FASHIONABLE

Data science
Retail

Medium difficulty
Interviewer-led case

This case is about an inventory reduction project at a fashion retailer. It includes a challenging structuring question.

This case is only suitable for candidates applying to data science consulting roles.

Problem definition

Our client, Fashionable, is a large online fashion retail company with a global presence. Every season, Fashionable sources items from independent designers around the world. The items are then stored in the warehouses distributed across 5 regions, and marketed to the local consumers through their website. As fashion changes quickly, the company has a double issue with finding the right inventory levels: for some popular items, they run out of stock very quickly, while other items don't sell quickly enough, and have to be sold at a discount to free up warehousing space. This has led to significant cost and potential lost sales opportunities.

Fashionable has hired you, a data scientist and consultant, to help them optimize their inventory to free up cash, reduce cost and increase sales.

Question 1 (Structuring)

What factors would you consider to answer the client's questions?

Additional information

If asked, please share that:

- The client has a global presence in 5 regions: Europe, East Asia, US, Australia and Middle East.
- Each region has one central warehouse and between 5 - 10 local warehouses.
- The items are mainly sourced from and produced in East Asia.
- It takes between 1-3 months from the start of sourcing until the day the products arrive in the central warehouse.

Possible answer

1. *Better predict demand, taking into account:*
   a. *Historical data*
   b. *Local trends and demand dynamics*
   c. *Broader fashion trends, through external data such as social media*

2. *Reduce order lead time: the shorter lead time would allow us to reduce the amount of stock needed and to react more quickly to forecasting error*

3. *Optimize inventory based on forecasted demand:*
   a. *Determine stock based on lead time, forecasted demand and logistic cost*
   b. *Find optimal tradeoff between out-of-stock risks and inventory cost*
   c. *Adjust inventory mix to match local taste and reduce transportation cost*

Question 2 (Numeracy)

Currently, the inventory for a product is set to twice the monthly sales forecast, with an additional buffer which is proportional to the forecast error rate, which is 40% currently. The annual revenue is around 1 billion USD, with 50% gross margin.

How much cash would you be able to free up should you be able to reduce forecast error by half? What other impact would you expect from the reduction in the forecast error rate?

Additional information

The candidate is encouraged to ask for more details about the forecast error rate. If asked, please share that:

- The 40% error rate is defined as mean absolute percentage error (MAPE) which equals to **mean of absolute (Actual sales - Forecasted sales) / Actual sales.** In this context, the error rate has a direct impact on inventory buffer.

Possible answer

The cash freed up by the reduction on inventory can be estimated numerically:

- *The monthly revenue is about $1,000M / 12 = $83.3M.*
- *If the monthly forecast is on average accurate, then the sales value of the inventory is 2 * 83.3 = ~ $167M.*
- *On top of that, we add a 40% of buffer, or $167M * 40% = $66.8M.*
- *If we were able to reduce the forecast error by half, then buffer will be reduced to $167M * 20% = $33.4M. This would yield a saving of $66.8M - $33.4M = $33.4M of inventory at sales value, solely from the reduction of the inventory buffer.*
- *With a gross margin of 50%, this will translate to a $33.4M * (1-50%) = $16.7M inventory reduction.*

Reducing the forecast error rate would bring additional benefits:

- *Increase of revenue by minimizing lost sales*
- *Decreases warehousing cost through reduction in inventory size*
- *Reduction of transportation cost*

Question 3 (Data Science expertise question)

Suppose that you have decided to construct a machine learning model to improve demand prediction and then reduce the inventory, and you have just started to collect data. **What data could be helpful and why?**

Possible answer

To build a demand prediction model, we need:

- *Historical sales data: it is starting point of the prediction model. It needs to be at the granularity of product level, at most aggregated at weekly and regional level. Daily and more local data are also welcome.*
- *Inventory data: This will complement the sales data and help us identify no-demand and out-of-stock items (both would show as zero in the sales date, but for very different reasons)*
- *Product information: it should contain the information on product category and SKU. More detailed information on the product character and sourcing origins will be helpful for sense checking and other types of analysis.*
- *External data: To capture the trend, we could collect external data, such as social network. However, this data is difficult to work with, especially at the local level. More relevant external data could be weather forecast, local average income, etc.*

To optimize inventory, we would also need:

- *Inventory data: it should contain weekly (or monthly) snapshots of inventory levels and inventory buffers for each product.*
- *Logistic information: It should contain locations of the warehouses, storage cost, transportation cost and lead time as well as capacity constraints for warehousing and transportation.*

Question 4 (Data Science expertise question)

You now are asked to construct the demand prediction model. Which methods and algorithms would you recommend using?

Guidance for interviewer

- Candidates are encouraged to give a more detailed description of the features and targets that they want to use. These can be previous monthly sales, marketing events, product category information, etc.
- It should be clear to the candidate that in feature-based ML methods, the forecast horizon is defined in the way that target and features are structured, i.e., if the sales of last month is used to project current sales, the forecast horizon is only 1 month.
- Candidates should recognize that the forecast horizon should be set according to the lead time.
- Candidate should never mix up logistic regression (predict binary result) with other regression methods (predicting quantities).
- Candidate should be able to explain why they prefer certain methods instead of others, and in which conditions some model will be preferable than others.

Possible answer

Several methods can be used:

- *Traditional time series technique for sales prediction: ARIMA models, including various de-trend and de-seasonality techniques (filters, linear regression, etc.), as well as its multi-dimension version (VARIMA).*
- *Feature based machine learning methods: Tree-based regression methods such as Random Forest, Gradient Boosting, Extra-Tree or else multi-linear regression.*
- *Deep learning models: LSTM and RNN, which handles sequence data. However, it may require a larger data set to avoid overfitting.*

Question 5 (Data Science expertise question)

If you were given a choice of models, how would you select the best one? Which metrics would you use? Suppose that the above showed the results of two models you have constructed, which one do you think is best?

Possible answer

I would consider the following factors when selecting the best model:

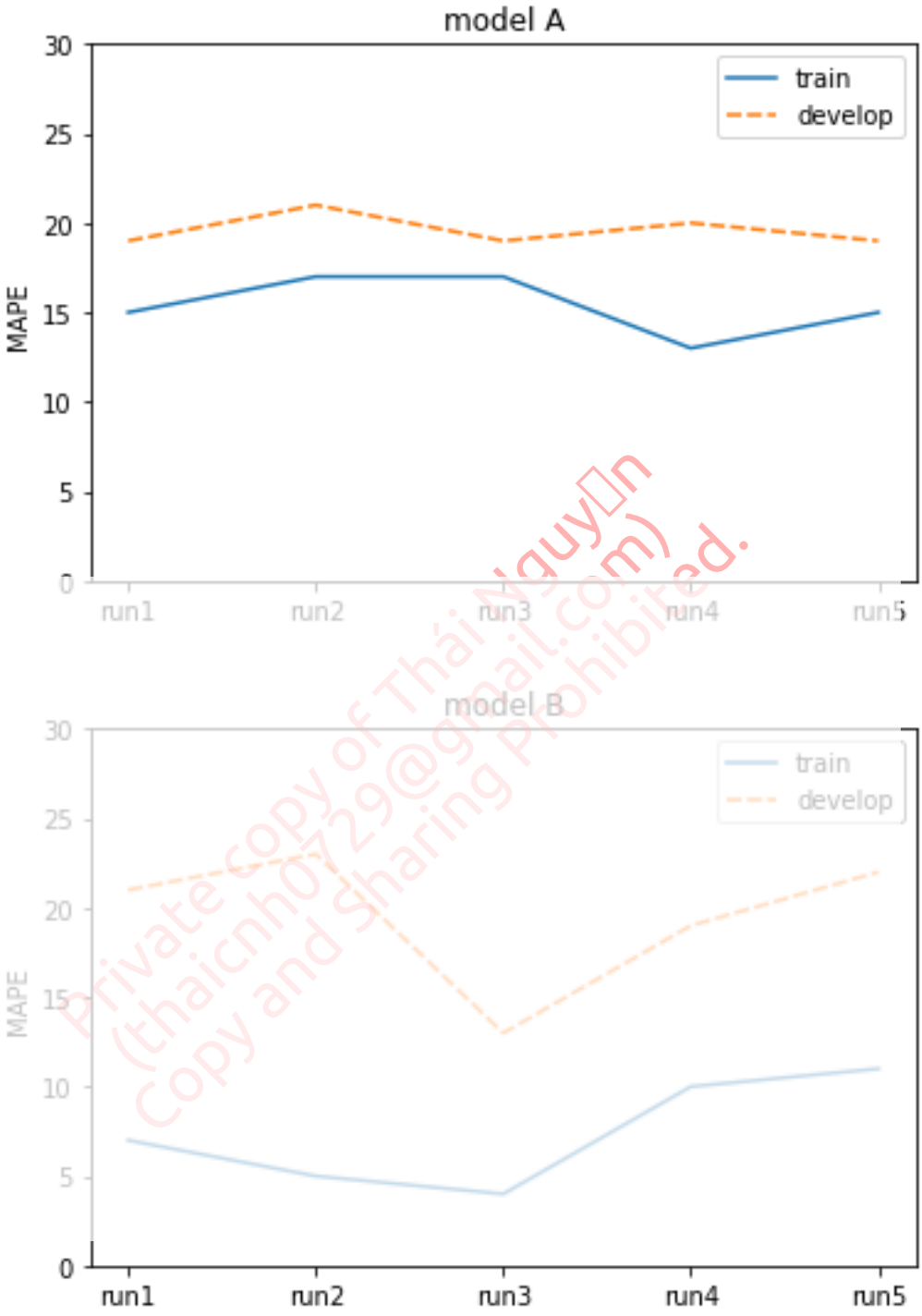
- *I would only select the model based on train and develop data sets, but never on test data set, which should only be used to report performance.*
- *I would consider the trade-off between model variance and bias, and how to prevent model from over-fitting.*
- *I would use metrics such as MAE, MSE, MAPE or other metrics for regression*

In addition:

- *We could use K-fold validation, grid search or other model selection methods which are appropriate for the chosen prediction model.*
- *We could use regularization methods, such as limit the tree depth in tree-methods and LASSO for linear regression.*
- *We should be careful not to split train, develop and test data sets randomly on time dimension (but they could be on product dimension). We should always make sure that future data are not used to predict past (i.e. model is trained on past data and validated on future data.)*

In this example, model A is best, because the difference between train and develop set performance is smaller, and because it shows less variance across different runs. These all indicates less over-fitting risks.

Exhibit 1



Question 7 (Data Science expertise question)

Now that you have built a successful demand forecast model, you decide to use it to optimize inventory levels. What are the main drivers you can use to optimize inventory?

Possible answer

- *Inventory should follow forecasted monthly sales, influenced by lead time and the forecast errors. The lower the forecast error and the shorter the lead time, the lower the inventory can be.*
- *The cost of holding un-sold inventory and the lost sales should both be evaluated in monetary terms. Inventory setting should consider the trade-off between them.*
- *Variations in demand and cost (sourcing, transportation, storage) in different regions and among different types of product should be considered as well.*
- *One can also consider warehousing and transportation capacity constraints.*

Question 8 (Data Science expertise question)

Assume that you were able to produce a very accurate forecast about demand, such that we can ignore the error rate. Now the center of your problem is to select the right items to stock in each regional warehouse and determine the right quantity.

Mathematically, could you formulate an objective function, such that by minimizing/maximizing it you will find the right product and the stock? What are the constraints? You can consider only one regional warehouse to simplify.

Additional information

Encourage the candidate to derive a symbolic formulation of the optimization problem. If asked, please share:

- Warehouse storage capacity is **M** cubic meters
- There are potentially **N** different items, each takes a certain amount of space (**v**) in the warehouse.
- We know almost exactly the how many items (**q**) will be sold monthly
- The lead time for all items is **L** months
- We know also the price (**p**), the cost (including fabric and transportation cost) (**c**) for each item

Possible answer

- *Suppose that the inventory quantity we are to choose for each item is **s**. Due to the lead time, inventory is only replenished every **L** month. As the monthly demand is **q**, the maximum amount that can be sold in **L** months is: $min(s_i, q_i \cdot L)$. The objective function should be the total net profit, which can be written as $\sum_{i=1}^{N} \{min(s_i, q_i \cdot L) \cdot (p_i - c_i)\}$. The optimisation algorithm will maximise the total net profit by finding the right value of s_i*
- *Under the constraints that the inventory volume stays below warehouse capacity $\sum_{i=1}^{N} s_i \cdot v_i \leq M$*
- *Note that this approach ignores warehousing costs.*

Question 9 (Data Science expertise question)

How would you change your approach if you found out that your data quantity is very big (terabytes), or instead very small (only a few megabytes)?

Possible answer

- *Different data engineering and machine learning tools should be used if data is very large. For example, one can use distributed data processing system like Hadoop or Spark, and opt for models which are fast, light-weight and parallelizable.*
- *If the data is very large, we can first work on a small sample of data to validate ideas for feature engineering and model selection. This is to reduce each iteration time in the model building. Once the idea is validated, we then deploy the model to the larger scale and monitor its performance closely.*
- *If data is small, one should opt for simpler models (models with fewer parameters) and use regularization methods to avoid overfitting.*

Question 10 (Synthesis)

Could you wrap up the case, by giving a summary of your approach and a few take-aways?

Possible answer

- *You have asked us to reduce Fashionable's inventory cost and opportunity cost of out of stock items.*
- *We found several drivers to solve this problem: improved forecast accuracy, shorter lead time and more precise inventory level setting methods. Given their current method to set inventory, we found that reducing demand forecast error by half would free up ~$17M of cash on its own.*
- *Data science methods leveraging historic data, local demand data, and external data such as weather would allow us to improve our forecast accuracy.*
- *Stock level would then be determined for individual products based on a profit maximizing function taking into the lead time, the forecast error, the tradeoff between inventory cost and missed sales, warehousing and transportation costs and capacity constraints.*
- *As for next steps, we will start collecting data to test several models and select the best one.*