Imperial College Business School | Executive Education

PROBLEM STATEMENT

T H E U N S E E N is a beauty startup that uses Material Science to change the conversation within the beauty industry, operating at the intersection of colour cosmetics for face & hair. Our aim is to create products and base strategic decisions on data and scientific evidence.

The company's operation started in October 2021 with the launch of a collection named SPECTRA eye colour. In June 2022, we launched our second collection called Colour Alchemy Hair in collaboration with Henkel/ Schartzkopft Professional. From this moment, we started to invest heavily in marketing activations, and ads and most importantly, we initiated a program to manage our data in a smarter way.

We are confident that in today's world, investing early in a data-driven environment is key to accelerating business growth as well as developing new technologies and products that excite our customers.

With this in mind, we have decided to create a strong forecasting model that can serve as an indicator of what is more relevant to be done at this stage.

The focus of this capstone project is to develop the initial part of this model. A sales forecasting model which predicts sales based on Website Traffic and Audience Growth data.

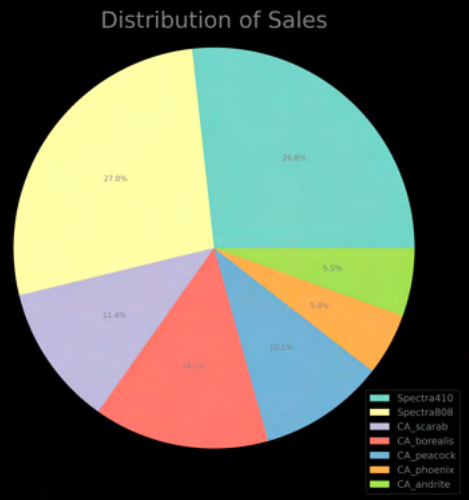
In the future, we aim to develop a second model to help us optimise marketing investments into projects that will maximise the two inputs of our sales model (website traffic and audience growth).

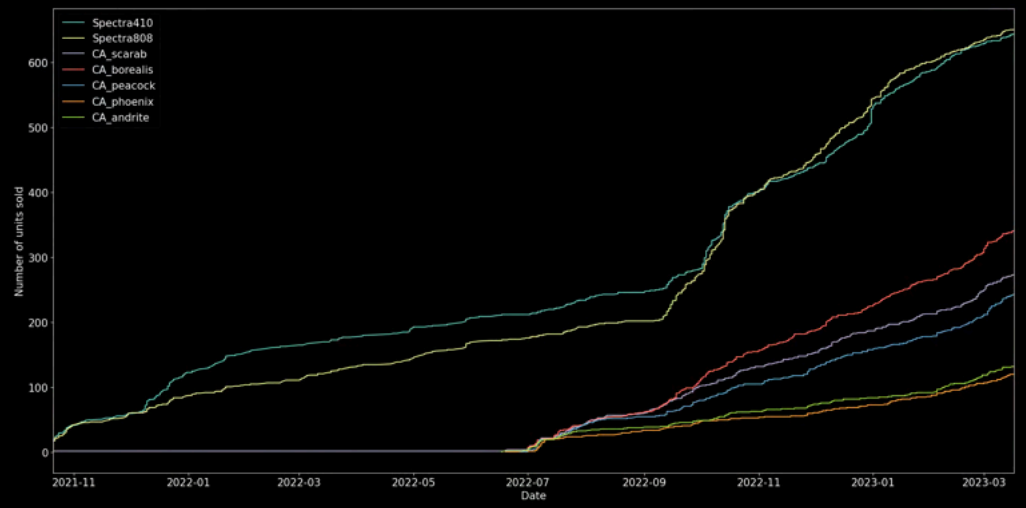
DATA ACQUISITION

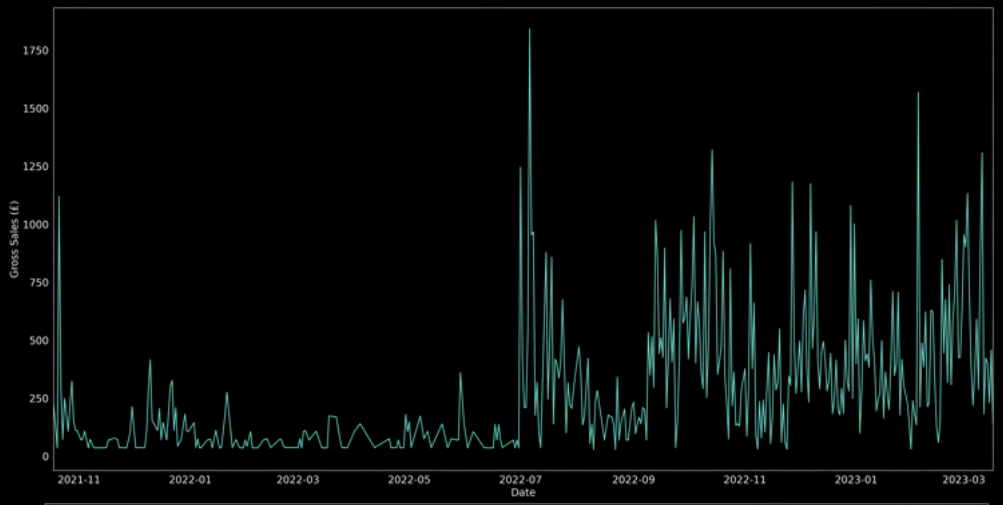
- The main data sources for this work are THE UNSEEN online shop platform (Shopify API)
- Google Analytics
- Meta
  - Instagram
  - Facebook
- TikTok

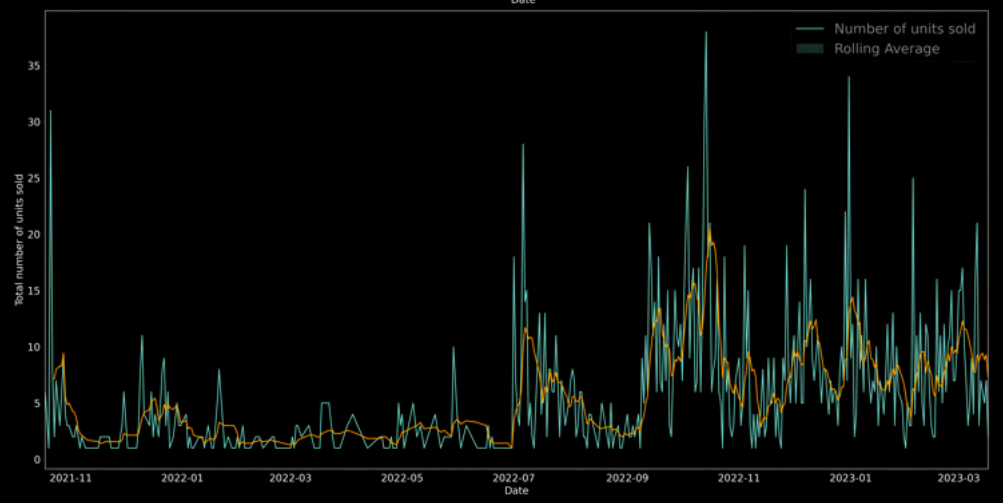
DATA ACQUISITION

Our sales have been slowly increasing and we have observed a strong link with specific marketing activations. The peak around October 2022 is not completely clear. We believe that is related to some content created by influencers. This type of anomaly is what we aim to uncover with the second model, purely dedicated to Marketing.









DATA ACQUISITION

The collaboration with Schartzkopft Professional was a significant milestone in the company's trajectory. From October 2022 our journey in data capture, analysis and decision making has significantly helped to guide the growth of the business.

This data was obtained using the API from our ecommerce platform (Shopify).

EXPLORATIVE DATA ANALYSIS

Since the data was retrieved from our e-commerce shop, we did not observe any anomalies such as the presence of Null or NaN values.

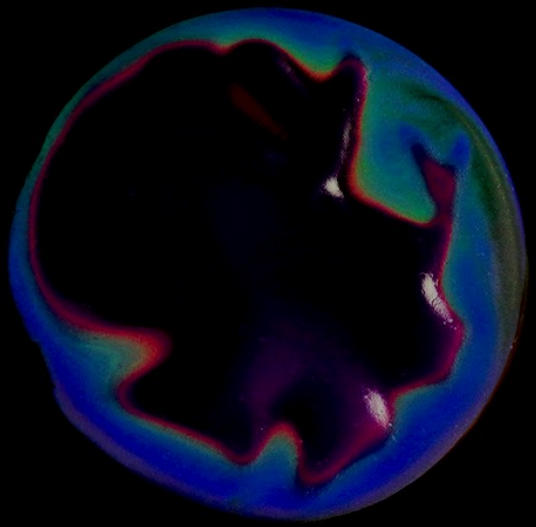
Null/NaN Values: 0
Duplicated Values: 0

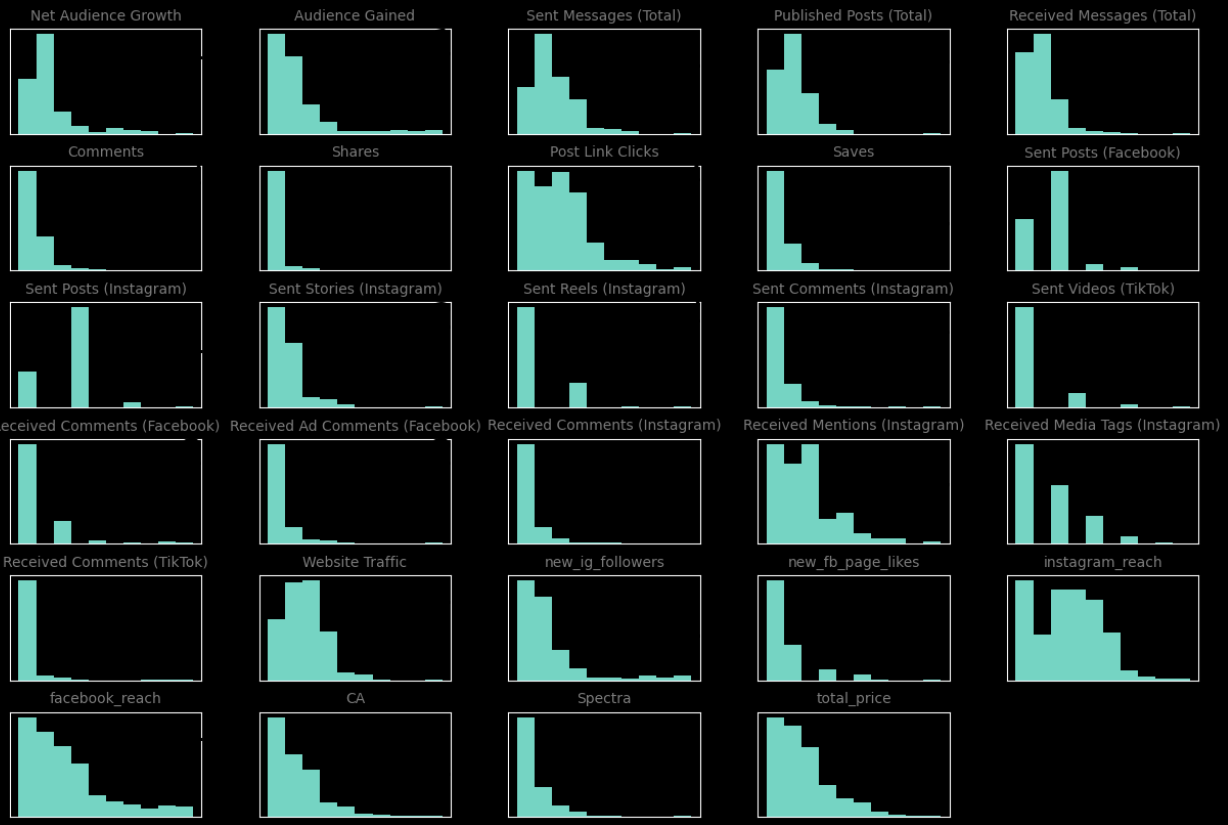
Float values were converted into integers to avoid misinterpretation during the ML model development.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 270 entries, 0 to 269
Data columns (total 37 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Date                            270 non-null    datetime64[ns]
 1   Net Audience Growth             270 non-null    float64
 2   Audience Gained                 270 non-null    float64
 3   Sent Messages (Total)           270 non-null    int64
 4   Published Posts (Total)         270 non-null    int64
 5   Received Messages (Total)       270 non-null    int64
 6   Comments                        270 non-null    float64
 7   Shares                          270 non-null    float64
 8   Post Link Clicks                270 non-null    float64
 9   Saves                           270 non-null    float64
 10  Sent Posts (Facebook)           270 non-null    float64
 11  Sent Posts (Instagram)          270 non-null    float64
 12  Sent Stories (Instagram)        270 non-null    float64
 13  Sent Reels (Instagram)          270 non-null    float64
 14  Sent Comments (Instagram)       270 non-null    float64
 15  Sent Videos (TikTok)            270 non-null    float64
 16  Received Comments (Facebook)    270 non-null    float64
 17  Received Ad Comments (Facebook) 270 non-null    float64
 18  Received Comments (Instagram)   270 non-null    float64
 19  Received Mentions (Instagram)   270 non-null    float64
 20  Received Media Tags (Instagram) 270 non-null    float64
 21  Received Comments (TikTok)      270 non-null    float64
 22  Website Traffic                 270 non-null    int64
 23  new_ig_followers                270 non-null    int64
 24  new_fb_page_likes               270 non-null    int64
 25  instagram_reach                 270 non-null    int64
 26  facebook_reach                  270 non-null    int64
 27  CA                              270 non-null    int64
 28  Spectra                         270 non-null    int64
 29  total_price                     270 non-null    int64
 30  dayofweek                       270 non-null    int64
 31  quarter                         270 non-null    int64
 32  month                           270 non-null    int64
 33  year                            270 non-null    int64
 34  dayofyear                       270 non-null    int64
 35  dayofmonth                      270 non-null    int64
 36  weekofyear                      270 non-null    int64
dtypes: datetime64[ns](1), float64(18), int64(18)
memory usage: 80.2 KB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 270 entries, 0 to 269
Data columns (total 37 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Date                            270 non-null    datetime64[ns]
 1   Net Audience Growth             270 non-null    int64
 2   Audience Gained                 270 non-null    int64
 3   Sent Messages (Total)           270 non-null    int64
 4   Published Posts (Total)         270 non-null    int64
 5   Received Messages (Total)       270 non-null    int64
 6   Comments                        270 non-null    int64
 7   Shares                          270 non-null    int64
 8   Post Link Clicks                270 non-null    int64
 9   Saves                           270 non-null    int64
 10  Sent Posts (Facebook)           270 non-null    int64
 11  Sent Posts (Instagram)          270 non-null    int64
 12  Sent Stories (Instagram)        270 non-null    int64
 13  Sent Reels (Instagram)          270 non-null    int64
 14  Sent Comments (Instagram)       270 non-null    int64
 15  Sent Videos (TikTok)            270 non-null    int64
 16  Received Comments (Facebook)    270 non-null    int64
 17  Received Ad Comments (Facebook) 270 non-null    int64
 18  Received Comments (Instagram)   270 non-null    int64
 19  Received Mentions (Instagram)   270 non-null    int64
 20  Received Media Tags (Instagram) 270 non-null    int64
 21  Received Comments (TikTok)      270 non-null    int64
 22  Website Traffic                 270 non-null    int64
 23  new_ig_followers                270 non-null    int64
 24  new_fb_page_likes               270 non-null    int64
 25  instagram_reach                 270 non-null    int64
 26  facebook_reach                  270 non-null    int64
 27  CA                              270 non-null    int64
 28  Spectra                         270 non-null    int64
 29  total_price                     270 non-null    int64
 30  dayofweek                       270 non-null    int64
 31  quarter                         270 non-null    int64
 32  month                           270 non-null    int64
 33  year                            270 non-null    int64
 34  dayofyear                       270 non-null    int64
 35  dayofmonth                      270 non-null    int64
 36  weekofyear                      270 non-null    int64
dtypes: datetime64[ns](1), int64(36)
memory usage: 80.2 KB
```

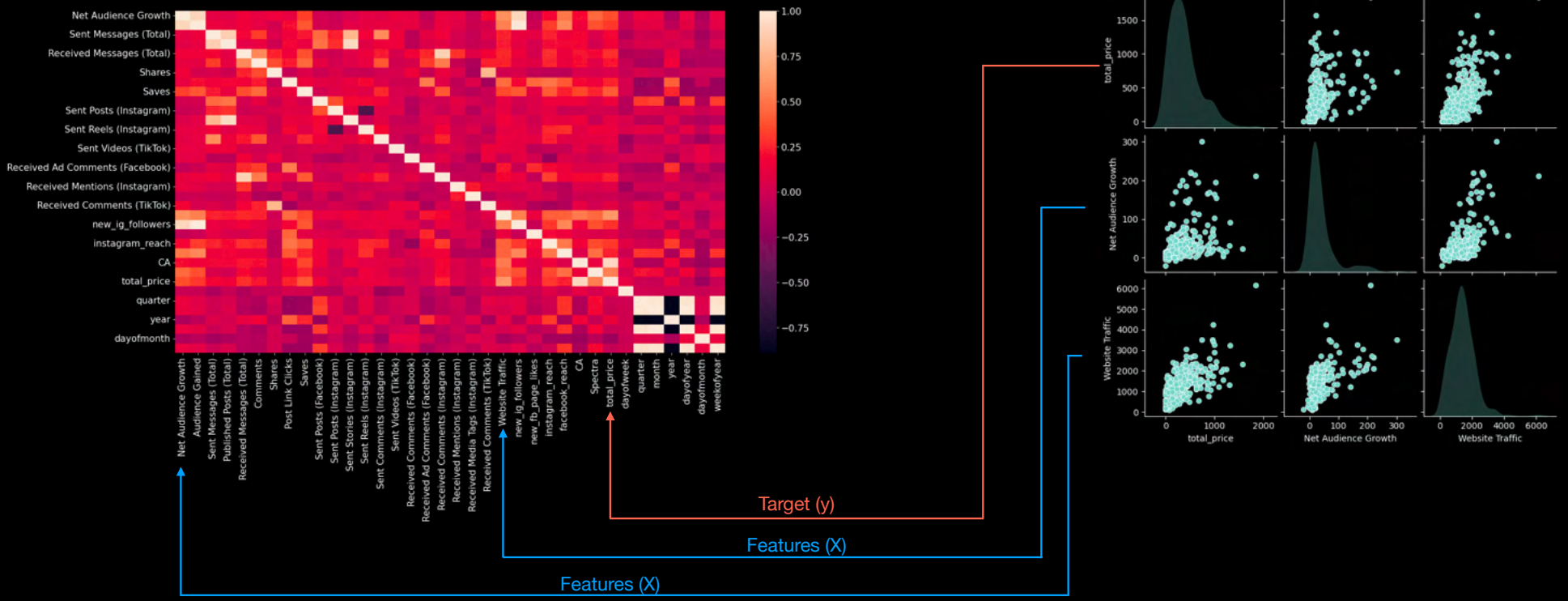
EXPLORATIVE DATA ANALYSIS





EXPLORATIVE DATA ANALYSIS

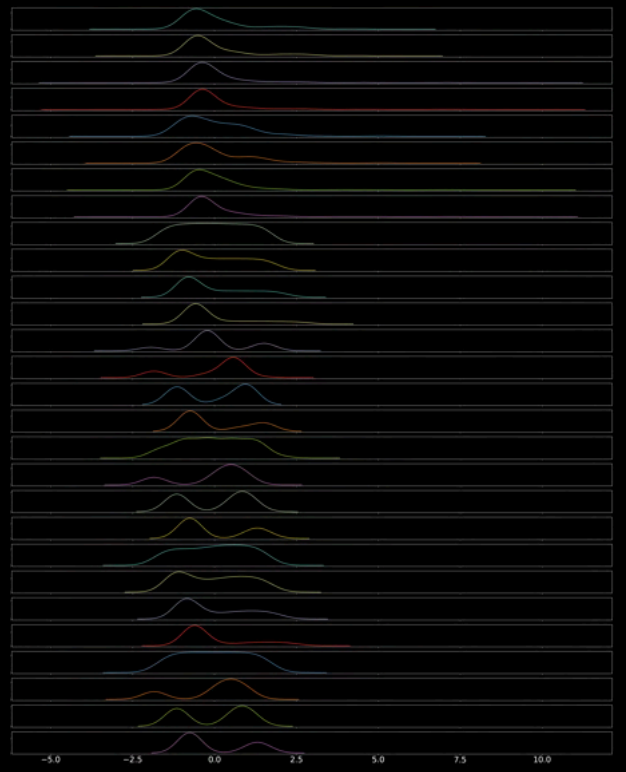
The way the Shopify API presents the gross sales is by the 'Total_Price' of the consumer cart. This information was defined as the Target (y). As we aim to develop a second model more focused in Marketing, it was decided to use Website Traffic and Net Audience Growth as the main Features (X).



DATA PREPARATION

Since this is a forecasting model, lag features were introduced with a forecasting horizon of 30 days (lag_1), 60 days (lag_2) and 90 days (lag_3).

Additionally, the Features were scaled utilising SciKit-Learn's Standard Scaler.

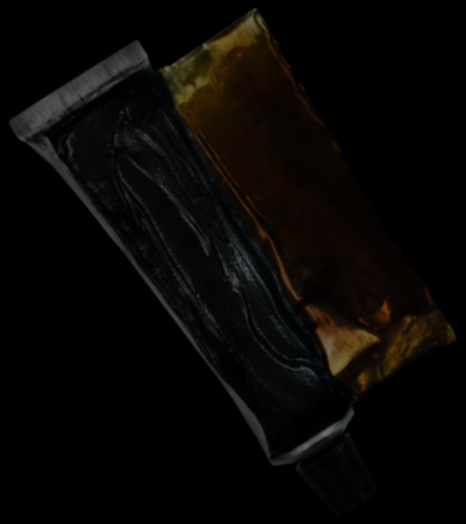


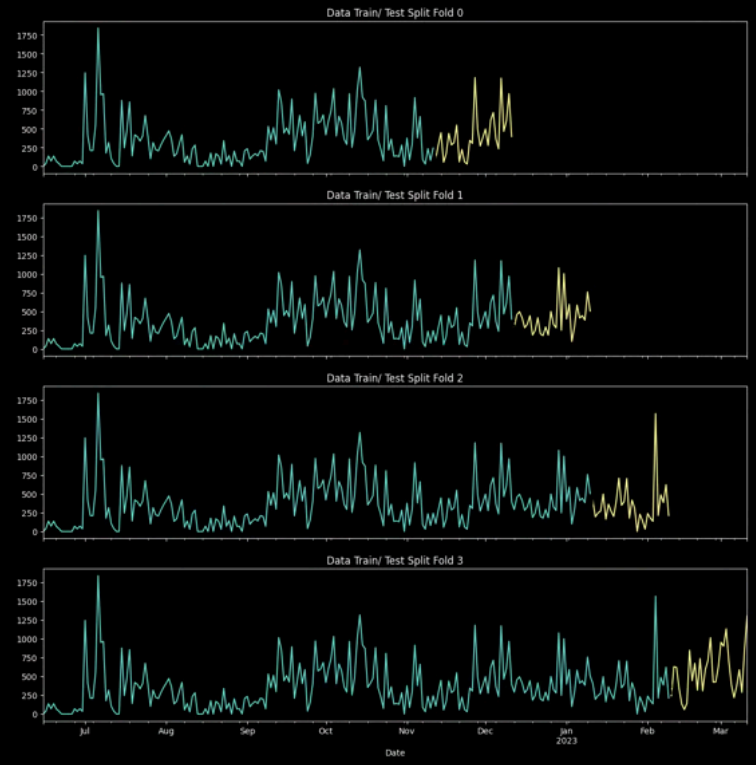
MODELLING

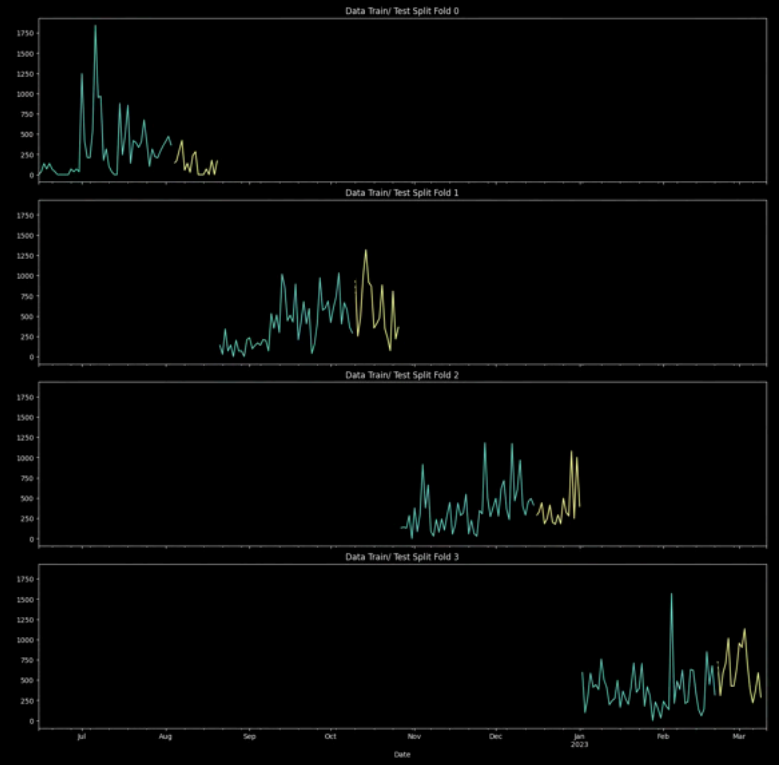
- The regressions models utilised were :
  - XGBoost
  - Random Forest
  - Elastic Net
  - Ensemble (Voting Regressor
- Two cross-validation time split series methods were used for the hyperparameter estimation of the ML models:
  - K-fold Time Series Split
  - Blocking Time Series Split
- Using the two different types of the split, a parameter search was conducted using two methods:
  - GridSearchCV
  - Iterative method
- From these two models, K-fold was chosen as it performed best in comparison to the blocking time series split.

CROSS-VALIDATION

- Two cross-validation time split series methods were used for the estimation of the ML parameters:
  - K-fold
  - Blocking Time Series Split







MODEL EVALUATION

MSE of XGBoost using GridSeachCV & k-fold Train Test Split: 362.37
MSE of XGBoost using GridSeachCV & Blocking Train Test Split: 354.13
MSE of XGBoost using GridSeachCV & Regular Train Test Split: 354.13

MSE of XGBoost using Iterative Method and K-fold Time Split Series: 265.32
MSE of XGBoost using Iterative Method and Regular Train Test Split: 277.91
MSE of of XGBoost using Iterative Method and Blocking Time Split Series: 117.60

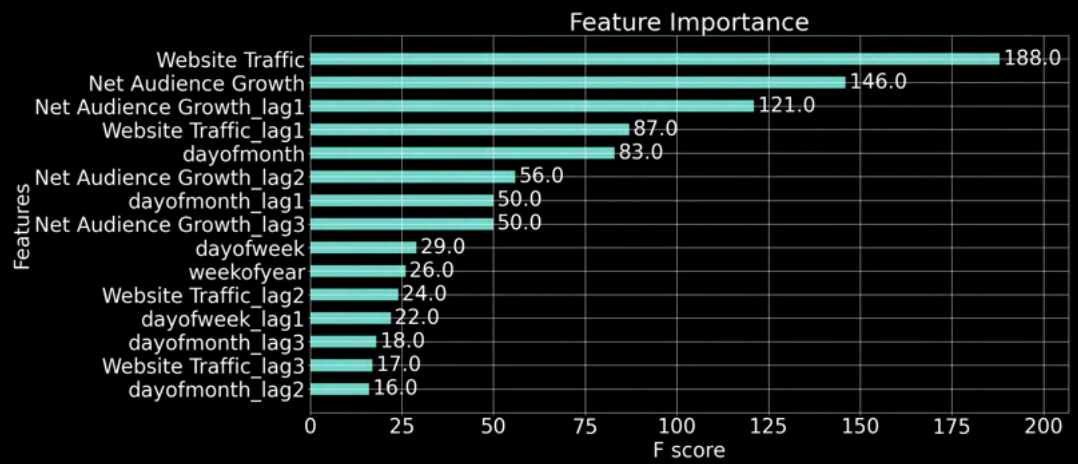
MSE of Random Forest using Iterative Method and Regular Train Test Split: 273.40
MSE of Random Forest using Iterative Method and K-fold Time Series Split: 211.81
MSE of Random Forest using Iterative Method and Blocking Time Series Split: 220.84

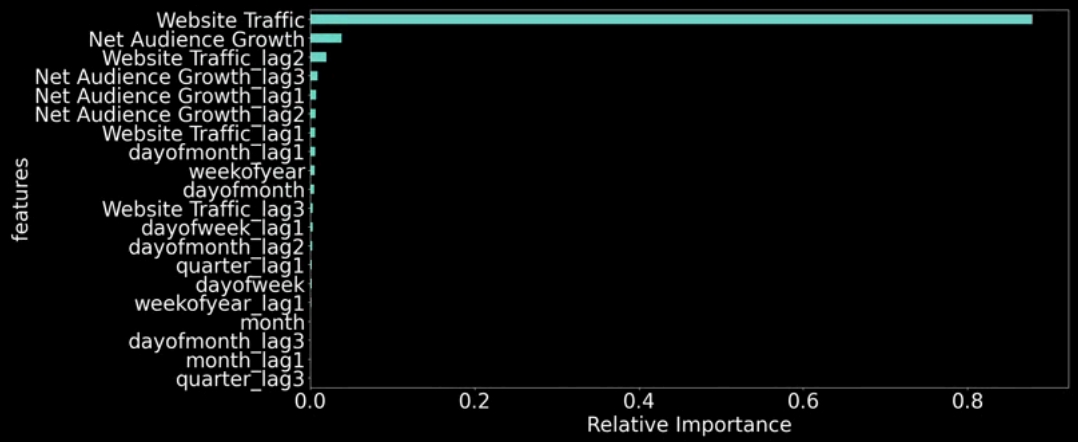
MSE of Elastic Net using Iterative Method and Train Test Split: 283.15
MSE of Elastic Net using Iterative Method and K-fold Time Series Split: 220.82
MSE of Elastic Net using Iterative Method and Blocking Time Series Split: 169.07

FEATURE IMPORTANCE

For XGBoost and Random Forest models, the feature importance was also assessed. As shown in the heat map, and as expected, Website Traffic has shown a strong positive correlation to sales.

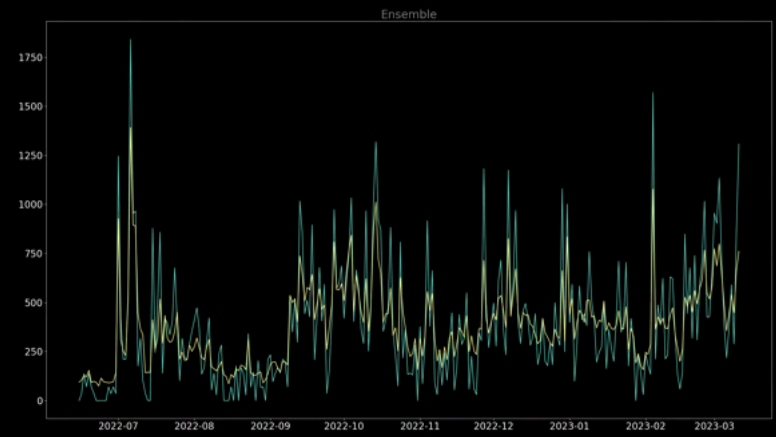
Since Elastic Net is a linear regression model, all parameters present the same importance and, therefore, are not shown here.

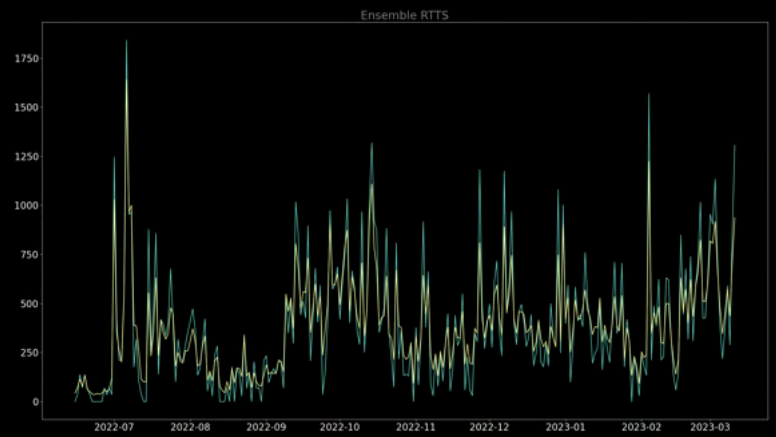




MODEL EVALUATION

Aiming to take advantage of the benefits of linear regression models coupled with the power of tree-based models, a Voting Regressor was also used. This model utilised the RMSE of each model previously discussed as a manner to weight the importance of each algorithm in the predictions.

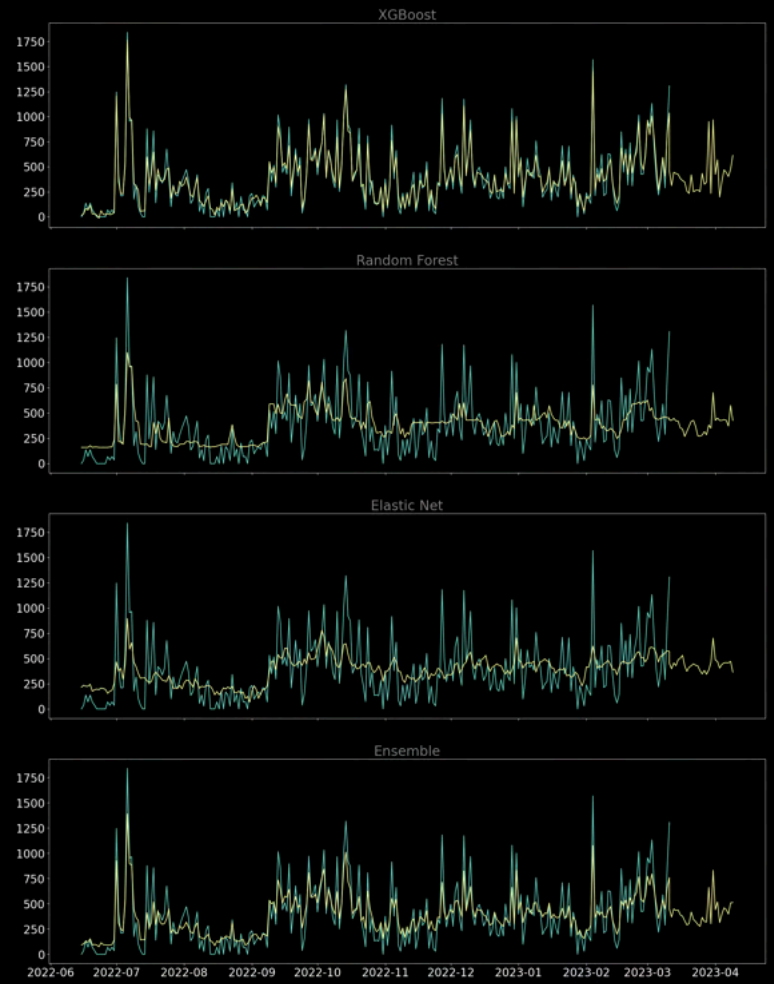


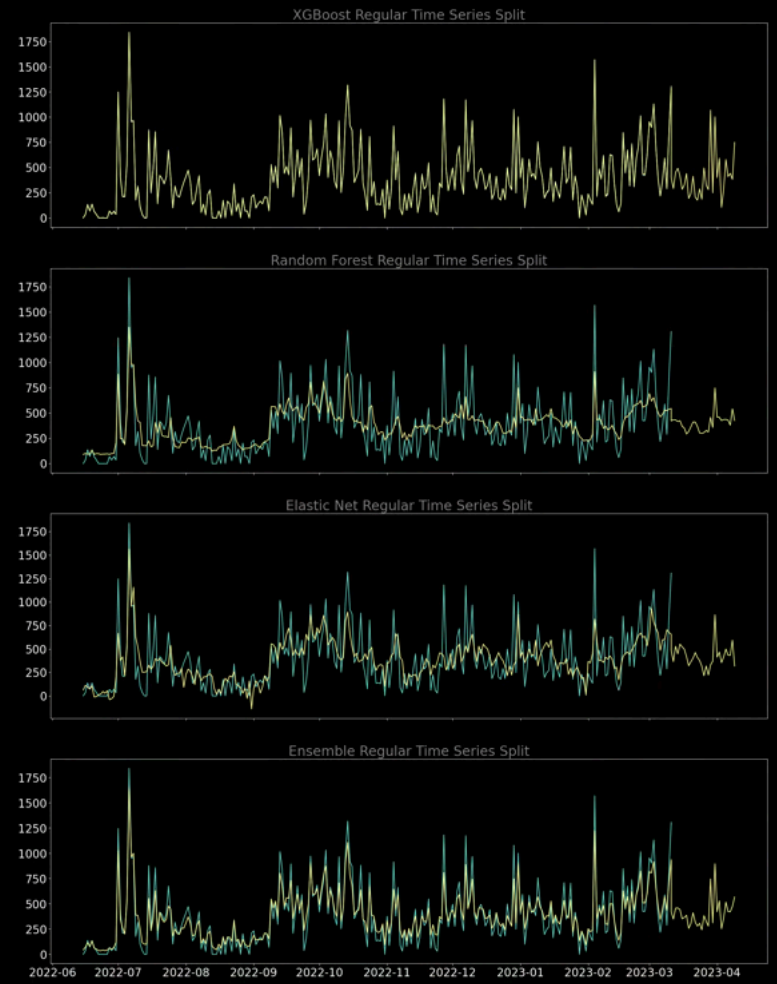


FUTURE PREDICTIONS

We can easily observe the benefits of adding a Voting Regressor. This model is capable of solving the bias–variance dilemma. For example, highly complex models, such as XGBoost, comes with low bias and high variance. On the other hand, simpler models like Elastic Net come with high bias and low variance.

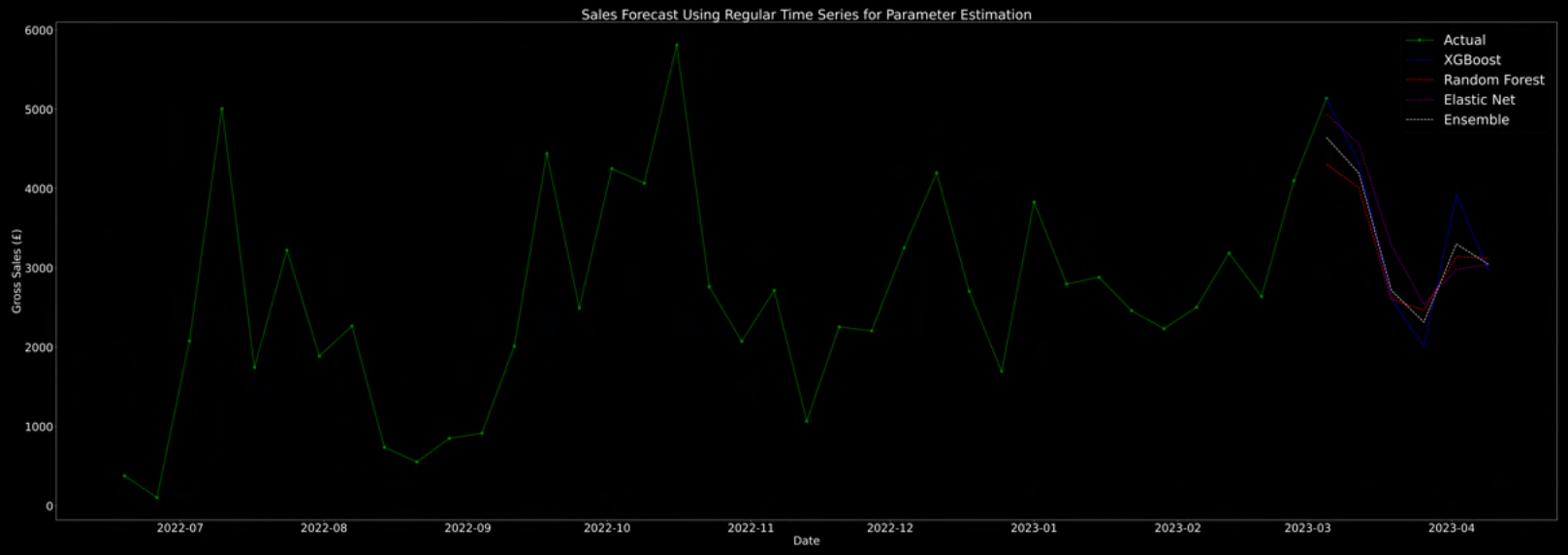
Introducing diversity by combining different approaches leads to a reduction of variance.





FUTURE PREDICTIONS

FUTURE PREDICTIONS




STREAMLIT APPLICATION
LINK


JUPYTER NOTEBOOK
LINK

NEXT STEPS

- Test additional scalers
- Develop a new ML classification model that analyses marketing activations (i.e best influencers to collaborate with, social events, popups), Ad trends and new R&D projects. The aim of this model will be to predict Net Growth Audience and Website Traffic (the forecast model inputs) aiming to optimise the best opportunities for new investments.