Data Science Screening

Demand Forecasting Test

# Introduction

Imagine you are the owner of the greatest Supermarket chain in the US: FairStorage. You want to maximise your profits by making sure that when someone comes to one of your shops, they buy what they were looking for (and some more).

How can you guarantee that? You could have an infinite stock of all of your products so that Mary's favourite kettle chips never go out of stock. However, warehouses are expensive and goods have expiry dates - making the whole idea of infinite stocking a little less feasible.

The most important factor that can help you plan is having an idea of what the demand will look like at any point in time, for every store and every item.

While some demand forecasting trends can be easily predicted (I am pretty confident red roses sales will be higher than usual on the 14th of February 2022), some others are not as obvious. Sales depend on a multitude of factors, including (but not limited to) promotions, price changes, holidays, market conditions, weather, supply-chain disruptions, weird Instagram trends… That is a lot of variables for a human to keep track of.

Your mission for this test is to come up with a model that takes the historical sales information and any additional feature, to forecast the next 21 days of sales for every product and every store in the training data.

# Your task

Your task is to provide a 21-day-forecast of the unit sales at the store, product level. Start with the train.csv file, which contains the historical data of the 1920 days prior to your forecast starting point. You can start with statistical endogenous methods (without external features). For example you can try different ETS or ARIMA models. You need to conduct exploratory analysis and explain a certain choice for the parameters that you are choosing in each case. I suggest you always include a naive and a seasonal naive model for comparison.

If you want to add additional features to improve the model, start by adding the price of each item found in the prices.csv file and the date features contained in the calendar.csv file. Feel free to add any other features you might find useful.

One of the models you choose for the forecast should be a machine learning model.

Your approach should include at least the following:

* Data preparation
* Data exploration
* Feature engineering
* Feature selection
* Cross validation
* Hyperparameter tuning
* Forecast using at least two models
* Error analysis of the forecast
* Blind period forecast
* Future Recommendation

Evaluate your models and forecasts extensively. Use different error metrics and explain the pros and cons of each from both a technical and business perspective.

If your computer is struggling with the computational load, consider using [Google colab](https://colab.research.google.com/?utm_source=scs-index).

### File descriptions

* **train.csv** - the training set
* **sample\_submission.csv** - a sample submission file in the correct format
* **calendar.csv** - calendar information related to the dates considered. You can modify this to include more relevant date features.
* **price.csv** - pricing information for the items by week.

### Data fields

* **id** - an id identifying the smallest data granularity for which the forecast will be assessed: by product and by store
* **item\_id** - the id of the item
* **subcat\_id** - the id of the subcategory the item belongs to
* **category\_id** - the id of the category the item belongs to
* **store\_id** - the id of the store the product is sold at
* **region\_id** - the id of the region the store is located in
* **wm\_yr\_wk** - week encoding

# Deliverables

### The Notebook

Analyse the given dataset. Write a tidy, informative, well documented Notebook exploring the data to find out its insights. The more you know about the data, the easier it will be to provide a good forecast.

Do some clustering analysis on the dataset. Are there products from different categories that behave similarly in terms of their demand? Keep in mind you might want to use their cluster association as an additional feature in your forecasting in the future. Who knows, maybe the clusters you find are even better than the given product categories to describe the difference in products? Showcase how you perform feature engineering and feature selection.

Prepare the data for forecasting. Start forecasting with simpler models, and slowly try more complex ones changing the parameters. Explain your parameter optimization logic. Explain your cross validation approach to test your model’s performance.

Compile your code in a readable format using Jupyter notebook(s) (that can call .py; files that run your models or other helper functions). A reader should be able to understand your logic through each step of the notebook.

Structure the notebooks based on the content (Intro with objectives, Data Prep, EDA, Feature Engineering & Feature Selection, Top Forecasting Models you select (explain how you got to the final parameters and feature set chosen for the submission), Models Comparison and Analysis, Blind Forecast Generation, Conclusion and Future direction). If you prefer, you can submit one notebook for the EDA and one for the forecasting + any `py files you need). If your output is more than one file, provide instructions on how to run the code.

### The Future Forecast

Your submission should follow the sample\_sumbission.csv format **exactly**. All the forecasted sales should be non-negative integers. If you were to encounter any negative integers in the training data, you should assume these are returns and ignore them in your modelling.

### The Presentation

Make a short deck (a .ppt file or .pdf file) explaining your winning approach, any insight you might have found, your approach and the final results. Keep in mind that your presentation should be insightful for a business audience that does not have much forecasting knowedge.

## Submission & Next Steps

Submit your notebook (.ipynb file), forecast (csv file) and presentation (ppt or pdf file) in a zipped folder named `Your\_Name\_Data\_Scientist.zip` and send it to this email by the deadline.

If your code and results are selected for the next steps, you will be asked to present your approach using the deck you prepared at this stage. We look forward to hearing about your approach and recommendations.