A Data Revolution for the Developing World

Domain and Problem

We selected a problem in international development at a macro- and micro-level. Approaching the problem at different levels allowed our team to identify an appropriate entry point for the development of a data product that can be implemented at scale.

## *Macro-level*

In 2014, United Nations Secretary-General Ban Ki-moon commissioned an independent group of experts to assess the degree to which the data revolution applied to sustainable global development. The Independent Expert Advisory Group noted that globally, while the volume of data in developing countries has increased significantly, the quality of that data varied significantly; concurrently, the demand for data in the developing world (to inform programming, policy, and investments targeted to improving the quality of life of the poor) was--and is--also increasing.[[1]](#footnote-0)

According to a report by McKinsey & Company’s Global Institute, increased digitization leads to a boost in productivity-related growth. In the same report, the agriculture sector **in the U.S.A.** is ranked 23rd out of 23 sectors that is sufficiently digitized, measured by digital assets, usage, and workers.[[2]](#footnote-1) Expanding to a global view, approximately three fourths of the world’s extreme poor live in rural areas, and most depend on agriculture for their livelihoods and as a means to feed their families.[[3]](#footnote-2) To what extent has the data revolution fostered increased agricultural productivity, considered necessary for economic growth and poverty reduction?

Since the 1980s, the World Bank catalyzed efforts toward a data revolution in the developing world through the inception of the Living Standards Measurement Study (LSMS) project, “...a household survey program focused on generating high-quality data, improving survey methods, and building capacity.”[[4]](#footnote-3) Over 20 years since the inception of the LSMS project, the dearth of agricultural data, and the lack of rigorous and standardized methodologies for collecting agricultural data in developing countries,

continued to impede poverty reduction efforts. As a result, the LSMS team worked with eight countries in Sub-Saharan Africa to implement the LSMS’s Integrated Surveys on Agriculture with the objective of, “...foster[ing] innovation and efficiency in statistical research on the links between agriculture and poverty reduction in the region.”[[5]](#footnote-4)

Due in large part to the global effort to support governments in the developing world to collect high quality data, a shift to the analysis of the troves of data now generated from various sources is taking place: translating data to actionable evidence is needed.

Our team was concerned with applying data science tools to answer policy relevant question using household survey data. At a macro-level, this led us to our first problem: Which policy-relevant questions can be answered utilizing high quality yet complex survey data, so that the right information can be accessed by the right people at the right time? Using the LSMS-ISA data, we explored how to create a data product that can be easily accessed by researchers and policymakers in the agriculture and food security sector.

## *Micro-level*

It quickly became apparent that tackling the broader problem of data utilization and predicting policy-relevant questions was an ambitious goal. Thus, our team decided to select one policy-relevant question that would allow us to learn about and apply data science tools.

In rural areas where a majority of world’s poor reside, several myths about agriculture persist, due in large part to the historical dearth of high quality data. Some of the pervasive myths about agriculture in Africa are that access to credit to finance modern inputs is limited, and that labor productivity in agriculture is low. With access to better and more frequent data from the LSMS-ISA, the World Bank and its partners (including the United Nations, Yale University, and the London School of Economics, to name a few) initiated the “Agriculture in Africa - Telling Facts from Myths” project to discern facts from what was considered conventional wisdom.[[6]](#footnote-5)

Our team started to investigate one myth pervasive in the agriculture sector in the developing world to focus on: the “fact” that women in Africa provide 60-80 percent of the labor in Africa.[[7]](#footnote-6) Using LSMS-ISA data, the myth that women contribute a majority of the labor in agriculture has been dispelled by researchers; in fact, women provide an average of 40 percent of the agricultural labor hours in crop production, based on LSMS-ISA data from six African countries.[[8]](#footnote-7) Why is this important? Recent innovations in the collection of individual-level data that examine intrahousehold dynamics in the agriculture sector suggests that women’s empowerment is associated with a higher proportion of children receiving a minimum acceptable diet, and in general improves maternal health and child nutrition outcomes.[[9]](#footnote-8)

Building on this area of research, we wanted to examine the factors that affected the number of hours a women worked on agricultural activities, a question relevant to policymakers insofar as the number of hours a woman spends on agriculture can have variable effects on reducing poverty and malnutrition, also dependant on the country context. The LSMS-ISA data was utilized to build a predictive model to address this question.

# Hypothesis

The micro-level problem discussed in the previous section will guide the rest of this report, although we will briefly revisit our macro-level problem in the application section. We posit that a set of socioeconomic features, access to resources, and control over key agricultural-related decisions affects the number of hours a female works on agricultural activities.

# Data Ingestion

The [LSMS-ISA](http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:23512006~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html) datasets contain household-, individual-, and plot-level data. Our instance is a household, though we also included individual-level data because our analysis focused on women. Initially, we split up into two country teams--Nigeria and Ethiopia--to develop a predictive model for our primary question, and ground-truth our model utilizing a similar dataset for another country. We also aimed to conduct a more

robust analysis by merging the country datasets and creating a country feature. Due to time constraints and challenges associated with data wrangling, we ultimately decided to only work with data from Ethiopia.

The LSMS-ISA datasets are separated into data files that correspond to different sections of the survey. The data dictionary describes the content in the data file; the number of cases and variables; the structure of the data (type and keys); and notes that explain how to merge the datasets. Related materials include questionnaires, technical documents and reports that describe the survey process and key results of the survey. In order to retrieve data, we created an account and had to log in every time we needed to download the original source data. The LSMS-ISA data for Ethiopia contained 57 data files and 2973 variables across these data files. The size of the dataset presented one of the earliest challenges for our team.

# Data Wrangling

Our team employed several methods to correctly and accurately wrangle the LSMS-ISA data. First, using python through jupyter notebook, we selected features that we hypothesized would affect the number of hours a woman worked on agricultural activities. Second, we needed to merge multiple data files to create one dataset; this step was employed because relevant features were contained in different data files that corresponded to different modules of the LSMS-ISA survey.

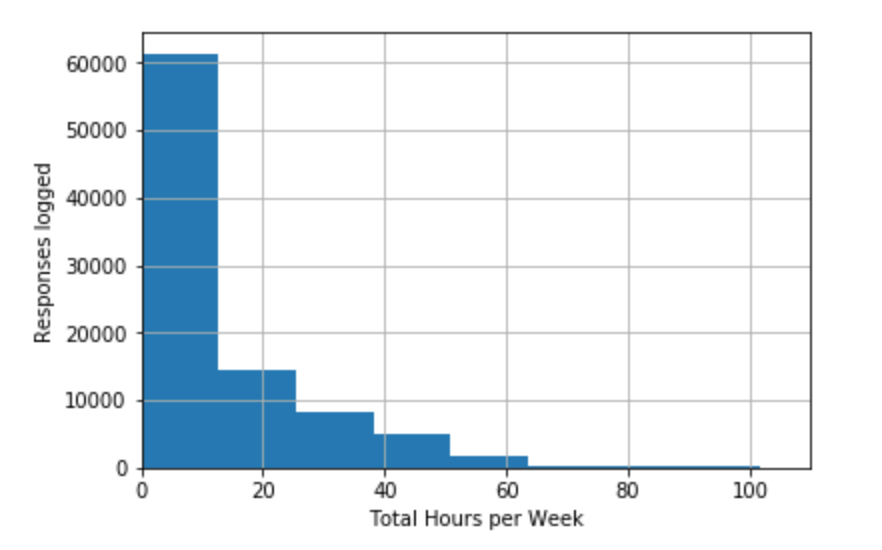
During the data merging stage we selected DB Browser for SQLite to store all of our raw data files, and to merge different tables with relevant features that might predict the number of hours a woman spends on agricultural activities. The volume of the data files posed a significant challenge when merging the different data files. Additionally, we had to rename all of the features in our dataset to names that were more intuitive. We have a large volume of data, over 100,000 instances, but also have a lot of missing data, including in our dependent variable. We narrowed down the features to create a final dataset but in preparing for modeling, the features had to be transformed again based on the model used. We also referred back to our original data files to perform analysis for some of our descriptive statistics.

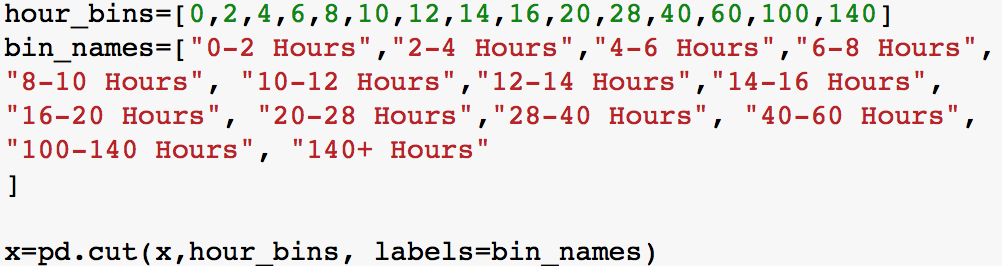
## Regression

Due to our data being a collection of surveys, there wasn’t a significant amount of numerical data involved in our dataset. Most of the fields in the dataset came pre-encoded with values of some categorical meaning. To get this data ready for regression we decided to create dummy matrices to represent the values within the survey questions. Also, the survey data contained a significant amount of null values from non-respondents. This lead us to drop a significant amount of variables that we thought might be useful for the analysis early on. We decided to keep some of the variables but we made the judgement call that if over half of the instances for a particular column as composed of null values that it was best to get rid of it. We imputed several variables that had over the threshold that we set. After getting through this we were left with 36 features and 91,000 instances. After creating the dummy variable matrices and concatenating them to the data frame we had closer to 100 features and 91,000 instances. We then ran Lasso, Ridge, and Elastic Net regularization on our data to determine which variables were significant. The answer was a resounding not much. We believe this to be due to the nature of the target variable that we were trying to predict. Over 50% of the instances of our target variable were zero values which in terms of regression shouldn’t generate much of a coefficient. We noticed that when we shifted the target variable we were able to get more predictive features so this confirmed our belief. However, due to the importance of answering this type of question we proceeded with our target variable. We believe that possibly an improvement in data gathering could help as we suspect not all of the zero instances may actually be zero instances. Also, we followed with a classification approach that we believed could assist with this.

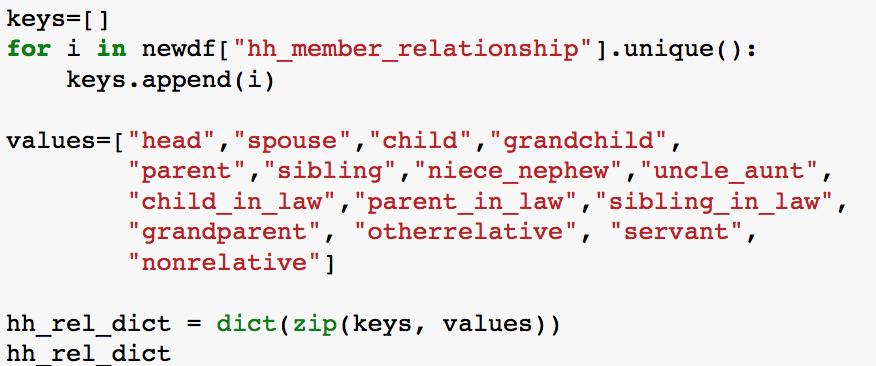
## Classification

In our regression approach we are attempting to create a model that can predict a value for the amount of hours worked per week for women in Ethiopia. We believed that an alternative approach of classifying which bin the hours value falls in could possibly help encompass the relationship between our output and input variables. In observing our hours distribution for the dataset, we saw that the data was heavily skewed towards the smaller hour outputs. Taking this into consideration, the bins created for this variable are smaller ranges and the bin ranges increase as we observe obvious decreases in the bin sizes.

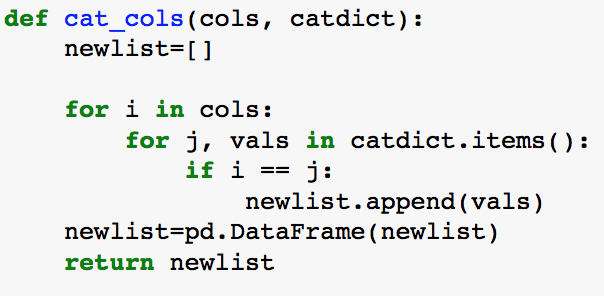


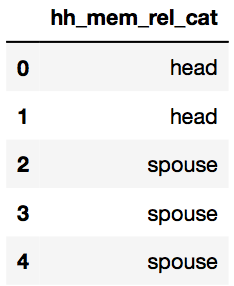
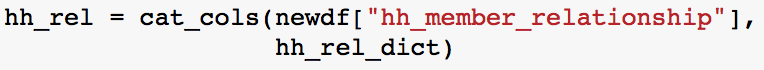


Though most of the data in the dataset was primarily categorical, most of the data had been converted to numerical outputs conveying a categorical meaning behind the number. Others did contain the string interpretation rather than being converted to a numerical meaning. To prepare this for the classification models, transformations were made to represent the numbers as the actual string meaning behind them. For example, the variable Household Member’s relationship to head of household was composed of instances that were given values one through 15 to represent their relationship with the head of household. Using these values as keys and creating a list of their string representations as values, a python dictionary was composed.



We wanted to use this dictionary and iterate through the column in the dataset to assign a value to it if it matched a key from our dictironary. Additionally, we wanted to create a dataframe object that we could later concatenate to our dataset. We composed the following approach:



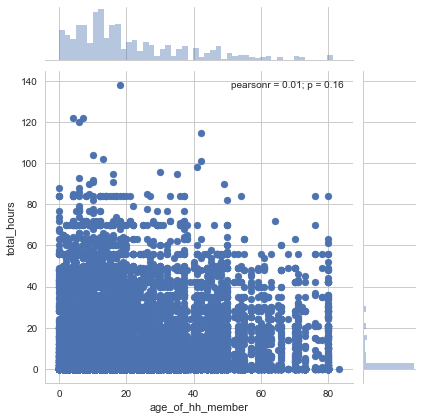


We followed the same conversion process for marital status, extension program, advisory program, regions codes, highest grade completed, types of planting, household sizes, household member ages, crop names, and who decides on the use of crops.

# Computation and analysis

We utilized visualization tools such as seaborn based on matplotlib to map out univariate and bivariate distributions of the relevant features (e.g., age\_of\_hh\_member, total\_hours, and highest\_education\_completed). We created histograms to examine the univariate distributions of the relevant variables. The results showed that the population comprised a large percentage of dependents such as children and the elderly. Eighty-four percent had at least high school education, approximately two percent had completed college, and one percent had some college. The rest was split between basic education and no education.

We also employed the jointplot function in seaborn to examine the bivariate distribution between total hours worked and age of the household member. The results indicated a Pearson correlation of 0.01 which identifies a significant relationship between total hours and age of the household member. However, the relationship is not linear which might also indicate that the Pearson correlation coefficient might not be the appropriate statistic measure of association.



## 

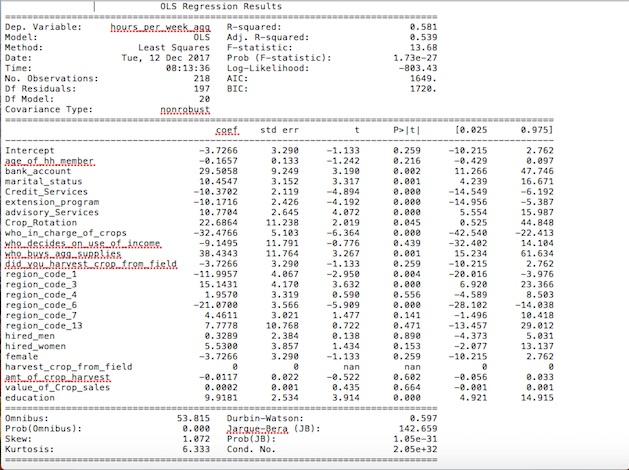
# Modeling

The full dataset contained too many features to include in our analysis. During the wrangling process, we selected a subset of features based on domain knowledge, though recognize that features contained in other data files may be relevant in predicting our target.

After receiving a complete LSMS dataset that was going to be fed into our models, further evaluation and manipulation of features was needed. It took time to determine which features were relevant in their raw form and which features needed to be converted in a manner that would be better predictor variables. We were trying to assess how particular features can predict the amount of hours females work on agriculture, and some features created noise and needed to be dropped (e.g., ID features). One of the models that we used was an OLS Regression. Many of the features were in binary and/or categorical form already, but many were not. Regression analysis treats all independent (X) features in the analysis as numerical, and numerical features are in interval or ratio scales and whose values are directly comparable (e.g. ‘10 is twice as much as 5’, or‘3 minus 1 equals 2’). Some of our features had multiple numerical categories, and to ensure that each feature had an intrinsic meaning to the model on their own, dummy variables were created to “trick” the regression algorithm into correctly analyzing feature attributes. We converted the following categorical features to new dummy features for the OLS regression model.

|  |
| --- |
| region\_code (Each region was created as a new binary feature) |
| marital\_status (1=Married 0= Other) |
| bank\_account (1=Bank Account 0= No bank account) |
| hired\_men(1=Men were hired 0=No Men Hired) |
| hired\_women ( 1=Men were hired 0=No Men Hired) |
| use\_fertilizer (1=Fertilzer was used 0= Not Used) |
| use\_manure\_fertilizer (1= Maure Fertilzer was used 0= Not Used) |
| chemical\_fertilizers\_used (1= Chemical Fertilzer was used 0= Not Used) |
| extension\_program (1= Extension Program was used 0= Not Used) |
| credit\_services (1= Credit Services were used 0= Not Used) |
| crop\_rotation (1= Crop Rotation was used 0= Not Used) |
| who\_in\_charge\_of\_crops (1=Female 0=Not Female) |
| who\_decides\_crops\_2\_sell (1=Female 0=Not Female) |
| who\_takes\_care\_of\_Crops (1=Female 0=Not Female) |
| who\_buys\_agg\_supplies (1=Female 0=Not Female) |
| female (1=Female 0=Not Female) |
| harvest\_crop\_from\_field (1=Yes 0=No) |
| did\_you\_sell\_harvest\_crop (1=Yes 0=No) |

Once the features were appropriately transformed for an OLS model, we used the StatsModelAPI OLS Regression model. We compared the features for weekly hours worked by females in agriculture to most of variables in our dataset. When the OLS Model was first ran our R-Squared value was approximately .08. After exploring the data further we had many observations (~ 17,000 instances) of zero hours of agricultural work by females. The zeros were dropped and the OLS model was re-run on the selected features; the R-squared value increased significantly. The model was then refined to only select the 20 features that had relevant or a statistically significant impact on weekly hours worked by females in agriculture. Below is the output of the refined OLS model that was run, with a reduction of only relevant and significant features.



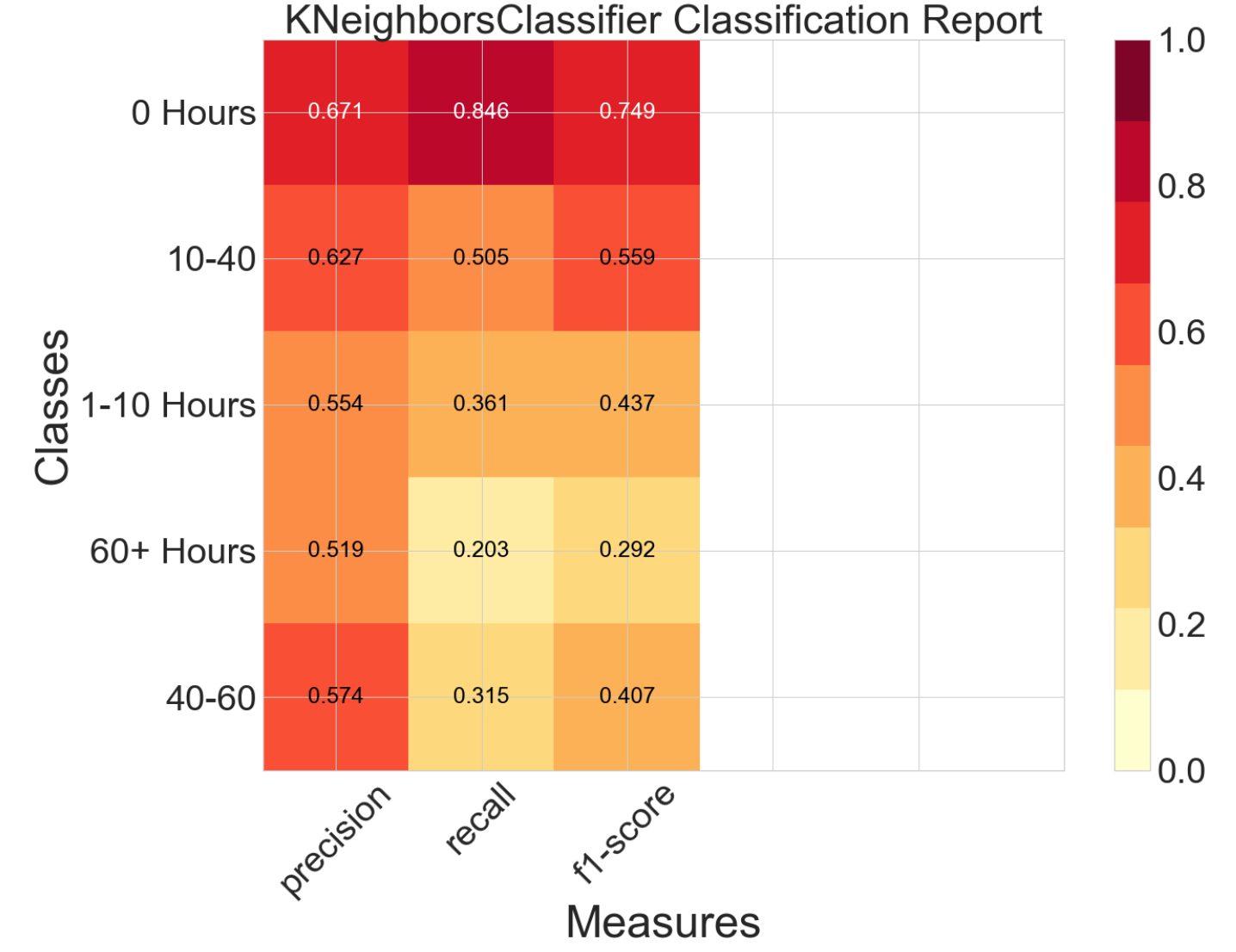
## *Results: Regression*

The results were interesting as to which features that would be better predictors from the OLS. The selected features gave us an R-Squared value of .58 which is pretty good. One of the features that was interesting was that if a female is the one who purchases agricultural supplies then she is likely to work approximately 38 hours more per week. This had a significant coefficient impact and was a statistically significant variable. Another variable that was found to be a good predictor is Region\_code\_6. In this region, women work fewer hours on agricultural activities. Is this because they are more efficient or are they more likely to be performing non-agricultural tasks? In general, we would have also liked to create dummy variables for particular crops based on their crop code. This was done in the Lasso and Ridge models but left out of the OLS to see how the features would perform without it. The selected features appeared to be good predictors of hours worked by females in agriculture, but there are certainly more sophisticated approaches to feature analysis and model selection that could be employed in the future.

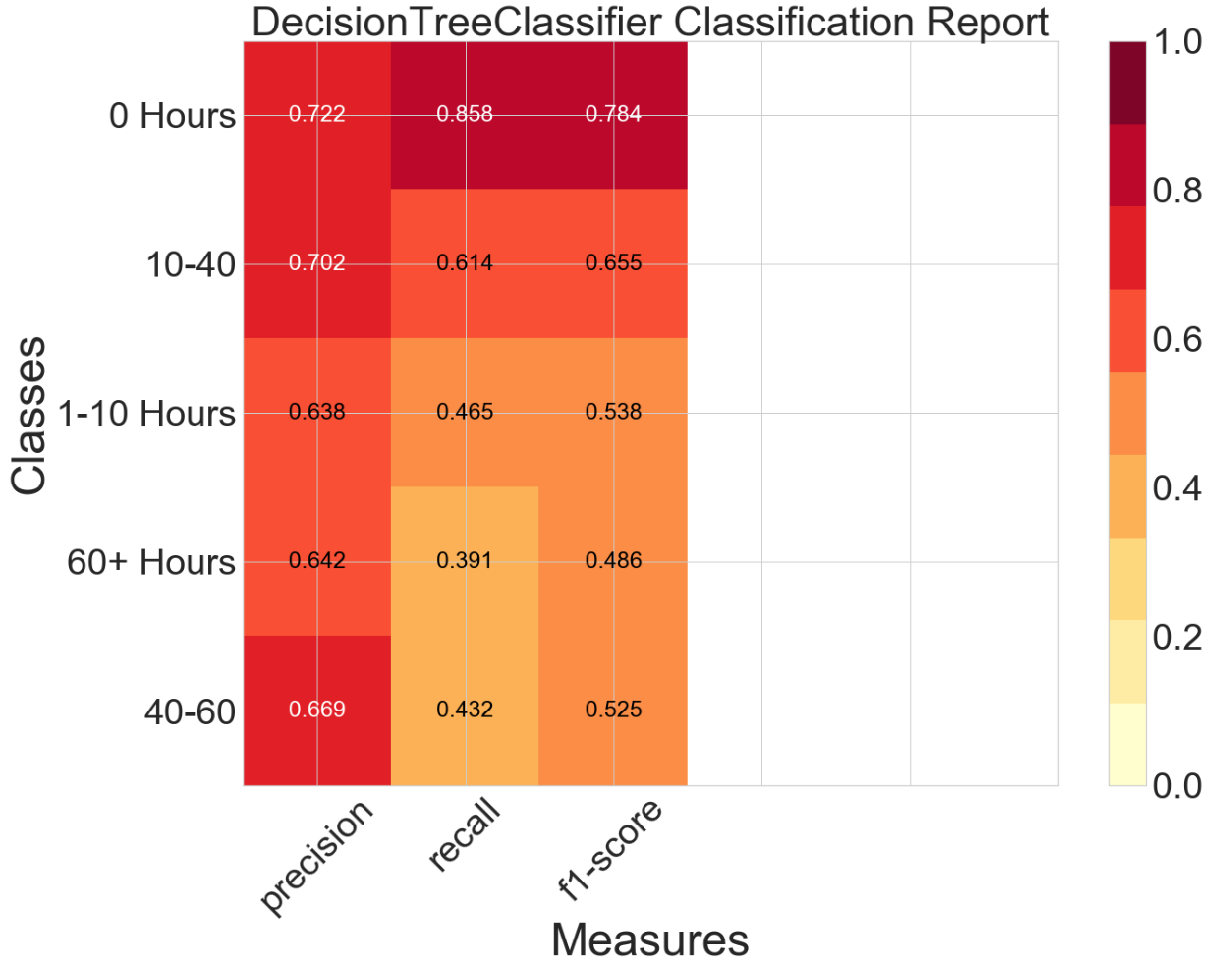
**Classification**

As mentioned earlier, a separate approach to modeling this problem that we decided to employ was a classification model. We suspected there may be no discernable difference between someone who worked 2 hours and someone who worked 4 hours. Additionally in classifying the amount of hours women worked, we were particularly interested in whether women did not have a place in participating in agriculture and harvest, whether they did participate, or if they were overworked or overstrained. So as mentioned in the wrangling section, variables were categorized as such. The models that we decided to use were K-nearest neighbors, Decision Tree Classifier, Bagging Classifier, Support Vector Machine classifier, and Random Forest Classifier. For the KNN model, we selected k = 6, implementing the number of classes plus one approach. Below are the results of the modeling efforts:

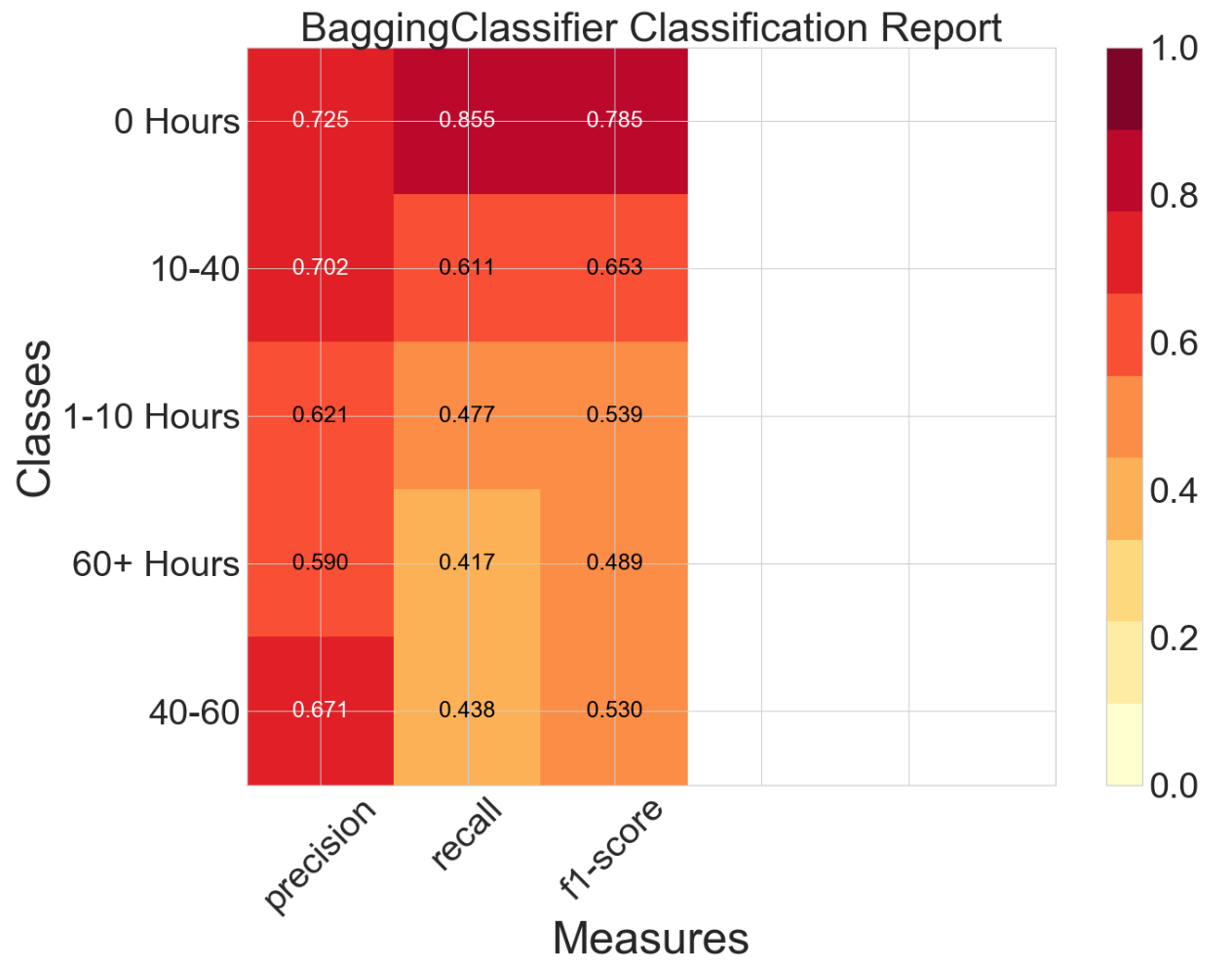
**K-Nearest Neighbors**

****

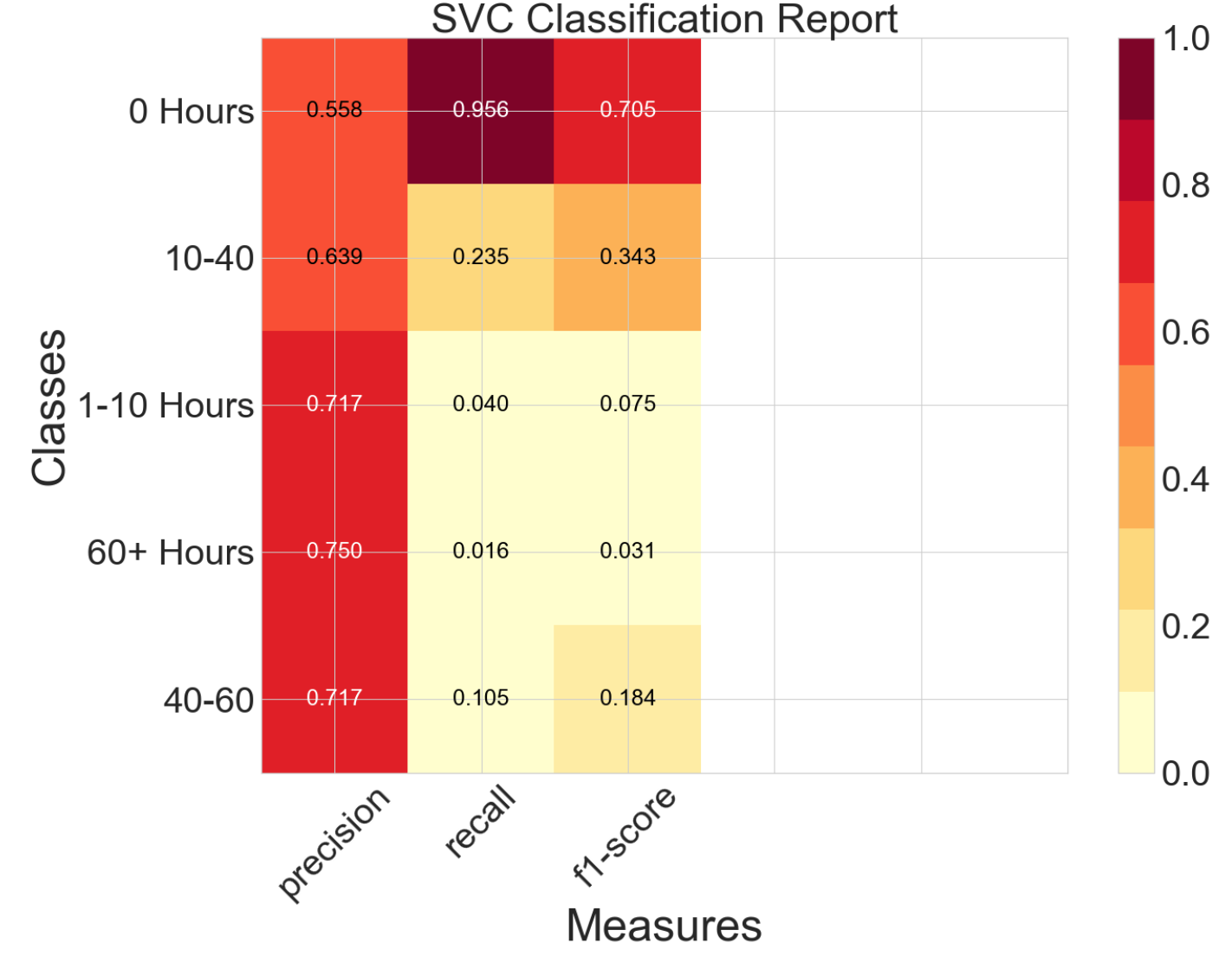
**Decision Tree Classifier**

****

**Bagging Classifier**

****

**Support Vector Machine**

****

**Random Forest Classifier**

# 

The results of the classification models ran are fairly consistent. Our KNN, Decision Tree, Random Forest, SVC - rbf, and Bagging Classifiers had r2 values of 0.6676, 0.7014, 0.7032, 0.5705, and xxxxx respectively. Due to zero values encompassing the majority of our predicted variable total hours, our model is able to handle predicting zeroes very well. However in the other ranges the performance can be a bit variable. Our Random Forest Model did the best a predicting the values in these other ranges. However, it did not perform admirably. We were able to select a few different bin sizes for the data and select the best one that we were able to generate. However, it also may be best to consult various algorithms for bin selection (outside of those that had a true domain impact such as the region bins) to maximize how well our data is able to predict the target variable.. Additionally, it may be best to encompass other data into the model, either from previous years or external data to the data source to train the model. Moreover, it is worth noting that we didn’t get to the stage to tune our hyperparameters to model the data better as we were still experimenting with the right bin sizes for the variable classification. Our data was not very separable when we visualized it which indicated to us that we still needed to find the right bins to classify the data into. One we had found the desirable bin sizes for our data it would have been more logical to tune the hyperparameters for our model. Lastly, building a pipeline to ingest the future years survey data would assist in providing the framework for our goal both at the micro and macro level.

# Feedback & Application

For the macro-level problem of predicting policy-relevant questions from the LSMS-ISA datasets, we would need to build predictive models for questions similar to the one included in this analysis. This would involve developing, training, and testing models for a variety of questions and understanding the questions that our users would want answer. The ideal framework that we have is that we could provide a search interface such as Google that returns results using Pagerank that may be of the most interest to the user.

**Lessons Learned**

As described above numerous times, dealing with survey data was a little tedious and unreliable. This is still a new initiative on LSMS’s part so it is reasonable to expect that data quality may improve as the survey standards continue to be standardized. Additionally, not that it is not possible to run regression models (it certainly is), but our efforts and time may have been better spent focusing on classification models due to there being a lot of similar responses to survey questions. We have also learned that it may have been better to narrow our scope earlier on than as gradually as we did. This may have also assisted in accomplishing a complete machine learning project with a data product.

# Future research

* To effectively deliver efficient and effective foreign assistance, information needs to be more accessible. collect a lot of data, but do little with it other than monitoring and reporting to stakeholders. LIttle in the way of understanding the underlying relationships (correlations? attribution??)
* Paul (State Dept.) described the USG as dinosaurs when it comes to harnessing innovations in the field of “big data” such as machine learning
* Companies like Cloudera want to get involved, but don’t know how
* Effort to strengthen national data systems

1. “A World That Counts: Mobilising the Data Revolution for Sustainable Development,” *United Nations Secretary-General’s Independent Expert Advisory Group*, November 2014, http://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf [↑](#footnote-ref-0)
2. “Digital America: A Tale of Haves and Have-Mores,” *McKinsey Global Institute, McKinsey & Company,* December 2015, accessed December 11th, 2017, https://www.mckinsey.com/industries/high-tech/our-insights/digital-america-a-tale-of-the-haves-and-have-mores [↑](#footnote-ref-1)
3. “Sustainable Development Goals,” Food and Agriculture Organization of the United Nations, accessed December 11, 2017, http://www.fao.org/sustainable-development-goals/goals/goal-1/en/ [↑](#footnote-ref-2)
4. <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:21610833~pagePK:64168427~piPK:64168435~theSitePK:3358997,00.html>. Date December 7th, 2017. [↑](#footnote-ref-3)
5. “Living Standards Measurement Surveys - Integrated Surveys on Agriculture,” World Bank, accessed December 14th, 2017, <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:23512006~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html> [↑](#footnote-ref-4)
6. “Agriculture in Africa - Telling Myths from Facts,” World Bank, accessed December 14th, 2017, <http://www.worldbank.org/en/programs/africa-myths-and-facts#1> [↑](#footnote-ref-5)
7. “Women, Agriculture and Work in Africa,” World Bank, accessed December 14th, 2017, <http://www.worldbank.org/en/programs/africa-myths-and-facts/publication/women-agriculture-and-work-in-africa> [↑](#footnote-ref-6)
8. Palacios-Lopez, A., Christiaensen, L., Kilic, T. (2015). How much of the labor in African agriculture is provided by women? (Policy Research Working Paper WPS7282) Retrieved from World Bank’s website: <http://documents.worldbank.org/curated/en/979671468189858347/How-much-of-the-labor-in-African-agriculture-is-provided-by-women> [↑](#footnote-ref-7)
9. Malapit, H. et al., “Measuring Progress Toward Empowerment: Women’s Empowerment in Agriculture Baseline Report.” (2014). Retrieved from Feed the Future’s website: <https://feedthefuture.gov/resource/measuring-progress-toward-empowerment-womens-empowerment-agriculture-baseline-report>

   [↑](#footnote-ref-8)