Thank you for inviting me to this interview and sending this case study.

It was a nice piece of data to work with and interesting problems so I glad to present my findings to you in this presentation today.

**The Brief**

The brief is that autistic spectrum disorder (ASD) is a neurodevelopment condition with significant healthcare costs.

To reduce these costs, it helps to identify this early but identifying it early is itself length and not necessarily cost effective.

The client has provided a dataset with patient characteristics and behavioural traits which could be used to vastly improve the efficiency of diagnosis.

Our job is to build a model which can predict using this features, whether someone is likely to have ASD or not.

Before we get on to the models, let’s start with a description of the sample.

**Listing data**

We have around 674 observations. Our target variable is whether there is a diagnosis of ASD or not.

**ASD diagnosis varies by gender, ethnicity, age**

Again, we have around 70,000 observations. Some of the observations have dropped off due to missing values but we still have a strong dataset.

**ASD diagnosis also varies by country of residence, jaundice status and if a family member has PDD**

Here we have, neighbourhood and room type.

As you might expect, boroughs within Central London show the highest price including city of London, Westminster and Chelsea. And borough on the outskirts such as Hillingdon, Bexley and Croydon show the lowest price.

Entire homes have highest price. Surprisingly a shared room is on average higher than a private room. Unclear why this is the case. However, we are not interacting this with other categories. So in fact, it could be that there are more shared rooms within more expensive neighbourhoods which leads to a higher mean price.

**Behavioural traits likely to be strong predictors of ASD**

Just to see how individual numerical variables correlate with price, I present this correlation table here. As we can see, even though there is some positive correlation between price and the property variables, it is very weak. There is also some multicollinearity between the variables on accommodates, bathrooms, bedrooms, and beds as expected. This informs some of the model tests as we will see in the next slide.

**Which model performs best and which features are most important?**

Finally, here we see the results from running multiple models. Of course, it made sense to use regression models given that our target variable is continuous. Running multiple models is useful as we can assess which model performs best as well as whether our model results are robust.

All models post an R2 on the testing sample of around 0.6 and an MSE on the testing sample of around 0.25. This means the model explains 60% of the prices and the mean squared error on the predictions is pretty low.

Out of which, the gradient boosting regressor performs best by a small margin. [continuously adds learners to minimise the loss function]

We also look at some of the feature importance aspects to this. Latitude and longitude have the strongest predicting power here even though we controlled for neighbourhood. This suggests that the location within a neighbourhood also matters.

Furthermore, the more people the listing accommodates and if it is an entire apartment, it seems to have a strong impact on the price.

So the advice to the client could be to maximise the number of beds and if possible, to rent out the entire home, apartment if they want to maximise their price.

**Model 2: Improving Customer Reviews**

Now for Model 2. As a reminder, here we are asked to maximise positive customer reviews to improve the listing number in search results.

Let’s start with the sample for this model.

**Sample**

Here, the sample size is even smaller mostly due to the review variables having even fewer observations. However, at around 50,000 it is still strong.

The target variable I chose for this is the overall review rating. On further research, I found that Airbnb uses this review rating as well as guest needs and trip details to determine the search ranking. And as you will see in the next slide, the overall review rating is determined by other review ratings with different strengths.

We include the same categorical features and numerical features as in the first model. But an important include in the list of features is the ratings for sub review categories. The overall review could be a result of strong communication, or accurate description, or good location and good value. Hence, these are important features to include. The inclusion of other categorical features help us sample the impact of other variables correctly. For example, for the same property type, an accuracy score is more comparable compared to different accuracy scores for different property types.

The brief also asked to include the description and neighbourhood overview variables. I thought of two options. The length of these variables could proxy for accuracy. Or one could run an unsupervised learning model to derive topics from these variables. Due to a lack of time, I opted for the former.

Again, the same cleaning was carried out as for the previous model.

**Do review sub categories affect the overall review score?**

The first piece of analysis I did was understand how the sub categories affect the overall review score. The prior is that these are important parts of the prediction model and if this is true, they would derive important insights into what hosts can do to maximise their ratings.

The model run here is an OLS to identify unbiased estimators. What we find all sub-categories significantly affect the overall review score except the location score. But together, they explain 99.7% of the variation. An increase in a score by 1 for communication leads to a marginal improvement in the overall review score by 11% with all else being equal.

Therefore, if hosts are to focus on one thing, they could focus on ensuring their communication with their users is great.

**Which prediction model performs best and which features are most important under the full specification?**

Of course, I then did the traditional train/test split and ran various regression models. All models gave strong R2 and MSE scores with most of them above 0.8 R2 scores.

The Gradient Boosting Regressor was the strongest here as well.

When we look at the feature importance of this model, we see a slightly different story. Whilst all review scores again score highly, the value category scores the highest and one above the communication score. This is probably a result of the training, testing split where unbiased estimators like in the previous model mean strong overfitting. Nevertheless, the communication aspect is a strong signal to hosts.

Interestingly, the latitude, price, number of amenities and unique words in the neighbourhood also scored in top 10 important features.

Therefore, the message to hosts are to provide good value relative to the competition where the pricing model could come in handy; provide good communication, be accurate in your description and maintain strong cleanliness.

**Potential improvements**

Of course, this work was done in a limited amount of time. It can definitely be improved in many ways.

**Questions to prep**

Why did you remove property type? – Too many categories and if I reduced the categories, it would be basically the same as room type.

Why did you not remove one of ‘accommodates’ or ‘beds’ if they are also multicollinear? Good point. Should’ve kept only one of them

What about the variable number of descriptions? Was that significant? No.

Why did you choose those models? Regressor models

Why do you think the Gradient Boosting Regressor gave the best result? Not sure.

What about the signs of coefficients? See below.

Why one-hot encoding? Features must be numeric. We can’t translate most vars into an order. So better off having separate binary vars for them.

Why Linear SVR, K Nearest Neighbours and Ridge CV?

* Linear svr: high prediction accuracy; excellent generalisation capabilities; effective in high dimension space. Normally, we try to minimise error rate, here it tries to fit the error within a threshold. Only take points with least error rates to get a better fitting model. Effect of outliers is less.
* K nearest neighbours
* Ridge CV: penalise coefficients if they are too far from zero.

**Model 1**

Name Coefficients

Latitude -0.16844378300848697

longitude 0.4892432255178091

accommodates 0.9482914416937972

bathrooms 0.8959122640314383

bedrooms 1.6090394635332546

beds -0.11657057816670169

minimum\_nights -0.19313533326434837

maximum\_nights 0.20163705724462958

**Model 2**

Name Feature Importance

review\_scores\_accuracy 0.39180948509394353

review\_scores\_cleanliness 0.297736524669986

review\_scores\_checkin 0.10268564574910274

review\_scores\_communication 0.2952850338468089

review\_scores\_location 0.05860809418580429

review\_scores\_value 0.4378451021926355

latitude 0.052506465338427386

price 0.04925965836641652

neigh\_unique -0.0010627441790932585

amenities\_length 0.006826957541614692