Prize recipient documentation guide

Congratulations! You've gone up against dataheads from around the globe and emerged victorious! Laugh, dance, brush your shoulders off. You demonstrated serious skills, and helped make this world a better place in the process. Awesome job. Now you've finished in one of the top spots of the private leaderboard, which makes you eligible to receive a monetary prize. You're almost there.

In accordance with the official competition rules, the DrivenData terms of use, and applicable State and Federal law, we both have some due diligence to take care of before we can announce winners and disburse prizes.

There are three steps left in the process which all involve you sending us materials:

I. **Legal identity verification.**
You send us documentation so that we can verify your legal identity. We verify your eligibility to participate and then review the specific laws and rules about giving out prizes based upon your nationality and our tax reporting obligations.
Note: we are required by US Federal Law to withhold 30% of prize winnings for non-US individuals who do not already pay US taxes, unless exempted under an applicable income tax treaty.

II. **Code submission and result reproducibility.**
You package up and send us your code, documentation of dependencies, and any other assets you used. We review your package and make sure that it works and that we can fully reproduce the workflow from raw data to a submission comparable to your best submission. We recommend our open source data science template as an effective structure for sharing code.

III. **Model documentation and write-up.**
You write up answers to our questionnaire, providing important context and documentation so that the beneficiary and the community get the most out of your work.

Please read this document carefully. Each section details exactly what is needed from you—the faster we can check all the boxes for our mutual responsibilities, the faster we can disburse your prize!

Thanks for your hard work, and congratulations for making it this far.

Best,
The DrivenData Team

I. Legal identity verification

Note: we are collecting this information in accordance with our competition rules and privacy policy. Virtually every government jurisdiction mandates that prize-awards are reported to tax authorities. It is our duty to collect this information from you, but **we will never share your personal information** except as legally required or specifically authorized by you. Please see the official competition rules and our privacy policy for more details.

1. **Basic information.**
   Please provide us with the following information, numbered and in order.
   a. Full legal name:
   b. Date of birth:
   c. Citizenship:[1]
   d. Residential address (where you actually live and what you list on tax forms):
   e. Mailing address:[2]

If for security reasons you would rather relay the following identification number over the phone or by other means, please let us know so that we can make arrangements.

   f. ***US only:*** Social security number (SSN) or taxpayer identification number (TIN)

   g. ***If not a US taxpayer:*** your local equivalent to a taxpayer identification number[3]

   h. ***If not a US taxpayer:*** Do you agree to have relevant tax reporting documents sent as an email attachment (our preference), or would you prefer to receive a paper copy by postal mail to the address above? *(Note: for added security, if we send as an email we will attach in a password-protected zip folder and send you the password in a separate email.)*

   ___ Email (our preference) ___ Postal Mail

2. **Documentation.**
   In your response e-mail, please attach a color scan or photograph of a currently valid legal identification document. Examples of acceptable forms of identification include:
   - Driver's license
   - Legal identification card
   - Passport

[1] From what country do you have a valid, current passport? Or, if you do not hold a passport, to which country would you file a request with the reasonable expectation that a passport would be issued?
[2] This is where we will send a check.
[3] If you are not a US citizen, what is the equivalent identification number that you would use to file your taxes? *Please notify us immediately if the laws of your country or locality do not permit companies to request or collect this information; in that case, we will figure out what information we are required to report to relevant tax authorities.*

3. **Basic information for winner announcement.**
   Please provide your preferred information for use in announcing the winners of the competition.
   - Name (first and last name or first name and last initial):
   - Hometown:
   - A recent picture of yourself or digital avatar:

II. Code submission and result reproducibility

You will need to submit a compressed archive of your code. You don't need to include the raw data we provided, but everything else should be included and organized. If the files are too large to be e-mailed, a Google Drive or Dropbox share (or other comparable method of transferring data) works.

Note: *please follow these instructions carefully*. The spirit and purpose of the competition (and the reason for offering prizes) is to give our beneficiary organizations the best possible solution *along with working code they can actually use*. In accordance with the competition rules, if we can't get your code working and reproduce your results with a reasonable effort, or if your entry is too disorganized to be practically usable, then your entry may be disqualified!

The overall concept is to **set up this archive as if it were a finished open source project**, with clear instructions, dependencies and requirements identified, and code structured logically with an obvious point of entry. Again, we have a data science project template which may be helpful.

At a minimum, **this means the inclusion of an extremely clear README** that details all of the steps necessary to get to your submission from a fresh system with no dependencies (e.g. a brand new Linux, Mac OS X, or Windows installation depending on what environment you choose to develop under) and no other data aside from the raw data you downloaded from us.

This will probably entail the following:

- Necessary tools and requirements (e.g. "You must install Word2Vec 0.1c" or "Install the required Python packages in `requirements.txt`").
  - **All requirements should be clearly documented**, for instance in either a requirements.txt file with versions specified or environment.yml file.
- The series of commands, in order, that would get a reasonably experienced and savvy user from your code to a finished submission.
  - **Ideally, you will have a main point of entry to your code** such as an executable script that runs all steps of the pipeline in a deterministic fashion. A well-constructed IPython notebook or R script meets this standard.
  - **The next best thing is a list of specific, manual steps** outlining what to do. For example, "First, open Word2Vec and set these parameters. [...] Take the output file and run the script `src/make_preds.R` with the following parameters [...]". *(The limitations of this approach should be clear to any experienced data scientist!)*
- **Make sure to provide access to all trained model weights necessary to generate predictions from new data samples** without needing to retrain your model from scratch. Note that model weights can be contained in your archive or shared via a cloud storage

service. The solution should provide clear instructions to perform inference on a new data point, whether or not it is included in the test set.
- Any other instructions necessary to end up with your winning submission file (or comparable — we understand that certain parts of model fitting are stochastic and won't result in exactly the same parameters every time).

III. Model documentation and write-up

Information included in this section may be shared publicly with challenge results. You can respond to these questions in an e-mail or as an attached file. Please number your responses.

1. Who are you (mini-bio) and what do you do professionally?

   If you are on a team, please complete this block for each member of the team.

2. What motivated you to compete in this challenge?

3. High level summary of your approach: what did you do and why?

4. Do you have any useful charts, graphs, or visualizations from the process?

5. Copy and paste the 3 most impactful parts of your code and explain what each does and how it helped your model.

6. Please provide the machine specs and time you used to run your model.
   - CPU (model):
   - GPU (model or N/A):
   - Memory (GB):
   - OS:
   - Train duration:
   - Inference duration:

7. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?

8. Did you use any tools for data preparation or exploratory data analysis that aren't listed in your code submission?

9. How did you evaluate performance of the model other than the provided metric, if at all?

10. What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?

11. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?