Ten questions an audience would ask with answers included:

1) Did you have to do many data transformations?
    - Initially there was a limited amount of data transformations, just dummy coding the categorical variables with >2 categories. However, SMOTE was used to address imbalance of the dependent variable – this is a much larger transformation.
2) What other types of analyses would you consider exploring with these data?
    - Tree based models can be explored to see if better results are obtained, whichever method that yields the best result should be used on a dataset with multiple years of survey data.
3) Have you considered expanding your variable selection from the Behavioral Risk Factor Surveillance System (BRFSS) and addressing gaps in your analysis?
    - Yes, this would be ideal, there are many questions available on this survey and this analysis was very limited in that regard. Of particular interest would be having a time variable to understand the length of time or the onset of the chronic diseases.
4) For your analysis, you mentioned initially going a different route, why did you change course?
    - I changed route because when I ran my ROC curve, I got a .5 line which was unfortunate, but it did make me go back and rethink my approach.
5) Do you think your model can truly predict who will have a stroke?
    - No, I do not but I think it is a starting point, and if we don't have a starting point in research, then what would we do? I acknowledge that additional work is needed and more data.
6) How dirty was your data?
    - The data was very clean which was good for me, but I am not sure if that meant that the kaggle.com source cleaned the data or the folks who run the BRFSS. Usually, I would expect the data to be a little less clean.

7) Did you consider regrouping certain variables like age and making fewer categories?
    - I had this consideration and I think it would be an excellent aspect to include in future analyses.
8) From your ethical considerations, can you explain more about sampling techniques for these surveys?
    - The sampling technique for surveys like the BRFSS rely on the following,

      "With technical and methodological assistance from CDC, state health departments use in-house interviewers or contract with telephone call centers or universities to administer the BRFSS surveys continuously through the year. The states use a standardized core questionnaire, optional modules, and state-added questions. The survey is conducted using Random Digit Dialing (RDD) techniques on both landlines and cell phones."
        - It is always important to consider the limitations of data collection using these methods as well as the willingness to participate, the timing (meaning you got the call or noted the communication) are all factors in terms of subject participation.
9) What other demographics do you think could be beneficial?
    - It would be good to have a more detailed picture on the participants in terms of Race or Income or Zip Code (though there could be issues using zip code), urban or non-urban living, level of education, closest grocery store with fresh vegetables, closest fast-food restaurant, closest park – all these variables would be interesting to further explore.
10) Did you consider creating a surrogate for time based on different research information related to diabetes and hypertension?
    - I considered trying to create a surrogate based on age however, I thought it would be too error prone and decided not to do it. It would be interesting to create such a variable and see once you had more information on time to the event of stroke – how the two compared.