Paypal

基本上大部分问题都是围绕简历以及工作中做过的项目。挑选一个最近做的项目，参与了哪些步骤，以及每一步参与的工作内容。
从项目需要解决的问题，到收集数据，数据分析，变量构造，变量选择，模型选择，模型对比，以及最后的模型实现。

1. 如何鉴别facebook上的fake user at login? What feature is available at the moment of login?
彼时我就蒙蔽了，面的又不是脸家，干嘛问别人家的问题。仓皇之下随意答了一波。
2. 如果categorial variable有很多值，应该如何处理？

去年面的，Paypal的DS主要是做risk和business相关的一些modeling，production有专门的码农帮忙搞。
电面的时候，问的问题比较简单，都是一些最基本的machine learning相关的基本概念，类似于如何prevent overfitting之类的。

Onsite一共有6轮，从早上10点面到下午4点，连吃午饭都在要求brain-storming，连喘口气的机会都没有，简单说一下大概面了些啥：

1. 一个印度大叔，面概率为主，N个数中有G个good number，问randomly sample n个数(without replacement)，有g个数是good number的概率，what is the expectation value of g?
2. 一个中国大姐，聊了聊background，问如何在用户刚刚sign-up的时候，做fraud detection？
3. 一个伊朗哥们问了一大堆feature engineering的问题，如何做feature selection
4. 一个印度小哥，也是问如何做feature engineering的，还有一个题是说，除了手机的GPS以外，还有什么方法可以locate user？
5. Hiring manager，聊background，以及一些behavioral quesiton
6. 一个中国大姐，面coding，不是很难，问两个list如何merge之类的

最近在paypal官网投了risk analyst职位，之后就收到了HR的screening，问了简历上基本的情况，还有一些behavioral questions，HR人还是比较nice，还提醒自己要对paypal最近的development有一些了解。。。

之后HR就帮我和Risk analytics Director约了phone interview，1小时15分钟，除了一开始的自我介绍和最后我提的questions，其他部分都是case interview。。其实HR之前没有提到interview会是什么样子，我面完了之后看了Director的linkedin才发现他在Capital One工作了N年，而C1家是最喜欢用case interview的，所以我的面试就和C1的很像。。。其实就是break-even类的case

说的是一个auto insurance的CEO，需要确定how much premium are we going to charge我们的customers，in order to break even，给的data是5%的customers会submit claims，全approve，然后average claim amount是5000刀

第二问是其他条件不变，我们现在为了acquire更多customers，开始发mail给potential customers，cost是$1 per email，response rate是4%，然后我们的approval rate是50%，还是要breakeven，现在premium需要多少？ n*4%* premium = n + n*4%*5%*50%*5000

第三问是得到第二问premium的基础上，我们target profit goal是$10,000，需要发多少mail才可以achieve？ N*4%*premium = n + 10000

基本上就是这样，我觉得三问我都是拿传统的方程来算，director觉得我的答案对，但是他就希望我能有一些shortcut的方法解，这样也方便和对方communicate，不要每次都设这个x，那个y来解，啊哈哈哈所以特来求教地里的亲们有没有什么简单一点的办法~

Resume + a fraud model case

What's the advantage of random forest over GBM?

- Less hyper parameters to tune
- RF shows very low variance
- Less likely to overfit. The more trees in RF, the better. That's because the multitude of trees serves to reduce variance. Each tree fits, or overfits, a part of the training set, and in the end their errors cancel out, at least partially. Random forests do overfit, just compare the error on train and validation sets.
- easier to parallelize

Fraud case: given transactions 2010 ~ 2017 and 10k fields, build a model to detect fraud

- Explore data, plot one way correlation
- Feature selection information gain
- Use PCA to remove correlated features
- Imbalance, downsample, how to determine the downsample ratio
- Metrics: AUC, TPR (recall), FDR = 1-precision
- Model selection: logistic regression, random forest, GBM, NN
- What would you do if you see your model perform bad in production

Onsite

Logistic regression

- Logistic regression. Is loss function convex? (yes, non-strict convex) Does Lasso/Ridge keeps convexity of the loss function? (yes, ridge is strictly convex, and lasso is non-strictly convex)
- For logistic regression, why does lasso results to multiple combinations of optimal beta (non-unique solutions)? How to pick one if there are non-unique optimal solutions? (group lasso)

Stats

- What's bias variance trade-off?

- What's boosting vs bagging? Boosting reduces bias, and bagging reduces variance
- If you need a sample to build a model, are you going to sample among a year, a quarter, or a month of data? A year to capture seasonality

TF-IDF

- What's the target of the 2nd/3rd tree in GBM? What's the range, is the target still binary?
- What's the expression for TF-IDF?
- Code TF-IDF, only use python dictionary, without using nltk, pandas dataframe, numpy ndarray, or collections

Python & SQL

- Find median in 2 sorted arrays in O(logN)
- SQL or python: paypal table - (trxn_id, sender_id, receiver_id, trxn_amt). How will you present user engagement? Count number of trxns per sender. What if there are users with no trxn, need an additional table of all user_id

第一轮问了ab testing的大致流程，sample size，run periods。然后问如果让你预测一个看病的人有没有病怎么做，features自己想

第二轮是take home。匿名数据不知道column name，依然是data preprocessing做好然后试几个model选个最高的，解释下参数就好了。虽然没有变量名，但是建议还是做下data visualization。

onsite被我拖了两周。。。。因为觉得自己没准备好，真的建议大家如果没准备好，宁肯拖也不要强面，一般大公司hr都很好说话，解释好原因就好

onsite五轮，感觉遇到的三哥人很好，然后中东小哥就很bug了，，，，

第一轮三哥问了概率题，强烈建议大家把学过的几个分布都看一下，然后典型例题写了，我太久没做概率题，这个忘记了。。。 然后好在三哥人很好，很耐心给我几个小例子一步步推出来，不过要是看过分布这个很好回答的。然后regularization，因为我做过NN也问了，后半程挺顺的，反正一定要准备好自己写在简历上的模型

第二轮中东小哥。。。一上来问我会什么，然后说既然你都会了就不问了，我们问问你不会的吧 然后让我推导back propagation，还问了很多实验设计的东西，反正被虐了觉得面完就挂了

第三轮中国小姐姐人特别温柔，做了个case study，掌握大概流程解释自己为什么要这样做的原因，唯一觉得稍微轻松的一轮

第四轮三哥人还是很nice，coding题，然后优化，难度leetcode medium，不会考linked list，dp估计也不会考，大家看看二分法还有单向双向双指针的题目。然后问了他自己前天做的一个modeling case让我debug，不过他问的东西都是基于你做过的东西，所以一定一定要把自己做过的model有什么优缺点了解的非常清楚。然后不会的时候要下hint不要自己一直想，我想了两个不对要了hint，然后做出来了。交流也是工作很重要的部分。