Hi Luke and Cole,

Sorry for not following up on this sooner; things have been quite hectic the last couple of days. But writing this email was on my to-do list, so thanks for getting the ball rolling.

Projects

Cole, as Luke has probably shared with you, when last we talked I suggested three projects you might work on this summer. I’m recapping them quickly below:

1. Build a predictive model of the likelihood that a borrower application gets completed (meaning the application is finished and all required documents are submitted);
2. Build a daily stability index for Jordan and other Middle Eastern markets using event data from GDELT
3. Conduct research to optimize our marketing efforts, by conducting A/B tests and/or mining our server logs.

Luke expressed interest in the first two. I agreed to share some info ahead of time to help you get started thinking about how to approach the problem, structure the work etc.

Judging from your email you seem to have narrowed in on Option 1), so I’ll focus on that here.

Technology

First a word on our data science stack. We have a reasonably complete scientific python setup on production, with scipy, numpy, pandas, scikit-learn and ipython. We can also add other libraries if you need them, especially if they’re available through pip.

Our business database is MariaDB/MySQL which we access on production code through SQLAlchemy (slightly modified to integrate nicely with pandas).

Our servers run Linux (Debian) and our developers work on Linux virtual machines (virtualbox) on their laptops (which are all Mac/OSX),. You could either get similar set-ups or work on your own machines, but if you want to work directly with production code we’ll need to set you up with the former.

I think for this particular project some knowledge of MySQL will be very very helpful, as I will explain below. If you are not familiar with SQL or MySQL from before it might be valuable to do some introductory reading here.

Data

So, the structure of the data: we have a so-called relational database where data is distributed over a number of different tables. When you arrive in Amman we’ll give you access to our VPN and you will be able to query this database directly using MySQL. As relates to the borrower application, the rough structure of the relevant tables is as follows:

Each borrower is associated with a row in a

(1) **user** table

which contains basic information about the individual. Users may have one or more businesses associated with them, stored in a

(2) **business** table.

Each business may have one ore more loan application associated with it, stored in a

(3) **funding campaigns** table.

Campaigns also have information about the purpose of the loan stored in a somewhat confusingly names

(4) **asset** table.

Each funding campaign also has an associated ticket which we store in

(5)**tickets**,

and each tickets has any number of ticket events associated with it, stored in

(6) **events,**

which include, for example, emails received, phone calls made etc.

So what we’re trying to predict here is the probability that a given user will have an associated ticket event signifying that his/her application was completed.

I’m attaching a list of the variables available in each of these tables. I hesitate to share more details about the structure, summary statistics etc in plain text over email for security reasons, as I hope you understand. Even the document I sent over is sensitive, so please don’t share it further or make it publicly available.

I’m happy to answer more specific questions about the data and the problem, of course.

Other resources

To get some subject area expertise on this issue, I highly recommend that you talk to our lovely Customer Service Representatives Doa’a and Sondos. They have personally dealt with thousands of borrower application and should be able to give you some insight into how our applicants think. Other members of the operations teams will have insights as well.

This can probably wait until you arrive in Amman, but if you find that it’s a priority I’ll set you up with Skype calls.

Finally, I also want to introduce you to Brian Marland, cc'd on this email, who is our highly talented Research Analyst. Brian is based in Amman, and you’ll find him a valuable resource for everything from MySQL to how our business processes work.

I hope this gives you enough to start planning and thinking about next steps. Again, I’m happy to answer more questions as they come up.

Very excited to work with you guys on this, and can’t wait to see with you come up with!

Best,  
David