Basics of Descriptive Analytics:

a. Can you tell me the difference between mode, median, mean?

Measures of center

- Mean – Average value
- Median – Middle Value
- Mode – Most frequent value

b. Can you explain range and standard deviation?

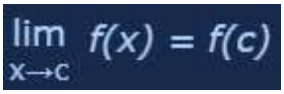
Measures of Spread

- Range – Difference between the largest and smallest value
- Variance – How far the values of the dataset are from the mean, on average
- Standard Deviation – How far the values of the dataset are from the mean, on average. Square root of the variance

Basics of Unconstrained Continuous Optimization:

a. You should haven understanding of continuous functions and the roles of the first, second derivatives

- A function f is continuous when, for every value c in its Domain: f(c) is defined and "the limit of f(x) as x approaches c equals f(c)



- **First derivative** – The first derivative of the function f(x) is the slop of the tangent line of the function at the point x. It tells us how whether a function is increasing or decreasing at the point x
  - f'(x) > 0 – Increasing
  - f'(x) < 0 – decreasing
  - f'(x) = 0 – Critical point; no new information obtained

- **Second Derivative** – The second derivative of a function is the derivative of the derivative of that function. It tells us if the first derivative is increasing or decreasing
  - f''(x) > 0 – Increasing; concave up / parabola open upwards/convex
  - f''(x) < 0 – decreasing; concave down/ parabola open downwards
  - f''(x) = 0 – Critical point; no new information obtained

b. Extreme Points (local vs global points)

- Second derivative test to find out whether a point is a local maximum or local minimum
  - Local minimum: If the first derivative is a critical point $f'(x) = 0$ and the second derivative is increasing $f''(x) > 0$
  - Local maximum: If the first derivative is a critical point $f'(x) = 0$ and the second derivative is decreasing $f''(x) < 0$
- Inflection points – Inflection point at x if the graph of the function goes from concave up to concave down at the point or if the graph of the function goes from concave down to concave up. It can only happen where at points where the second derivative is 0 $f''(x) = 0$. It tells us where there might be an inflection point, but not definitely
- If second derivative changes signs within a domain, that will show inflextion point

Basics of Probability:

a. Calculate probability for Multiple Events & Bayesian Statistics

- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  - When A & B are mutually exclusive, $P(A \text{ and } B) = 0$
- $P(A \text{ and } B) = P(A) \times P(B|A)$
  - When independent, $P(B|A) = P(B)$
- Conditional Probability – $P(B|A) = P(A \text{ and } B) / P(A)$
- Total Probability – $P(A) = P(A|B) \times P(B) + P(A|B^C) \times P(B^C)$

b. The Common Types of variables in statistics. Independent, covariate, un-correlated random variables, etc.

- Independent – variable that is not affected by anything but might affect other variables
- Dependent – variable whose value changes based on other variables
- Confounding – extra variables that have a hidden effect on your results
- Random – associated with random processes and give number to outcomes of random events
- Covariate – has an effect on dependent variable but is usually not the variable of interest. Similar to independent variables
- Un-correlated random variables – if covariance $cov[X,Y] = E[XY] - E[X]E[Y] = 0$

d. Simple combinatorics

- Permutations – way to combine elements where order matters
  - ${}_nP_r = n!/(n - r)!$

- Combinations – combine elements where orders don't matter
  - ${}_nC_r = n!/r!(n-r)!$

Not covered: measure theoretic probability, moment generating functions, exponential families, inequality results (e.g., Chebychev, Markov, Jensen); multivariate pdfs and cdfs, Jacobian transformation, determining the pdf of a transformation of a single or multiple random variables, knowledge of the theoretical relationships between various random variables (e.g., gamma with alpha=1 is an exponential distribution, gamma with alpha=p/2 and beta=2 is a chi-squared distribution, etc.).

Basic probability distributions and limit theorems:

a. Can you compare Probability Density Function vs Cumulative Distribution Function?

- PDF – probability that a random variable takes a certain value. Easier for discrete random variable. For continuous variable, we can't use the PDF directly, since the probability that x takes on any exact value is zero. Got to use integral to calculate PDF
- CDF – probability that a random variable takes on a value less than or equal to x. CDF is always increasing i.e. non-decreasing
- PDF is the derivative of CDF

b. Understanding differences between basic continuous and discrete random variables

- Continuous Random Variables
- Discrete Random Variables

c. You should be able to explain mean/variance? And be able to talk about covariance

- Mean - Average value
- Variance – Spread of the dataset. How far the values of the dataset are from the mean, on average
- Covariance – measure of how two random variables vary together
  - $Cov(X,Y) = \Sigma E((X - \mu) E(Y - \nu)) / n-1$ where:
  - $E(X) = \mu$ is the expected value (the mean) of the random variable X and
  - $E(Y) = \nu$ is the expected value (the mean) of the random variable Y

e. Understanding differences between Central Limit Theorem and Law of Large Numbers.

Samples need to be random and independent

- Central Limit Theorem – sum or average of many independent samples of a random variable provided they samples are sufficiently large is approximately a

normal random variable even though the population distribution might not be normal

- Law of Large Numbers – the average of many independent sample is closer to the mean of the underlying distribution with a high probability. This also states that the histogram will resemble the underlying distribution

Basics of Estimation and Hypothesis Testing

a. Understanding differences between population parameters and sample statistic

- Population Parameters
- Sample Statistics

b. Hypothesis Testing and scenarios

- Hypothesis Testing

c. Type 1 error, Type II error

Type 1 Error = P(False positive)
= P(Rejecting Null when null is true)
= P(Actual Negative, Prediction Positive)

Type II Error = P(False Negative)
= P(Not rejecting Null when null is not true)
= P(Actual Positive, Prediction Negative)

d. P-values, R^2

P-values = P(test finding the results by chance | H0)
= P(test resulting at least as extreme results as observed| H0)
= P(rejecting null when null is true)

Null – No difference

Alternative – Better or Worse

Null not true = Actual Positive/Negative Class

Confusion Matrix

	Predicted (Decision about Null)		
Actual (Null)		+	-
	+	True Positive (TP) (1- β) Reject null, Null not true	False Negative (FN) β, Type II Error Dont reject null, Null not true
	-	False Positive (FP) α, Type I Error Reject Null, Null True	True Negative (TN) (1- α) Don't reject null, Null True

Actual	Predicted	Conclusion
+	+	TP
+	-	FN (Type II Error, β)
-	+	FP (type I Error, α)
-	-	TN

Accuracy: Total Correct Predictions = TP + TN / (TP + TN + FP + FN)

Precision: How many predicted + did we get right? = TP / (TP + FP)

Recall: True Positive Rate (TPR): How many actual + did we get right? = TP / (TP + FN)

False Positive Rate (FPR): How many – did we misclassify? = FP/(FP + TN)

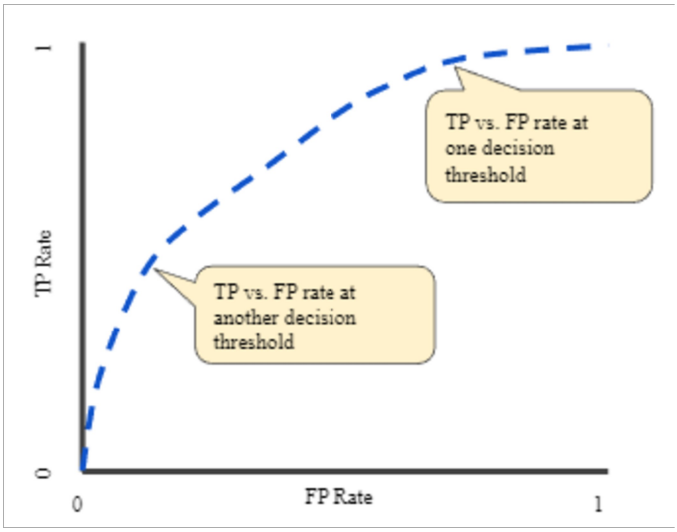
Specificity: How many – did we get right? = TN / (FP + TN) = 1 – FPR

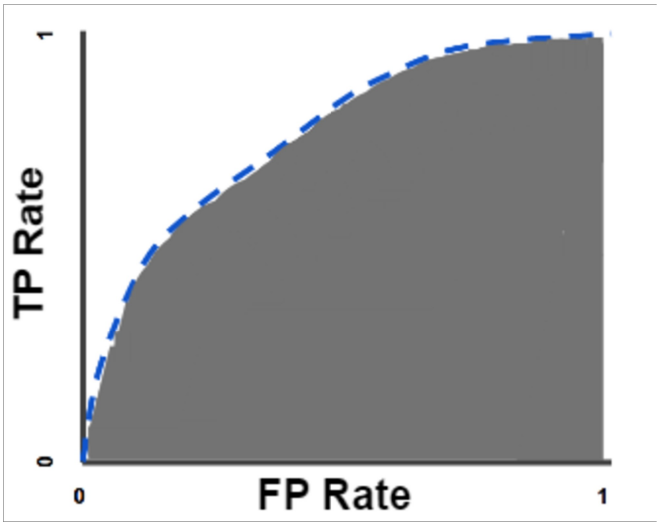
F1 Score: Hybrid metrics for imbalance classes = 2TP / (2TP+FP + FN)

F1 Score = 2 X Precision X Recall / (Precision + Recall)

ROC Curve (receiver operating characteristic curve) = Graph showing a performance of a classification model at all classification thresholds. Its plots FPR along the x-axis and TPR along the y-axis

AUC – Area under the ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.





T-Tests

- When we don't know the population variance
- Sample size is small n < 30

Z-Tests

- When the population variance is known or sample size is n >= 30

Basics of Regression:

a. General concepts of regression and the different types

b. Simple linear regression model, assumptions, Ordinary Least Squares (OLS)

Linear Regression Assumptions:

1. Linear Relationship
2. Independent Variables |
3. Variance is Constant – Homogeneity of variance that the variance within each of the populations is equal
4. No Multicollinearity – Highly correlated independent variables. This doesn't reduce the power of prediction or reliability or goodness of fit, but it makes the coefficients erratic
5. Normality of residuals, constant variance of residuals and Independent residuals
6. No strong outliers

Linear Regression function : $y = X \beta + \boldsymbol{\varepsilon}$

$$\beta = (X'DX)\text{-}1\ X'\ Dy$$

$$= [1\ x_1\ x_2\ \ldots\ x_k]\ [\beta_0\ \ \beta_1\ \ \beta_2\ \ldots\ \ \beta_k]^T + \varepsilon$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x};$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

$SSR=\sum(y\hat{}-\bar{y})^2$
$SSE=\sum(y-y\hat{})^2$
$SSTO=\sum(y-\bar{y})^2$
$SSTO = SSR + SSE$

R^2 = SSR / SSTO
R^2 = Is the percet of the variation in independent variables explained by the variation in the predictor variables

Loss Function =

Mean Absolute Error = $1/n\ (\sum(y - y\hat{}))$

Mean Absolute % Error $= 1/n\ (\sum(|y - y\hat{}|/y)* 100)$

Mean Squared Error = $1/n(\sum(y-y\hat{})^{2)}$

Logistic Regression Assumptions:

1. Linear Relationship of independent variables and log odds
2. Large sample size (at least 10 cases with the least frequent outcome for each independent variables)
3. Independent Variables |
4. Variance is Constant
5. No Multicollinearity – Highly correlated independent variables. This doesn't reduce the power of prediction or reliability or goodness of fit, but it makes the coefficients erratic
6. Normality of residuals, constant variance of residuals and Independent residuals
7. No strong outliers

Logistic Regression is a sigmoid function

$y = 1 / (1 + e^{-Z})$

$y' = 1 / (1 + e^{-Z})$

log odds = $Z = b + w_1x_1 + w_2x_2 + ... + W_nx_n$

log odds Z = log(y / (1- y))
odds = exp(Z)
probability p = e^Z / (1 + e^Z)
Odds Ratio = p / (1 - p)

Loss Function =

$$\text{Log Loss} = \sum_{(x,y)\in D} -y\log(y') - (1-y)\log(1-y')$$

How does logistic regression calculate the coefficients?

8

You are right that although you should be able to calculate the OLS coefficient estimate in logit space, you can't do it directly because the logit, $g(y) = \log \frac{p}{1-p}$, goes either to $-\infty$ for $y = 0$ or ∞ for $y = 1$. An added difficulty is that the variance in this model depends on x.

The likelihood for logistic regression is optimized by an algorithm called **iteratively reweighted least squares** (IRLS). There is a nice breakdown of this in Shalizi's Advanced Data Analysis from an Elementary Point of View, from which I have the details below:

- To deal with the infinite logit problem, make a first-order Taylor approximation to $g(y)$ around the point p such that $g(y) \approx g(p) + (y - p)g'(p)$. Since $g(p)$ is by definition $\beta_0 + \beta x$, put that in there instead of $g(p)$ and say that your **effective response** is $z = \beta_0 + \beta x + (y - p)g'(p)$.
- Calculate the variance $V[Z|X = x] = V[(Y - p)g'(p)|X = x] = g'(p)^2 V(p)$. Use this to **weight your samples** so that you can simply do a weighted regression of z on x.

At this point you might ask yourself how you can use the regression coefficients you're trying to estimate to calculate your effective response, z. Of course you can't. And what is p anyway? That's where the iterative part of IRLS comes in: you start with some guess at the βs, for instance to set them all to 0. From this you can first calculate the **fitted probabilities** p, and second use these fitted probabilities and your current coefficient estimates to calculate z.

All this and you get a new estimate for your βs, and it should be closer to the right one, but probably not the right one. So you iterate: use the new coefficients to calculate new fitted probabilities, calculate new effective responses, new weights, and go again. Sooner or later, unless you're unlucky, the βs will converge to a nice estimate. Says Shalizi:

> The treatment above is rather heuristic, but it turns out to be equivalent to using Newton's method, only with the expected second derivative of the log likelihood, instead of its actual value.

So in summary: you never use the logit directly because, as you point out, it's impractical. You can certainly calculate the logistic regression coefficients by hand, but it won't be fun.

Time Series Analysis & Forecasting:

1. Why not standard regression in time series setting?
   - Autocorrelations which violates independence assumptions
2. Understanding of integrated processes and cointegrated processes
   - Stationary – no seasonality on the data i.e. one whose properties do not depend on the time at which the series is observed

- Process to convert time series to stationary processes
- Cointegrated – forms a synthetic stationary series from a linear combination of two or more non-stationary series

3. Univariate time series methods
4. Multivariate time series methods
5. Understanding stationary and tests for unit roots
6. Moving Averages
7. Serial correlation and seasonality
8. AR, MA, ARMA, ARIMA and VAR

Problem Solving:

a. Identify, own, and brainstorm intrinsically hard problems (e.g., highly complex, ambiguous, undefined, with less existing structure, or having significant business risk or potential for significant impact). This can be done through an example business type Case Study. For example

b. Ability to provide solutions to complex and/or ambiguous problems.

c. Can visualize and translate well-defined problems into data science problems.

Basic ML:

a.Supervised vs unsupervised learning

a. Supervised learning – machine learning approach when the data is trained on per-defined labeled datasets and use the information gained to predict outcomes accurately. Linear regression, classification are supervised learning
b. Unsupervised learning – machine learning approach to analyze and cluster unlabeled datasets. The algorithms discover hidden patterns in data without the need for human intervention. Clustering, association/recommendations, dimensionality reduction,

b.Bias-variance trade off

a. Bias error – Erroneous assumptions in the learning algorithm
b. Variance error – Error from sensitivity to small fluctuations
c. As bias is reduced variance increased and as variance decreases, bias increases i.e. a simple model has low variance, but high bias and as model complexity increases, variance goes up and bias goes down. Always a balancing act between these two

c.Overfitting and Underfitting

a. Overfitting – When a model does very well in training, but does horrible in testing. This happens because the noise and random fluctuations in the training data is picked up and learned as concepts by the model, but these concepts don't generalize to the new unseen data. This happens when the model is too complex. Need to simplify it more. Hold a validation set to minimize overfitting
b. Underfitting – Model does horrible in testing and training. This can be solved by making the model more complex

d.Cross-validation

a. Splitting a dataset into train and test sets to validate the performance of the model. Use the train datasets to train the model and test dataset to test the model.

e.Use of logistic regression for classification problems

f.Confusion matrix, precision, recall, F1 score, AUC

g.Basics of text data mining (bag-of-words data structure)

a. Feature extraction within text data is called bags-of-words model of text
b. representation of text that describes the occurrence of words within a document

h.Boosting

a. Technique used to reduce errors in predictive data analysis by training multiple weak learners into a single strong learning models. Train week learners one after another
b. This is used in mainly decision trees. Boosting creates an ensemble model by combining several weak decision trees sequentially. It assigns the weights to the output of the individual trees and it gives incorrect classifications from the first tree a higher weight and input to the next tree. After number cycles, the boosting method combines these weak rules into a single powerful prediction rule

i.Decision Trees, Random forest (bagging)

a. Decision Trees
    a. Supervised machine learning method that can be used to classify or predict a value
    b. It includes root node which is the base of the decision tree and splits them into decision node. When a decision node does not further split

into additional sub-notes this is the leaf node and it represents possible outcomes

c. Utilize pruning by cutting down some nodes to stop overfitting
d. Entropy is the uncertainty in our dataset or measure of disorder

$$E(S) = -p_{(+)}\log p_{(+)} - p_{(-)}\log p_{(-)}$$

Here p_+ is the probability of positive class

p_- is the probability of negative class

S is the subset of the training example

e. Entropy measures the impurity of a node. Impurity is the degree of randomness. A pure sub-split means that either you should be getting "yes", or you should be getting "no". When the purity is 100%, it is a leaf node. Higher the Entropy, lower the purity and higher impurity
f. Information gain measures the reduction of uncertainty given some feature and is a deciding factor for which attribute should be selected as a decision node or root node. It is the entropy of the full dataset
g. Information gain = E(Y) – E(Y|X)

b. Random Forest

- Combines multiple decision trees to one – ensemble method
- Trained using CART (Classification and Regression Tree) algorithms
- Metrics such of Gini Impurity, Information Gain or MSE can be used to evaluate the qualify of the split
- prone to overfitting
- Utilizes bagging and feature randomness to create an uncorrelated forest of decision trees. It fits multiple model on different subsets of training dataset and then combines the prediction from all models
- Bagging – Random forest generates a random subset of features which ensures low correlation among decision trees
- Gini Impurity – Measurement of the likelihood of an incorrect classification of a new instance of a random variable if that new instance was randomly classified according to the distribution of class labels

- Gini Impurity = 1 – sum(P_i^2) where Pi denotes the probability of an element being classified for a distinct class

j.Regularization methods

a. L1 (Lasso) – Least absolute Shrinkage and Selection Operator. Penalty term that is the absolute sum of coefficients. This takes the coefficients to zero and is used in dimension reduction. This decreases the value of coefficient to reduce loss. Struggles with a lot of predictors and multicollinearity
b. L2 (Ridge) – Adds a penalty term square of the magnitude of coefficients. Decreases the complexity of the model but doesn't reduce the number of variables. It is not good for feature selection

k.K-means (k-medians); Distance Measures

a. Goal of clustering is to divide items into similar groups
b. Similarity in clustering is defined by distance metrics
c. Euclidean Distance = $\sqrt{((x2 - x1)^2 + (Y2 - Y1)^2)}$
d. Manhattan Distance = (X2 – X1 + Y2 – Y1)
e. K means is a centroid based clustering. It takes user supplied amount of clusters – k and divides the data into many partitions
f. Initially, k centroids are initialized
g. Next distance between a given data point and each of the three centroid is calculated. This is done for all data points
h. Each data point is assigned to the cluster of the centroid that it is closest to
i. The centroids are updated by recalculated by averaging the coordinates of each data point in the respective clusters
j. This process of reassigning points and updating centroids continues until the centroids no longer move

K Median Clustering

1. Centroid is determined by calculating the median rather than the mean. This minimizes error overall cluster with respect to the 1-norm distance metric as opposed to the squared 2-norm distance metric

l.Support vector machines

a. SVM works by mapping data to high-dimensional feature space so that the data points can be categorized even when the data points are not otherwise linearly separable

b. separator between the categories is found and the data is transformed in such a way that the separator could be drawn as a hyperplane
c. best hyperplane is the one that maximizes the margins from both tags
d. characteristics of new data can be used to predict the group t which a new record should belong
e. Functions used for transformation is known as kernel function and can be linear, polynomial, radial basis function, sigmoid function etc...

m.K-nearest neighbors

a. supervised learning method used for classification or regression
b. assumes similar things exist in close proximity
c. choose k to chosen number of neighbors
d. class label is assigned on the basis of a majority vote i.e. the label that is most frequently represented around a given data point is used
e. for each example example in data, calculate the distance between all other points
f. add the distance and index of the example to an ordered collection
g. sort the ordered collection of distances and indices from smallest to largest by the distances
h. pick the first k entries from the sorted collection
i. get labels of the selected K entries and return if classification; if regression, return mean of the k labels

n.Different loss functions

Data Manipulation:

a.Ability to work with real data in an efficient manner, usually manifested by mastery of languages such as SQL, R, or Python. This is the ability to execute the mechanics of generating data, creating metrics/models.