1. What is the process of analyzing a dataset and making predictions?

- Data cleaning and pre-processing: check for skewness (box-cox transformation), check for number of categories, missing data, outliers, interactive effects, check the base rate (if distribution is significantly unbalanced – what to do?), etc.
- Check the relationship between predictors and outcome: using scatter plot matrix. Also, to check for nonlinear relationship, could plot: (1) observed – predicted plot; (2) predicted – residual plot (some curvature).
- Check multicollinearity (between-predictor correlations): (1) Check correlation matrix. To diagnose multicollinearity in the context of linear regression, the variance inflation factor can be used (Myers, 1994), which is computed for each predictor and a function of the correlation between the selected predictor and all of the other predictors. Remove highly correlated predictors. (2) Perform PCA (check scree plot and variance explained by each component). But challenge with this it PCA does not consider any aspects of the response when selecting components. (3) PLS (Jing's note: idea similar to relative weights): PCA (unsupervised), linear combinations are chosen to maximally summarize predictor space variability; the PLS (supervised), linear combinations of predictors are chosen to maximally summarize covariance with the response (a compromise between the objectives of predictor space dimension reduction and a predictive relationship with the response).

  Other remedy may include PLS, or ridge, lasso, or elastic net.
- May scale (standardize) the dataset to put them on the same scale
- Feature selection is needed
- Choose the right model
- Training set: used to tune the parameters, estimate the models, and determine initial estimates of performance using repeated 10-fold cross-validation.
- Testing set: used for a final characterization of the models of interest.
- Use validation data (or cross-validation) for choosing the tuning parameters
- Avoid the temptation for over-fitting
- Apply on real data
- Plot and visualize the data
- Check Chapter3: Data Pre-processing.

2. What are some ways to make my model more robust to outliers?

- Use a model that's resistant to outliers. Tree-based models are generally not as affected by outliers, while regression-based models are. If you're performing a statistical test, try a non-parametric test instead of a parametric one.
- Use a more robust error metric. As Peter Mills mentions in his excellent answer, switching from mean squared error to mean absolute difference (or something like Huber Loss) reduces the influence of outliers. The mean is the default measure of central tendency, but one main problem is that it can be overly influenced by outliers. This is why distributions like household income or average house value is usually summarized by the median rather than the mean.
- Here are some changes you can make to your data:
  - Winsorize your data. Artificially cap your data at some threshold. SeeWhen are some applications of winsorization?

- Transform your data. If your data has a very pronounced right tail, try a log transformation.
- Remove the outliers. This works if there are very few of them and you're fairly certain they're anomalies and not worth predicting.

3. In which cases is the mean square error a bad measure of the model performance?

- MSE will be a bad measure is those cases where linear regression's (parametric) assumptions are violated, like skewed distribution, nonlinear relationship, outliers, etc.
- MSE is good for decomposing sums of squares into meaningful components like "between group variance" and "within-group variance."
- MSE is a derivative measure, which is easier to solve for finding the minimum: Often we want to minimize our error. When the error is a sum of squares, we are minimizing something quadratic. This is easily accomplished by solving linear equations.

4. What error metric to use for evaluating how good a binary classifier is?

- Classification Accuracy (misclassification error):
  - Weaknesses: (1) Not cost-sensitive – one of several ways to see the problems with proportion classified correctly is that if the overall proportion in one category is 0.9 you will be correct 0.9 of the time by ignoring the data and classifying every observation as being in that category. Works poorly when the signal in the data is weak compared to the signal from the class imbalance. (2) Also, you cannot express your uncertainty about a certain prediction.
- Area under the curve (AUC): Definition (direct) - The area under the ROC curve.
  - ROC: In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. (The true-positive rate is also known as sensitivity in biomedical informatics, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as 1 - specificity). The ROC curve is thus the sensitivity as a function of fall-out.
  - Definition: A reliable and valid AUC estimate can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example (binary outcome). (Intuitive: given a random positive instance and a random negative instance, the probability that you can distinguish between them.)
  - Strengths - Works well when you want to be able to test your ability to distinguish the two classes.
  - Weaknesses: (1) In general, it is not a good idea to compress such a curve into one number. One recent explanation of the problem with ROC AUC is that reducing the ROC Curve to a single number ignores the fact that it is about the tradeoffs between the different systems or performance points plotted and not the performance of an individual system. (2) You may not be able to interpret your predictions as probabilities if you use AUC, since AUC only cares about the rankings of your prediction scores and not their actual value. Thus you may not be able to express your uncertainty about a prediction, or even the probability that an item is successful.
- Confusion Matrix (sensitivity and specificity): Sensitivity (also called the true positive rate, or the recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition),

and is complementary to the false negative rate. TPR = TP/P = TP/ (TP+FN). Specificity (sometimes called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. SPC = TN/N = TN / (FP + TN).

5. What error metrics to use for multiclass (or multinomial) classifications?

- Misclassification error.
- Confusion Matrix. For N classes you have to consider either a single N×N table or N 2×2 tables each of them comparing one of the classes (A) against all other classes (not A).
- AUC: As in several multi-class problem, the idea is generally to carry out pairwise comparison (one class vs. all other classes, one class vs. another class. Product ROC Curves using One-Vs-All Approach

6. How to deal with unbalanced prior distribution (base rate)?

- Prior probability, cost estimate.
- When the classes are unbalanced, the baseline is not 50% but the proportion of the bigger class. You could add a weight on each class to balance the error. Let Wy be the weight of the class y. Set the weights such that $\frac{1}{W_y} \sim \frac{1}{n} \sum_{i \leq n} 1_{y_i = y}$ and define the weighted empirical error.

7. Linear Regression:

- Ordinary Linear Regression: minimizes the sum-of-squared errors (SSE) between the observed and predicted response.
- Partial Least Squares (PLS): linear combinations of predictors are chosen to maximally summarize covariance with the response (a compromise between the objectives of predictor space dimension reduction and a predictive relationship with the response). One tuning parameter: number of components to retain, use cross-validation to tune. (Similar to relative weights, better for handling multicollinearity problem).
    - Weakness: still assume linear relationship.
- Penalized Models: such as Ridge Regression, the LASSO, and the Elastic Net.
- Similarity: decompose (expected) MSE into components of model variance, model bias (close to true relationship), and irreducible variation.
    - Model Variance: refers to the amount by which f would change if we estimated it using a different training data set. In general, more flexible statistical methods have higher variance.
    - Model Bias: refers to the error that is introduced by approximating a real-life problem. Generally, more flexible methods result in less bias.
    - So there is bias-variance trade-off (the danger of over-fitting): if you want to minimize the bias and mimic the pattern in the data as closely as possible, your model will have high variance – any small change in data will significantly impact the model fit. [Complex model -> high variance -> over-fitting.] May usually results in a U-shape where MSE first decrease because of increase in flexibility, but then increase because of large increase in variance.

- Difference: along the spectrum of the bias-variance trade-off. OLR, at one extreme, finds parameter estimates that have minimum bias, whereas ridge regression, the lasso, and the elastic net find estimates that have lower variance.
- Advantage of linear regression model: (1) highly interpretable; (2) their mathematical nature enables us to compute standard error of the coefficients. These standard errors can then be used to assess the statistical significance of each predictor in the model.
- Limitation of linear regression model: (1) the relationship would need to fall close to a flat hyperplane (linear). (2) It is prone to chasing observations that are away from the overall trend of the majority of the data to minimize SSE – sensitive to large outliers (limitation of parametric approach). Remedy for this – use absolute errors when residuals are above a threshold (Huber function).

Variations on ordinary least regression:

- Poisson regression: used to model count data.
- Why do we want to use some regularized least-squares regression methods:
  - Increase prediction accuracy: by constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias, which will result in smaller MSE.
  - Model interpretability: less coefficients, easier to interpret a model.
- Regularized least-squares regression (LASSO): more restrictive (set some coefficients to zero), more inflexible, but more interpretable (because of smaller number of coefficients). We will often obtain more accurate predictions using a less flexible method.
  - Helps when you have too many predictors by favoring weights of zero. The Lasso is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients (values s: a tuning parameter – smaller s has a larger shrinkage).
- Ridge regression: can help with reducing the variance of your weights and predictions by shrinking the weights to 0.

How to reduce model variance in linear regression:

- Subset Selection: best subset of p, forward/backward stepwise selection, choose the optimal model (using cross-validation).
- Shrinkage (or Regularization): Ridge, Lasso.
  - Ridge regression: instead of minimizing RSS, ridge regression minimizes $\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$

    This added shrinkage penalty has the effect of shrinking the estimates of b towards zero, and lambda is a tuning parameter that control the impact of the penalty.).
    - Note: it is best to apply standardization before ridge regression.
    - Limitations: will reduce the magnitudes of the coefficients but will not result in exclusion of any of the variables (i.e., will not set any of the predictors as 0).

- LASSO: minimizes $$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

  The penalty will set some parameters to be 0 when lambda is sufficiently large, so LASSO performs variable selection.

- Neither ridge nor lasso will universally dominate the other, depending on whether predictors are all related to outcome, one might be better than another. So it is important to use cross-validation to determine which approach is better.

- Dimension Reduction: PCA (principle components regression), PLS (partial least square).
  - Step 1: obtain the M transformed predictors.
  - Step 2: fit eh model using these M predictors.
  - Note: PCA is not a feature selection method! PCR is very closely related to ridge regression.
  - M: the number of components is a tuning parameter that could be chosen from cross-validation.

8. Metrics for evaluating linear regression:

- RMSE: interpreted as either how far on average the residuals are from zero or as the average distance between the observed values and the model predictions.
- MSE: mean squared error is sometimes used to refer to the unbiased estimate of error variance.

  MSE = RSS (residual sum of squares) / N-P (the number of degrees of freedom) or N??
- R^2 (coefficient of determination): proportion of the information in the data that is explained by the model. Must remember – this is a measure of <u>correlation, not accuracy</u>.

  The most general definition of the coefficient of determination is

  $$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

  - Problem: R^2 is dependent on the variation in the outcome.

9. Parametric vs. Non-Parametric Method:

- Parametric Method:
  - 2 steps – (1) make an assumption about the functional form, or shape of *f* (the model); (2) use training data and some method to find the estimates of the (p+1) parameters. This model-based approach reduces the problem of estimating an arbitrary p-dimensional function *f* to one of estimating a set of parameters.
  - Weakness: (1) the model we choose will usually not match the true unknown form of *f*. (2) But if we want to fit a more flexible model, then it requires estimating more parameters, which will lead to over-fitting (chasing the errors/noise too closely).
  - Advantage: more interpretable.

- Non-parametric Method:
    - Do not make explicit assumptions about the functional form of *f*. Instead they seek an estimate of *f* that gets as close to the data points as possible without being too rough or wiggly. So the fitting can be accurate for a wide range of possible shapes of *f*.
    - Weakness: as they don't reduce the problem to estimating a small number of parameters, a large number of observations is required to obtain an accurate estimate for *f*.

10. When p is very large, even p > n, forward stepwise selection is a viable subset method while background stepwise would not work.

11. What metrics to use when choosing the best model among a collection of models?

- Couldn't use RSS or R^2: because these quantities are related to the training error.
- Metric 1: make an adjustment to the training error to account for over-fitting.
    - $C_{p,}$ AIC, BIC, Adjusted R^2 (which adjust for number of predictors).
    - These metrics are not appropriate in high-dimensional setting.
- Metric 2: test data set or cross-validation (preferred).
    - 1-SE rule: first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one SE of the lowest point on the curse. The rational is that if a set of models appear to be more or less equally good, then just choose the simplest model with the smallest number of predictors.

12. What is the curse of dimensionality?

- The test error tends to increase as the dimensionality of the problem (i.e., the number of features or predictors) increases, unless the additional features are truly associated with the response.
- High dimensionality makes clustering hard, because having lots of dimensions means that everything is "far away" from each other. It's hard to know what true distance means when you have so many dimensions. That's why it's often helpful to perform PCA to reduce dimensionality before clustering.
- High dimensionality is also a curse when one is trying to do rejection sampling. With a higher dimension probability distribution, it becomes increasingly harder to find an appropriate enveloping distribution since the acceptance probability will keep shrinking with dimensionality.

13. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

- This is called dataset shift, which could lead to inaccuracy of the model.
- As pointed out by Justin Rising, this is a problem of dataset shift. I looked online and find this slide, which summarizes some reasons as follows
    - Covariate shift: training and test input follow different distributions, but functional relation remains unchanged.

- Sample selection bias: the training examples have been obtained through a biased method, such as non-uniform selection.
- Non-stationary environments: Training environment is different from the test one, whether it's due to a temporal or a spatial change. One typical scenario is adversarial classification problems, such as spam filtering and network intrusion detection.

14. How to choose among different classification algorithms:

- How large is your training set? If your training set is small, high bias/low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., kNN or logistic regression), since the latter will overfit. But low bias/high variance classifiers start to win out as your training set grows (they have lower asymptotic error), since high bias classifiers aren't powerful enough to provide accurate models.
- You can also think of this as a generative model vs. discriminative model distinction.
- Advantages of some particular algorithms:
  - Advantages of Naive Bayes: Super simple, you're just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. And even if the NB assumption doesn't hold, a NB classifier still often performs surprisingly well in practice. A good bet if you want to do some kind of semi-supervised learning, or want something embarrassingly simple that performs pretty well. NB is good for text data.
  - Advantages of Logistic Regression: Lots of ways to regularize your model, and you don't have to worry as much about your features being correlated, like you do in Naive Bayes. You also have a nice probabilistic interpretation, unlike decision trees or SVMs, and you can easily update your model to take in new data (using an online gradient descent method), again unlike decision trees or SVMs. Use it if you want a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're unsure, or to get confidence intervals) or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.
  - Advantages of Decision Trees: Easy to interpret and explain (for some people -- I'm not sure I fall into this camp). Non-parametric, so you don't have to worry about outliers or whether the data is linearly separable (e.g., decision trees easily take care of cases where you have class A at the low end of some feature x, class B in the mid-range of feature x, and A again at the high end). Their main disadvantage is that they easily over-fit, but that's where ensemble methods like random forests (or boosted trees) come in. Plus, random forests are often the winner for lots of problems in classification (usually slightly ahead of SVMs, I believe), they're fast and scalable, and you don't have to worry about tuning a bunch of parameters like you do with SVMs, so they seem to be quite popular these days.
    - Advantages of SVMs: High accuracy, nice theoretical guarantees regarding over-fitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm. Good for large training dataset. Memory-intensive and kind of annoying to run and tune, though, so I think random forests are starting to steal the crown.
- To go back to the particular question of logistic regression vs. decision trees (which I'll assume to be a question of logistic regression vs. random forests) and summarize a bit: both are fast and scalable, random forests tend to beat out logistic regression in terms of accuracy, but logistic regression can be updated online and gives you useful probabilities. Also, having probabilities

associated to each classification might be useful if you want to quickly adjust thresholds to change false positive/false negative rates, and regardless of the algorithm you choose, if your classes are heavily imbalanced (as often happens with fraud), you should probably resample the classes or adjust your error metrics to make the classes more equal.

- But, recall, though, that better data often beats better algorithms, and designing good features goes a long way. And if you have a huge dataset, your choice of classification algorithm might not really matter so much in terms of classification performance (so choose your algorithm based on speed or ease of use instead).
- And if you really care about accuracy, you should definitely try a bunch of different classifiers and select the best one by cross-validation. Or, to take a lesson from the Netflix Prize and Middle Earth, just use an ensemble method to choose them all!

Another answer:

- Logistic regression: no distribution requirement, perform well with few categories categorical variables, compute the logistic distribution, good for few categories variables, easy to interpret, compute CI, suffer multicollinearity.
- Decision Trees: no distribution requirement, heuristic, good for few categories variables, not suffer multicollinearity (by choosing one of them).
- NB: generally no requirements, easy to understand, good for few categories variables, compute the multiplication of independent distributions, suffer multicollinearity.
- LDA (Linear discriminant analysis not latent Dirichlet allocation): require normal, not good for few categories variables, compute the addition of Multivariate distribution, compute CI, suffer multicollinearity.
- SVM: no distribution requirement, compute hinge loss, flexible selection of kernels for nonlinear correlation, not suffer multicollinearity, hard to interpret.
- Lasso: no distribution requirement, compute L1 loss, variable selection, suffer multicollinearity.
- Ridge: no distribution requirement, compute L2 loss, no variable selection, not suffer multicollinearity.
- Bagging, boosting, ensemble methods (RF, Ada, etc): generally outperform single algorithm listed above.

15. What is p-value?

- The P value or calculated probability is the estimated probability of rejecting the null hypothesis (H0) of a study question when that hypothesis is true.
- For dummy: When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- In statistics, the p-value is a function of the observed sample results (a statistic) that is used for testing a statistical hypothesis. Before performing the test a threshold value is chosen, called the significance level of the test, traditionally 5% or 1% and denoted as α. If the p-value is equal to or smaller than the significance level (α), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the alternative hypothesis is accepted as true. When the p-value is calculated correctly, such a test is guaranteed to control the Type I error rate to be no greater than α.

- The p-value is calculated as the lowest α for which we can still reject the null hypothesis for a given set of observations. An equivalent interpretation is that p-value is the probability of obtaining the observed sample results, or a "more extreme" result, when assuming the null hypothesis is actually true (where "more extreme" is dependent on the way the hypothesis is tested). Since p-value is used in Frequentist inference (and not Bayesian inference), it does not in itself support reasoning about the probabilities of hypotheses, but only as a tool for deciding if to move from the null hypothesis to the alternative hypothesis.

16. What is maximum likelihood estimation?

- Maximum likelihood estimation begins with writing a mathematical expression known as the Likelihood Function of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimates or MLE's.
- In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function.
- Intuitively, this maximizes the "agreement" of the selected model with the observed data, and for discrete random variables it indeed maximizes the probability of the observed data under the resulting distribution. Maximum-likelihood estimation gives a unified approach to estimation, which is well-defined in the case of the normal distribution and many other problems. However, in some complicated problems, difficulties do occur: in such problems, maximum-likelihood estimators are unsuitable or do not exist.
- The method of maximum likelihood corresponds to many well-known estimation methods in statistics. For example, one may be interested in the heights of adult female penguins, but be unable to measure the height of every single penguin in a population due to cost or time constraints. Assuming that the heights are normally (Gaussian) distributed with some unknown mean and variance, the mean and variance can be estimated with MLE while only knowing the heights of some sample of the overall population. MLE would accomplish this by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probable (given the model).
- Under normal distribution, maximum likelihood is equivalent to least square estimation.
- It may not exist.

17. What is confidence interval?

- (1) There's some quantity you want to know. This number has a definite value. One that doesn't change. Like the average salary of, say, working mothers in the US. (2) You use some method to compute an interval (two numbers, one higher than the other). (3) If you use that method a huge number of times, 95% (or some other percentage) of the intervals you generate would contain the quantity you are looking for.
- Note this, though: It is NOT "the probability that the quantity you're after is in that particular interval". That quantity is fixed, and that interval is fixed, so the quantity is either inside the interval, or it isn't. The probability part comes from the fact that you do the procedure a gazillion times, generating new intervals each time. That's the part that confuses people.

- A 95% confidence interval does NOT make the statement that the population mean falls in between a lower limit L and an upper limit U 95% of the time. It instead says that 95% of confidence intervals constructed from various samples will contain the population mean.
- 95% is not for the interval constructed, but for the method of constructing the interval, usually sample statistics plus minus margin of error. Once the interval is constructed from the sample, the parameter is either in, with probability of 1, or out with a probability of 1.