Question for Candidate

Question:

We've collected a bunch of data from just-eat users with a survey. The surveys are then manually typed into a computer by our office in a foreign country. Those guys are not super reliable. We have internal data on these users and have already separated these into two classes:

- great_customers: those who are very profitable to our company,
- no_great_customers: and those who are not.

Data:

- The CSV file attached called final_dataset.csv.
- The first variable is a unique user identifier.
- The final variable great_customer_class is the "class" variable of whether the customer is a great customer or not.
- The other variables are features extracted about the user.

Instructions

- Use Python (preferably 3) please, that is what the team works in and we can't accept submissions written in other languages.
- There is some existing code (only a couple classes) that we would like you to use. Use the abstract classes provide for your models and data loader.
- Write a unit test for something. We don't really mind what, even if you just write a test for the checking the file exists or loading it, we'll be happy.
- Predict the binary class using two prediction models. Use Random Forests, then afterwards a second exciting approach of your choice. This could be something to show your knowledge about problem datasets like support vector machines, or could be something fancy like NN.
- Determine ways to evaluate your prediction model.
- Outline any concerns with the data and how you approached them.
- Explain the dominant features for classifiers if it's possible.

During the panel, we'll be going through this code with you and asking questions about your approach. We're looking for how you might handle this type of problem in a real life scenario, and demonstrate some maturity at software engineering.
Feel free to write me at aiml@just-eat.com with any questions. If something in the test is vague or not well defined, please feel free to use your best judgement.