Practice Interview Questions

## Machine Learning

### Fundamentals

* Can you make the distinction between an algorithm and a model? *Model is a function representing a data set, algorithms are a way to obtain that function*
* What’s the difference between supervised and unsupervised learning? *Labeled vs. unlabeled data*
* What’s the difference between a regression and a classification problem? How about clustering?
* Why do we use a train/test split? *Avoid overfitting and better generalization*
* What is cross-validation used for? What types of cross-validation do you know?
* What is generalization error?
* What is the bias-variance trade off?
* What is the difference between overfit and underfit?
* **Video:** What accuracy metrics do you know, both for classification and for regression? When would you use one metric vs the other?
* What is the curse of dimensionality?
* Why do you need to set the random seed prior to running certain ML algorithms?

### Regression

* Can you explain the difference between Linear and Logistic Regression?
* How are the coefficients in a Linear Regression interpreted?
* How is the intercept in a Linear Regression interpreted?
* Can the coefficients in a Logistic Regression be directly interpreted?
* What is the Adjusted R-Squared? What range of values can it take?
* Why is the Adjusted R-Squared a better measure than the regular R-Squared?
* How does Logistic Regression work “under the hood”? Can you explain Gradient Descent?

### SVM

* Can you explain how a Support Vector Machine works?
* What is the kernel trick?
* Where does the “support vector” term come in the SVM name?
* What kind of kernels exist for SVMs?

### Trees

* How do Decision Trees work?
* What criteria does a tree-based algorithm use to decide on a split?
* How does the Random Forest algorithm work? What are the sources of randomness?
* How is feature importance calculated by the Random Forest?

### Other Supervised Learning

* What is the difference between Bagging and Boosting?
* How do Gradient Boosted Machines work?
* What is Regularization, and what types do you know? *Avoid overfitting, L1 and L2 regularization. L1 used as dim reduction, L2 better for overall generalization, bonus points for ElasticNet*
* Can a Random Forest and a GBM be easily parallelized? Why/why not?

### Unsupervised Learning

* How does PCA work? What are the uses cases for it?
* How can you determine the optimal number of principal components?
* How does the K-Means algorithm work? What are its limitations?
* What other clustering algorithms do you know?
* How can you assess the quality of clustering?
* Can you explain in detail any other clustering algorithms besides K-Means?