**About Dataset**

**Context**

E-commerce has become a new channel to support businesses development. Through e-commerce, businesses can get access and establish a wider market presence by providing cheaper and more efficient distribution channels for their products or services. E-commerce has also changed the way people shop and consume products and services. Many people are turning to their computers or smart devices to order goods, which can easily be delivered to their homes.

**Content**

This is a sales transaction data set of UK-based e-commerce (online retail) for one year. This London-based shop has been selling gifts and homewares for adults and children through the website since 2007. Their customers come from all over the world and usually make direct purchases for themselves. There are also small businesses that buy in bulk and sell to other customers through retail outlet channels.

The data set contains 500K rows and 8 columns. The following is the description of each column.

1. TransactionNo (categorical): a six-digit unique number that defines each transaction. The letter “C” in the code indicates a cancellation.
2. Date (numeric): the date when each transaction was generated.
3. ProductNo (categorical): a five or six-digit unique character used to identify a specific product.
4. Product (categorical): product/item name.
5. Price (numeric): the price of each product per unit in pound sterling (£).
6. Quantity (numeric): the quantity of each product per transaction. Negative values related to cancelled transactions.
7. CustomerNo (categorical): a five-digit unique number that defines each customer.
8. Country (categorical): name of the country where the customer resides.

There is a small percentage of order cancellation in the data set. Most of these cancellations were due to out-of-stock conditions on some products. Under this situation, customers tend to cancel an order as they want all products delivered all at once.

**Inspiration**

Information is a main asset of businesses nowadays. The success of a business in a competitive environment depends on its ability to acquire, store, and utilize information. Data is one of the main sources of information. Therefore, data analysis is an important activity for acquiring new and useful information. Analyze this dataset and try to answer the following questions.

1. How was the sales trend over the months?
2. What are the most frequently purchased products?
3. How many products does the customer purchase in each transaction?
4. What are the most profitable segment customers?
5. Based on your findings, what strategy could you recommend to the business to gain more profit?

**Project background:**

**Business problem:**

An e-commerce store with a vast inventory of thousands of SKUs faces challenges in deciding when to expand storage for specific items or when to promote products to decrease stock levels. If the company fails to maintain an efficient inventory management, it may miss sales opportunities or incur higher inventory costs, thereby reducing profits.

**Use scenario:**

We plan to use a machine learning model to predict future demand quantity for listed products. By comparing the predicted sales quantities with current inventory levels, we will be able to decide whether to increase our inventory or launch promotions to reduce it.

**What precisely is the predictive modeling (ML) problem?**

We need to know how much of a SKU to order for the next month.

**What is a data instance?**

A data Instance is the performance of products over a month. This enables us to gain insights around seasonality, pricing trends, complement/substitute products, and in other ways build a large model with diverse features.

**What exactly would be the target variable?**

  The future demand quantity for listed products each month is what we would get after putting it through the decision logic, but the model would output a performance score which would be used by the decision logic to identify if more or less will be sold next month and by how much.

**How would you go from a prediction to a decision to an action?**

If the predicted demand quantity at a given price surpasses our current stock—a hypothetical scenario, as we lack actual stock data—we must increase our prices and/or inventory to satisfy demand based on strategic resource limitations. Conversely, if the predicted demand is below our current stock levels, we should promote the product and/or lower the price to accelerate sales.

All of this sounds like a good start.  Don’t worry too much about having actual stock data.  You can make stuff up.  You should focus on designing a “business story” that is realistic, even if it’s not real.  My feeling is that you’ll adjust this as you see what you can and cannot actually do with the prediction part.  That’s fine.  The goal is not to do what you proposed to do, but to have a solid and interesting final project report.  [By the way, to me this is realistic – not the making stuff up part (ok, that’s somewhat realistic early in such an effort), but the continual pivoting on the business side based on learning more about the technical side of things.  There are lots of ways to get business value from AI/ML, and you can’t always “predict” that well at the beginning.]

**What features would be useful?**

Record date, created revenue, product category, region…etc.

A more though list is: Sales per item per month (one feature per item per month considered), Price point (1 feature per month considered), Geolocation (1 feature per major location), Repurchase (range of 0-1 about percent of customers that repurchase the product in a given year or month or whatever is significant), a bunch of features surrounding other products so that compliments and substitutes are captured

This is what I would dig into right away.  My advice: at least one of you should be working on a “non-ML” prediction algorithm.  Create trends and seasonality measures and just estimate the next month (and/or however far in advance you feel is appropriate).  The first predictor would be: last month’s sales.  Then you could add a trend.  Then you could add seasonality.  This ought to be straightforward, but if it is not, you will want to find that out now – as that would also give you problems with an ML-based solution.

**How exactly would predicting the target quantity solve the business problem?**

If we can predict demand at a given price, we can use these forecasts to adjust product pricing in real time and even promote sales for products that are overstocked. Additionally we can make predictions about how much we need to order for next month

**Can you walk me through the AI Flows and tell me what you would be doing for each stage?**

Business understanding-

Define what is useful for features based on our understanding of the business

Data understanding - learn the available features and which ones are useful, We will learn to create useful features out of the available data. This is a repeating cycle with the previous step.

Data preparation - the data available is largely focused on individual order so those need to be converted into product centric data and from there into useful features.

Modeling:

Analyze the sales data → build a prediction models for demand, this will be doen several times in different ways to give us enough to evaluate latter.

Evaluation- We will then evaluate it against training data, and if we can find it other data. We will optimize feature definitions, model type, complexity etc.

Ah, wrong framework :-)  The “AI Flows” is Framework #3 from the first class.  Probably a good thing to update in your neural networks!

In any case, that sounds like a good start – getting going on doing it is really the most important part, as you’ll (normally) revise everything you’re thinking about as you actually try to make it work.

I would suggest seeing what you can do with one or a few products with a lot of data to start out.  I have not looked through the data sets, but if there are products that both have a decent amount of history and a lot of sales, that should give you something good to work with.

**Targeted Dataset:**

[Online Business Sales 2017-2019](https://www.kaggle.com/datasets/tylermorse/retail-business-sales-20172019) - small and has limited useful information

[E-commerce Business Transaction](https://www.kaggle.com/datasets/gabrielramos87/an-online-shop-business) - medium and has example code

<https://www.kaggle.com/datasets/rishikumarrajvansh/marketing-analytics/data> - huge and more  complicate data

One other thing.  Next week, ask me for a book particularly on using ML for forecasting like this.  I either have it in my office, or I can find it on Amazon.  The one I’m thinking of is not a thick book :-)  but it gives lots of tricks for how to make features and stuff like that.  I’m in France now, but will be back in the office on Tuesday and have made a note to see if it is on my shelf.