT*	The historical norm which takes the temporal aspect into mind.
Model 4	*RO*	The most recent observation for a livelihood zone with the months taken into account is a good prediction for the future.

Network (2019) (which are a derivative of their actual IPC values) and the forecast that *Famine Early Warning Systems Network* (2019) have made for the change events for these months. It is important to note that this is only an indication, since their model slightly differs from ours since the benchmark that we have created, using their predicted and true values, is based on data that is updated on the months 2, 6 and 10 while our model is based on monthly data. Nevertheless it should at least give us an indication how our model performs compared to theirs.

6 Results

6.1 Results Data

In order to get a better understanding of the patterns of the change events we have first visualized the mean change events for all livelihood zones in Ethiopia for the existing time scale. Since our model can basically predict for different monthly prediction intervals, we for now have decided to only predict 4 months ahead, since otherwise the study would get to large. Nevertheless we will give an indication of the performance of the model for other prediction intervals than 4 months in the temporal insight section. To first get an idea if there is any seasonality in these change events we will first analyze how the change events fluctuate over the years. In order to do this we will take the mean change event per month and plot this

which resulted in figure 6. The image reveals that from 2010 until 2011 there were less improvements due to the mean not peaking as high as the other years. Moreover there seems to be some sort of seasonality with the change events. As figure 6 reveals, in Q1 of each year there seems, on average, to be more deterioration compared to no change of the food security state and improvements of the food security. In the middle (Q2 and Q3) of each year there are more improvements of the food security states on average compared to deteriorations and no change events. The state slowly declines at the end of year (Q4) from, on average, more improvements to more deteriorations. It is however interesting to see that the seasonality does change especially if you look at the year 2015 until 2017. The image makes clear that usually the peak is around Q3 and sometimes between Q2 and Q3. However for the year 2015 until 2016 this peak is not present. Subsequently, this change in seasonality could potentially make the readability of the target variable harder for our model.

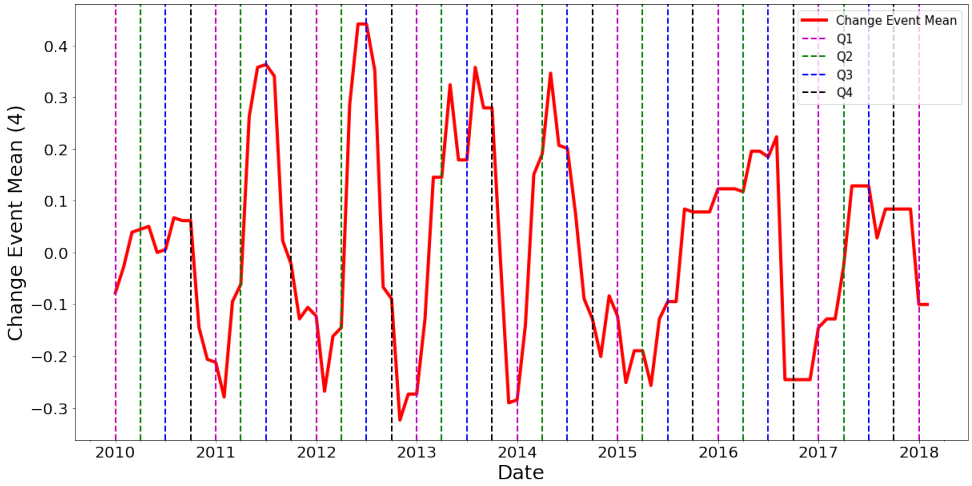


Figure 6: This image visualize the aggregated mean (of the change event) per month for Ethiopia from 2010 until 2018. This reveals that there is some periodicity. It is however also the case that these fluctuations can differ from year to year. More specifically, if you look at the year 2015 until 2016 you can see that there is some change in the fluctuations with regard to the change event. The peak that is usually in Q2 or (beginnning of) Q3 is not present between the years 2015 and 2016.

Since this was just an overview of the fluctuations over the whole country we will also take a look to the fluctuations per livelihood zone. In order to visualize this properly we have decided to calculate the standard deviation per livelihood zone and map this on the chart of Ethiopia (see figure 7). This reveals that livelihood zones in the north western part of Ethiopia look pretty stable and its standard deviation is close to zero. However, the livelihood zones in the mid and eastern part of Ethiopia change more frequently. Moreover the livelihood zones in the mid-northern part of Ethiopia seems to change most often.

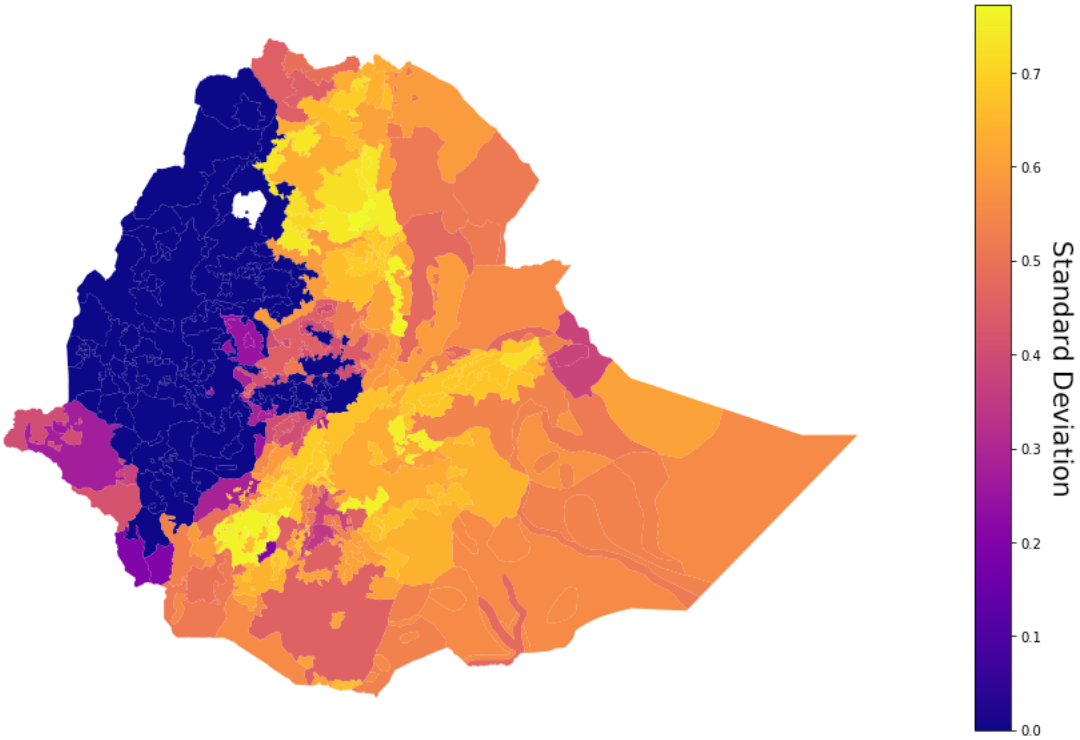


Figure 7: This image visualizes the aggregated variability (standard deviation) of food security changes per livelihood zone, with a time window of 4 months for all livelihood zones in Ethiopia from 2010 until 2018. Livelihood zones in the north western part of Ethiopia almost never change when we look at the short term change event. Other livelihood zones are in the mid northern part change more frequently when we look over the whole time period

6.2 Modelling

6.2.1 Optimizing the model

The first step which might lead to improvement of performance of the model is feature engineering and resampling. In order to test this we will run 100 different Xgboost models (by randomizing the seed) using a time series split with 3 number of splits and average the scores over our train set (2010-01-01 until 2016-01-31). We will do this for 4 different data sets. One data set contains the original data without any resampling strategy or feature engineering (called ORG). The second data set contains the original data but resampled using the ADASYN algorithm (called RES). The third strategy that we will test is the data set including the engineered features (time lags, cumulative sum, binary food market prices and rolling window) and also resampled which is called the RESFE strategy. The last strategy is GS, in which we have tuned the hyper parameters for the dataset of the RESFE strategy using a grid search time series cross validation (for the exact steps see appendix A.4). The best parameters that we have found during the grid search and have used for measuring the performance of this strategy are listed in table 7. In order to identify which strategy is the most effective we have first calculated the descriptive statistics for the f1 macro score and individual f1 scores for each class.

Table 7: Summary of the best hyper parameter combination that we have found in our grid search time series cross validation (GS strategy).

Hyperparameters	
max depth	4
min child weight	7
gamma	0.1
subsample	0.45
colsample by tree	0.75
n_estimators	400
learning rate	0.01
reg_alpha	0.00001
reg_lambda	10

As figure 8 visualizes, the strategy RESFE is more effective (M = 0.534; SD = 0.005) compared to the RES (M = 0.488; SD = 0.005). The difference between

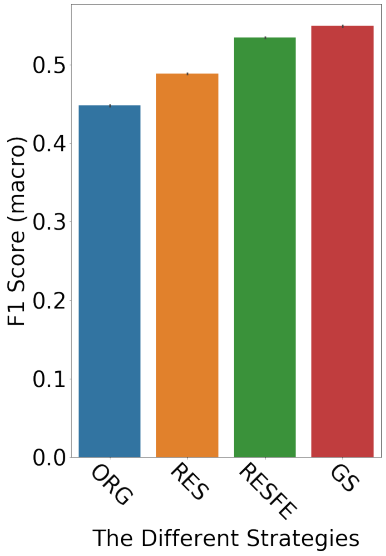


Figure 8: This figure shows the performance of the f1 macro score for the different strategies, including the grid search test performance score (GS). For each strategy we have run a bootstrap of 100 models. This figure indicates that the each strategy increased the performance more than the ORG strategy. However the leap from RES and RESFE in performance is larger than the performance increase from the RESFE to GS.

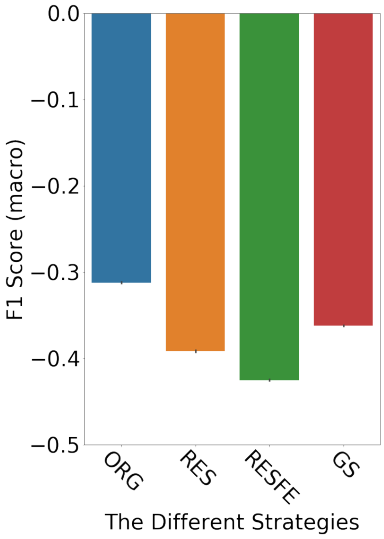


Figure 9: This figure shows the performance of the train test difference score (with regard to the f1 macro) for the different strategies, including the grid search test performance score (GS). For each strategy we have run a bootstrap of 100 models. Even though the performance of the f1 macro did not increase, the train test difference with regard to the f1 macro score did decrease. This in turn indicates that the model is now overfitting less after the GS strategy.

strategies is even more apparent when we compare the RESFE strategy with the ORG strategy (M = 0.448; SD = 0.006). Moreover when we compare the GS strategy (M = 0.549; SD = 0.004) with the RESFE a small improvement takes place. Next to this improvement, using the GS strategy also resulted in a smaller difference between the train and test score (M = -0.362) compared to the RESFE strategy (-0.425) for the f1 macro score (see figure 9). In other words the GS strategy controlled the complexity of the model a bit better which as a result lead to slightly less overfitting.

Table 8: The descriptive statistics of the f1 score for each of three different and for each strategy. ORG is the dataset withouth any feature engineering or resampling. RES is the ORG dataset but resampled using the ADASYN algorithm. RESFE is the RES dataset but includes extra features that have been created through feature engineering. Lastly the GS strategy is the RESFE strategy but of which the hyperparameters have been tuned.

Group	**Class**	**M**	**SD**	**CI (95%)**	**n**
ORG	Deterioration	0.132	0.013	0.129 - 0.134	100
ORG	No Change	0.835	0.003	0.835 - 0.836	100
ORG	Improvement	0.378	0.009	0.376 - 0.379	100
RES	Deterioration	0.174	0.011	0.172 - 0.176	100
RES	No Change	0.800	0.005	0.799 - 0.801	100
RES	Improvement	0.491	0.006	0.490 - 0.492	100
RESFE	Deterioration	0.274	0.012	0.271 - 0.276	100
RESFE	No Change	0.829	0.003	0.828 - 0.829	100
RESFE	Improvement	0.502	0.006	0.501 - 0.503	100
GS	Deterioration	0.347	0.009	0.346 - 0.349	100
GS	No Change	0.792	0.002	0.792 - 0.793	100
GS	Improvement	0.508	0.004	0.507 - 0.509	100

Not surprisingly, if we look at the individual f1 scores the scores for the GS for the class *Deterioration* (M = 0.347; SD = 0.006) and *Improvement* (M = 0.485; SD = 0.002) are higher than all the other strategies (see table 8). The class *No Change* however, has the lowest f1 individual score (M = 0.792; SD = 0.002) compared to the other strategies. It is however interesting to see that the ORG strategy has the highest f1 score for the class *No Change*. Looking at the individual scores, shows

that as we push to the more difficult classes the majority class (*No Change*) gets a lower performance.

All in all, the results indicate that compared to the other strategies GS strategy (which is basically the RESFE but with a grid search) was the most effective in terms of performance, by pushing to the minority classes (*Deterioration* and *Improvement*). Even though this resulted in a worse score for the majority class (*No Change*), it also resulted in less overfitting since it has a lower difference between the train test score.

6.2.2 Validating

Using the best found parameters (which are listed in table 7) we will now retrain the model on the train set (2010-01-31 until 2016-01-31) and test it on our hold-out set (2016-01-31 until 2018-06-31). In order to make sure that the values of our performance are as reliable as possible we will bootstrap the model 100 times and create confidence intervals. We will also calculate a train score of the model and a test score so we can identify how well our performance generalizes between training and testing. Nevertheless the results show that our Xgboost model (XGB) scores 0.526 on the f1 macro.

Table 9: Summary of the performance (f1 macro score) of the XGB model test, train and also the grid search time series cross validation (GSTSCV) test scores for a distribution of 100 models for each. The table shows that our model does overfit slightly, since there is a difference between the train and test score. Between the grid test score and our test score on the hold out the difference is smaller. The f1 score of class 1 is even beter for the test score than the grid test score.

	Performances								
	GSTSCV			**Train**			**Test**		
	M	*SD*	*CI*	*M*	*SD*	*CI*	*M*	*SD*	*CI*
f1 score ***Average***	0.549	0.004	0.548 - 0.550	0.722	0.002	0.721 - 0.722	0.526	0.003	0.526 - 0.527
f1 score ***Deterioration***	0.347	0.009	0.346 - 0.349	0.645	0.003	0.644 - 0.645	0.302	0.008	0.300 - 0.304
f1 score ***No Change***	0.792	0.002	0.792 - 0.793	0.845	0.001	0.845 - 0.846	0.772	0.002	0.771 - 0.772
f1 score ***Improvement***	0.508	0.004	0.507 - 0.509	0.675	0.003	0.647- 0.675	0.506	0.005	0.505 - 0.507

Before validating the model against the baselines and benchmark we will first compare the test score of the XGB with the train score of the XGB to get an idea if the model overfits (see table 9). The results indicates the the model over fits slightly since the train f1 macro score is higher (M = 0.722; SD = 0.002) compared to the test score (M = 0.526; SD = 0.003). An option to reduce could be that we would push the XGB to the lowest train test difference score even further however we would most likely underfit and under perform. As so for now this seems to be an acceptable difference between the macro scores, which again is most likely also the effect of the (historical) data limitation. Table 9 also contains information about the test score from the grid search time series cross validation (GSTSCV) that we have conducted on the train set in the previous section. The score of the grid search (M = 0.549; SD = 0.004) is pretty close to our hold out score which is a good sign since this means that we have been able to identify a set of parameters that generalizes (compared to the grid search) pretty well. Notably the f1 score of class *Improvement* is almost identical (M = 0.506; SD = 0.005) for our hold out test compared to the grid test score (M = 0.508; SD = 0.004).

In order to further validate the performance of the model we have to also look at the different baselines that we have set up (see table 10). We have only run our baseline ones, since most of the baselines will score the same score every time (with exception of the DCS model).

Table 10: Summary of the performance (f1 macro score) of the XGB model and the baselines on the hold out set. From the table it becomes clear that our model performs better than the baselines.

	f1 macro average	**f1 score Deterioration**	**f1 score No Change**	**f1 score Improvement**
XGB	0.526	0.302	0.772	0.506
DCS	0.337	0.124	0.754	0.132
RO	0.329	0.115	0.762	0.111
HNTEMP	0.321	0.082	0.808	0.072
FeP	0.289	0.000	0.868	0.000
HN	0.289	0.000	0.868	0.000

Nevertheless when we compare our model to these baselines the Xgboost model scores higher (M = 0.526; SD = 0.003) when comparing it to the baselines, of which

the highest one (DCS) scores a f1 macro score of 0.337. If we look at the individual classes, not surprisingly, classes *Deterioration* and *Improvement* score substantially better than the baseline. As an example the highest score the baselines got for the *Deterioration* class is 0.124, while our model scores higher (M = 0.302; SD = 0.008). The Xgboost model even performs beter for the class *Improvement* (M = 0.506; SD = 0.005). This is quite impressive since the highest scoring baseline, is 0.132. Our Xgboost model however does under perform for the class *No Change* (M = 0.772; SD = 0.002) compared to three different baselines (FeP, HN and HNTEMP). Nevertheless this part shows that Xgboost model can predict food security to some extent. Finding *No Change* events, looks to perform worse than the baselines. This could however be a side effect of us focusing so much on optimizing the model for the minority class and making the data set more balanced. Moreover the scores indicate that the model finds improvements of food security better than deterioration. As you may have noticed the scores for the baseline HN and FeP are identical. This is because the historical norm is that usually livelihood zones food security does not change, hence it will always be the class *No Change*. This same rule applies for our third baseline. If we assume that the future equals the present, tomorrow is the same as today which means there is no change, hence it will always be the class *No Change*. It is however important to note that this reasoning, has a chance to not always hold. In other words, there can be a situation in which the historical norm might be different than the FeP model for a region in the future, thus we still included it in the table.

Table 11: Summary of the performance (f1 macro score) of the XGB model on the hold out set and the benchmark that we have created using the predictions of *Famine Early Warning Systems Network* (2019).

	f1 macro average	**f1 score Deterioration**	**f1 score No Change**	**f1 score Improvement**
XGB	0.526	0.302	0.772	0.506
FEWSNET	0.637	0.564	0.848	0.498

To also give an indication how the performance of the Xgboost model compares to the prediction of *Famine Early Warning Systems Network* (2019), we have also calculated the performance of their prediction of the change events (which is a derivative of their IPC prediction) compared to the real change events (which is also

a derivative of but in this case of the actual IPC values at that current moment). This will serve as a benchmark. Moreover, we have only run the benchmark once, thus we don't have a distribution of the scores. That being said, the results shows that our Xgboost model, with a f1 score average score of 0.526, does not perform as well as the benchmark which scores 0.637 on the f1 macro (see table 11). Our model also performs worse compared to the benchmark for the classes *Deterioration* score and the *No Change* on the f1 score. It is however interesting to see that the f1 score of the *Improvement* class does perform better for our model compared to the benchmark.

6.2.3 Feature insight

In order to identify which features are the most important for our model we have saved the feature name and its corresponding feature importance when we had bootstrapped the XGB model over a distribution of 100 models (see figure 10, for the full list of features and their importances please refer to the appendix and figure 20). Afterwards we have aggregated the mean values of each feature over these 100 runs. Not surprisingly, the features that hold either information about the IPC class (named CS in the graph) or information about how many times a livelihood zone had a deterioration previously are important to the model. Moreover features with regard to the soil moisture (ssm_mean and smp_mean), food market prices and the features holding information about the season are also relative important. Lastly the precipitation and population density also have a respectable feature importance compared to the rest of the features of the model.

6.2.4 Spatial Insight

In order to understand for which livelihood zones the model predicts best and which were the most difficult, we will map the different findings from our prediction and compare this to the real values. To elaborate, we will compare the predictions against the actual value for each of 100 models that we have run. In order to do this, we will calculate the accuracy score per livelihood zone (which is the correct number of predictions divided by all predictions). This in turn should give us an indication of how many times the model found a livelihood zone correctly. In order to identify these regions we will first map the accuracy score for each region over the geometry of Ethiopia (see figure 11).

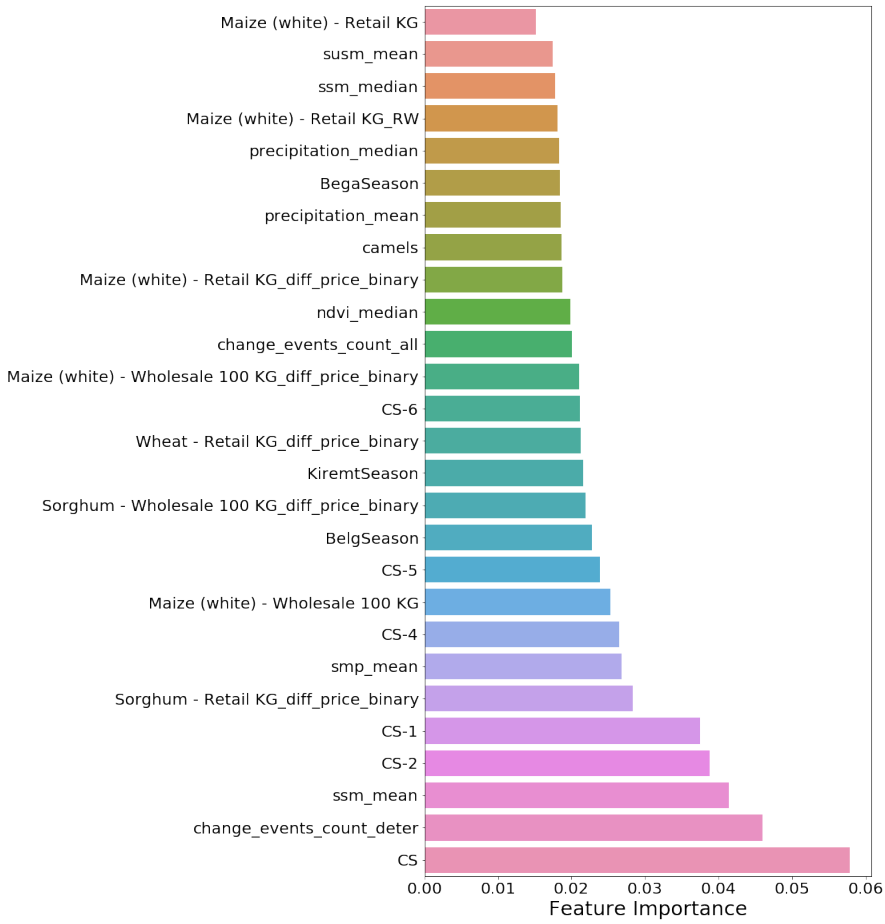


Figure 10: This image shows the average feature importances of the bootstrapped model that are larger than 0.015. It makes clear that that CS (IPC class), the CS time lags (CS-1 up to CS-6) and the number of times that a *Deterioration* change event happened before in a region are important. Moreover the soil moisture, seasonality, food market prices, precipitation and NDVI also have a relatively high importance.

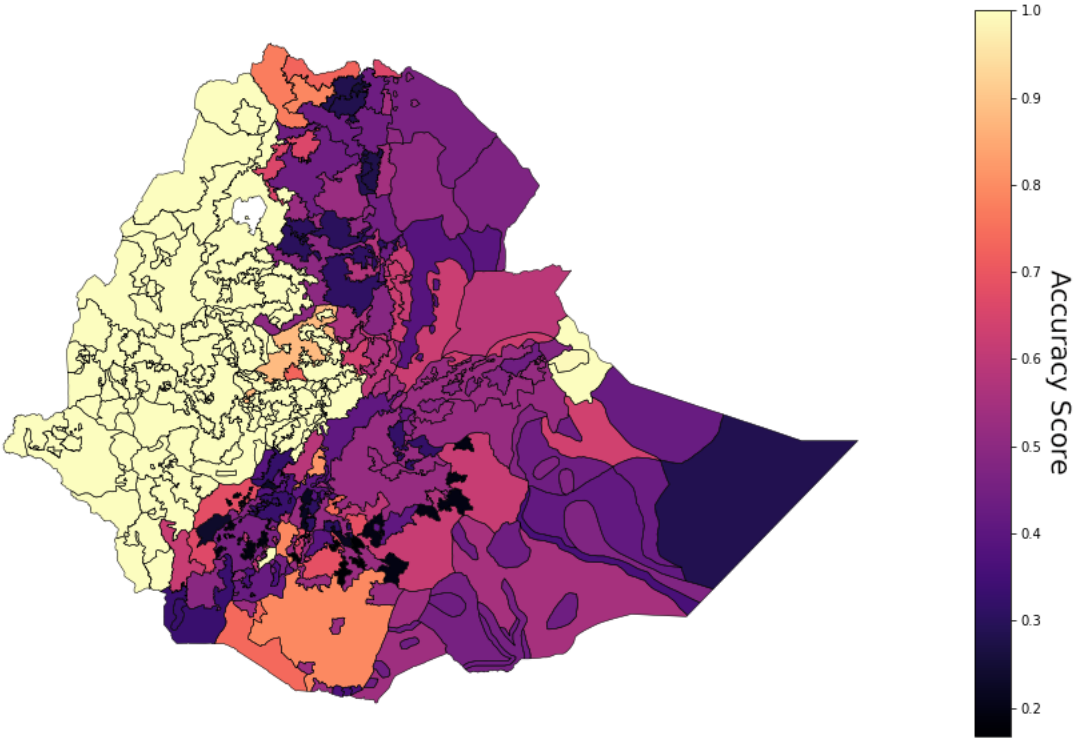


Figure 11: This map gives an indication which region were predicted correctly and which weren't. The regions in the north western were the easiest for the model to identify (thus an higher accuracy). The regions that are darker are more difficult to identify compared to regions that have a lighter colour(higher accuracy).

The result shows that especially regions in the western part of Ethiopia were more often correctly identified compared to the other regions. They even had an accuracy score of 100 procent. The most difficult livelihood zones (that have the darkest colors) are clustered around the southern part of Ethiopia and also at the northern part of Ethiopia. All in all, even though the model doesn't find each livelihood zone with a high accuracy, it does find each livelihood zone at least once over a distribution of 100 models. Moreover these results also show there seems to be strong spatial dependency for food security in Ethiopia as the performance to identify livelihood zones correctly looks to be clustered in some areas.

6.2.5 Temporal Insight

Even though we specifically tuned the hyper-parameters of our model in order to predict change events in food security four months into the future, we will also test how the model performs when we change this prediction interval. It is however important to note that the performance might be higher if we would do grid search optimization for each different prediction window. However, if we would do this the runtime would increase substantially, we would have to do the grid search for 12 different models. Nevertheless this section should at least give an indication of the performance of the model for different prediction intervals.

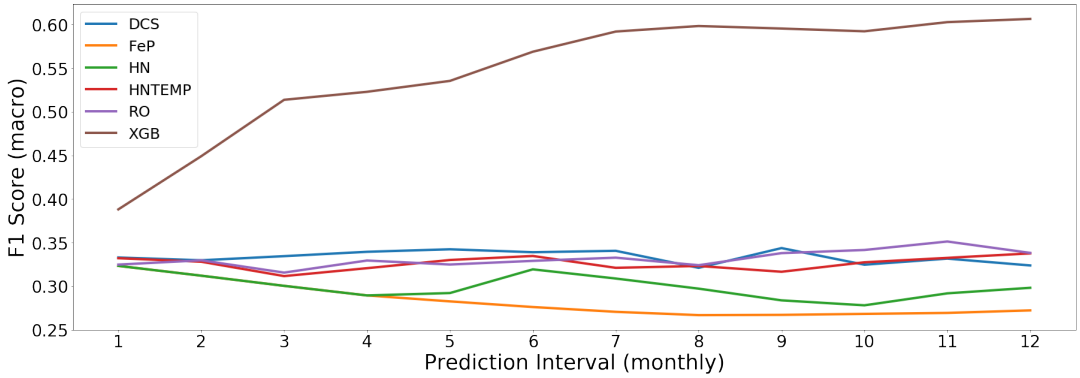


Figure 12: This image shows the performance for each of the different models that we have tested with regard to the f1 macro score. The Xgboost classifier (XGB) performs better for each prediction window when compared to the baselines. You can even see that the performance increases as the prediction interval expands.

In order to give an indication of the performance for different prediction intervals, we will also test our XGB model against the different baselines. We couldn't test the benchmark against the different prediction intervals, because the benchmark that we have created is only available for a prediction interval of 4. As figure 12 shows, our XGB model outperforms the different baselines that we have created if we compare these on the f1 macro score. You can see that as the prediction interval increases the performance of the XGB also increases.

In order to get more insight we will also have a look at the performances of the individual classes. These figures show that firstly the performance of the individual f1 scores are better for the class *Deterioration* and *Improvement* compared to the baselines. Moreover the same pattern exist as with the f1 macro average score, as the prediction interval increases the performance increases (figures 15 and 13). It is however interesting that if we look at the performance of the class *No Change* (figure 14). The performance decreases as the prediction interval expands. If we compare the XGB model with the baselines it performs worse when the prediction interval is smaller than 4 months. However after the prediction interval increases beyond 4 it performs better than the dummy stratisfied classifier (DCS) and the recent observation (RO) baseline. At prediction interval 6 it also outperforms the historical norm with temporal aspect in mind (HNT). It is interesting that the class *No Change* under performs when the prediction interval increases, because it shows that as the prediction interval increases and the minority classes get more samples the majority class gets less samples and thus decreases in performance. In order to validate this argument we have plotted the count for both our train data for each class in figure 16. This figure shows that the number of counts (for the train data without resampling) for the classes *Improvement* and *Deterioration* increases as the prediction interval expands. The counts for the class *No Change* decreases over time while the counts for the other two classes *Deterioration* and *Improvement* increases.

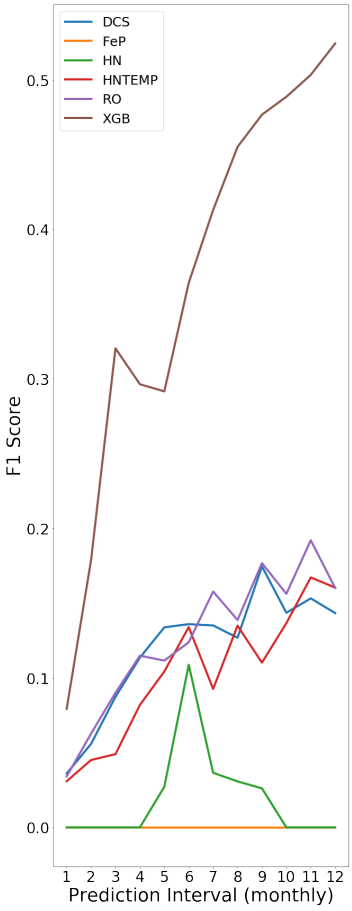


Figure 13: This image shows the f1 performance for the class *Deterioration* for each of the different models that we have tested. The Xgboost classifier (XGB) performs better for when the prediction interval expands compared to the baselines.

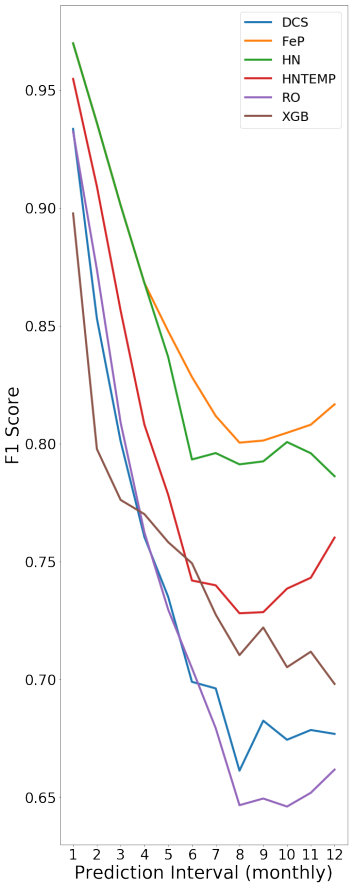


Figure 14: This image shows the performance for the class *No Change* for each of the different models that we have tested. The Xgboost classifier (XGB) performs worse than most of the baselines and only performs better compared to the other models after prediction interval 4. The performance decreases as the prediction interval expands.

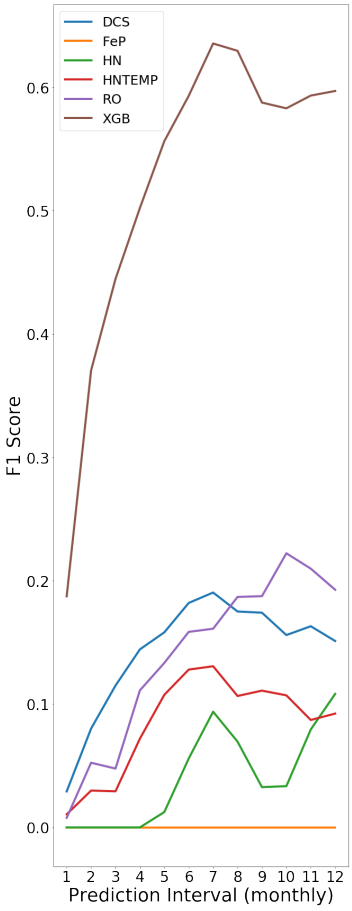


Figure 15: This image shows the f1 performance for the class *Improvement* for each of the different models that we have tested. The Xgboost classifier (XGB) performs better for each prediction window when compared to the baselines. You can even see that the performance increases as the prediction interval expands.

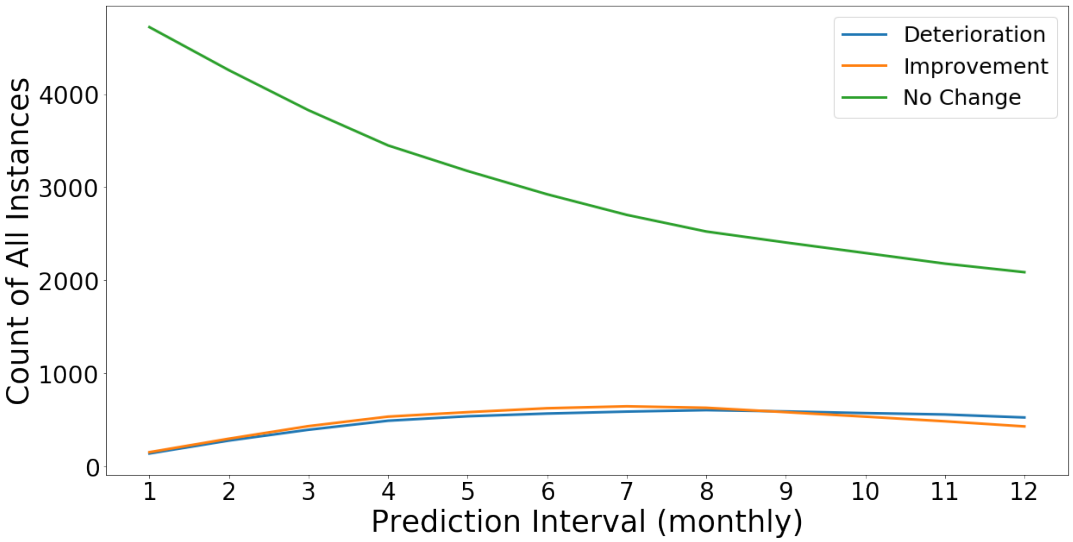


Figure 16: This image visualizes the count of the samples for each class of the train data. You can see the same pattern, as with the test data. As the prediction interval increases the number of instances for the minority classes (*Deterioration* and *Improvement*) increase while the majority class (*No Change*) decreases.

7 Discussion

The goal of this study was to create a model that can predict the change events (transitions in food security) for the livelihood zones in Ethiopia using scalable features. After collecting and processing scalable features, selecting our machine learning algorithm (Xgboost), metric (f1 macro score), the type of cross validation and optimizing we finally reached our goal. Even though this model is not perfect yet, the results show that our Xgboost model does have predictive value. It performs better than the baselines when we look at the f1 macro average score. The results also showed that our Xgboost model performs better than our baselines when predicting *Deteriorations* of the food security (negative change events) and *Improvement* in the food security (positive change events). Even though the f1 score of the *No Change* is higher than the other two classes, it does not perform better than most of our baselines. This is an example of why a baseline can be valuable: a high score alone is not enough to conclude whether the performance of a model can be considered as sufficient. We have also compared our Xgboost model with a benchmark that we have created using the predictions of *Famine Early Warning*

Systems Network (2019). The results showed that our Xgboost model is not as good as the predictions of *Famine Early Warning Systems Network* (2019) yet, however we did get close to the benchmark. Moreover when we look at the individual classes we even beat the benchmark for the class *Improvements*. However our model did not perform as well for the class *Deterioration* and *No Change* compared to the benchmark.

The results of the feature insight section showed that the importance's of our features (IPC, change event history, soil moisture, food market prices, seasonality, rain precipitation and population density) mostly reflects the study of (Misselhorn, 2004), in which climate, environmental stressors were one of the most cited drivers. Even though you can argue that the cumulative *Deterioration* variable that we have created through feature engineering (which is basically the history of a livelihood zone: how many times has there been a deterioration of the food security), contains some indirect information about whether a livelihood zone is stable or not. It is still surprising that fatalities did not have a big impact as that we would have expect. Social and political unrest or war was cited quite some time in the study of Misselhorn (2004) as a indirect driver. Moreover, like we have discussed before, it has been argued by organizations like the FAO, IFAD, UNICEF, WFP and WHO (2017) and the Food Insecurity Information Network (2018) that conflict is one of the main drivers of the rise in food insecurity. Nevertheless, apart from conflict, most features that got a relative high importance correspond with the study from (Misselhorn, 2004). With regard to features, this study also succeeded in using features that are scalable in our model. These features that we have collected and the tools that we have created should be relatively easy to implement for other countries. The features that we have created through feature engineering also resulted in a better performance for the Xgboost model.

The results also showed that there is a spatial dependency with regard to the performance of the model. To elaborate the livelihood zones in the north western part of Ethiopia were more easier to identify correctly compared to the livelihood zones in the other areas. The areas in the north western part of Ethiopia even got a score of 100 percent. In order to identify why this happened we checked the change events for these regions for the time frame of our hold out set (which is from 2016-01-31 until 2018-06-01). We realized that during this period these regions never changed. The same rule applies for the other three regions that have close to 100 percent of accuracy that are at the eastern side of Ethiopia and the small area in the

southern side of Ethiopia. In other words these areas never changed with regard to a food security state. This can be a result of either not having enough historical data about this region so we can test on a larger time interval, or maybe these regions really don't have that many transitions, and thus are more stable. Nevertheless we would have preferred to have more train data and more test data so that we would have a more complete image of the historical trend and seasonality of these regions.

Lastly our results also showed that the Xgboost model overall performs better for longer prediction intervals. There are probably several reasons why this effect took place. First, the smaller the prediction interval, the smaller the chance that the Xgboost model can actually predict a change within that window. In other words if the prediction interval is larger this means that the Xgboost model is less likely to make a mistake, to correctly identify the change event at an exact time moment. For example if the model would predict one day ahead, the chance that the model would be able to find the change event exactly on a daily scale would be almost impossible. In order to make this easier to understand we will compare this with an analogy between two soccer teams. Assume there is a soccer game between two teams: team A and team B. Team A is presumably way better, since they have better players and a better goal difference compared to team B. In other words, you are pretty sure that Team A is more likely to win from Team B within a 90 minute game. However if we would have to predict in which time window of 5 minutes team A would score this would be harder (since there are 18 possibilities), compared to when we would want to predict within a time window of 45 minutes when team A would score against team B (2 possibilities). This is the same for our model. If would predict for a longer time interval it is easier to predict (since there are less possibilities) whether there is a transition in food security compared to when there is a shorter time interval. Second, the train data set gets more balanced. The reason for this is that there are more likely to be more change events in the long term (longer prediction interval) compared to the short term (smaller prediction interval). Subsequently because of the larger train data, the resampling strategy could also be more effective since it has more examples per class that it can synthetically create. Lastly, it could also be the case that our features that we are using to predict the change events are more effective in identifying changes in the food security state in the long term compared to the short term.

7.1 Further Research

During the study some suggestions and questions have come to light. In order to validate the Xgboost model further and find out whether it actually predicts food security some different steps have to be taken, which unfortunately were not possible in the time frame that this study had. First, questions have to be raised with regard to the IPC variable (which we have used to create the change event) from *Famine Early Warning Systems Network* (2019) in combination with our Xgboost model. Specifically, in order to test how well this model generalizes and performs in other countries, it might be a good idea to test its performance in a different country like Kenya. We would suggest Kenya because figure 21 (in the appendix A.5) shows that the fluctuations in Kenya are less volatile with regard to the change event mean (which means that there are less moments in which the food security state improves or deteriorates) compared to Ethiopia. Testing the Xgboost model on data that is presumably more stable, according to data of *Famine Early Warning Systems Network* (2019), might be interesting to further validate the generalizability of the model. Another way to validate the Xgboost model, would be to use an alternative target variable. Talking to an expert at the Red Cross it became clear that they might get access to the woreda hotspot clarification, which is an alternative metric for food security. Using a different target variable that should roughly express the same concept could also validate the model further. A severe drawback however, using the hotspot woreda's, would be that this variable is only available for Ethiopia while the IPC is available for most countries.

This study also has some suggestions how the performance of the model can be increased. First, there are limitations to using open data, not every feature that might be important to predict these change events in food security might be open data. Thus it might be a good idea to also try to incorporate non open data for the model so it can improve its scores. To give an example, we expected conflict (the data set from Uppsala Conflict Data Program (2018) to have a bigger impact on the model than our results show. The data set from Uppsala Conflict Data Program (2018) had a minor effect on our model (see figure 20). A issue with this data set is that it contains the death tolls due to conflict which is collected through news outlets. News outlets however can be biased, since there is a chance that not every news event gets reported, thus the reliability of this set can be questioned. In order to overcome this, an option would be to find a different data set or source which is more reliable The problem however with a concept like conflict, is that there is not a lot of

open data available about this sensitive topic. So a suggestion for a non open data source for this information is the tool developed by DHL called Resillience360. This tool gives organizations the option to visualize and track relevant early warnings and emergency incidents near-real time around the world (*DHL*, 2019). Another potential data set would be through the International Committee of the Red Cross. However, the problem with this would be that this data set is probably not open data and might thus be difficult to get, since it contains sensitive data.

Secondly, adding even more features, not surprisingly, could also improve the performance of the model. Features that could potentially increase the performance of the model are land cover, soil quality and types and poverty. The problem with satellite imagery of land cover is that we could not find a data set that has been updated recently. The most up to date data set of of land cover went up to 2016. Moreover, we could also not find a data set that had information about the soil quality and types in the GEE. A different feature that could potentially also increase the performance of the model is poverty. Since poverty is also a sensitive topic, the questions remains whether we could express this concept through open data. However, a potential candidate to express poverty to a certain extent is using city lights as a proxy for poverty. Composite light data sets have been used, by studies like that of Yu et al. (2015), to predict poverty. However in our case we did not include this, since the data from the composite light data set of the GEE started started from 2012. Subsequently, this would mean that we would have need to throw away 2 years of data (since our target variable started from 2010). In hindsight we could have also treat these composite light data sets as a constant factor. Nevertheless it would still be interesting to see how city light satellite imagery would influence the models performance.

Another way of achieving more performance for the model is to get in contact with *Famine Early Warning Systems Network* (2019) in order to get more information on what kind of scale their variables are (what is the granularity that they use). To be more specific, it could potentially be the case that the data that *Famine Early Warning Systems Network* (2019) uses to create the IPC is not as detailed as satellite imagery. Clarification about the granularity could increase the performance of the model, since we than know on what kind of scale they operate and what kind of variables we have to search for.

Another option to increase the performance of the model is to spend even more time on feature engineering. Techniques like the condition index from the study

of Kogan (1990) could potentially increase the performance of the model. This is an index that compares the current value at a given location with the historical data. This can be benefcial since according to Sannier, Gilliams, Ham, and Fillol (2015) there is a need to compare the current value at a given location with the historical data. Sannier et al. (2015) argues that it is not possible otherwise to determine whether for example, the vegetation conditions are better or worse than normal. Next to condition index, it might also be more beneficial to get more domain knowledge, or get in contact with more experts, to determine what features and feature engineering could potentially increase the performance of the model.

Another suggestion is using a different aggregation level. For now the livelihood zone areas seem to be the most relevant from a more theoratical view. However from a more practical view the admin level 3, which basically divides Ethiopia in geographical provinces, divides the country in smaller regions compared to the livelihoodzone. Subsequently, this could potentially increase the performance of the model since the aggregation level is on a more granular scale than the livelihoodzone, and thus results in more data per region.

As the spatial insight and the data exploration results showed, there seems to be some spatial dependency and relationship. There seems to be a strong clustering with regard to the change events of food security. Using more spatial information could potentially lead to an increase of performance. To give two examples we have only tested the model on Ethiopia and presumed that It is an isolated country. However deriving features based on neighboring regions (to better take network effects into account, or to see how food insecurity spreads) might increase the performance of the model.

Moreover, even though our model can predict what kind of transition of the food security states takes place, for now it can't differentiate between different transitions (transition from IPC 1 to 3 or IPC 1 to 2) and what kind of IPC value these transitions had. Due to time, we only decided to focus on the first step, a model that can predict the transitions. However, expanding the model to give more information back would be beneficial if the Netherlands Red Cross wants to use this model for decision-making, but also get more insight when and why these transitions happen.

Another suggestion for further research is building a model that can predict the more extreme cases (transition from IPC class 4 to 5). We couldn't make this for Ethiopia, or test for the simple reason that IPC class 5 has never been measured in Ethiopia. Moreover, these extreme cases almost never happen and our thus super

imbalanced. Building a model to detect these transitions will be a challenge.

There were also several issues that this study ran into that might need some more clarification in further research. First, the question can be asked on how much data the XGB model we train on. To put it differently, how much years of historical data do we need to train in order to predict in the future. On the other hand it could also be the case that there is not enough historical data and that we simply need more data. Unfortunately this is of course limited by the target variable which started in 2010. Secondly, there are many different type of models that you can use. We chose the Xgboost because it is flexible and performs well. However the question remains how other type of models perform. That being said, this study did try to use a Random Forest as well, however since the performance of the Xgboost was better at default settings we opted to continue our analysis with the Xgboost model.

8 Conclusion

This study aimed to create a model using scalable features that can predict changes in food security per livelihoodzone in Ethiopia. As the result and discussion have shown, our Xgboost model can predict transitions of food security states for livelihood zones in Ethiopia better than the baselines. And even though we don't perform better (on average) yet than the benchmark (which we created using the predictions of *Famine Early Warning Systems Network* (2019)), we believe that further improvements of this model can increase the performance so that it performs better than or similar to the benchmark. Overall our Xgboost model can identify regions that improve with regard to food security better than areas that deteriorate or stay the same. Nevertheless even though this tool might not optimize the decision making of the Netherlands Red Cross right away, it does show great potential. Moreover we have also gained more insight and a better situational awareness of what kind of features contribute in predicting food security, that there is a spatial dependency in livelihood zones that needs more investigation and that performance of predicting change events might differ for different prediction intervals. Even more we have also succeeded in collecting and using scalable features that can be used for other countries. This can be highly valuable for the Netherlands Red Cross since they can use this tool to extract satellite data more easily for other countries and potentially even use it for their own Community Risk Assessment Dashboard.

All in all this study showed that using Artificial Intelligence and state of the art techniques can be relevant in the humanitarian sector and potentially, in the future, get us closer to optimizing the humanitarian assistance so that we can support and relieve food insecure regions more efficiently.

References

Tropical Rainfall Measuring Mission (TRMM). (2011). *Trmm (tmpa/3b43) rainfall estimate l3 1 month 0.25 degree x 0.25 degree v7* (Vol. 7). Goddard Earth Sciences Data and Information Services Center. (data retrieved through the Google Earth Engine, `https://disc.gsfc.nasa.gov/datacollection/TRMM_3B43_7.html`) doi: 10.5067/TRMM/TMPA/MONTH/7

510. (2018, July 06). *Mission and vision.* Retrieved 26-7-2018, from `https://www.510.global/video/`

Alvera-Azcárate, A., Sirjacobs, D., Barth, A., & Beckers, J.-M. (2012). Outlier detection in satellite data using spatial coherence. *Remote Sensing of Environment*, *119*, 84 - 91. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0034425711004469` doi: https://doi.org/10.1016/j.rse.2011.12.009

A Quinn, J., Okori, W., & Gidudu, A. (2010, 01). Increased-specificity famine prediction using satellite observation data. *Proceedings of the 1st ACM Symposium on Computing for Development, DEV 2010*. doi: 10.1145/1926180.1926203

Barrett, C. B. (2002). Chapter 40 food security and food assistance programs. , *2*, 2103 - 2190. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1574007202100272` doi: https://doi.org/10.1016/S1574-0072(02)10027-2

Barrett, C. B. (2010a). Measuring food insecurity. *Science*, *327*(5967), 825–828. Retrieved from `https://science.sciencemag.org/content/327/5967/825` doi: 10.1126/science.1182768

Barrett, C. B. (2010b, 02). Measuring food security. , *327*, 825-8.

Barron, J., Tharme, R., & Herrero, M. (2013, 08). Drivers and challenges for food security. In (p. 7-28).

Bolten, J. D., & Crow, W. T. (2012). Improved prediction of quasi-global vegetation conditions using remotely-sensed surface soil moisture. *Geophysical Research*

Letters, *39*(19). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL053470 doi: 10.1029/2012GL053470

Bolten, J. D., Crow, W. T., Zhan, X., Jackson, T. J., & Reynolds, C. A. (2010). Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *3*(1), 57–66. doi: 10.1109/jstars.2009.2037163

Bora, a., Ceccacci, I., Delgado, C., & Townsend, R. (2011). *Food security and conflict* (Background Paper). World Bank. Retrieved 07-9-2018, from http://web.worldbank.org/archive/website01306/web/pdf/wdr\%20background\%20paper_bora\%20et\%20al.pdf

Center for International Earth Science Information Network (CIESIN). (2016). *Gridded population of the world, version 4 (gpwv4): Population count adjusted to match 2015 revision of un wpp country totals.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Retrieved 2019-01-10, from https://developers.google.com/earth-engine/datasets/catalog/CIESIN_GPWv4_unwpp-adjusted-population-count doi: https://doi.org/10.7927/H4HX19NJ

Champion, N., Le-Hir, E., Massera, S., & Bellaiche, N. (2017). Automatic production of large-scale cloud-free orthomosaics from multitemporal satellite images. *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, 1-3.

Cheeseman, J. (2016). 7 - food security in the face of salinity, drought, climate change, and population growth. In M. A. Khan, M. Ozturk, B. Gul, & M. Z. Ahmed (Eds.), *Halophytes for food security in dry lands* (p. 111 - 123). San Diego: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780128018545000078 doi: https://doi.org/10.1016/B978-0-12-801854-5.00007-8

Chen, T., & Guestrin, C. (2016, 08). Xgboost: A scalable tree boosting system. , 785-794. doi: 10.1145/2939672.2939785

CIESIN. (2016). *Gridded population of the world, version 4 (gpwv4): Population density adjusted to match 2015 revision of un wpp country totals.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Retrieved 2019-01-10, from https://developers.google.com/earth-engine/datasets/catalog/CIESIN_GPWv4_unwpp-adjusted-population-density

doi: https://doi.org/10.7927/H4HX19NJ

Connolly-Boutin, L., & Smit, B. (2016, Feb 01). Climate change, food security, and livelihoods in sub-saharan africa. *Regional Environmental Change*, *16*(2), 385–399. Retrieved from https://doi.org/10.1007/s10113-015-0761-x doi: 10.1007/s10113-015-0761-x

Danielson, J. J., & Gesch, D. B. (2011). *Global multi-resolution terrain elevation data 2010 (gmted2010).* U.S. Geological Survey. doi: 10.3133/ofr20111073

Denny, R. C. H., Marquart-Pyatt, S. T., Ligmann-Zielinska, A., Olabisi, L. S., Rivers, L., Du, J., & Liverpool-Tasie, L. S. O. (2018, Mar 01). Food security in africa: a cross-scale, empirical investigation using structural equation modeling. *Environment Systems and Decisions*, *38*(1), 6–22. Retrieved from https://doi.org/10.1007/s10669-017-9652-7 doi: 10.1007/s10669-017-9652-7

Dhl. (2019). DHL. Retrieved from https://www.resilience360.dhl.com/

E. Mladenova, I., D. Bolten, J., Crow, W., Anderson, M., Hain, C., Johnson, D., & Mueller, R. (2017, 01). Intercomparison of soil moisture, evaporative stress, and vegetation indices for estimating corn and soybean yields over the u.s. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *PP*, 1-16. doi: 10.1109/JSTARS.2016.2639338

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., ... Zyl, J. V. (2010, May). The soil moisture active passive (smap) mission. *Proceedings of the IEEE*, *98*(5), 704-716. doi: 10.1109/JPROC.2010.2043918

Famine early warning systems network. (2009, Nov). Famine Early Warning Systems Network. Retrieved from http://fews.net/east-africa/ethiopia/livelihood-zone-map/november-2009

Famine early warning systems network. (2019). Famine Early Warning Systems Network. Retrieved from http://fews.net/IPC

FAO. (2003). Trade reforms and food security. *Rome: FAO*.

FAO, IFAD, UNICEF, WFP and WHO. (2017). The state of food security and nutrition in the world. *Rome: FAO*.

FFSSA. (2004). Achieving food security in southern africa: Policy issues and options. *FFSSA Synthesis Paper, Forum for Food Security in Southern Africa*. Retrieved from http://www.odi.org.uk/food-security-forum

Food Insecurity Information Network. (2018). Global report on food crises 2018.

Food Security Cluster. (2016). Core indicator handbook. *Rome: Food Security Cluster*.

Frelat, R., Lopez-Ridaura, S., Giller, K. E., Herrero, M., Douxchamps, S., Djurfeldt, A. A., ... et al. (2015). Drivers of household food availability in sub-saharan africa based on big data from small farms. *Proceedings of the National Academy of Sciences*, *113*(2), 458–463. doi: 10.1073/pnas.1518384112

Geda, N., & Stoecker, B. (2011, 12). Household food insecurity and hunger among households in sidama district, southern ethiopia. *Public health nutrition*, *15*, 1276-83. doi: 10.1017/S1368980011003119

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., ... Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. *Science*, *327*(5967), 812–818. Retrieved from `http://science.sciencemag.org/content/327/5967/812` doi: 10.1126/science.1185383

Google. (2019a). *Classification: Accuracy.* Retrieved 2019-02-12, from `https://developers.google.com/machine-learning/crash-course/classification/accuracy`

Google. (2019b). *Classification: Precision and recall.* Retrieved 2019-02-12, from `https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall`

Grillo, J. (2009). *Application of the livelihood zone maps and profiles for food security analysis and early warning guidance for famine early warning systems network (fewsnet) representatives and partners.* Retrieved 23-7-2018, from `http://fews.net/sites/default/files/documents/reports/Guidance_Application_of_Livelihood_Zone_Maps_and_Profiles_en.pdf`

Gubert, M., Benicio, M., Padilha da Silva, J., Rosa, T., Santos, S., & Santos, L. (2010, 06). Use of a predictive model for food insecurity estimates in brazil. , *60*, 119-25.

Guimarães Nobre, G., Davenport, F., Bischiniotis, K., Veldkamp, T. I., Jongman, B., C. Funk, C., ... C.J.H. Aerts, J. (2018, 10). Financing agricultural drought risk through ex-ante cash transfers. *Science of The Total Environment*, *653*. doi: 10.1016/j.scitotenv.2018.10.406

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322-1328.

HEA. (2018). *Baseline assessments.* Retrieved 20-7-2018, from http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm

Headey, D., & Barrett, C. B. (2015). Opinion: Measuring development resilience in the world's poorest countries. *Proceedings of the National Academy of Sciences*, *112*(37), 11423–11425. Retrieved from https://www.pnas.org/content/112/37/11423 doi: 10.1073/pnas.1512215112

Hesselberg, J., & Yaro, J. (2006, 09). An assessment of the extent and causes of food insecurity in northern ghana using a livelihood vulnerability framework. *GeoJournal*, *67*, 41-55. doi: 10.1007/s10708-006-9007-2

Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal avhrr data. *International Journal of Remote Sensing*, *7*(11), 1417-1434. Retrieved from https://doi.org/10.1080/01431168608948945 doi: 10.1080/01431168608948945

Holloway, A. (2003). Disaster risk reduction in southern africa: hot rhetoric, cold reality. *African Security Review.*.

IPC Global Partners. (2012). Integrated food security phase classification technical manual version 2.0. evidence and standards for better food security decisions. *Rome: FAO*.

J. Weiss, D., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A., ... W. Gething, P. (2018, 01). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, *553*. doi: 10.1038/nature25181

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, *3*(3), 262–267.

Kogan, F. (1990, 08). Remote sensing of weather impact on vegetation in non-homogeneous areas. *International Journal of Remote Sensing - INT J REMOTE SENS*, *11*, 1405-1419. doi: 10.1080/01431169008955102

Mbukwa, J. (2013, 01). A model for predicting food security status among households in developing countries. , *22*, 2168-8662.

Misselhorn, A. (2004, 04). What drives food insecurity in southern africa? a meta-analysis of household economy studies. *Global Environmental Change*, *15*, 33-43. doi: 10.1016/j.gloenvcha.2004.11.003

Okori, W., & Obua, J. (2011). Machine learning classification technique for famine prediction. *Proceedings of the World Congress on Engineering 2011 Vol II*.

O'Neill, P. E., Chan, S., Njoku, E. G., Jackson, T., & Bindlish, R. (2016). Smap

l3 radiometer global daily 36 km ease-grid soil moisture, version 4. boulder, colorado usa. nasa national snow and ice data center distributed active archive center.

Osipenko, A. (2018, Aug). *Backtesting time series models - weekend of a data scientist.* medium. Retrieved from https://medium.com/cindicator/backtesting-time-series-models-weekend-of-a-data-scientist-92079cc2c540

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, *12*, 2825–2830.

Python api reference. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/latest/python/python_api.html

Rooyen, J. V. (2000). Regional food security and agricultural policy in southern africa: A challenge of policy conversion in diverse settings. *Development Southern Africa*, *17*(1), 7-22. Retrieved from https://doi.org/10.1080/03768350050003389 doi: 10.1080/03768350050003389

Sannier, C., Gilliams, S., Ham, F., & Fillol, E. (2015, 06). Use of satellite image derived products for early warning and monitoring of the impact of drought on food security in africa. , 183-198. doi: 10.1007/978-1-4939-2602-2_12

Sassi, M. (2017). Understanding food insecurity: Key features, indicators, and response design. *Cham: Springer*.

Swindsdale, A., & Bilinsky, P. (2006). Develop-ment of a universally applicable household food insecurity measurement tool: Process, current status, and outstanding issues. *Journal of Nutrition*, 1449-1452.

the Red Cross Red Crescent Climate Centre, GRC and 510. (2018, July). *A guide to trigger methodologyfor forecast-based financing.* (http://fbf.drk.de/fileadmin/user_upload/FbF_Manual_-_A_guide_to_trigger_methodology.pdf)

Uppsala Conflict Data Program. (2018, 07 20). *Ucdp conflict encyclopedia: www.ucdp.uu.se.*

Upton, J. B., Cissé, J. D., & Barrett, C. B. (2016). Food security as resilience: reconciling definition and measurement. *Agricultural Economics*, *47*(S1), 135-147. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/agec.12305 doi: 10.1111/agec.12305

World Food Programme. (2019). *Global food prices database.* Retrieved 2019-01-10,

from https://data.humdata.org/dataset/wfp-food-prices

Yu, B., Shi, K., Hu, Y., Huang, C., Chen, Z., & Wu, J. (2015, March). Poverty evaluation using npp-viirs nighttime light composite data at the county level in china. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(3), 1217-1229. doi: 10.1109/JSTARS.2015.2399416

A Appendix

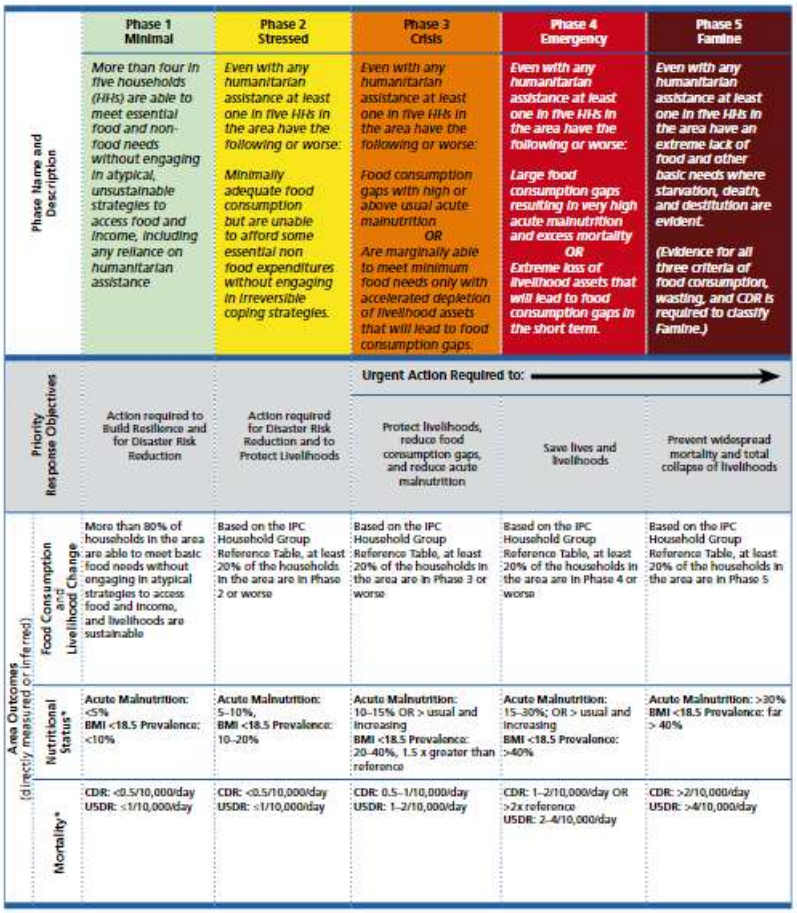
A.1 Reflection

During this study I've utilized my coding, critical evaluation, statistical and team skills, which I've acquired during my master Artificial Intelligence, to create a model that predicts the change events in food security. Even though a part of this study was collecting data and processing it, I have learned a lot about making sure that the model learns the correct patterns and how easy it is to make mistakes with regard to information leakage in time series models (processing data correctly is very important). Next to this, I've also learned a lot about choosing the right metrics, optimizing the model (in this case a Xgboost) itself and validating the model properly. Next to cross validation, having both a baseline and benchmark is of utmost importance to determine how well the performance of the model actually is. Lastly, I have also learned a lot about how to translate numbers to understanding. Making figures, like in the spatial aspect section, can give a lot of insight where the model performs strongly and where it doesn't.

A.2 Why is Artificial Intelligence relevant for this study?

If you think about Artificial Intelligence, you will probably not think directly about the humanitarian sector, but big companies like Google or Tesla with its self driving cars. However this study has shown that using Artificial Intelligence in the humanitarian sector has a lot of potential. To be more specific, by creating a model that predicts the transitions of food security, humanitarian organizations can better estimate which areas are the most in need of help. By using the state of the art techniques and Artificial Intelligence in this study, we are a step closer to the purpose of the 510. Namely, using data to positively impact faster and more (cost) effective humanitarian aid. This in turn can lead to timely finance prior to a disaster which can be more cost-effective than post-disaster expenditures (Guimarães Nobre et al., 2018). Doing research, with regard to Artificial Intelligence, can in the future get us closer to optimizing the humanitarian assistance, so that organizations can support and relieve food insecure regions more efficiently.

A.3 IPC Definitions



	Phase 1 Minimal	Phase 2 Stressed	Phase 3 Crisis	Phase 4 Emergency	Phase 5 Famine
Phase Name and Description	More than four in five households (HHs) are able to meet essential food and non-food needs without engaging in atypical, unsustainable strategies to access food and income, including any reliance on humanitarian assistance	Even with any humanitarian assistance at least one in five HHs in the area have the following or worse: Minimally adequate food consumption but are unable to afford some essential non food expenditures without engaging in irreversible coping strategies.	Even with any humanitarian assistance at least one in five HHs in the area have the following or worse: Food consumption gaps with high or above usual acute malnutrition OR Are marginally able to meet minimum food needs only with accelerated depletion of livelihood assets that will lead to food consumption gaps.	Even with any humanitarian assistance at least one in five HHs in the area have the following or worse: Large food consumption gaps resulting in very high acute malnutrition and excess mortality OR Extreme loss of livelihood assets that will lead to food consumption gaps in the short term.	Even with any humanitarian assistance at least one in five HHs in the area have an extreme lack of food and other basic needs where starvation, death, and destitution are evident. (Evidence for all three criteria of food consumption, wasting, and CDR is required to classify Famine.)
Priority Response Objectives	Action required to Build Resilience and for Disaster Risk Reduction	Action required for Disaster Risk Reduction and to Protect Livelihoods	Urgent Action Required to: Protect livelihoods, reduce food consumption gaps, and reduce acute malnutrition	Save lives and livelihoods	Prevent widespread mortality and total collapse of livelihoods
Area Outcomes (directly measured or inferred) — Food Consumption and Livelihood Change	More than 80% of households in the area are able to meet basic food needs without engaging in atypical strategies to access food and income, and livelihoods are sustainable	Based on the IPC Household Group Reference Table, at least 20% of the households in the area are in Phase 2 or worse	Based on the IPC Household Group Reference Table, at least 20% of the households in the area are in Phase 3 or worse	Based on the IPC Household Group Reference Table, at least 20% of the households in the area are in Phase 4 or worse	Based on the IPC Household Group Reference Table, at least 20% of the households in the area are in Phase 5
Area Outcomes (directly measured or inferred) — Nutritional Status*	Acute Malnutrition: <5% BMI <18.5 Prevalence: <10%	Acute Malnutrition: 5–10%, BMI <18.5 Prevalence: 10–20%	Acute Malnutrition: 10–15% OR > usual and increasing BMI <18.5 Prevalence: 20–40%, 1.5 x greater than reference	Acute Malnutrition: 15–30%; OR > usual and increasing BMI <18.5 Prevalence: >40%	Acute Malnutrition: >30% BMI <18.5 Prevalence: far > 40%
Area Outcomes (directly measured or inferred) — Mortality*	CDR: <0.5/10,000/day U5DR: <1/10,000/day	CDR: <0.5/10,000/day U5DR: <1/10,000/day	CDR: 0.5–1/10,000/day U5DR: 1–2/10,000/day	CDR: 1–2/10,000/day OR >2x reference U5DR: 2–4/10,000/day	CDR: >2/10,000/day U5DR: >4/10,000/day

Figure 17: Definitions of the different IPC classes (IPC Global Partners, 2012)

A.4 Hyperparameter Tuning

The first step in optimizing the model was to initialize a data set containing all features and our target variable (change event within the next 4 months). Afterwards we will split the data set like we have discussed in the methods so that the train set consists of data from 2010-01-01 until 2016-01-31 and the holdout set consists out of data from 2016-01-31 until 2018-02-01. We in turn use the train set to find the best set of parameters using a self built grid search time series cross validation. The first step was to determine the best optimal number of estimators. We have chosen to test 10, 20, 40, 80, 100, 200, 300, 400, 500, 800, 1000, 1500 of estimators, while keeping the rest of the parameters as default, with the exception of the learning rate (0.01), colsample by tree (0.8) and subsample (0.8).

Figure 18 shows that the score of the f1 macro increases as the n_estimators increases up to around 400 estimators (f1-macro = 0.547) and afterwards slowly decreases again. You can see the same pattern if we look at the individual f1 scores of each class (see table 12). Nevertheless, increasing the n_estimators also increases the chance to overfit. We have visualized this in figure 19, in which you can clearly see that as the n_estimators increases the difference between the f1 macro train score and test score of the grid search also increases. We will however choose 400 n_estimators since it has the largest score for this intersection of parameter space. However, it is important to note that the optimal n_estimators might change when we have optimized the other parameters in the next sections.

Using 400 n_estimators we will first optimize the tree based parameters. More specifically we will iterate over different combination of parameters which are listed in table 13. Our goal is to find the parameters parameters that give a high f1 macro score while keeping a check on imbalanced results between the individual class scores. With the chosen parameters (which are underscored in table 13), we have been able to increase the f1 macro score to 0.551 which is an small improvement over the score that we got when we increased the n_estimators to 400 (0.543). If we look at the individual classes, the f1 score of class *Deterioration* increased from 0.326 to 0.348. Class *Improvement* also increased from 0.489 to 0.509. However class *No Change* slightly decreased from 0.814 to 0.795.

The next step is to tune the regularization reg_alpha and reg_lambda to increase the performance, reduce the chance of overfitting by reducing the complexity of the model itself. In this case we will use the found tree based parameters that have been identified in the previous section. For reg_alpha we have tested 5 different values

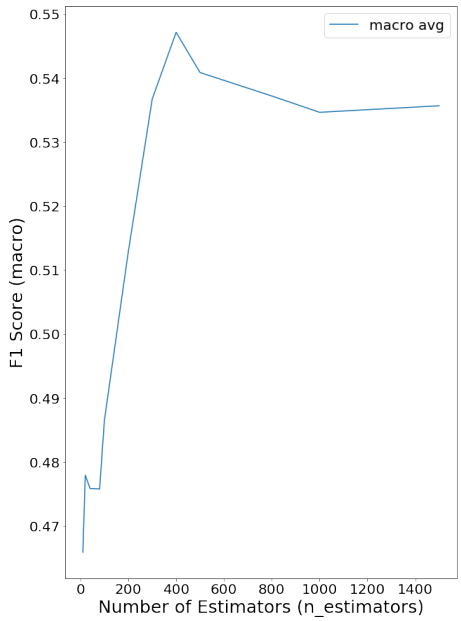


Figure 18: The performance increases as the estimators are increasing. However after 400 estimators the performance slowly decreases again. The f1 score represents the averaged f1 macro score over 3 time series splits for the f1 macro average,

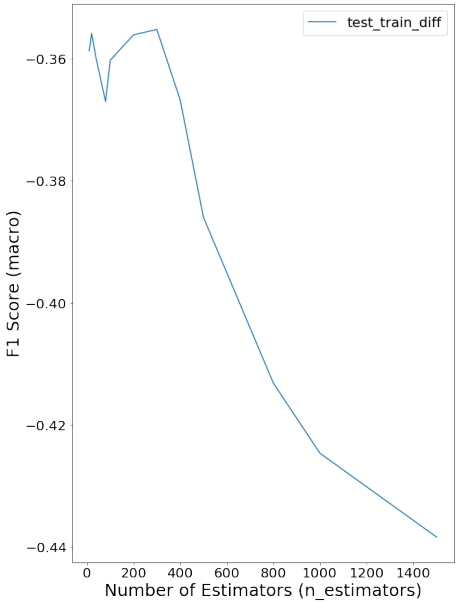


Figure 19: The difference in train and test macro scores in the grid search. As the number of estimators increases the difference between the train and test score also increases. Thus, this shows that increasing the number of trees results in overfitting with the current learning rate (0.01)

Table 12: Summary of the different f1 scores of the individual classes and macro average over all classes. From the table it is apparent that as the number of estimators is increasing the performance also increases up to 500 number of estimators.

n_estimators	**f1 score Deterioration**	**f1 score No Change**	**f1 score Improvement**	**f1 macro average**
10	0.310	0.629	0.458	0.466
20	0.286	0.667	0.480	0.478
40	0.304	0.649	0.474	0.476
80	0.285	0.664	0.478	0.476
100	0.295	0.671	0.494	0.487
200	0.324	0.721	0.494	0.513
300	0.345	0.765	0.500	0.537
400	0.330	0.794	0.518	0.547
500	0.306	0.812	0.505	0.541
800	0.287	0.824	0.500	0.537
1000	0.273	0.831	0.499	0.535
1500	0.277	0.830	0.500	0.536

Table 13: Summary of the found tree based hyper parameters in our grid seach time series cross validation. The hyperparameters that gave the best score for our chosen metrics in this iteration are underscored.

Hyperparameters				
max depth	3	5	7	
min child weight	1	3	5	7
gamma	0	0.1	0.3	
subsample	0.5	0.7	0.8	
colsample by tree	0.5	0.7	0.8	
n_estimators	400			
learning rate	0.01			

Table 14: Summary of the different parameter used in order to tune the regularization parameters and reduce the chance of overfitting. The parameters that we have chosen are underscored.

Hyperparameters					
max depth	4				
min child weight	7				
gamma	0.1				
subsample	0.45				
colsample by tree	0.75				
n_estimators	400				
learning rate	0.01				
reg_alpha	<u>0.00001</u>	0.01	0.1	1	100
reg_lambda	0.5	1	5	<u>10</u>	

(0.00001, 0.01, 0.01, 1 and 100). For reg_lambda we have tested three different values, namely 0.5, 1, 5 and 10. The grid search had identified that a reg_alpha of 0.00001 and a reg_lambda of 10 lead to the best f1 macro score which was increased to 0.555 from 0.551 (score of previous step in the grid search). Moreover looking at the individual f1 scores, class *Deterioration* further increased to 0.356 from 0.348. Class *No Change* decreased to 0.794 from 0.795 and class *Improvement* increased to 0.514 from 0.509. However in order to identify whether overfitting decreases we will calculate the difference between the f1 macro scores between the train and test sets. Or in other words we will identify how big the difference is between the train and test phase of the model. After we had optimized the tree based parameters the train and test score was -0.363. However after we had tuned the alpha and lambda score of the Xgboost the difference score improved to -0.357. In other words using these regularization parameters we could slightly improve the performance score but also reduce overfitting.

The last step is tune the n_estimators in combination with the learning_rate to further reduce overfitting and the complexity of the model. For the grid search we have used 500, 1000, 2000, 4000, 8000 and 10000 n_estimators. For the learning rate we have used 0.01, 0.005, 0.003, 0.001 and 0.0001. For the other parameters we have used the parameters that we have identified during the previous steps in our grid search. During this grid search 400 estimators in combination with a learning rate

of 0.010 lead to the most optimal performance compared to the other combination of hyper parameters. This combination of estimators and learning rate is the same as the previous step. The performance is thus almost the same (slightly variations due to randomness), so we won't go into depth about the exact scores.

A.5 Extra Figures

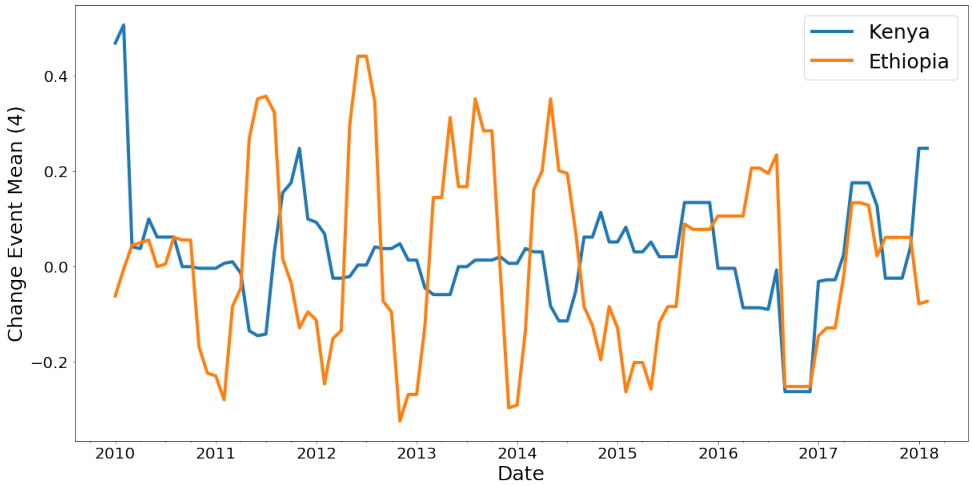


Figure 20: This image shows that the mean score for both Ethiopia and Kenya of the change events. This thus shows that there are more fluctuations for Ethiopia (on average) compared to Kenya.

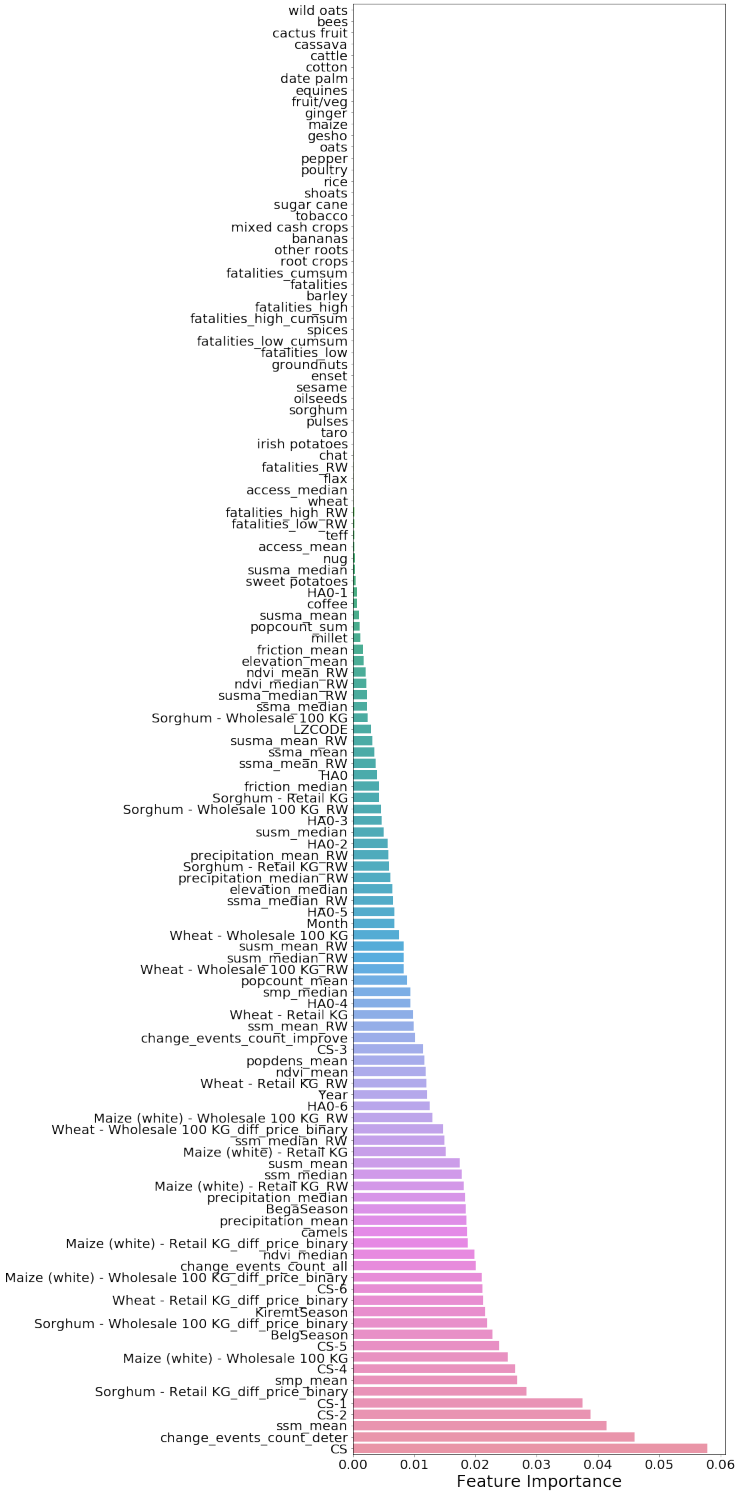


Figure 21: Feature Importances for all features.