Interview Questions: Modeling

Signal Data Science

- What is the bias vs. variance tradeoff?
- What is the difference between supervised and unsupervised learning? What is an example of each?
  - Supervised learning has a target variable, unsupervised doesn't. Examples: linear regression vs. K-means clustering.
- How would you explain a linear regression to an engineer with no statistics background?
- What are some ways I can make my model more robust to outliers?
  - Answer on Quora
  - Use a model resistant to outliers (tree models over regression based models), use a more resilient error metric (mean absolute difference instead of mean squared error), cap data at a certain threshold, transform the data, remove outliers manually.
  - Can also project all points onto unit sphere in parameter space.
- What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?
  - Answer on Quora
  - Minimizing squared error finds the mean whereas minimizing absolute error finds the median. The former is easier to compute and the latter is more resistant to outliers.
- What are the differences between the following pairs of terms: Type I error and Type II error, sensitivity and specificity, precision and recall?
- What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?
  - The standard cost function is the log loss. You can also use sensitivity / specificity for a binary classifier; due to class imbalance you can't just look at a single loss metric. If you have to, however, area under

the ROC is a good single metric to use. Finally, the multinomial log loss (= cross entropy) can be used to evaluate the performance of >2 group classification.

- What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What's the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)
  - Answer on Quora
  - Logistic regression, especially when regularized, is fairly robust and can be used as a baseline estimate. Its output can also be interpreted as probabilities, unlike some other algorithms. You can also add interaction terms to model nonlinearity.
  - SVMs can take quite a long time to train, even if they work well in practice. Also, the standard formulation of SVMs is only applicable to binary classification (although one can use multiple SVM models for multinomial classification).
  - Ensemble tree-based models are a good "standard" nonlinear technique to apply (random forests and gradient boosted trees) which usually don't overfit too much; they handle high dimensionality, big datasets, and multinomial classification well. Gradient boosted trees perform marginally better than RFs but need more tuning and can be a little more prone to overfitting as a result.
  - Naive Bayes is easy to compute and used for cases where you have many binary features. It makes an assumption of independence – feature A being present provides no information about feature B being present and vice versa.
- What is overfitting and how would you detect if your model suffers from overfitting?
- What is regularization and where might it be helpful? What is an example of using regularization in a model?
  - In the most general sense, regularization is the addition of a term to a model's cost function which measures model complexity. It helps prevent overfitting (learning the background noise instead of the generalizable patterns); it increases model bias but decreases model variance.
  - In the context of linear regression, regularization refers to adding on a norm (usually L^1, L^2, or a mix) of a linear model's coefficients to the cost function.
- What's a hyperparameter? What are the hyperparamters for regularized linear regression?

  - A hyperparameter is a parameter of the model itself. Regularized linear regression has λ (penalization strength). (*Elastic net* regularization also has α, the degree of mixing between L^1 and L^2 norms.)

- Why might it be preferable to include fewer predictors over many? How do you select which predictors to use?
  - With too many predictors, the model will overfit and perform worse on test data. You can use regularization for feature selection. Alternatively, you can regress against a subset of the principal components or look at the correlation matrix for the features and remove those highly correlated with others.

- How can you determine which features are the most important in your model?
  - Look at magnitude of regression coefficients or the variable importance measure of a random forest model. Look also at which ones are the most highly correlated with the response variable.

- You have some variables which are all positively correlated with your target variable. In the linear regression, one of them has a negative coefficient. Does this make sense in light of the positive correlations? What is the interpretation of this result?
  - When accounting for the other variables, the negative coefficient variable has a marginal negative effect.

- You run your regression on different subsets of your data, and find that in each subset, the coefficient for a certain variable varies wildly. What could be the issue here?

- If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse? Why?
  - You can fit a (cross-validated, properly tuned) regularized linear model with the models' predictions to the response variable. This will perform no worse than the best of the individual models and quite possibly better. It's known as stacking; different models have different strengths and deficiencies, so they compensate for one another.

- What's the difference between a decision tree and a decision forest?
  - A decision forest is an ensemble of many different decision trees, averaged together. In a random forest specifically, at each branch of each tree only a random subset of the predictors is used.

- How would you give different weights to points in a linear regression?
  - In the cost function to be minimized, which is the sum of squared errors, multiply each squared error by the corresponding weight.