大概是你是一个电厂power plant的CEO，你这个电厂在Boston附近，现在要考虑从natural gas转到renewable energy。
Q1：哪些factor你会cosider
Q2：目前用的是natural gas，然后告诉你了每年最多发电量（大概是这个意思）是8.8m megawatt-hour per year
目前，Land $5m /month, fix cost: $25m/year, variable cost $20 MWh,
Sale price: $40 MWh
Investor 最开始投了$ 400M 然后第一年期待10%的return
问：minimum需要generate 多少电，算出来是6.25M，然后他问了你觉得怎么样，我说大概是71%的capacity还是很合理的。他说make sense
（这里我刚开始把Investor的$400M 也算进去了然后他提示了一下我就改过来了。也可能是这里第一次算错了导致挂掉了吧。唉）
Q3：现在考虑renewable energy
第一种是太阳能，初始投资和installation fee：12.5M
产出要考虑天气：75%的晴天产出是150,000 MWH /year 25%阴天的产出是500,000MWH/year
第二种是XXX（实在没听清），初始投资和installation fee：2.5M
不用考虑天气，产出是100,000 MWH/year 还有variable cost是$30 MWH，问你这两种方案要多长时间能收回初始投资。
楼主算出来回收初始投资的时间是一样的，都是2.5 year,然后他问那现在投资人问你选哪种，你怎么推荐。
回答完这里他就说 case closed了。

x cost : $\boxed{5\times12+25}+20x$

profit : $40x$

$$40x-(60+25+20x)=10\%\times400$$
$$20x=40+85$$
$$x=125/20=6.25\text{ M}$$

$$\frac{6.25}{8.8}=71\%$$

①设 y年收回

$$\underbrace{(150{,}000+500{,}000)}_{\times10^{-6}}\times y=12.5$$

$0.15+0.5$

0.65

$0.65\overline{)12.50}$

y

$$12.5=75\%\,y\times150{,}000\times40\times10^{-6}+25\%\,y\times500{,}000\times40\times10^{-6}$$

$$12.5=0.75y\times0.15\times20+0.25y\times0.5\times20$$
$$=y(0.75\times3+0.25\quad)$$

$$\frac{12.5}{0.25}=y(9+1)\times2$$
$$50=y(10)\times2$$
$$y=50/20=2.5$$

$$2.5=y\times100{,}000\times10^{-6}\times(40-30)$$
$$=y\times0.1\times10\qquad y=2.5$$

Power Day
1. Technical Interview
2. Statistical Role Play
3. Case Interview
4. Job Specific Interview

Technical Interview
国人大姐 techinical，很实际的一些问题，多半都涉及到scaling和time expense，比如你要predict用户下一次transaction的消费类别（dining, entertainment, ...），要确实在production server上deploy，要注意哪些问题，你的建议。大概思路就是要考虑有时候transaction会很密集，有可能会造成很大负荷（当然我是被提示了才知道要考虑这种问题），所以需要设置一个frequency阈值，比如把未来一小时的transaction全部看做同一个next transaction来预测。还有个问题是先问了你用过哪些分类方法，评价他们的表现，然后hypterparameter怎么选，deploy的时候有哪些问题，如果database更新了，有了新的feature，该怎么办（需要定期更新model）。有个很奇怪的问题是给定一个data，存在服务器上。以前有个员工pull下来做了分析，后来离职了没法联系到，你想比较一下结果，pull同一个源data，但发现你pull下来的samples只有之前那人报告里所说的一半多，你如何确认你的pull query是不是正确。这题不知道咋答，糊过去的

第一轮tech: 给一个case让你建模。前面是关于处理missing data, categorical features, 选择模型，讲一讲几个模型之间的区别，有很多的追问一定要清楚你讲的方法具体适用的情况和优点缺点；后面着实没想到问了超多production的问题，数据量超大怎么处理，怎么让模型每天自动跑，更新，监测模型有没有出错，github流程。这轮感觉并不简单，不少延伸开去的问题不知道答的怎样

Technical phone interview is about a case problem about predicting the duration of phone calls from call centers of a cable company to approximate the case complication and use it as an index to allocate representative’s workload.
Questions related to business:

- What's the impact of incorrect prediction?
- Same prediction, but the calls are related to different types of services. How will you do?

Questions related to big data:

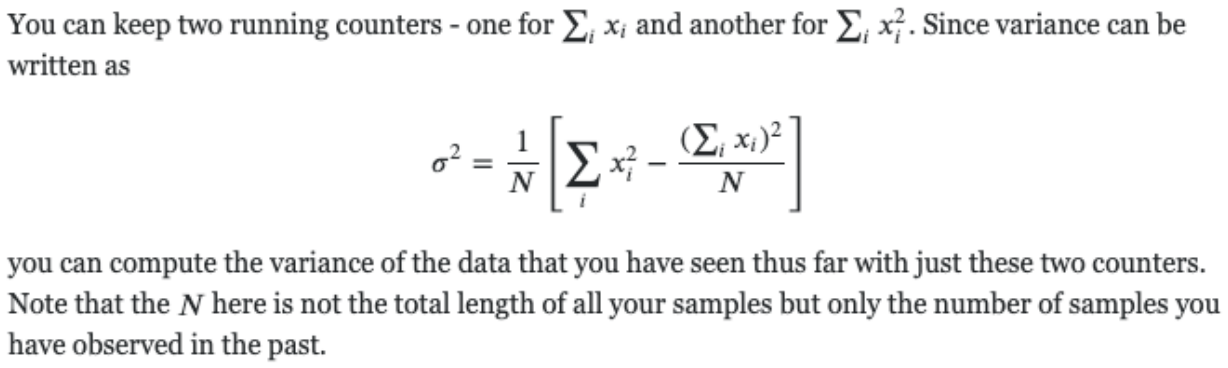
- how will you process 4 TB?
- 4 GB? It can load into the memory + no distributed computing available

Questions related to model design:

- What does the data look like to you? X, y? Data types. How are the data structures?
- deal with NA in X, y; how to impute?
- what if you don't know the names of columns.
- What variables will you add to the model?
- What kind of models will you try?
- What toolkits? Can you specify the code a bit more?

然後開始case study：情景是customer service 如何分配較長的call給有經驗的agent 較短的call給沒有經驗的agent
基本上就是預測這通call會是長的還是短的

Interviewer提供大概會有什麼樣的features (including missing values)
然後開始 問問題
1. 談談你想要的approach
2. 哪些feature覺得有用.
3. 如何清理data（handle missing values).
4. 你會想用哪一個model 會問細節
5. 如果data其實是big data 你該怎麼應對
6. 最後問如果把較短的call分配給有經驗的agent 然後較長的call分配給沒有經驗的agent 會怎樣 （基本上就是說你的prediction發生錯誤時 對於business的影響）



You can keep two running counters - one for $\sum_i x_i$ and another for $\sum_i x_i^2$. Since variance can be written as

$$\sigma^2 = \frac{1}{N}\left[\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{N}\right]$$

you can compute the variance of the data that you have seen thus far with just these two counters. Note that the N here is not the total length of all your samples but only the number of samples you have observed in the past.

technical的问了大数据相关的和传统的machine learning，以及一个如何有效计算mean和var基于streaming data 写code那种

就是和一些mapreduce的idea相关的，看面试官想不想问你深入了，技术面和onsite我都被问到了，还有就是如果数据量很大，要怎么建模型，你可以去借鉴一下online learning的思想。machine learning会问的很细，比如有哪些hyperparameter，要怎么train之类的
我没有问到，理论的问你如何处理，也会让你写个代码，比如找unique id满足一定条件。train就是问有哪些parameter，有什么用，一般设置什么value之类的

题目是call center predict customer call time。感觉是有固定的背景，固定的小题，面试官小哥人很nice，照着在读题，clarification也很耐心。小问主要关于
1. missing data怎么处理
2. 哪些possible features会影响
3. 用什么模型predict，一些基本的量比如conditional mean怎么算之类的
4. 怎么guarantee fast enough出predict结果
5. as always, 大数据怎么处理。颠来倒去问了很多，我不够了解ds，有些问题不太明白区别在哪。
6. 如果错误predict对公司和customer会有什么影响
7. 一些各种business sense的问题

1:这么多数据怎么处理?
从data review到feature extraction讲一遍，没什么追问。
2：什么feature觉得合理， 什么model 觉得合理
一般就是啥啥的mean， median，std之类的， classification 算法呗，列举几个例子
3：如果数据量变得特别大，需要各种retrain model，怎么办?
spark，而不是hadoop，因为要real time 更新results。
4： 如果给你好几年的数据，怎么办?
那就改用hadoop呗，毕竟不着急。
5：如果有个limit 5000的俩用户，一个用了100% 一个用了2%，模型给俩人呢分到 一类怎么办?
我答得不好，我后来也问了面试官怎么办，他支支吾吾的也没说好，就过了。
6：怎么在hadoop上实现RF
我说我没用过，不知道。
7： 怎么评价模型
AUC ROC
8: 怎么做调整?
K fold CV。
9： imbalance 怎么解决
然后轮到我问他问题，但是把他稳住了，最后跟我说sorry，我们俩陷入了沉默。然后我就说我可能现在做的project要用，只是好奇问问的。

有没有用过hadoop spark这些 cloud用过没有用的什么cloud
然后HR总结说他们公司现在都是AWS hadoop spark 语言用python这样子
解释以下spark和hadoop

貌似是处理一个data stream 算mean 和variance的，不过nyc的tech跟其他地方不一样，如果在其他地方面试，就没有什么参考价值了。hr跟我讲是做coding challenge的slides，结果面试时候根本没问。

第五轮： tech面，挑一些take home里面的东西，考了hypothesis testing，model为什么选这个不选那个，说说clustering的区别，怎么实现的，SQL statement，leftjoin casewhen什么的，然后一些casual inference的东西。

case：credit card churn model to predict whether user will close the account and will increase cash back reward for those who may close. A total of 4 million rows of data including date, purchase amount, type of store...
1. Q: How would learn about the data(好像是）?
（瞎说的) Exploratory Analysis?...
2. Q: How would you handle missing data?
Either Delete(row wise/ column wise and assumptions) or Impute(categorical: mode, as a level, logistic ; Continuous: mean, median, linear,0...)
3. feature engineering
4. (big data) what would you do if the data is updated hourly
后来问了面试官 好像意思要注意computation complexity
5. What if we have several years of data?
Hadoop MapReduce
6. What if only one computer or a single node is given
上个回复补充说了 但是不确定

7. What model would you use?

GBDT, RF, but also try other methods like LR

8. what language/lib will you use if data is like 1G

python/sklearn (稍微提了下 特别大会用spark 不知道对不对)

9. How to evaluate

AUC ROC

10. 如果模型结果不太好对business会有什么影响

11. 然后就是的5000 credit limit， 使用率相差很大的问题， 100% 和2%

一上来是讲一讲简历上一个project 具体问了问Linear regression的 assumption, cross validation 然后是用plain english解释一些stat concept, 有confidence interval, correlation coefficient, variance, marginal error

A confidence interval indicates the range that’s likely to contain the true population parameter

Understanding confidence levels

The probability that the confidence interval encompasses the true value is called the *confidence level* of the CI. You can calculate a CI for any confidence level you like, but the most commonly used value is 95 percent. Whenever you report a confidence interval, you must state the confidence level, like this: 95% CI = 114–126.

In general, higher confidence levels correspond to wider confidence intervals, and lower confidence level intervals are narrower. For example, the range 118–122 may have a 50 percent chance of containing the true population parameter within it; 115–125 may have a 90 percent chance of containing the truth, and 112–128 may have a 99 percent chance.

Recently, I got asked about how to explain confidence intervals in simple terms to a layperson.

We can never recover the exact value of a population characteristic from our data. We can only estimate it, and estimates are always wrong to some degree. Thus, we will usually also want to quantify the uncertainty in our estimate, which means that we wish to express how large the error in our estimate might be. One way to do this is using *confidence intervals*.

A confidence interval is a form of *interval estimate*, meaning that instead of presenting our findings as a single number, which we know is not exactly correct, we present our findings as an interval.

This interval is constructed so that it has a high chance (probability) of containing the true value of interest, e.g. the true unemployment rate in the population

Instead of producing an interval that is guaranteed to contain the value of interest, we may aim to construct an interval that has a defined probability of containing this value. This probability is called the *coverage probability*, and is most often taken to be 95%. Thus, we aim to construct an interval that is just wide enough so that 95% of the time, the interval contains the value of interest. Occasionally we may see people using a 90% or 99% confidence interval, which are intervals that either 90% of the time or 99% of the time cover the value of interest.

> http://dept.stat.lsa.umich.edu/~kshedden/introds/topics/confidence_intervals/

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s. **Pearson’s correlation** (also called Pearson’s *R*) is a **correlation coefficient** commonly used in linear regression. If you’re starting out in statistics, you’ll probably learn about Pearson’s *R* first. In fact, when anyone refers to **the** correlation coefficient, they are usually talking about Pearson’s.

Correlation is a measure of relation between variables, but cannot prove causality between them.

In practice, **variance** is a measure of how much something changes. For example, temperature has more **variance** in Moscow than in Hawaii.

The margin of error expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter. To be meaningful, the margin of error should be qualified by a probability statement (often expressed in the form of a confidence level).

For example, a pollster might report that 50% of voters will choose the Democratic candidate. To indicate the quality of the survey result, the pollster might add that the margin of error is ±5%, with a confidence level of 90%. This means that if the survey were repeated many times with different samples, the true percentage of Democratic voters would fall within the margin of error 90% of the time.

然后问两个method response rate, 如何比较
然后问如果有一个很大的dataset 1w数据吧，100个variable，问buid model的一些想法. 1

2：问简历上的项目，顺着项目问了些算法，比如RF和GBDT的区别，哪个更容易over-fitting, 为什么
3：怎么预测用户是否会点击一个广告？如果有A, B两广告，点A一次平均盈利50，点B平均盈利60， 怎么去判断哪个盈利多?
4：怎么对数据流提取特征？ 如果10分钟内有1T的数据，怎么取提取特征建模，用什么模型什么技术比较好?

只有一个可能是临时想到问的：
如果big data不给我hadoop/spark 在single node或者一个电脑上怎么弄
我答的把dataset分开 一部分一部分处理 因为RAM可能不够大 然后面试官说make sense （也不知道是不是真的）

然后问简历， 你最近做的项目是啥，方式他也会跟你探讨，所以项目具体细节你要准备清楚。 问我map reduce的工作流程， spark跟map reduce的区别？ 用过KNN吗（简历有写用）
最后问白板，让我写个gradient descent 的pesudo. 我学DL的时候写过，给忘了，只能说了个大概。。。为啥这个题我没看任何地里的人写过，我光是准备传统的ML了？ 完美避开所有准备的考点。。

下午是两轮tech，第一轮2点开始，问了multinomial distribution，结合不同的模型谈这个分布的应用，然后白板写sampling from multinomial distribution，之后问了variance 和 bias， 解释和如何检测，最后聊了聊如何根据不同分布生成fake data。我是这一轮答的不好，其实挺基础，但我之前并不常用这个分布也没准备到，最后也就挂在了这一轮。

A **multinomial distribution** is the probability distribution of the outcomes from a multinomial experiment.
Note: A binomial experiment is a special case of a multinomial experiment. Here is the main difference. With a binomial experiment, each trial can result in two - and only two - possible outcomes. With a multinomial experiment, each trial can have two *or more* possible outcomes.

> https://stattrek.com/probability-distributions/multinomial.aspx

给你sample input data 问你怎么样output 特定时段的最高点击率；基本leetcode med难度；剩下coding题就是考逻辑的。given a str 需要怎么样detect 具体的pattern 和计数；

(onsite) How would you explain the multinomial distribution and write python code on a whiteboard to represent this distribution

how do you detect multicollinearity?

Technical- asked some machine learning questions about random forest, regularization, map reduce, and derive some equations by hand. This was the only interesting part.

Explain the bias-variance tradeoff.
Write pseudocode for map reduce
What does regularization do?
Difference between random forest and gradient boosted tree.

Are false positives or false negatives more important?
It depends on the application. In medical tests False Positive (FP) is preferred to False Negative (FN). Because if a person is labeled to be positive (sick) then additional tests can be taken to confirm if it is really sick or not. But if the person is labeled negative (not-sick) while this is false, then no action will be taken and this can be harmful.
Alternatively, in other applications FN can be preferred to FP. For example in email spam classifier, FP will result in missing some important email but FN will be better because user won't miss such valuable information

What is VIF (in regression output)?
Interpret this ANOVA table.
Treatment effects are most often analyzed using ANOVA, which is short for “Analysis of Variance”.
R^2 is the percentage of variation in the response that is explained by the model. The higher the R^2 value, the better the model fits your data. R^2 is always between 0% and 100%.
A high R^2 value does not indicate that the model meets the model assumptions. You should check the residual plots to verify the assumptions.

How would you address the overfitting problem?
By Cross-validation. A fraction of the dataset would be earmarked for Cross-validation apart from the training and the test data sets. Various techniques such as Leave-one-out and K-fold would be used to deal with this problem.
Certain machine learning techniques such as Random Forests also address the overfitting problem.

How to build up a model to predict credit card fraud?
Couple things to keep in mind regarding fraud:

1) you're dealing with an imbalanced data set (your fraud cases may be 3-5% of all your data). So, consider either oversampling, or giving higher weight to your fraud cases.
2) you data may not have all the true fraud cases - in other words, there maybe actual fraud cases not captured in your data. So, some form of anomaly detection may be needed.

How would you structure a basic MapReduce problem?

Statistical Role Play
(Given 20 minutes to look through several printouts of data, charts and slides) How would you communicate the findings from this model to a non-technical executive?
Answer: 先简要总结现有模型

Role Play: flight delay
情景是你是某航空公司，报纸写了篇文章批评你们delay rate（指起飞延误）是几家公司里最高的，你们公司请了一个外包分析了下航班数据，但不知道做得怎么样，就请你来check一下quality，并且讲给non-technical的manager听。 看完材料以后跟客户提建议，之后客户化身专业人士深入聊。
一是时间没安排好，15分钟里面刚刚看完材料，根本没时间总结，另外在展示上，我没有很好的区分给客户讲和给专业人士讲的分别，因为之前看了很多经验都是讲第二部分模型问题的，所以注意力全在挑毛病上面了，没能够先简要总结现有模型。

就是假设你是一个数据咨询公司的咨询师，面试官是你的客户，一位business manager，假设他不懂统计和模型，他给你提供另一个数据咨询公司做的分析，是大概10几页slide，里面有各种分析图表和一个预测模型，让你给他讲一下这个分析都做了什么，根据它提供一些解决delay的思路，同时评价一下这个分析做的好不好，不好的地方提出改进思路。给你15分钟自己看材料，然后25分钟给他讲。讲的时候我是把材料一页一页都过了一遍，以咨询师的角度，抓住几个重点，1是讲解材料内容解释数据图表和模型，数据中不合理的地方要指出（如异常值）；2是发现问题提出改进，分析做的不好的地方，没意义的图表，模型的缺陷等等，提出改进办法；3是要时刻为客户着想，通过手上的材料，客户可以采取哪些行动和尝试来减少delay。

具体来说模型的问题在于，模型选择不当（应为分类器而非回归），数据清理，week of day处理不当，correlation严重，起降地没有被考虑，温度这个变量没啥用等。另外问了你r square 和adjusted r square是啥，p value，为什么r square低，共线性怎么处理等等。另外有一些现象出现的原因你要会和实际结合起来。

role play 主要是要记住时刻为客户着想，present的时候不要一直想着自己说，多问问他有什么问题，听懂了没有， 这样可以显示你为客户着想的一面，我觉得这个挺重要的。

AdjustedR square, P value, VIF, Correlation
The **adjusted R-squared** is a modified version of **R-squared** that has been **adjusted** for the number of predictors in the model. The **adjusted R-squared** increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.
Every time you add a independent variable to a model, the **R-squared** increases, even if the independent variable is insignificant. It never declines. Whereas **Adjusted R-squared** increases only when independent variable is significant and affects dependent variable.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

Multicollinearity generally occurs when there are high correlations between two or more predictor variables. In other words, one predictor variable can be used to predict the other. This creates redundant information, skewing the results in a regression model. Examples of correlated predictor variables (also called multicollinear predictors) are: a person’s height and weight, age and sales price of a car, or years of education and annual income.
An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables. If the correlation coefficient, r, is exactly +1 or -1, this is called perfect multicollinearity. If r is close to or exactly -1 or +1, one of the variables should be removed from the model if at all possible.

Multicollinearity makes it difficult to gauge the effect of independent variables on dependent variables.

模型选择Regression / Tree: 关心delay超过8分钟与否

1. 模型选择不当（应为分类器而非回归）：公司只关心delay<8分钟与否，不care具体时间
2. missing data: 一部分不是missing at random，only missing in one category，所以不可以直接impute。
3. misinput value / negative value: Min of No. of seats = -1
4. categorical variable: Day of week should be categorical variable, not continuous variable
5. multicollinearity: ground attendants, gate attendants, passengers on plane, seats on plane, plane type are highly correlated feature之间有correlation，还有就是根据图标判断哪些feature对于predict response没太大用
6. 缺少departure location的变量，两个城市没有分开来算之类的, 起飞地点他没有区分，后面有一个图可以看出不同起飞地方增加地面服务人员影响不一样
7. 没有加入飞机上面实际座位的影响，加这个进入predictor就可以了。
8. dataset是imbalance的，如何选取evaluation indicator

然后他是放在一页有四个类似于confusion matrix的ppt里所以大家看到那页的时候记得注意一下。
但其实讲的过程中还是被问了很多technical的问题，比如他的regression做得怎么样，要不要添加或者修改自变量，anova怎么interpret。最后根据分析的结果给manager提建议，如何降低delay率。只有一个回归表格里的VIF不知道什么东西, 注意解释一下VIF，最后是anova table，注意解释一下F test significant 但R2很小这种情况。
the significant P value indicates that you can reject the null hypothesis that the coefficient equals zero (no effect). The good news is that even when R-squared is low, low P values still indicate a real relationship between the significant predictors and the response variable.

注意1: 开始是要假装对方是你的client，要简短而且容易理解的方式讲一下这个model
注意2: 最后需要说出自己对改进步骤对意见和建议

遇到的问题，multi collinearity，what else features will be helpful，这里r-square 极低，大家看到一定要看到，给出建议。
weekday encoding的问题，还有intercept的问题，你要会解释一下

Correlated features, in general, don't improve models (although it depends on the specifics of the problem like the number of variables and the degree of correlation), but they affect specific models in different ways and to varying extents:
For linear models (e.g., linear regression or logistic regression), multicolinearity can yield solutions that are wildly varying and possibly numerically unstable.
Random forests can be good at detecting interactions between different features, but highly correlated features can mask these interactions.

1. 报告里面给你的delay 其实有俩种target, 一个是binary的， delay or not; 一个是continuous的， 是delay的多少时间， 你做的第一件事情，是看看variable list里面到底是binary那个continuous. 因为我直接犯了低级错误， 我看他建的linear regression模型， 直接默认他用得continous target, 结果他用binary target, 建立的linear regression 模型。我没有第一时间发现， 肯定扣很多分。
2. 在模型结果那一页，他会要求你用plain language 解释各种metrics
3. 在variable relationship那些图里面， 有的variable 是高度线性相关的， 他会问你到底选哪个variable入模型， 为啥 （我觉得我答得并不好， 地里的朋友们可以来答一答）
4. 有一个环节， 是让你keep 要的图表， table, 我漏一张图表， 怎么说， 是个很隐含的考察你对regression 模型的假设，就是说linear regression, 一个基本前提就是（有好几个基本前提）predictor 和target的关系要线性， 其中一个表， 一个variable 明显跟target有关系， 但不是线性关系， 你得指出来， 并且说说如何解决那个问题.

我实际遇到的问题：
* 在共线性上拘泥太久，但实际上面试官的意思是airplane type基本可以涵盖很多共线变量，所以#passenger | #gate attendance 等等都可以不要

Assumptions of Logistic Regression
Logistic regression does not make many of the key assumptions of linear regression and general linear models that are based on ordinary least squares algorithms – particularly regarding linearity, normality, homoscedasticity, and measurement level.
First, logistic regression does not require a linear relationship between the dependent and independent variables. Second, the error terms (residuals) do not need to be normally distributed. Third, homoscedasticity is not required. Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale.
However, some other assumptions still apply.

1. First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
2. Second, logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.
3. Third, logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.
4. Fourth, logistic regression assumes linearity of independent variables and log odds. although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
5. Finally, logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10*5 / .10).

第二个是Role play，还是飞机晚点问题，但午饭的时候我问那个人他说可能马上就换了。这个问题我因为见过，所以以为自己应该挺顺利的，但是当我说到有multicollinerity的variable应该从model里移除时，那个面试官一直在追问我为什么，然后我说correlation会导致variance增加，p-value不显著等等，bias estimate【难道不会影响estimate么？】，而且这些variables are telling same story。但是感觉面试官对这个回答并不满意，一直在追问，但是我学过的都是vif>5就移除啊。。。不是这样么？难道我要用PCA？但这个模型显然并不需要用PCA消除correlation这么麻烦啊。。。有什么别的方法么？求教。
我觉得第二个题的他想听到的应该是，存在multicollinerity会影响到我们对model的interpretation。
楼主可以看看

https://onlinecourses.science.psu.edu/stat501/print/book/export/html/346
Effect #1和#2
对于multicollinerity的问题，我才机器学习的角度看，是觉得添加了不必要的变量，会使得整个model overfitting。

然后就是feature selection啊， 他还都问我了如何做feature selection吧， 我记得我说的用了Lasso，而且还说了这个步骤很重要，而且还要注意bias variance trade off----他还专门指出 我说的这一步其他面试的人都没提到，他觉得这是个good point. 然后又问了些Lasso啊 最后要检验模型好坏啊 验证啊 CV啦之类的。

经典题，flight delay问题。我选的是用decision tree分析，可能在role play的时候条理不够清晰吧。关于模型的问题，主要是variable definition，还有misinput value，比如说# seat 不可能有-1这种数值。还有一点是target variable是delay的时间，因为manager关心delay是否大于8分钟，所以在define target avariable的时候，个人觉得应该把delay按照8分钟来分0,1两个数值。最后的模型ROC的值只有0.58，不到0.7，这就比较低了。

一轮是统计分析，给你一些数据，是航空公司的，让你用问你什么模型来估计迟飞率，其中coefficient代表什么，pvalue怎么看，R2 怎么解释，correlation table怎么看。最后问你有什么方法提高

第３个是role play.分析如何把航空公司晚点的概率降低。给了一个regression的数据和图表，从correlation matrix看里面有些东西有strong correlation,然后他们直接去fit regression，然后你自己我去看这个东西有没有问题以及如何用它来指导business unit，我分析的问题有。１，它把day of the week(mon,tue...)当成了continuous variable,肯定不行，改成categorical,２。公司只关心delay<8分钟与否，不ｃａｒｅ具体时间，所以这个就把ｒｅｇｒｅｓｓｉｏｎ改为了logistic regression.３。它把温度作为contiuous ｖａｒｉａｂｌｅ，其实没有必要，你只要看是不是温度很差就可以了，所以自己设threhold把温度也变成categorical vraible. ４。起飞地点他没有区分，后面有一个图可以看出不同起飞地方增加地面服务人员影响不一样，所以可以分开做两个logistic regression ｍｏｄｅｌ，因为一个是la,一个是nyc，他们的温度变化会很不一样。５。没有加入飞机上面实际座位的影响，加这个进入predictor就可以了。６，还有就是他的图表有可能有些数字不合理，比如座位书＝－１，你有空就看一下。然后就照着上面的这些东西自己重新建ｍｏｄｅｌ就完了。然后他还顺便问了一下有那个strong correlation你咋办，都是标准答案，然后有补充了一句如果你一定要把拿两个factor都放在Ｍｏｄｅｌ里面你该如何搞和如何解释，都可以从书上找到答案的，就不贴出来给大家一点悬念了。

温度这个变量没啥用等。另外问了你r square 和adjusted r square是啥，p value，为什么r square低，共线性怎么处理等等。另外有一些现象出现的原因你要会和实际结合起来。被问到model各个parameter的意义，auc什么意思怎么处理correlation，如何improve model。

问了correlation，p-value，VIF
当我提出不合理的， 他让你解释概念， 比如p-value, r2, multicollinarity, 部分graphs.

Remove correlated features:
1. Save training time and grid search time
2. Reduce the complexity and variance / overfitting
3. Improve performance
4. Detect interactions between features

If your dataset has perfectly positive or negative attributes then there is a high chance that the performance of the model will be impacted by a problem called — "Multicollinearity". **Multicollinearity** happens when one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy. This can lead to skewed or misleading results. Luckily, decision trees and boosted trees algorithms are immune to multicollinearity by nature. When they decide to split, the tree will choose only one of the perfectly correlated features. However, other algorithms like Logistic Regression or Linear Regression are not immune to that problem and you should fix it before training the model.

How Can I Deal With This Problem?
There are multiple ways to deal with this problem. The easiest way is to delete or eliminate one of the perfectly correlated features. Another way is to use a dimension reduction algorithm such as Principle Component Analysis (PCA).

Multicollinearity is mostly an issue for multiple linear regression models. There, it can cause a variety of issues, including numerical instability, inflation of coefficient standard errors, overfitting, and the inability to accurately isolate and understand the effects of individual features.
For nonparametric models such as decision trees, it's possible that there might be some danger in overfitting the model if the level of correlation between features in the training set doesn't generalize to unseen data. This is less of an issue for tree ensembles, especially when feature bagging (independent random samples of feature subsets used to build each tree) is employed, e.g., Random Forests - which always employ this mechanism - and most modern Gradient Boosting implementations. But for the most part, multicollinearity, or any other feature correlation structure, is really only an issue for trees when estimating individual feature importance. But if you only care about model performance, it's generally not a concern.

It depends on what you need from a decision tree.
If you want to use a decision tree to extract the feature importance for further analysis, Multicollinearity may have some bad effects, it makes the coefficients (feature importances) unstable and incorrect.
If you just want to have the prediction, then multicollinearity does not affect the result.

Decision trees follow the non parametric approach.
As the decision at each node of the tree is made based on the single feature ; Mutlicollinearity doesn't affect in decision trees.Though single tree leads to greedy algorithm if the data is skewed or imbalanced, ensemble learning methods as random forests and gradient boosting trees make the prediction robust to the multi collinearity .
So , no the multi collinearity will not be a problem in prediction using decision trees.

Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch. Pruning is useful because classification trees may fit the training data well, but may do a poor job of classifying new values. A simpler tree often avoids over-fitting.

Case Interview
Online bank: 面试内容主要就是Capital One要併购一家online bank，好处跟坏处是什麼。然后就是让你看部分代码，搞清楚这代码是干什麼的，怎麼改善。c1早期和另外一个网上银行合并的好处和坏处是什么。这个问题我在glassdoor上看到好多次，所以就提前准备了一下。这个网上银行就是现在的c1 360。我就说了record怎么match啊，consistency啊什么的。 第二部分就是看这个网上银行以前的代码（三个boolean parameter，glassdoor上也提到了），然后简化，并且告诉对方如果要继续用这个代码，建议怎么做。这个case我准备过，很快就讲完了，接着就是尬聊...

case: 跟analyst 做的case interview 很不一样...先告诉你, 公司买了一家做别的服务的公司，这样business 和 technical上有什么好处. Merge 两家公司的system 会有什么challenge.
然后给一堆奇怪的code, 一个判断客户能不能开账户的function, 要找出bug/problem, 简化它。对这code 你会给什么suggestion 给你的manager。然后说现在其中一个input不再是true/false，而是有多种可能，那怎么修改刚刚的code。写完最后问如果你到一个新的team，what is the first thing you will focus on (in the first few months/weeks).

第三轮 面试官先说假如说capital one现在只有credit card的服务，现在买了一家银行带有其他的服务，这对银行和技术方面各有什么好处。然后拿出一张纸上面是一堆code，大概意思是一种判断对方能不能开账户。然后让你解释一下。接下来问你能不能优化，怎么优化。最后问，因为里面所有input都是boolean，现在其中一个不能直接用boolean来表示，因为存在更多的可能，你觉得应该用什么data structure来存。最后问现在假如你已经写完了这个code，你需要和哪些人去沟通交流。 我解释的有点乱，我感觉最后一个主要考验你怎么去处理问题。

mobile app of game: 第一轮是case 因为当时我们program已经有两位同学拿到了C1家实习的offer，而且据说面的都是life insurance的case

所以当时很心大的基本只准备了这个。。。结果一上来就是一道app的题。。。然后就只能硬着头皮上了。一上来先问app有什么赚钱方式，我说的两点是买app的钱和in-app purchase，他问还有呢,我思考了一小会, 面试官很好心的提醒了说 “想想instagram” 然后我就马上反应过来是广告费。然后他给了两列数字: 一列是 free app的，一列是paid app的。数字包括user number，ad rev per user，cost per user之类的大概六七行数字让算profit (注意这里是profit 不是breakeven哦) 挺简单的。 然后问应该选哪个。 因为记得free app的profit是400,000， paid app是600,000，所以肯定选paid 啦. 然后面试官说ok 假设我们现在只有200,000个free用户 (之前是200,000个free用户 和50,000个paid用户)，现在要多少个人去paid才能达到之前600,000的profit。当时有些不太理解就问了面试官 “可是只要paid app用户少于之前的50,000，那肯定会比600,000的profit少啊” 然后我忘了他说了句啥我就反应过来：要找的是两种app在新的用户数的情况下，加起来的profit是600,000的那个点，算两个app的用户分别要有多少。这边因为紧张算错了几次… 而且因为是视频面试, 脑子里一直想着不能安静太久不能让气氛变尴尬之类的所以就出错了。 但是面试官最后很nice说没关系不用担心，所以大家一定要记得不用太紧张～ 算完之后关于insight他就会问有什么提高profit的方法。我只是很框架的从增加revenue和减少cost的角度去分析。但是面试官其实想要说的是 paid app user到了free app以后 他们的purchasing behavior不会变，所以其实free app 的in-app purchase那块的revenue会比原来的数字更高 以此来到达比600,000更高的profit。他会慢慢引导你说出答案，所以也不用担心一定要一次答对

1. 什么因素衡量app是否值得发布, 跟上边的差不多，所有case都是revenue，cost，再扯点其他的就可以了
2. 给了#users, download price, in -app purchase per user, Ad rev per user, cost..将近10行的数据吧，2组产品free 和 paid的，问哪个更盈利。这个就是revenue-cost的题啊，很简单吧，别着急，复杂的在后面
3. 如果要同时发布两个产品，会发生什么。 当然是用户flow到另外一个组而影响profit了
4. 如果同时发布，问多少user从paid APP flow 到free app才break even， 这个计算的假设是什么
5. 难题来了，现在就是三个组了，free, transfer to free from paid, stay in paid 这三个组。 先计算free组的 ad rev per user, 然后更新每个组的profit，然后重新计算上题的break even point
6. 根据你的计算，如果要创造最大的profit, 那你有什么建议。f

这个题从第5题开始，LZ就蒙了，而且由于是video面试，增加了面官给我纠错的难度。虽然最后在面官的引导下回答出了正确答案，想想还是心慌慌的

是要达到一个数，我记得是$60K，然后有X的user从paid app跑到free app， 所以最后的公式是 profitOfFreeApp * (#originalFreeAppUser + X) + profitOfPaidApp * (#originalPaidAppUser - X) = $60K, 其他具体的数我就不记得了。

Auto Loan: Case Interview: Auto Loan
主要问了以下几个问题.
- Auto loan的revenue & expense:
Revenue主要是interest, collateral, 各种feeExpense主要是Bad loan
- 你会选什么样的模型predict bad loan
- 计算一: 已知P(true bad), P(predicted bad), P(true bad | predicted bad), 求P(predicted bad | true bad)，这个值对应confusion matrix的哪个值，是越大越好还是越小越好
- 计算二: 已知不变，求P(true bad | predicted good)，这个值对应confusion matrix的哪个值，是越大越好还是越小越好
- 如何改进模型：只需要说的比较笼统，比如增加/减少feature，考虑数据的imbalance等

Case interview：autoloan，会问revenue和cost有包括哪些，然后告诉你有一个已有的模型，让算算条件概率，然后问你这个模型好不好，然后如果让你重新设计一个模型，该怎么做。其实这个case已经用了好多好多遍了，但是感觉每次大家被问到的小问题还是不太一样
这个问题之前是 假设P(True bad customer) = 3%, P(predicted bad) = 5%, P(True bad|predicted bad) = 15%, 先求P(True bad|predicted good) = ?. 然后求出来的P跟3%差不多，所以觉得这个model不好，然后就说该怎么重新设计model。其实我当时也是有点懵的，就说了一堆feature engineering，试什么model，cross validation, ROC之类的，但感觉没答到点子上，那个interviewer就一直问我 还有呢还有呢，我就很懵逼 -。- 对了 条件概率那个地方，她还问了 说如果我们random猜一个customer是good还是bad，那P(True bad|predicted good) =?

Case Interview:
the case is the car finance loan.
- what are revenues and expenses
- given a model that predicts when a customer is good (loan should be approved) or bad (loadn should be decline) find out: 1. the probability that the customer is good given the model predicts good 2. the probability that the customer is bad given the model is good 3. given a pentile graph of # of checked off loans / # of loans what is a better model than the current; what is the best model.

Credit Card: 第三面business case：这个case是假设你在一家巴西的类似沃尔玛的公司工作，他们想发行Store Credit Card。如何决定要不要发行这个卡，根据盈利率和客户群体等考虑。这个还蛮简单的，面试官很nice，一直在帮忙引导我。C1也会给准备材料。这种类型的问题没有标准答案，因为公司主要考察business sense。只要你说得有道理，面试官就会跟着你继续聊下去。
但基本上是interviewer led 的profitability 的case，找均衡点的。第一问都是问你busienss sense，比如需要考虑哪些方面等等。这部分我建议用case interview 的profitability framework，先说考虑revenue 和 cost， 然后里面再说的细一点，尽量做到MECE。第二问基本上都是math，分析均衡点。数学比较简单，但要心细。最重要的是尽量不要用short cut，要把interviewer 当傻逼，一点点地喂。我的做法是把问题分成一步步，每一步算一个变量。比如求profit，你先告诉interviewer我要用revenue-cost，然后告诉他我要算revenue，算完了再告诉他我要算cost，总之就是把每一步都解释清楚。不要自己在脑海里憋着算。如果你练习过management consulting 的case，capital one 的要求类似，但程度应该是没那么严格。毕竟面你的人不是consultant，对softskill的要求没那么高。一般第三问往往就是和你讨论各种alternative situation，比如把这个变量变成2倍会如何等等。有时候会让你画图。最后会让你出个receommendation，你就照着consulting 的标准就行了。附件里是我网上找的题，我觉得帮助挺大的，尤其是如何计算信用卡收入支出这块，很有帮助。
最后是behavior， 你找找glassdoor的例题就行，但要注意interviewer会问的比较深入，所以如果你编了一个例子，最好把细节想好。。。

第三轮CASE, 巴西超市要发credit card了， CEO问咨询师你觉得我们要不要发呢？ 是marketshare市场考察咨询题。 第一题考察你对credit card的商业了解，revenue是啥， cost是啥来热个身。一定要自己Google好。
计算 profit，revenue，cost，market share，计算超市信用卡的 market share。 注意他直接问你market share， 不会告诉你任何条件，你要自己问城市的人口总数，成年人比例，和人均信用卡持有数三个数据这些可以帮助你计算的。然后算breakeven。算出来market share 5%，问你那要不要开发这个项目呢？ 如果开发的怎么吸引客户呢？ （NMD， cash reward啊！我居然没想起来。） 两个面试官, 一个主要面试官低等级一点，另一个全程姨母笑的高级面试官，calibratiing前一个问我问题的面试官的。。为什么给我这种令人紧张的节奏（嗯？嗯？？？）

基本是。我面的时候第一个case是分析银行推出一个新的服务，做了一个实验，分treatment group 和 control group， population 分两类人，根据给的数据分析出每类人消费变化。第二题是冰淇淋，关于找到最优数量的制冰机和搅拌机。影响因子有机器费用和客户需求cap等等。算是profit 问题的一个变体吧。根据我的体验，附件pdf的很有代表性，我觉得你把附件里的case研究透了，on-site的时候再心细些，就ok了。剩下的就是注意表达和交流，这些你可以看看consulting的case 材料。

第一轮business case，场景是超市发放private credit card，有一些上一年的历史数据，问题涉及计算 profit，revenue，cost，market share，计算 market share 的时候要先计算全城有多少信用卡，面试官不会一下子把数据都给你，你要想计算时需要什么数据，考虑多种情况，同时问面试官某些数据有没有，比如说计算全城有多少信用卡就需要全程人口总数，成年人比例，和人均信用卡持有数三个数据，这些都需要问面试官才会得到。另外最后会有开放性的讨论，就是计算出一些结果，问你根据这个结果要采取什么样的行动，这个就比较靠business sense，要讲出道理。

一个巴西超市要发credit card, 首先问credit card可能的revenue/cost来源， 之后问要break even的话， market share是多少，面试官不会给你全部数据， 你需要自己定义market share是什么， 有可能是营业额， 有可能是多少张卡， 不同的定义需要的数据不同， 我说我选# of cards，然后我说我需要全巴西有多少张卡，巴西有多少成年人， 成年人中有多少人是credit worthy, 之后他就给了我， 然后就很快算出来，之后又有很多follow up，只记得一个是怎么提高revenue云云， 反正第一面面试的很愉快。

Short-term loan: 第一个Case是很火的Short-term loan，因为最近的面经说有考到所以我着重准备了一下，所以问到consideration的时候回答的非常顺利，我刚美滋滋的时候第一个数就算错了，面试官神神秘秘的说，it’s a bit off，重新算了一下算对了，-11M，这个数我记得很清楚，因为面试的时候也不知道自己到底算没算对，出来之后和印度姐姐对答案她说她也是-11，松了一口气。然后问为什么亏这么多，怎么办，我就说rate低，没有collateral之类的。第二部分大同小异，改变了几个条件，再用同样的方式计算，边算边说自己的思路，算出来的结果是4.5M，然后算了breakeven year。第三部分，给出了另外的两个option，对每个求breakeven然后画出他们profit，问你觉得哪个选项最好，言之成理即可。这一轮的interviewer全程都保持着神秘微笑，也是我心里最没底的一轮，后来和一起面试的印度姐姐聊到他，发现这个面试官就是这个神秘风格，听说会有interviewer故意不给回应，给你营造个有压力的氛围，看看你怎么应对。

第一个case是经典的90天short-term loan。面试官简单的介绍了case的背景，有一家公司需要90天的贷款。
第一问：如果我们想要给这家公司发行这个90天的贷款，需要考虑什么factor?
第二问：给了customer base, loan amount, operating cost, marketing cost, 需要pay 5% interest rate的客户的百分比，3% interest rate 客户的百分比，以及charge-off 客户的百分比，让算profit，我算出来是-11m
第三问：问为何profit是负的
第四问：给了新的数据，重新计算profit
第五问：有一个a公司和b公司，我们和他们合作能挣钱。给了a公司和b公司第一季度和第二到第七季度给我们带来的profit，问选择和哪个公司合作，或者是不和两个公司合作，需要阐述原因。
整个case就结束了，计算很简单，最后一个阐述合作原因我觉得我有点语无伦次，面试官也有分享他自己的想法。

一家小企业想申请90day 短期贷款，公司目前没有这种短期贷款，问你是否应该开拓短期贷款的新市场。计算就是给你每个quarter的总贷款额，利率（分为两种，一种是到期偿还的利率，另一种是提前偿还），然后告诉你x%到期偿还，y%提前偿还，z% default。计算quarterly profit。算出来为负数。
然后加入新的计算条件，大概就是说另一家企业也想申请贷款，给出新的利率，计算profit和breakeven。
最后就是现在有两个公司和你谈合作，告诉你合作之后每个季度的营业额变化，要你选一个。没有正确答案，分情况讨论。

Venmo: 第二轮Case也是很常考到的Venmo，直接计算，告诉你fee structure计算第一个月的profit，这一问计算方法和上一题一样，注意一下年化的问题就好了。算出来也是负的，问原因。第二部分是你的朋友提出跟你合作，算一下新的profit。第三部分和前一个case一样，公司两三年之后要进行IPO，而且有两个新的option，和上一轮几乎一样，先画图再问选哪个，我先明确了一下了time horizon，说IPO的话需要present positive financials，三个都在IPO之前break even，之后就看这三个是不是exclusive，如果不是就都投，是的话就选一个。这一轮的面试官风格就很不一样，他会在计算的时候一直夸，然后每一个部分结束都会对回答进行一点评价，他会经常说I like your point that

Life insurance:（这个好像以前有人发过。。但是每次问的问题好像有点不同）
先问了一下你认为有哪些人我们不愿意轻易issue insurance给他们 （老人啊，有严重疾病的人，高危职业从事人员， etc）
再问了一个death probability的问题，算一下在这个prob在什么时候我们可以考虑给保险。最后一个问题没记太清楚，大概是说各种dead prob的group，怎么样选择能使得profit最大。

case interview 是关于life insurance的，之前地里有人提过。先问你对insurance了解有多少，如果你是manager，为了考虑到顾客死亡的概率你会收集哪些顾客的信息。然后就是计算的部分了，很简单小学数学题。算达到break－even的死亡的概率是多少。然后面试官画了一个柱状

图，有四个组，分别代表high risk low risk，median high 和 median low， 柱状图给出每组的死亡概率，然后有四种方案，1:全包括，2:只包括low risk，3:包括low 和median low 4:包括low, mdiean low 和 median high 问你会选哪个方案（假设每个组的人数一样）。如果想更赚钱你会怎么做。最后是opening question，如果你想去predict 死亡的概率 你会用什么方法，为什么。

Issue 一个 life insurance 要考虑哪些因素 premium, term, death rate, target customers, marketing and operation cost, competitor
然后大概给了上面的revenue 和 cost， 5、6个数据的样子，计算一年和6个月的death rate to break even.
给了 ABCD四个组的 death rate, 问应该issue给哪个group combination, 并以x-axle 为 groups (A,AB,ABC,ABCD) y-axle为profit绘折线图，注意slope 和 最后profit要落在0以下
如果硬要issue给高death rate 的人，你有啥建议to make profit

The **charge-off rate** is the amount of **charge**-offs divided by the average outstanding credit card balances owed to the issuer. **Charge-off** is actually an accounting term that means a company has decided it has no chance to collect a debt and charges it **off** its books.

revolver 和transactor: 客户分成两组，revolver 和transactor, revolver不能及时还清bal, transactor可以。
问题1， revenue stream是什么?
问题2，算revolver的profit,给了一组数字，但是记不得了。计算非常简单。
问题3，给transactor的一组数字，问transactor的annual balance是多少能和make same amount profit as revolver
问题4，什么样的情况会影响revolver和transactor?
中间还有一些小的business conceptual quesition一句带过的那种，就记不得了

我也遇到了transactor v.s revolver的case， 但是我完全不懂finance里的APR，charge off rate， delinquency rate这种term，而我是个ds，真心不懂考这个的意义，而且楼主竟然这个case写出来了很佩服， 楼主也是给了一推rate然后算profit吗， 至今不懂那个cost fund rate （cost of default rate）跟average balance怎么联系起来

另一个business case是关于信用卡计算transactor 和reserver的profit以及提出建议的。数学比较多，但是我之前在地里看到过类似的case
就是所有的信息都需要你自己想，问一张信用卡的profit和cost有哪些，对reserver和Trans actor有什么不同，然后reserver的话，每个月都有balance x的前提下去计算funding cost是多少，以及后面的gross profit之类的

Coding: 你是个manager, 你手下3个人，coder, tester, deocumentor. 现在你们两周生产1000 lines of codes, 有个新的客户，这个客户想和你签个合同，两周要再多加1000行codes.
问题1， 要不要签这个合同，考虑什么因素?
问题2: 给你这三个人的效率， coder写15 lines/hr, tester 2 lines/5mins, documentor 1 line/2 mins. 问你两周的capacity 是多少? 你发现了什么问题?
问题3: 现在2周生产2000 lines, 要多少小时over time?
问题4：算成本，假设over time钱24/hr, 平时base pay rate 16/hr, 原先2周生产1000lines和现在2周生产2000lines的成本各是多少? per line的成本是多少?
问题5: 为什么1000 lines的per line成本高?
问题6： 你觉得之前算的overtime合理吗? （coder 2周要over time 53.33小时，不合理）， 不合理怎么办? （再招一个人）
问题7: 再招一个人招ft还是contractor? （算成本， 两个一样）那你觉得招ft还是招contractor好?
然后面试官说可以考虑cross train,其实他一开始说假设不能cross train了呵呵呵呵

Amusement park: 第二个case是关于amusement park的 感觉地里应该有的 关于卖三种不同票 一种是一天票一种五天票一种全年票 怎么计算全年票的人每年去几次... 然后再各种revenue cost 算profit 面试官很耐心 有哪里算错了会马上帮你指出来让你改正

第一部分是一个case interview 关于一个amusement park 说今年的revenue减少， 问你可能是什么原因．然后给了market share的一个公式， 还有计算一下如果改变门票价格 如何能breakeven 还有问你有什么建议
要注意的是 改变价格的时候 之间给的market share 公式不再适用 要自己想其他办法来计算 ，期间Interviewer 有correct 我几次， 最后也是把所有问题做完了

第一轮CASE, amusement park， 地里有详细的我只记得框架。 这公园的CEO问你（你是consultant）建议和策略。 公园有三种票， 每种价格不同， 第一题问游乐园的revenue和cost。这属于profit类的case，网上很多分析框架， 看完可以加入自己的脑洞。开放题，除了计算。 第二题， 让你算breakeven。 第三问如果三种票价变了，怎么breakeven， 第四问给你FINANCIAL STATMENT和不同的assumption再算一遍，问你照你现在的计算结果，CEO要不要扩张土地。

记得使用margin作比较而不是profit

Profit margin is calculated by dividing the net profits by net sales, or by dividing the net income by revenue realized over a given time period.

Gross margin (%) = (Revenue – Cost of goods sold) / Revenue

Wine Store: 关于local的一个卖酒的商店的。会给很多的图表。每一年的gross revenue，average profit margin 以及基于channel的不同的profit margin。以及other costs
先算今年的profit。
然后说怎么样可以increase profit，前提是有一个更influential的competitor
然后给一个competitor的财务表之类的，让你看里面有什么数据有问题，就是profit margin偏低，然后问你为什么
最后是给一个不同的marketing expense增加的sale，问你哪一个点是optimal

Job Specific Interview

印度小哥 job fit (team match?) 他们做fraud detection的，互相介绍完问了简历的一些细节，然后就是bq，挺多奇奇怪怪的问题。比如问你碰到过什么hard machine learning problem最后怎么解决的，比如一个银行想用一个摄像头加人脸识别来判断进来的客户是不是危险分子，你觉得这个计划有哪些潜在的问题，还有啥我记不清了。现在想想其实答得很一般，不太会答这种奇怪的问题。。。

Closed circuit television systems (CCTV)
We propose a CCTV based theft detection along with tracking of thieves. We use image processing to detect theft and motion of thieves in CCTV footage, without the use of sensors. This system concentrates on object detection. The security personnel can be notified about the suspicious individual committing burglary using Real-time analysis of the movement of any human from CCTV footage and thus gives a chance to avert the same.

According to the "CCTV Operational Needed things Manual 2009", the job of the CCTV operator is to watch (for changes, unusual things, etc.) and detect, control, recognize, watch/ notice/ celebrate/ obey, and identify situations and people that are possibly harmful to other people and their property.
A solution to these problems is to apply image-processing and algorithms to the CCTV footage, which will replace the human operatives and alert the security officials if a terrible situation is found

In the previous few years, deep learning and Convolutional Neural Networks (CNNs) have accomplished best results to all the classical machine learning methods in image detection, classification and clustering, etc. Deep learning CNNs automatically discover more and more higher-level features from data, instead of manually selecting features. Using CNN's, we focus at building a good weapon detector in real time

A great deal of research has been performed for automatic object detection and curbing the situations such as bank robberies and ATM tampering.
1. Real-time Processing
CNN contains a lot of interconnections and complex mathematical computations which takes a lot of processing power and computation time. The accuracy of the video dataset is directly dependent on its computation time. But concerning the system to be a real time we will need to compensate the accuracy for better computation time.
2. Video Resolution
The currently available video datasets from YouTube and other social networking sites have lower resolution to be qualified as a good dataset. So, object detection and classification is quite difficult from such a dataset.
3. Object Orientation and Positioning
4. Non-reusability
5. New Dataset

Coding

问3个LC简易编码问题，＃155，＃347，＃482
Leetcode easy and medium level, two questions, plenty of time.
1. string manipulation.
2. min number of moves to sort a list of int, making even numbers on the left and odd numbers on the right. (use two pointers)

1. LC死要而
2. LC贰零叁
3. LC雾散

Behavior interview

- Describe a situation in which you were able to use persuasion to successfully convince someone to see things your way.
- Describe a time when you were faced with a stressful situation that demonstrated your coping skills.
- Give me a specific example of a time when you used good judgment and logic in solving a problem.
- Give me an example of a time when you set a goal and were able to meet or achieve it.
- Tell me about a time when you had to use your presentation skills to influence someone opinion.
- Give me a specific example of a time when you had to conform to a policy with which you did not agree.
- Please discuss an important written document you were required to complete.
- Tell me about a time when you had to go above and beyond the call of duty in order to get a job done.
- Tell me about a time when you had too many things to do and you were required to prioritize your tasks.
- Give me an example of a time when you had to make a split second decision.
- What is your typical way of dealing with conflict? Give me an example.

- Tell me about a time you were able to successfully deal with another person even when that individual may not have personally liked you (or vice versa).
- Tell me about a difficult decision you have made in the last year.
- Give me an example of a time when something you tried to accomplish and failed.
- Give me an example of when you showed initiative and took the lead.
- Tell me about a recent situation in which you had to deal with a very upset customer or co-worker.
- Give me an example of a time when you motivated others.
- Tell me about a time when you delegated a project effectively.
- Give me an example of a time when you used your fact-finding skills to solve a problem.
- Tell me about a time when you missed an obvious solution to a problem.
- Describe a time when you anticipated potential problems and developed preventive measures.
- Tell me about a time when you were forced to make an unpopular decision.
- Please tell me about a time you had to fire a friend.
- Describe a time when you set your sights too high (or too low).

(1) A project that you are proud of;
(2) Time about helping other people in your team
(3) Time about failure and mistakes.

第二部分是行为面试部分，问了三个问题，第一个是你最大的成就是什么，第二个是讲一个你帮助别人的经历，第三个就是你最引以为豪的项目经历（因为楼主有过实习经历，因此就讲了一个自己记得细节比较清楚的）

三场都有 behavioral question，非常标准，面试官几乎就是照着念出来题目，按照 STAR - Situation, action, result 原则回答就好，准备的时候可以想下这个问题问的是那个方面，在 result 里突出就好了（楼主是按照 brief，situation，action，result 准备的，有个 brief 是为了抓眼球）。地里可能都是大神，但是还是想建议，尽量减少 umm 或者手势，让自己看起来自信的叙述故事和结局，楼主以为自己以前做的不错，对着镜子练过和录音以后，简直想打脸。最后建议 behavioral 一定卡下时间，别有冗余说清楚就行，把时间留给 case。碰巧都准备到了所以答得很顺，后买呢follow-up就是问一些这些事情对于你以后做事有什么样的影响。而且面试官也有强调最好用STAR原则

BQ：achievement， help people/teamwork，failure
三个 behavioral questions：
Tell me a time about -
1. recent failure or mistake you had in work
2. help others
3. a project that you are most proud of