Data Science Interview

THE BRIDGE DIGITAL TALENT ACCELERATOR

Interview Questions - ML

- **Can you explain cost function of decision trees?**
- **How will you explain deep neural network to a layman?**
- **What do you mean by model regularization and how will you achieve regularization in linear models?**
- **What is the difference between supervised and unsupervised machine learning?**
- **What is bias, variance trade off ?**
- **What is vanishing gradients ?**
- **What is a confusion matrix ?**
- **Explain how a ROC curve works ?**
- **Explain Decision Tree algorithm in detail.**

Interview Questions - ML

- **What is Ensemble Learning ?**
- **What is Random Forest? How does it work ?**
- **What is logistic regression? Or State an example when you have used logistic regression recently.**
- **What is a Box Cox Transformation?**
- **What is TF/IDF vectorization ?**
- **What is the difference between Regression and classification ML techniques.**
- **What is p-value?**
- **What is 'Naive' in a Naive Bayes ?**
- **How do you detect if a new observation is outlier?**
- **Explain the steps for data wrangling and cleaning before applying machine learning algorithms.**

Interview Questions - coding

- Merge k (in this case k=2) arrays and sort them.
- Three friends in Seattle told you it’s rainy. Each has a probability of 1/3 of lying. What’s the probability of Seattle is rainy?
- Find the maximum of sub sequence in an integer list.
- Create a function that checks if a word is a palindrome.
- How do you find percentile? Write the code for it.
- Given a list of tweets, determine the top 10 most used hashtags.
- Given a log file with rows featuring a date, a number, and then a string of names, parse the log file and return the count of unique names aggregated by month.
- How do you handle missing or corrupted data in a dataset?

Take Home Challenges (I)

TEST: PREDICTING BAGGAGE LIKELIHOOD

OVERVIEW

At eDreams Odigeo we're always looking for ways to improve customer satisfaction. With this objective in mind, we would like to predict whether a new customer is interested in buying additional baggage in order to speed up the booking process.

GOAL

The goal of this task is to predict which new customers are going to purchase additional baggage for their trips using historical from previous customers. The code can be developed in any of the following languages: R, Python or Java.

DATA DESCRIPTION

Two files are attached with the training and test datasets. The training dataset contains 50000 bookings and the test dataset 30000 bookings. The data fields are the following ones:

TIMESTAMP: *(date)* Date when the booking was bought.

WEBSITE: *(string)* Website where the trip was purchased. It is composed of a prefix that stands for the website ("ED" = Edreams, "OP" = Opodo, "GO" = Go Voyage) and a suffix for the country (for example: ES = Spain)

GDS: *(integer)* Number of flights bought through the Global Distribution System

NO GDS: *(integer)* Number of flights bought though other channels.

DEPARTURE: *(date)* Departure date

ARRIVAL: *(date)* Arrival date

ADULTS: *(integer)* Number of adults

CHILDREN: *(integer)* Number of children

INFANTS: *(integer)* Number of infants

TRAIN: *(boolean)* Whether the booking contains train tickets or not

DISTANCE: *(float)* Distance travelled

DEVICE: *(string)* Device used for the purchase

HAUL TYPE (string): Whether the trip was "Domestic", "Continental" or "Intercontinental".

TRIP TYPE: *(string)* Trips can be either "One Way", "Round Trip" or "Multi-Destination"

PRODUCT: *(string)* Bookings can contain only travel ("Trip") or the travel and a hotel ("Dynpack").

SMS: *(boolean)* Indicates if the customer has selected a confirmation by SMS

EXTRA BAGGAGE: *(boolean)* Variable to predict, only in the train dataset. Indicates if the customer has purchased extra baggage for the trip or not.

EVALUATION

The evaluation will be based on the quality and explanation of the source code as well as the prediction score. For the prediction score, the metric will depend on the output: in case a binary output is submitted, the evaluation method will be the F1 Score, in case a probability is given, the evaluation method will be the AUC ROC.

SUBMISSION FORMAT

The submission must contain the source code and the predictions for the 30000 bookings in CSV format, for instance:

ID	EXTRA_BAGGAGE
0	True
1	False
[…]	

Or the probabilities:

ID	EXTRA_BAGGAGE
0	0.35
1	0.78
[…]	

Additionally, the submission should include comments or a document file explaining the solution or any insights the candidate might have found.

Take Home Challenges (II)

Airbnb Data Science Inference Data Challenge

Airbnb is a two sided marketplace which matches guests to hosts. The booking flow at Airbnb is as follows: a guest finds an available room (listing) that she likes and sends an inquiry to the host. There are two[1] ways to send an inquiry: 'book_it', 'instant_book' - detailed at the bottom of this document. Upon receiving the inquiry, the host decides whether or not to accept the request (for the 'book_it' method).

Hosts can reject a guest for many reasons. Some might be logistical, e.g. dates do not work, and some may be more personal, e.g. the guest seems risky. Our goal is to help guests maximize their likelihood of being accepted by the hosts they contact.

Questions

Please use the attached (fabricated) contacts data and experiment assignments data to answer these three questions.

1. We would like to better understand the reasons behind whether a guest ultimately is able to make a booking and what makes a guest more likely to end up with a booking. Build a model that can be used to explain the booking rate. Based on the model, what product feature do you most recommend for improving booking rate?[2]

2. We have run an experiment where we require the guest to write a message that is at least 140 characters long to explain why they are interested in staying with the host, and we run this as a 50/50 experiment (50% in treatment, 50% in control) to see the effect on booking rate. Should we launch this product feature to everyone?

3. Consider a different hypothesis: *a 12 hour time limit on a host's first response time will increase the number of bookings* (when a host does not respond by this time limit the booking request is auto-rejected). You are working with a Product Manager to test this hypothesis. Please answer the following questions:
    a. How might the new time limit change the guest and host behavior?
    b. What is the size of the potential business impact?
    c. In designing an experiment to test this change:
        i. What are the subjects of the experiment and how will you randomize them into control and treatment?
        ii. Which key metrics will you monitor?
        iii. How will you decide when to stop the experiment?
        iv. How does this experiment design account for good user experience of the Airbnb product and spillover effects?

[1] In reality there are three methods: 'contact me', 'book it', and 'instant book' but for the purposes of this challenge we simplify and assume only the latter two methods.
[2] Predictive models should not be built on data with different treatments but for this question you can assume the data is all generated from one treatment.

Important instructions

- Aim to spend no more than 6-8 hours on this challenge and roughly equal time on each question.
- Please attach both your write-up and your code, in its entirety (please pdf into one document).
- Aim to have a writeup that is easy to understand in 20 minutes. One way you might help your reader is by including a summary, perhaps the length of a tweet, at the top of your report that could be sent to the entire Airbnb product team.
- Please include at least one visualization to communicate your results.
- You are free to use whatever tools you are most comfortable with to work through the analysis.
- We suggest that you browse the Airbnb website and look at listings to see the different ways that you can message a host ('book_it', and 'instant_book').
- Please note that data provided here has not been cleaned or sanity checked.

Data provided

Experiment Assignments - contains a row for every time that a user gets assigned to an treatment group.
* `id_user` - random id of the user.
* `ab` - The experimental group the user is assigned to.

Contacts - contains a row for every time that an user makes an inquiry for a stay at a listing in San Francisco.
* `id_guest_anon` - id of the guest making the inquiry.
* `id_host_anon` - id of the host of the listing to which the inquiry is made.
* `id_listing_anon` - id of the listing to which the inquiry is made.
* `ts_interaction_first` - UTC timestamp of the moment the inquiry is made.
* `ts_reply_at_first` - UTC timestamp of the moment the host replies to the inquiry, if a reply has been made.
* `ts_accepted_at_first` - UTC timestamp of the moment the host accepts the inquiry, if the inquiry has been accepted.
* `ts_booking_at` UTC - timestamp of the moment the booking is made, if a booking has been made.
* `ds_checkin_first` - Date stamp of the check-in date of the inquiry.
* `ds_checkout_first` - Date stamp of the check-out date of the inquiry.

* `m_guests_first` - The number of guests the inquiry is for.
* `m_interactions` - The total number of messages sent by both the guest and host (at the data was pulled)
* `m_first_message_length_in_characters` - Number of characters in the first message sent by the guest, if a message was sent
* `dim_contact_channel_first` - The contact channel through which the inquiry was made. One of {book_it, instant_book}. *See bottom of page for more detail*
* `dim_room_type` - indicates whether the room is an entire home, private room, or shared room
* `dim_total_reviews` - the total number of reviews of the listing at the time the data was pulled.
* `dim_person_capacity` - this is the stated total capacity of the room as indicated by the host
* `dim_guest_language` - language used by the guest the last time the user visited the website
* `dim_host_language` - language used by the host the last time the user visited the website

Some things to keep in mind

There are two ways to book a place on Airbnb:
1. book_it - The guest puts money down to book the place directly, but the host has to accept the reservation request. If the host accepts, the booking happens automatically.
2. instant_book - The guest books the listing directly, without any need for the host to accept or reject actively (it is an autoaccept by the host).

Take Home Challenges (II)

Take home case Data Scientist

CoE Financial Crime and Regulatory Technology

Congratulations, you have made it to the next step of our recruitment process!

This step includes a take home case, which provides you with a problem to solve and corresponding data. We ask you to deliver a working Python code with a rationale of your solution and a presentation.

This document should provide you with the description of the problem and of the accompanying data, as well as of our requirements which we ask you to solve as a part of your application procedure.

We wish you good luck and are looking forward to welcoming you onsite!

Business Case

A core business of our bank is the execution of various kinds of transactions. Unfortunately, our clients are occasionally victims of different kinds of fraud. Although we already have very effective counter fraud systems and processes in place, some fraudsters keep coming forward with new ways to commit this crime.

Some fraud is relatively easy to verify but still requires some manual processing. These transactions are typically not executed (criminals actually do not succeed to collect any money from our clients' accounts). In practise, there is an alert raised upon the transaction execution request, the transaction is put on hold and an analyst confirms if it is a fraud or not.

More advanced fraud involves three times as much manual processing time of an analyst (with 3x higher operational costs for the bank) and in such cases the transaction gets executed because the bank can't hold it back that long. In these cases clients money get withdrawn from his/her account and the bank compensates the client's loss fully. In addition to those costs the bank suffers a reputational damage.

Data Description

reported_frauds.csv	**transaction_amounts.csv**	**transaction_features.csv**
Transaction IDs of the fraudulent transactions of the first described kind	Amounts of each transaction in our database along with the transaction ID	Engineered features at transaction level.[1]

[1]Features are results of domain knowledge and long counter fraud practice and are a given in this assessment. The exact meaning of the features is not disclosed as for security reasons.

Objective

We would like to develop a machine learning model that helps us in two ways:.

- It identifies fraudulent transactions so that we can remove the manual processing part of the 'easy to verify' kind of fraud.
- It identifies the more hard to verify kind of fraud so that we can remove even more processing costs and also the direct reimbursement costs.

Remarks

- The provided data includes hidden fraudulent transactions → fraudulent transactions whose IDs are not provided here.
- These transactions are part of the second hard to verify kind (as described earlier). Targeting such transactions has in fact higher value for the bank.
- We know, however, that the incidence of the two kinds of fraud is roughly the same.

What are you required to deliver?

- A **Jupyter notebook** with your code
  - Try to structure your code to be as modular as possible
  - Make sure to comment your code so that others can quickly assess it
  - Feel free to use the markdown cells if needed
- A **presentation** of your solution
  - You can either use a PowerPoint file or a generated PDF file

You are kindly requested to send your solution (notebook and presentation) to us at least one day before your onsite presentation.

Additional Remarks

- Your are allowed to make assumptions regarding e.g. cost of processing of alerts, reputational damage incurred per fraud instance caught too late etc. if needed.
- Note also that we are not looking for s specific solution as there is a multitude of solutions.
- We also do not expect a solution that works well enough to be put in production, from any perspective.
- You are allowed to borrow bits and pieces from public sources, but the overall solution must be of your own.
- The information supplied here is sufficient to solve this assessment successfully.

What will be assessed?

- **Your business to analytics reasoning**: What fundamental choices did you make and why? How dies that dress the posed problem?
- **Your design and modelling choices**: How do you measure performance in your model? How did you make your choices? Link to the posed business problem? How did you tweak your model hyper-parameters?
- **Your coding skills**: Modularity, comments and readability
- **Your presentation style and interaction with the audience**: Mainly, you will be asked questions in a peer to peer style but on occasion we may ask you to frame an answer as for a non-technical stakeholder.