WorldCover

Data Scientist

MINI PROJECT

Please return your findings by **Monday, February 24, 2020** midnight (local time).

Background

This mini project is based on a real-life use case using a sample of WorldCover transactional database and customer surveys.

A particular business challenge for our Data Science team is to measure how well our weather-index insurance products perform for farmers. A well-performing index is a good predictor of (aka. is well correlated with) farmers' harvests season after season. In other words we expect insurance indemnity payouts to trigger when farmers experience low crop yields on their farm, and to not trigger when harvests are good.

To help us measure product performance (also referred to as ***basis risk***), WorldCover conducts phone surveys at the end of every cropping season, where we ask farmers to share their experience with recent harvests on their farm.

In this assignment your role as a Data Scientist is to:

1) Using data collected in Ghana at the end of Y2019 cropping season, calculate summary metrics that can inform our Research Team about **product performance**.

2) Present this information in a **synthetic and visual manner** (you might be asked to go over some of these findings orally during our final interview)

Some sample questions that are of particular interest:

1. Do crop yields (in kg/acre) vary significantly from region to region, from district to district, from farmer to farmer? What seems to be a normal, vs. low, vs. high yield for maize in Ghana?

2. Do other factors seem to influence crop yields, such as a farmer's gender, literacy, phone ownership, farm size, use of fertilizer, proximity to a larger town, etc.?

3. Out of the survey sample (`cryield` table), what is the proportion of farmers who reported bad/good crop yields and received (or did not receive) an insurance payout?

4. At the district level, can we say that districts with poorer (or better) harvests received higher (or lower) insurance payout amounts?

5. Can we trust the quality of our Y2019 sampled data?

Instructions

To investigate these questions we provide you with table dumps in CSV format and with a data schema that shows table fields and relationships. You are by no means limited to these datasets and you can use any external sources you think can help bring additional insights about our customers' experience.

You have 48-hr to complete this project. You can dedicate as much or as little time as you wish/need (but really we don't expect you to spend any longer than 2 to 4-hr). Reach out to melanie.bacou@worldcovr.com or Skype (mbacou) if you have any questions.

- We expect the main deliverable to be a notebook or a static report using any tool or platform of your choice. A static document, static presentation, or spreadsheet are all acceptable, as long as your document your process.

- Document and explain problems you encounter along the way and choices you make, what pitfalls, if any, you see with our current data. Also tell us what other questions around product design and customer experience you would like to investigate given more time and more data.

- Finally, provide any **queries/code files** that you used to generate all artifacts.

- Your answers to this project should live in a **publicly accessible Git repository**.

Don't worry! This is an open-ended exercise and is meant to test whether you can compile a few quantitative and visual insights under a tight deadline.

Good luck!

Definitions

`season`	Agricultural cropping season (each season lasts about 3 months). Some regions have 2 seasons in a year, a major or long rainy season from March to August and a minor or short rainy season from July/August to November/December
`premium`	Cost of drought insurance (amount paid by a customer over cash or USSD)
`payout`	Indemnity payment made by the insurer to the policyholder when drought events occur.
`USSD`	Unstructured Supplementary Service Data, also called "Quick Codes" or "Feature codes", is a text-based communications protocol used by GSM cellular phones to communicate with the mobile network operator's computers. USSD can be used for prepaid callback service, mobile-money services, location-based content services, and menu-based information services.

Data Schema

4 tables in CSV format are included:

`customers`

`cust_id`	Unique customer identifier
`date_reg`	The date when customer first signed up with WorldCover
`gender`	M (male) or F (female)
`has_mobile_money`	Customer has access to a mobile money account on his/her phone
`farm_size`	Customer farm size in acres
`literacy`	Customer literacy level
`ussd_created`	Customer was first registered through USSD (mobile phone)
`cht_season`	Customer belongs to this season's cohort
`cht_channel`	Customer was acquired through this channel
`cht_phone`	Customer has a mobile phone
`type`	Type of customer (simple customer or Contact Agent)
`amount_usd`	Total insurance premiums collected from customer since signup in USD

contracts

cntr_id	Unique contract identifier
cust_id	Unique customer identifier (foreign key)
loc_id	Unique location identifier (foreign key)
product_code	Crop (code) that is insured by this policy
season	Cropping season. In certain regions there are 2 cropping seasons in a year (major S1 & minor S2)
date_issued	Date when the customer first purchased this policy
date_planted	Date when the farmer planted his/her crop
date_planted_in	Date when farmer communicated his/her planting date
status	Current contract status[1]; one of `pending, planted, priced, active, triggered, expired, payout due, payout initiated, paid out, dispute, refunded`
amount	Premium amount in local currency (Cedis)
amount_usd	Premium amount in USD
payout	Indemnity payout in local currency (Cedis) if any
payout_usd	Indemnity payout in USD if any
mm_paid	Premium received over mobile money transfer

locations

loc_id	Unique location identifier (this is a village in Ghana)
loc_nm	Name of the location (this is the name of a village in Ghana)
date_reg	Date when location was created
iso3	3-letter country code
country	Country name
reg_nm	Top-level subdivision in the country (region)
dist_id	ID of the 2nd-level subdivision in the country (district)
dist_nm	2nd-level subdivision in the country (district)
cust_N	Number of customers at location
visit_N	Number of times WorldCover agents visited location
amount_usd	Total premiums collected at location since first signup in USD
X	GPS longitude in decimal degree
Y	GPS latitude in decimal degree

[1] Contracts that paid out have one of 3 status flags `payout due | payout initiated | paid out`

cryield

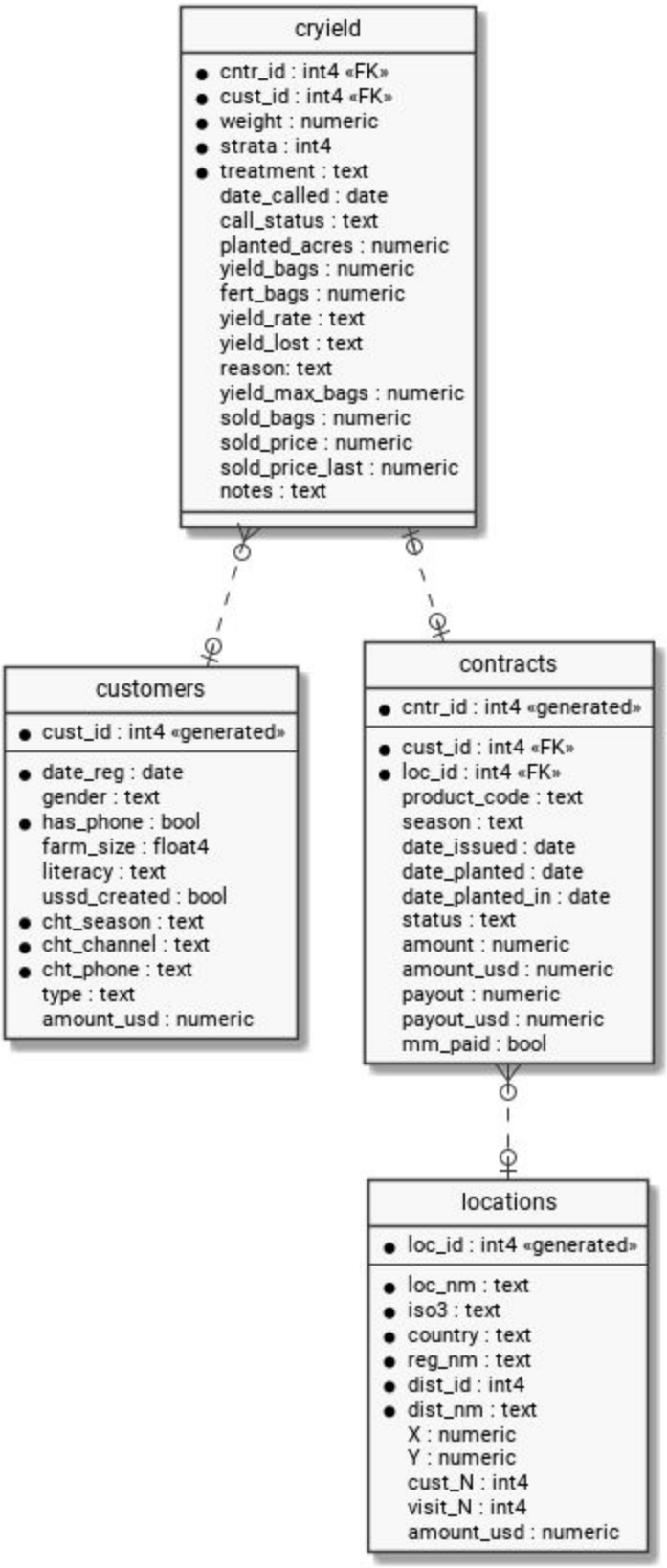
cust_id	Unique customer identifier (foreign key)
cntr_id	Unique contract identifier (foreign key)
weight	Sampling weight[2] (probability)
strata	Sampling strata (regions and treatment groups)
treatment	Customer treatment group grp_none | grp_paidout | grp_msg
date_called	Date of last call attempt (successful or not)
call_status	Did the call connect?
planted_acres	How many acres of that crop did you plant this season?
yield_bags	How many bags[3] of that crop did you harvest this season?
fert_bags	How many bags of fertilizer did you apply?
yield_rate	What is YOUR OWN assessment of this season's harvest? very poor | poor | average | good | very good
yield_lost	Approximately what fraction of crops was lost due to lack of chemicals or calamities? none | 1/4 | 1/3 | 1/2 | 2/3 | 3/4 | all
reason	What was the primary reason for yield losses this season? failed start[4] | drought before flowering[5] | drought after flowering | excess rain | no fertilizer | no labor/tractor | pest/disease | other
yield_max_bags	How many bags do you harvest of that crop in a very good year?
sold_bags	How many bags did you sell (or are you planning to sell) to market or to an offtaker?
sold_price	What unit price did you receive for 1 bag sold this season?
sold_price_last	What unit price did you receive for 1 bag sold in the prior season?
notes	Any extra notes (e.g. ask farmer if he/she intercrops his primary crop with other crops, details of pests/diseases)

[2] Sample of 328 customers out of 2,166 customers for Y2019 major and minor seasons in Ghana North and South. The sample is stratified by region and by treatment. In each region we oversampled customers who received a payout (grp_paidout) and who received planting advisories (grp_msg), hence the uneven sampling weights.

[3] bag of maize = 100kg, bag of groundnut = 82kg, bag of rice = 100kg, bag of sorghum = 100kg

[4] failed start: means the seed did not germinate and emerge after planting during days 1-10, typically because of insufficient soil moisture or unviable seeds or pests.

[5] flowering: silking stage for maize (about day 40-60 after planting)



WorldCover Data Schema (Ghana, extracts)