**Stats**

**Explain p-value**

**[Answer 1]**

* For hypo testing: p-value indicates statistical significance
* For statsmodels: p-value indicates whether the feature is statistically significant (higher than 0.05 means the feature can be dropped)

**What is power analysis?**

**[Answer 1]** Power is the probability that a statistical test correctly rejects the null hypothesis (i.e. probability of a true positive result). Power analysis can estimate how much power a test has, given the effect size, sample size, and significance level. Power analysis can also be used to estimate any of these four parameters given values for three.

**Explain t-test (explain what data you need, H0 vs H1, etc)**

**[Answer1 - SHF]**

'Student's' t Test is a common technique for testing if two groups of data are significantly different from each other, typically to show if there is a difference as the outcome of an experiment e.g. when testing the efficacy of a drug on controlling tumor growth.

* H0, the null hypothesis that there is no difference ie. mean difference is zero. The alternative hypothesis H1 is that there is a difference, i.e. the mean difference is different from zero.
* First calculate a t-statistic which is the difference between the means of the 2 groups of data standardized by the group variances.
* Secondly, we will calculate a p-value, which if done by hand, is by using the t Table using the t-statistic and degrees of freedom, the latter being sample size - 1.
* The p-value indicates the probability that the difference was due to random chance. For example, a p-value of 0.4 would mean that there is a 40% likelihood that we measured a difference even though there is actually no difference in tumor sizes of the group that took a drug compared to that who took a placebo.
* Typically, if we calculate a probability of 0.05 or less, then we can reject the null hypothesis, that is, we can conclude that the tumor sizes of the two groups do differ.
* Generally, z-tests are used for testing for differences when we have large sample sizes (n > 30), whereas t-tests are used a smaller sample size (n < 30) or when we do not know the true standard deviation of the population.

**Explain chi2 test (explain what data you need, H0 vs H1, etc)**

**[Answer1]**

A chi2 test is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis (H0) is true. See if you want know if there is a difference between two population and you collect sample from these two groups. Then calculate the expected value if there is no difference between these two groups. The chi2 value is a single number that adds up all the differences between our actual data and the data expected if there is no difference. If the actual and expected data (there is no difference) , the chi2 value is 0. A bigger difference gives a bigger ch2 value. The larger the chi2 value, the greater the probability that there is a significant difference.

**When can you use t-test and when chi2?**

**Explain what is a confidence interval to a non technical audience**

**[Answer 1]:** A range of values provided that is likely to contain the true value.

**[Answer 2]:** A range of values which might contain an unknown population parameter (e.g. mean).

**Explain ab testing to a non technical audience**

**What is the Central Limit Theorem?**

**[Answer 1]:** Distribution of the sample means will approach a normal distribution as the sample size increases

**[Answer 2]:** The distribution of sample mean from a population will approximate a normal distribution if large enough.

**What is a “normal distribution”?**

**Explain a/b testing to a non technical audience**

**What is the difference between type-1 and type-2 error?**

**[Answer 1]:**

Type-1 Error: Rejecting null hypothesis when it is in fact true (False Positive);

Type-2 Error: Not rejecting the null hypothesis when it is in fact false (False Negative)

**Machine Learning**

**What is a baseline? What baselines can you use in different problems (regression, classification, time series, etc.)**

**[Answer 1]:** You need to spot-check algorithms on the problem to see if you have a useful basis for modeling your prediction problem. To do this, you need a basis for comparison of results. This would be the baseline result which is the simplest possible prediction.   
For Classification - you can select the class with the most observations and use that class as the result for all predictions  
For Regression - you can use central tendency measure as the result for all predictions, such as the mean or the median

For Time Series - you can use the overall average number or break it down by hour or day of the week or something to that effect

**How do you select k for k-means clustering?**

**[Answer 1]:** 1. Using Silhouette Analysis, 2. Elbow (plotting Sum of Squared Errors against number of clusters) 3. Manual method of deciding k (with domain knowledge)

**Regression: what are the steps to evaluate the results of a linear regression?**

**[Answer 1]** The model evaluation metrics for linear regression(or regression in general) are 1) Mean Absolute Error(mean of absolute value of errors)

2) Mean Squared Error(mean of squared errors)

3) Root Mean Squared Erro(square root of mean of squared errorrs).

**What are the assumption of a multi linear regression?**

**Explain recursive feature elimination**

**[Answer 1]:**

Recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef\_ attribute or through a feature\_importances\_ attribute. Then, the least important features (or more important) are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

**How can you find outliers?**

**[Answer 1]**

* Use a boxplot to visualise (if there are only a few features)
* If there are a lot of features, can use descriptive stats and use the value of Mean +/- 3\*std to filter the outliers

**Explain the difference between bias and variance**

**Explain AUC to a non technical audience**

**[Answer 1] - ROC context**

The AUC is the area under the ROC curve. This score gives us a good idea of how well the model performances. A higher score would mean that good job of distinguishing the positive and the negative values.

**When would you use accuracy, precision and recall?**

**[Answer 1]**

- Use Accuracy to determine the ratio of TP and TN compared to overall result. Useful if data is balanced

- Precision is used to determine ratio of TP to TP + FP. Useful if data is balanced

- Recall is used to determine ratio of TP to TP + FN. Useful if data is unbalanced

**[Answer 2]**

* Accuracy:
  + Proportion of model’s predictions that are accurate regardless of class. Calculated with (TP + TN) / Total No. of Observations.
  + Useful when you need one score to summarise the model’s overall performance (provided the classes are balanced)
* Precision:
  + Proportion of the model’s prediction for a class that was accurate.
  + Useful when it is important to minimise False Positives (i.e. increase Precision). E.g. Filtering videos for children. It is more important to minimise instances where the model classifies a video as kid-safe when it is actually not
* Recall:
  + Proportion of the actual class that the model predicted correctly.
  + Useful when it is important to minimise False Negatives (i.e. increase Recall). E.g. flagging positive for cancer.
  + Also important to check if recall is low when precision is very high, as that could be indicative of a model that might be too stringent (threshold to classify as positive is very high → all the positives classified by the model is accurate, but we end up with a lot of False Negatives that weren’t classified positive because the model was too strict)

**Regression: explain pros/cons of mean absolute error vs. mean squared error vs. median absolute error**

**[Answer1]:**

\* Mean absolute error (MAE): measures the average absolute deviation between the predictions and the true values. It has better interpretability.

\*Mean squared error (MSE): measure the average of the squared difference between prediction and actual observation. The errors are squared before they are averaged and thus gives a relatively high weight to large errors. It is more useful when large errors are particularly undesirable.

\*Median absolute error is useful because it is essentially insensitive to outliers and it is a more robust metric.

**[Answer 2]:**

They are all ways to measure error and serve as a loss function to minimise.

MSE: Penalises outliers more by amplifying large errors

MAE: more robust to outliers

Median AE: only corrects for outliers in the response/target variable, not for outliers in the predictors/feature variables.

[Answer SHF]

MAE is the most intuitive, it is the average absolute difference between the data and the model's predictions.

MSE has mathematical properties (continuously differentiable) which makes it easier to minimize than MAE. Because of the square, large errors have relatively greater influence on MSE than do the smaller error. Therefore, MAE is more robust to outliers. On the other hand, MSE is more useful if we are concerned about large errors whose consequences are much bigger than equivalent smaller ones

**Make an example of a problem where you want high accuracy and a model where you want high recall**

**How can you know the features you are using are the best ones?**

**[Answer 1]** To identify the best features, you need to understand the business problem that you’re trying to solve, make sure the features are independent of each other and run a correlation matrix to find the features with the highest correlation to the target.

**Explain how SVM works**

**Explain how random forest works**

**[Answer1]:**

Supervised Ensemble learning method for both classification and regression tasks.

Supervised – Labelled, Ensemble – made up of many decision trees.

Takes into account the results from multiple decision trees and output is the average.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Advantages:

Hard to overfit, due to averaging.

Easy to use, parameters easy to understand. n\_estimators, max\_features

**Explain the coefficients of a logistic regression**

**[Answer 1]:** The coefficients of a logistic regression can be interpreted a the log odds ratio that each unit of the independent variable contributes

**Explain how gradient boosting works**

**What is the kernel trick in SVM?**

**Explain how PCA works**

**What’s the difference between the feature importances of a random forest and the coefficients of a logistic regression?**

**SQL**

**What is a nested query?**

**[Answer 1]**

* Nested query is used to do a cascading query (inner query can be used to do 1st level filtering of data) and treat as a table to select data from

**What is the difference between left/right/inner/outer join?**

**[Answer 1]**

* Left join will take the left table and merge the values of the right table (whether it exists or not) to the left
* Right join is similar to Left join (except that the right table is the main reference)
* Inner join will take the intersection of both left and right tables
* Outer join will simply take both tables and merge them together

**What’s the difference between where and having?**

**[Answer 1]**

* Where is used to put condition on individual fields of the table
* Having is used to put condition on aggregates

**Time series analysis**

**What is the I of ARIMA models?**

**[Answer 1]**

‣ “i” stands for integrated: we are combining AR and MA techniques into a single model: ARIMA.   
‣ Autoregressive (AR) models are those are that use data from previous time-points to predict the next time-point. These are very similar to previous regression models, except as input - we'll take some previous outcome.  
‣ Moving average models, as opposed to autoregressive models, do not take the previous outputs (or values) as inputs, but instead take the previous error terms. We will attempt to predict the next value based on the overall average and how incorrect our previous predictions were.  
‣ Integrating the two tactics results in us selecting some differencing term, d, where we are now predicting the DIFFERENCE between one prior period and the new period, rather than predicting the new period’s value itself.

‣ Integrated – To model only the differential in a time series.

**[Answer 2]**

I stands for Integrated, which refers to the use of differencing of raw observations (e.g. subtracting the value of an observation from its previous value) in order to make the time series stationary. The larger the I, the higher the degree of differencing.

**Explain SARIMAX(3,2,0)(2,0,0,order=12)**

**[Answer 1]**

The (p,d,q) order of the model for the number of AR parameters, differences, and MA parameters. d must be an integer indicating the integration order of the process, while p and q may either be an integers indicating the AR and MA orders (so that all lags up to those orders are included) or else iterables giving specific AR and / or MA lags to include. Default is an AR(1) model: (1,0,0). In this case, p=3,d=2,q=0

The (P,D,Q,s) order of the seasonal component of the model for the AR parameters, differences, MA parameters, and periodicity. d must be an integer indicating the integration order of the process, while p and q may either be an integers indicating the AR and MA orders (so that all lags up to those orders are included) or else iterables giving specific AR and / or MA lags to include. s is an integer giving the periodicity (number of periods in season), often it is 4 for quarterly data or 12 for monthly data. Default is no seasonal effect.Here, P=2, D=0, Q=0

**Model 1 has AIC=85, Model 2 AIC=89. Is it enough to say that model 1 is better? What other things would you look at?**

**[Answer 1]:** No, we should also plot the residuals

**EDA & data analysis**

**Explain the basic steps of an EDA for a classification problem**

**Explain pros/cons of different missing values imputation techniques**

**[Answer 1]**

* **Mean**
  + Pros: Simple to compute.
  + Cons: Reduces the variance of the data.
* **Median**
  + Pros: Simple to compute
  + Cons:
* **Single Model Imputation**
* **Multiple Imputation**

#### 

**Business acumen & co**

#### **How could you use GPS data from a car to determine the quality of a driver?**

**How many chicken rices are sold today at lunchtime in the orchard area? (discuss about assumption and methods you’re using to get to the solution)**

**[Answer 1]**

- Determine the boundary of Orchard

- Google how many eatery establishments are there within the Orchard boundary

- Select or 2 or 3 eateries and perform survey from 11am to 2pm on the number of chicken rice sold there

- Perform (hypo testing) T tests to see whether this sample represents the entire population (Orchard)

- If Yes, take the average of the median of the samples and multiply it to the number of eateries in Orchard

**Final answer:** average \* number of eateries

Add on: (alternative)

Assumption: lunch hour ~2 hour

There are 20 malls in orchard area and every mall has a chicken rice store

Each store can serve 5 min/ serving

(2\*60/5)\*20= 480

**How much money the students from dsi5 will spends on laptops and PCs in the next 5 years? (discuss about assumptions and methods you’re using to get to the solution)**

Assuming DSI5 students will change their laptops thrice over the next 5 years, considering:

(1) Every student already has an existing laptop in use with an average age of 1 year old

(2) Laptops / PCs to be changed every two years to take advantage of performance improvements (given Moore’s law)

And if the average price of the laptop / PCs purchased will be approximately 1.8k, and a class of 24 students, DSI students will spend:

24 x 1,800 x 3 = ~ 130,000

**You are Facebook and you want to test the hypothesis that showing the “these are your highlights with Alice (i.e. one of your friend)” videos improves the user interaction with Facebook. How would you test this hypothesis? What metric would you use?**

[SHF]:

* H0: Highlights video has no effect on user interaction
* H1: Highlights video has a positive effect on user interaction
* Randomly split users into 2 groups: Shown highlights (experimental group) video vs No highlights video (control group)
* Metric: User interaction defined as average session duration
* Use t-test to see if there is a statistically significant difference between the mean session duration between the experimental and control group
* If p-value is lower than the predetermined significance level (e.g. 0.05), then we reject the null hypothesis that highlight videos have no effect on user interaction

A/B testing - User interaction can be measured via (1) whether user watches the video, (2) whether user navigates to the friend page / message the friend, and (3) how long the user stays on FB. Show the video to a test group and keep a select group as control. Track the metrics and measure if there is statistical difference between test and control group using the chi-squared test.

**NLP**

**Explain tf-idf**

**How does sentiment analysis work from a technical point of view?**

**Explain sentiment analysis to a non technical audience**

**Explain topic modeling and pros/cons of LDA**

**[Answer 1]:**

* Topic modeling is a technique in Natural Language Processing where ‘topics’ are discovered through the statistical distribution of certain words across many documents.
* LDA is one example of topic modelling.
* Pros: Easy to apply on large body of texts. Easily classify documents by topics.
* Cons: Unsupervised → response topics are ‘generated’. Model does not tell you what each topic means.

**[Answer 2]:**

Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents.

Latent Dirichlet Allocation (LDA) is a topic model that generates topics based on word frequency from a set of documents. LDA is particularly useful for finding reasonably accurate mixtures of topics within a given document set.

Pros:

Don’t have to come up with topics beforehand

Can see the frequency of keywords

Cons:

Might be hard to interpret results, as each ‘topic’ has to be derived from a group of text.

#### **How would you come up with an algorithm to detect plagiarism in online content?**