Machine Learning Interview Questions Part 2

1. What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a subset of machine learning that is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structured data.
In that sense, deep learning represents an unsupervised learning algorithm that learns representations of data through the use of neural nets.

2. Which is more important to you– model accuracy, or model performance?

There are models with higher accuracy that can perform worse in predictive power — how does that make sense?
Well, it has everything to do with how model accuracy is only a subset of model performance, and at that, a sometimes misleading one.
For example, if we wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were fraud.
However, this would be useless for a predictive model — a model designed to find fraud that asserted there was no fraud at all! Questions like this help we demonstrate that we understand model accuracy isn't the be-all and end-all of model performance.

3. What's the F1 score? How would we use it?

The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst.
We would use it in classification tests where true negatives don't matter much.

4. How would we handle an imbalanced dataset?

An imbalanced dataset is when we have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if we have no predictive power on the other category of data! Here are a few tactics to get over the hump:
1- Collect more data to even the imbalances in the dataset.
2- Resample the dataset to correct for imbalances.
3- Try a different algorithm altogether on your dataset.
What's important here is that we have a keen sense for what damage an unbalanced dataset can cause, and how to balance that.

5. When should we use classification over regression?

Classification produces discrete values and dataset to strict categories, while regression gives we continuous results that allow we to better distinguish differences between individual points.
We would use classification over regression if we wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If we wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)