# Data Science Test

## 1. Airline Use Case

Suppose we build a recommendation engine for a budget airline that predicts online purchases of pre-booked items for each passenger, such as priority boarding, extra leg room, exact seating, and food and beverages.

### **Question 1.1**

**There is a confusion at the client about how to set up the machine learning task. What are the training examples (X) and what are the predicted outcomes (y) ? What kind of algorithm to use? What should be the evaluation metric of the model? Please share your suggestions with them for each question.**

The training examples are the passangers’ details (age, price of ticket, traveling alone, travel distance), the outcomes are the items priority / extra leg room / exact seating / food and beverages. Maybe we should consider the “purchased nothing” option aswell.

We have more than two outcomes so it’s a multi class classification. We could use Decision Tree or Random Forest, Support Vector Machine but the best could be to use a few models, and decide which is better.

For evaluation we can use accuracy score.

### **Question 1.2**

**We know that the popularity of the products are varying. E.g. purchase rate for priority boarding is 20 % while for food is 2 %. How would this influence the recommendations? Do we need to handle it somehow?**

I wouldn’t know without testing it. It could happen that the natural distribution of the data is enough, so at evaluation it turns out the model could make a difference.

If it doesn’t help, I could oversample the food class, or undersample the priority boarding.

### **Question 1.3**

**We settled to use one year’s data of online pre-booked purchase behavior for model training, which we split into 70% training and 30% evaluation sets randomly. Our final model is ready and it performs well on both sets. The plan is to retrain the model (no hyperparameter-tuning, just re-run) every day at 1 am based on data of the previous 30 days.**

**A data scientist from the client’s team expresses concerns that the production system will not perform as well as indicated by our training setup. Is this concern valid? How would you address his concern? Write an email to him.**

Dear Tim!

I understand your concerns of the daily retraining. I advise you to try it for 2-3 days, and let’s see if the performance decreases. We should see how it works on new data, and if it’s not good, we will tune it.

Please tell me, if it’s good to you.

Thanks,

Balazs