Data Glacier

Your Deep Learning Partner

Forecast Project – Final Results

Virtual Internship

19-8-2023

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Model results

Final Results

Executive Summary

This project involves forecasting sales through a systematic approach encompassing tasks such as data cleaning, variable relevance sorting, exploratory data analysis (EDA) with a time series focus, feature engineering, and model selection, training, error estimation, and tuning. The main goal is to predict sales accurately for informed business decisions. Data cleaning ensures data quality, followed by sorting variables by their impact on sales. EDA emphasizes time series patterns, aiding feature engineering. Model selection involves choosing the best-fit model, which is trained, evaluated for accuracy, and fine-tuned. The project aims to construct a reliable sales forecasting model for effective resource allocation and decision-making based on y target ‘Sales’.

Problem Statement

The Client:
The large company who is into beverages business in Australia. They sell their products through various super-markets and also engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality. They needed forecast of each of products at item level every week in weekly buckets.

Objective:

1. Build at least 4-5 multivariable forecasting model which included ML or Deep Learning based Model in PySpark leveraging parallel computing techniques.
2. Demonstrate best in class forecast accuracy (Forecast Accuracy = 1 - Wt. MAPE where Wt. MAPE = sum(Error)/sum(Actual).
3. Write a code in such a way you run the model in least time.
4. Demonstrate explainability in the form of contribution of each variables
5. Leveage Feature Engineering concepts to derive more variables to gain accuracy improvement

Approach

What's the key feature to boost sales with discounts during drops?

Which top 3 products need investment due to high sales (SKU1, 3, 6)?

When does SKU1 excel with >10% sales during holidays?

How do In-Store promos affect SKU1-5 sales instantly?

Do Catalogue promos increase sales for all products next day?

Does Store-End promo spike SKU4 sales on the next day?

What's the daily sales range (95%) for each SKU (1-6)?

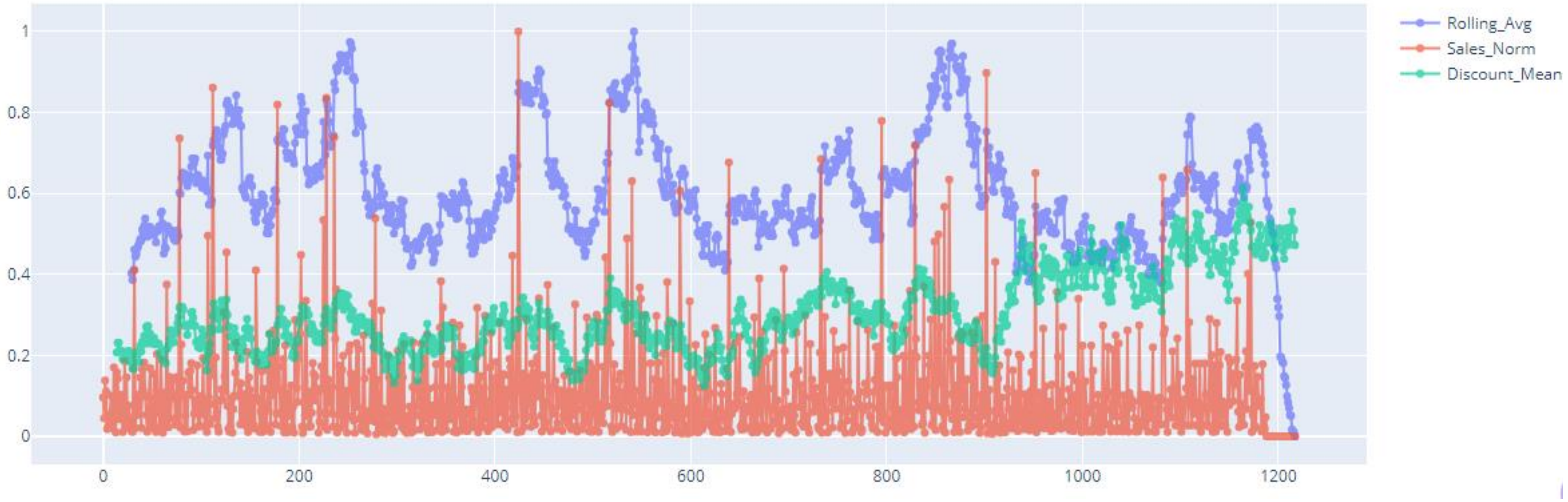
Does Google Mobility impact sales; should sales be closed?

What are optimal discount ranges for each SKU (1-6) to boost sales and profit balance?

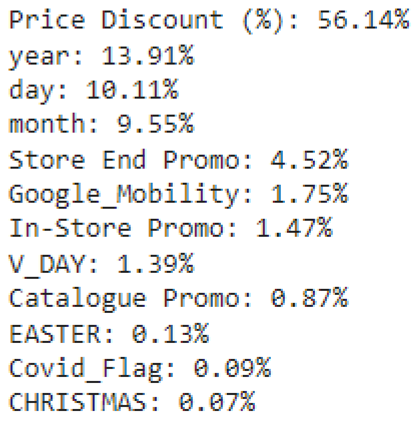
Result 1

EDA

Price Discount is the most impactful feature within the dataset to optimize sales for every product, whenever sales are dropping low, it would be the best to offer discounts.

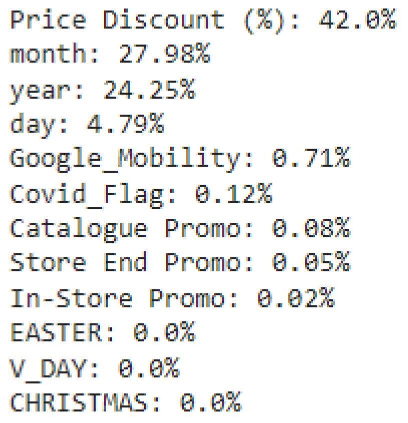


Random Forest:



```
Price Discount (%): 56.14%
year: 13.91%
day: 10.11%
month: 9.55%
Store End Promo: 4.52%
Google_Mobility: 1.75%
In-Store Promo: 1.47%
V_DAY: 1.39%
Catalogue Promo: 0.87%
EASTER: 0.13%
Covid_Flag: 0.09%
CHRISTMAS: 0.07%
```

Gradient Boost:

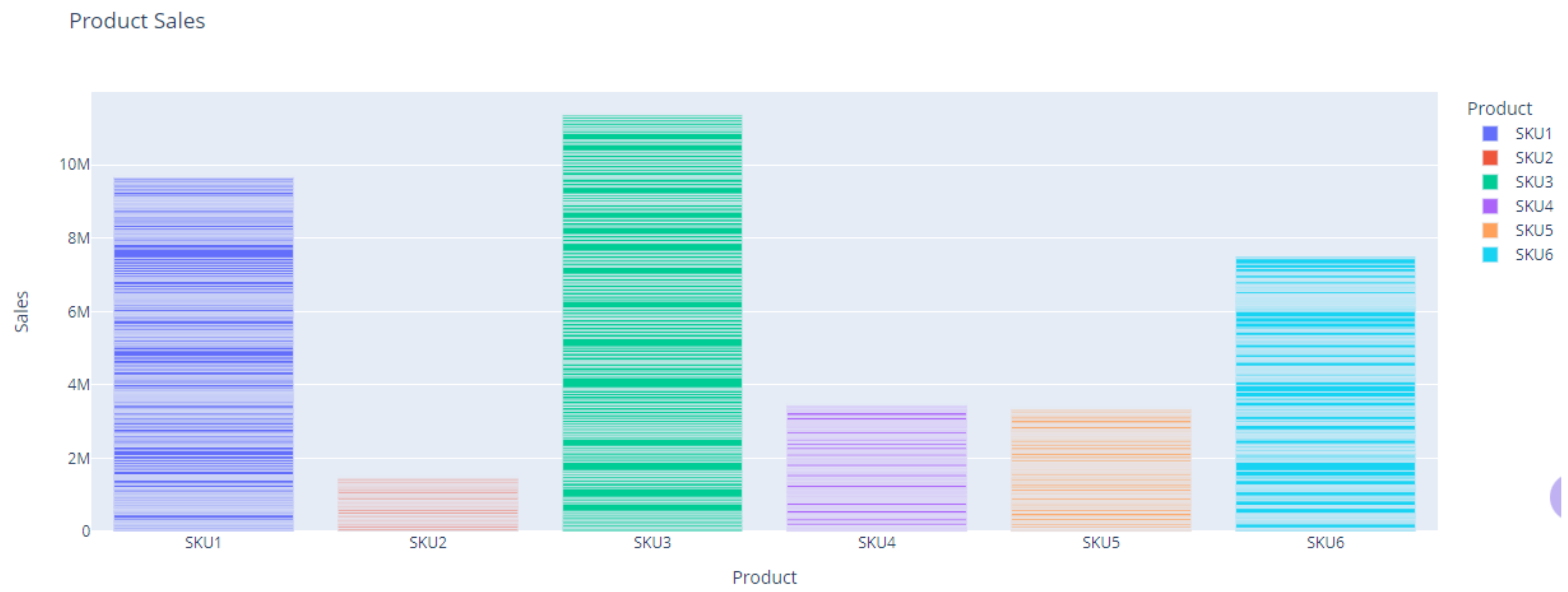


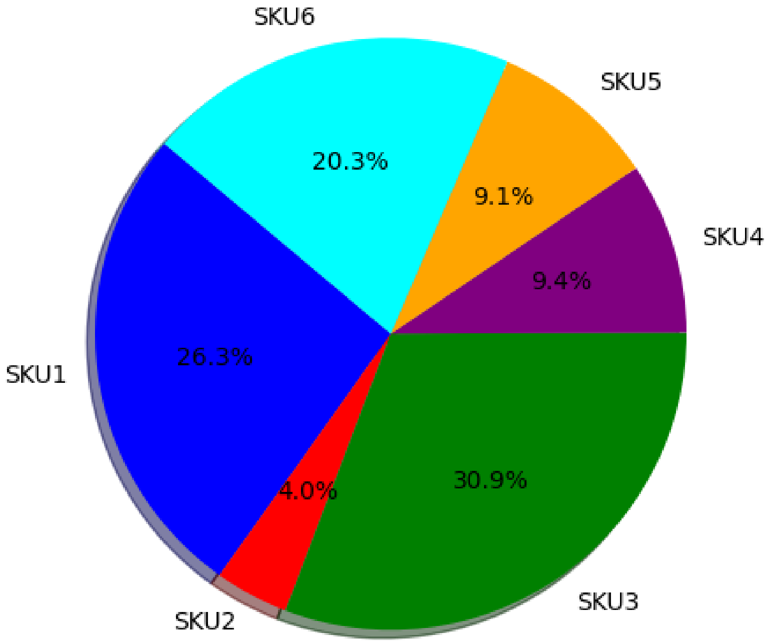
```
Price Discount (%): 42.0%
month: 27.98%
year: 24.25%
day: 4.79%
Google_Mobility: 0.71%
Covid_Flag: 0.12%
Catalogue Promo: 0.08%
Store End Promo: 0.05%
In-Store Promo: 0.02%
EASTER: 0.0%
V_DAY: 0.0%
CHRISTMAS: 0.0%
```

Result 2

EDA

SKU1, 3 and 6 have the majority of the sales, investments should prioritize these 3 products.



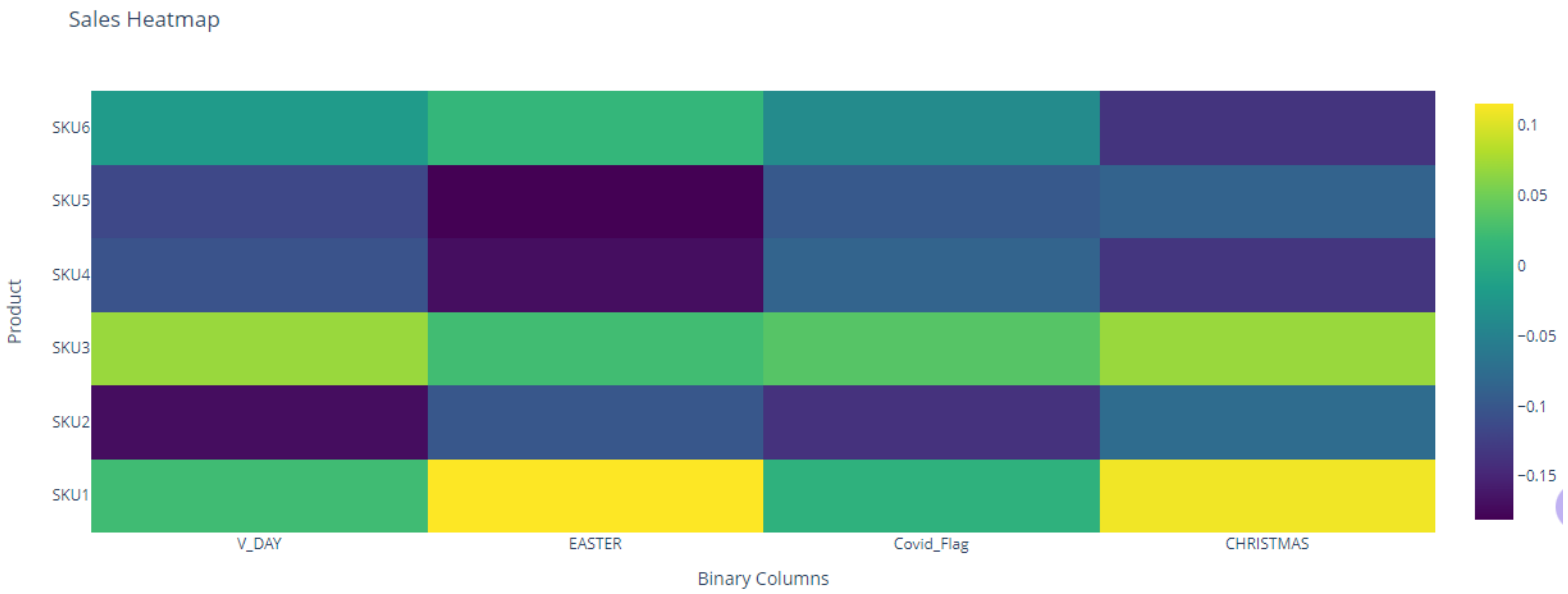


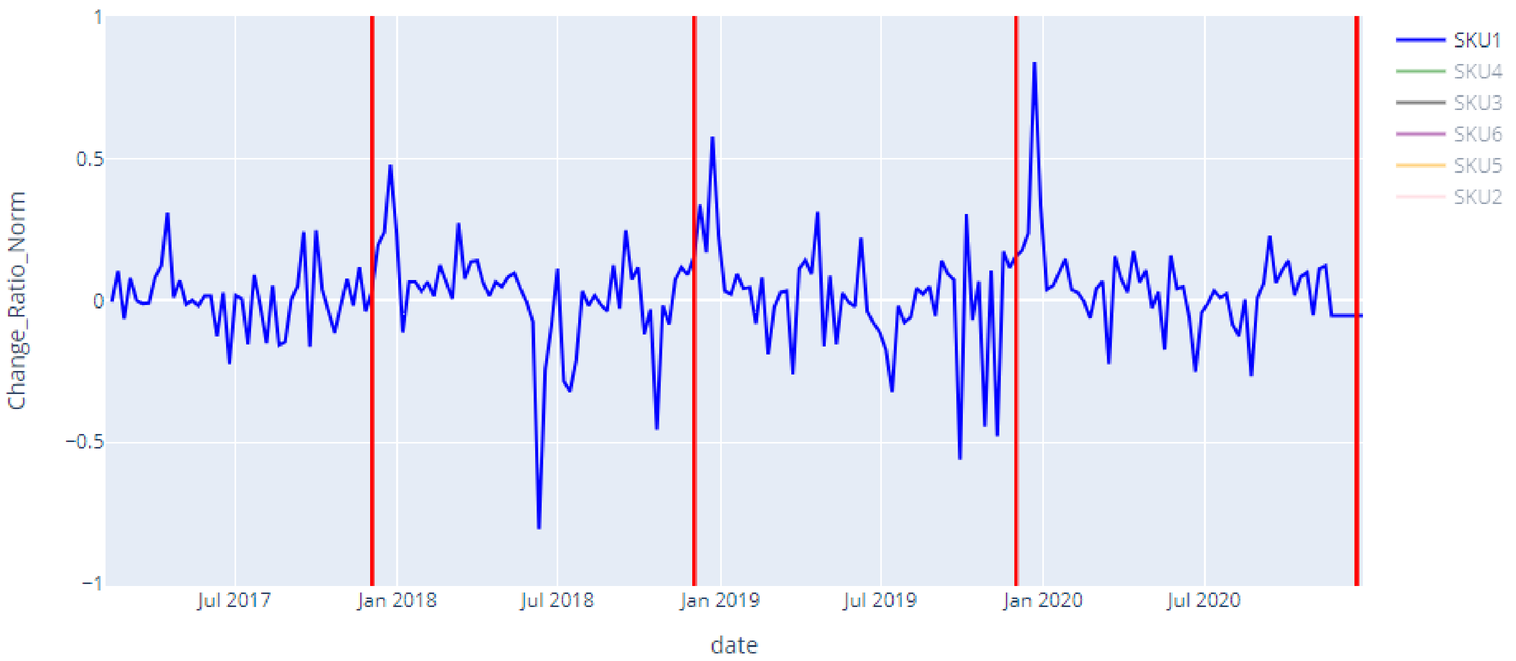
Result 3

EDA

During Christmas and Easter, SKU1 is the optimal product for sales, it would have > 10% of it's standard sales rate during that day..

Data Glacier
Your Deep Learning Partner

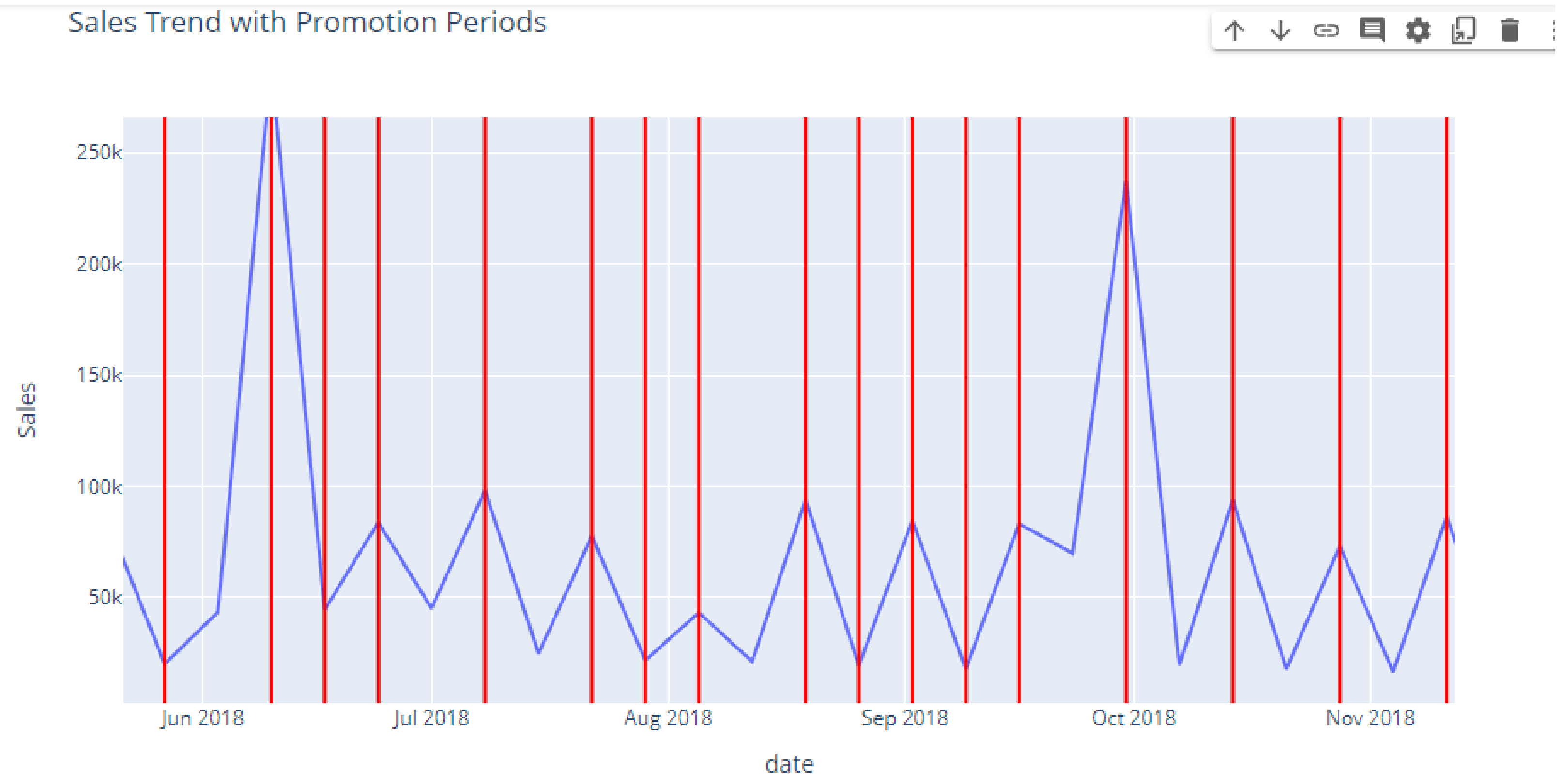


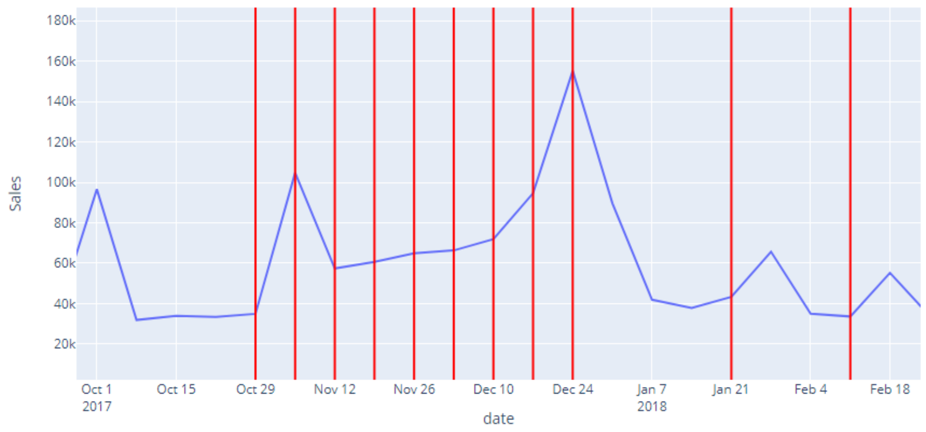


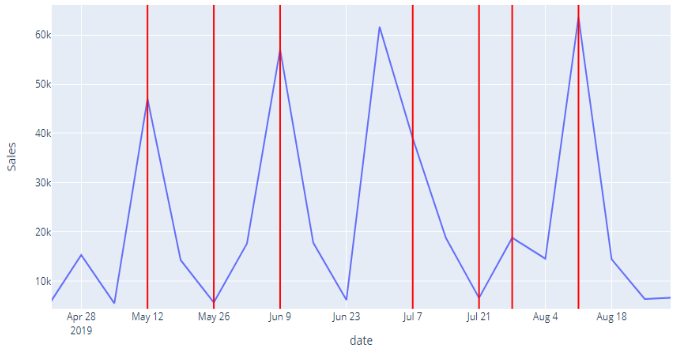
Result 4

EDA

For products from SKU1 to SKU5, offering In-Store promotions would increase sales the same day or the day after the promotion.



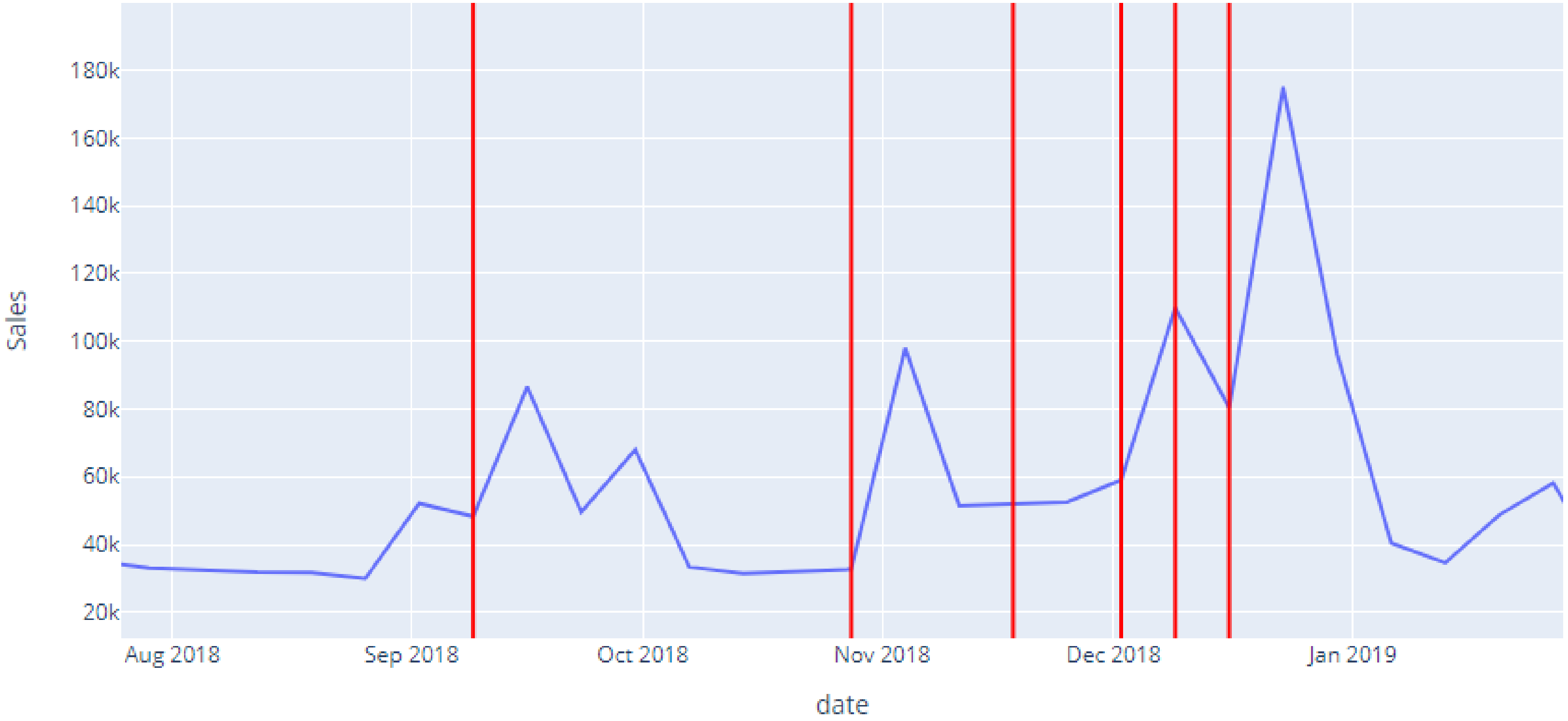




Result 5

EDA

For all products, offering Catalogue promotions would increase sales the day after the promotion.

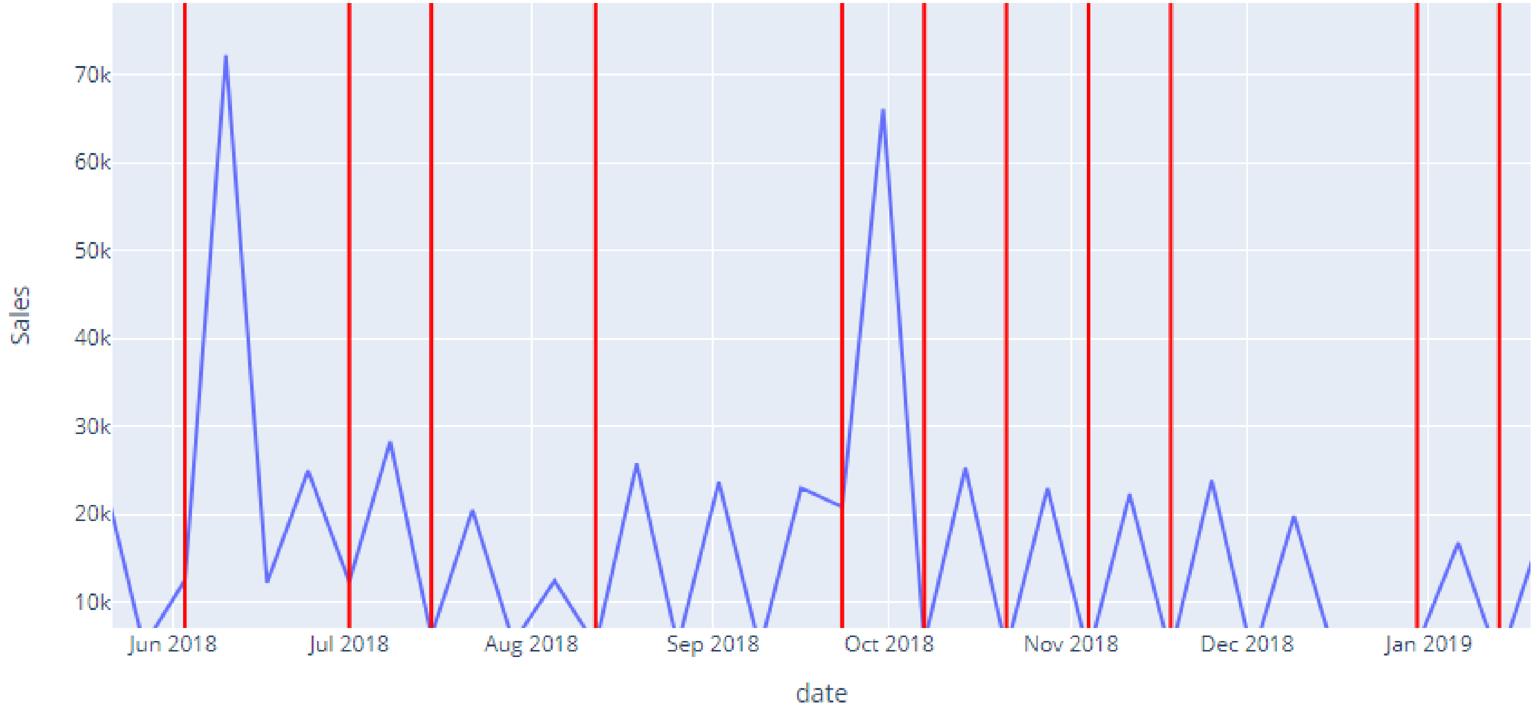


Data Glacier
Your Deep Learning Partner

Result 6

EDA

For SKU4, there is a huge increase on sales the day after a Store-End promotion was chosen.



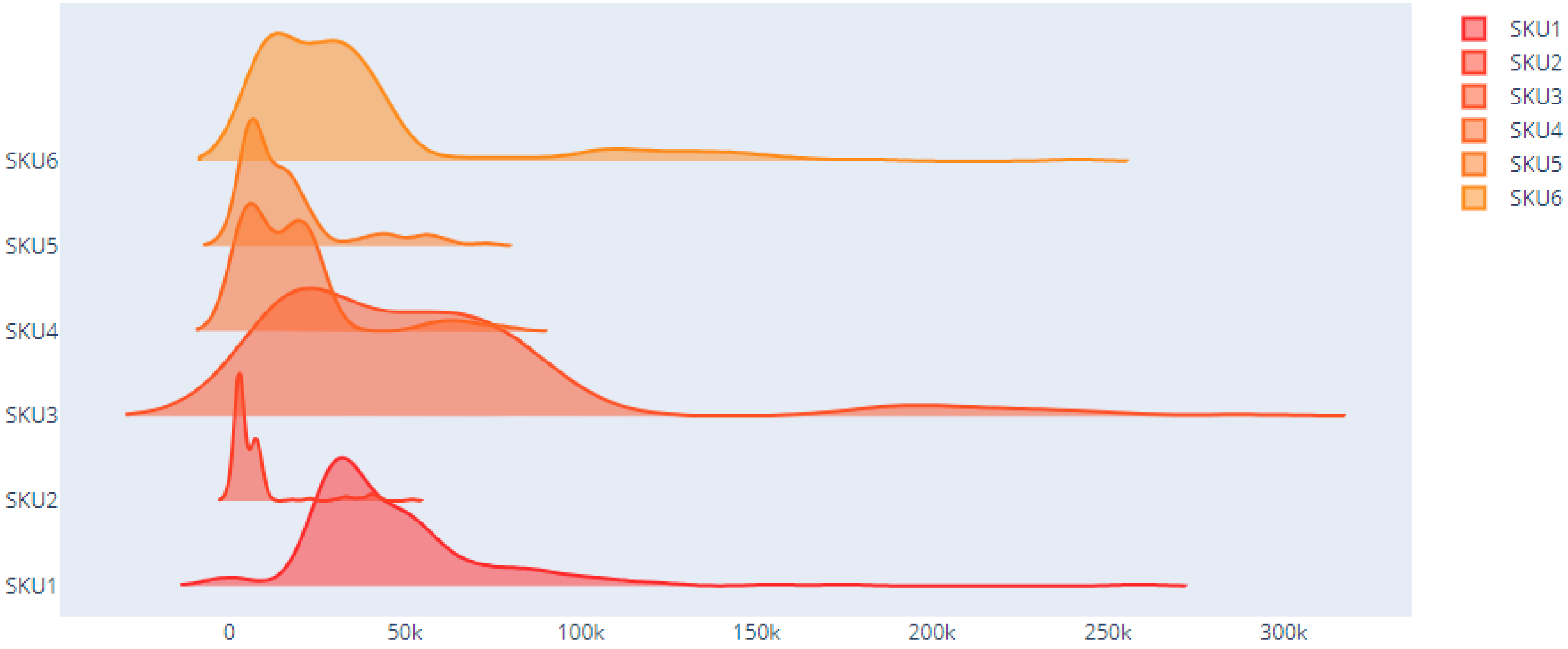
Data Glacier
Your Deep Learning Partner

Result 7

EDA

Each product has a 95% probability of generating these amount of sales daily (to prepare budget before hand):

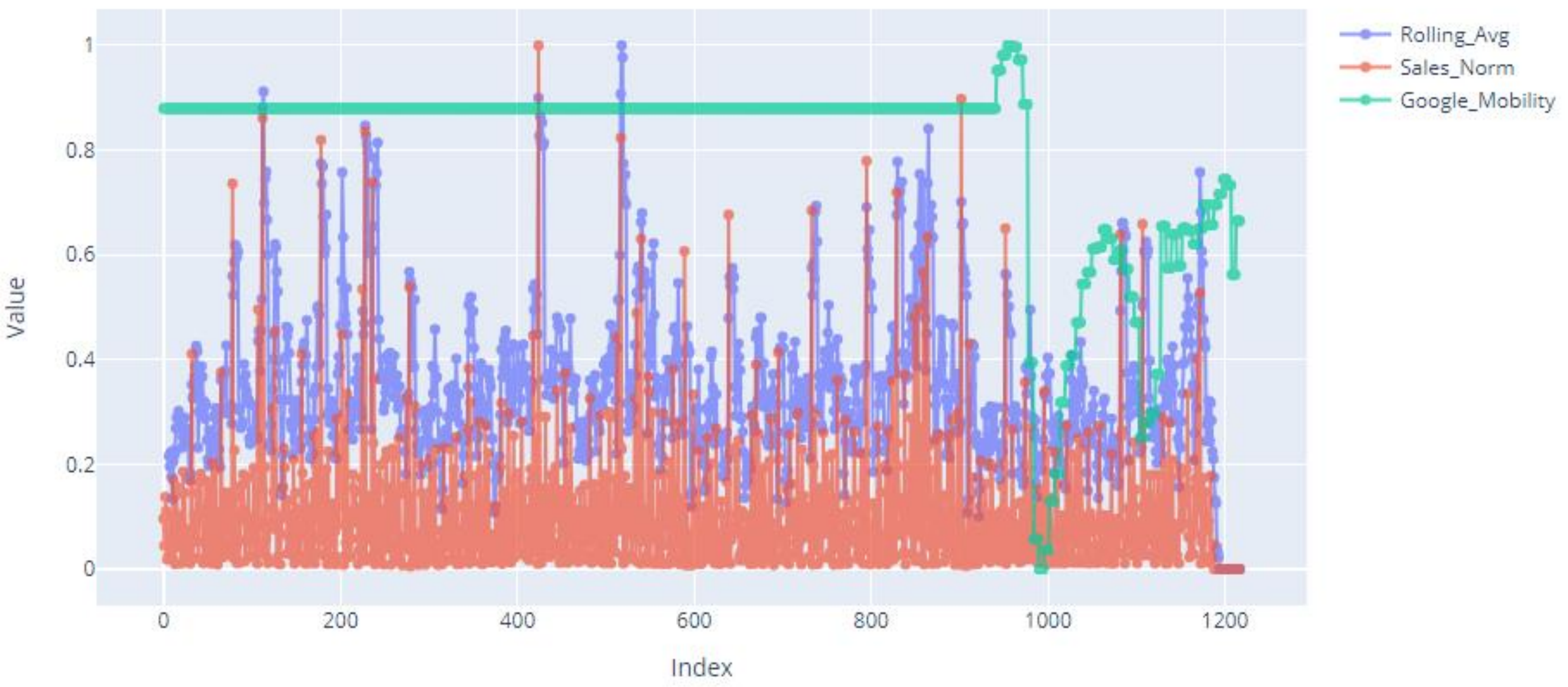
SKU1: [43574.16, 52112.24]
SKU2: [6107.43, 8654.14]
SKU3: [49403.63, 63337.02]
SKU4: [14931.67, 19247.07]
SKU5: [14460.09, 18752.23]
SKU6: [32916.42, 43159.34]

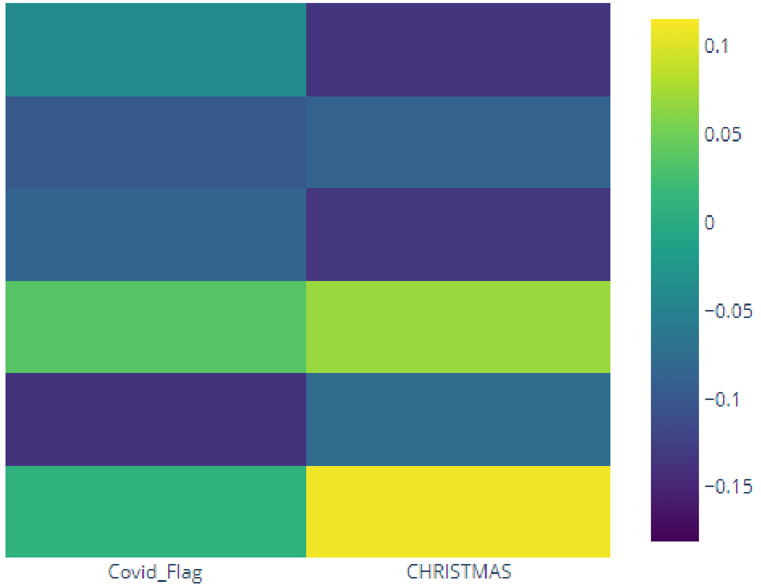


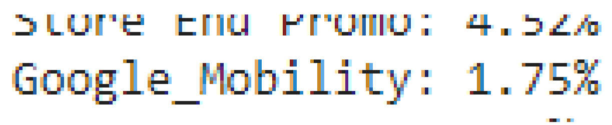
Result 8

EDA

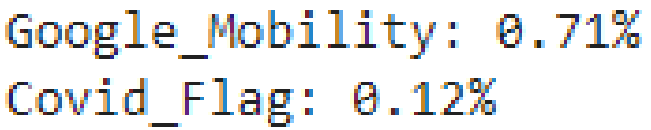
Google Mobility has no effect over the sales of any product, so there should be no need to close off sales given this situation.







```
Store End Promo: 4.52%
Google_Mobility: 1.75%
```



```
Google_Mobility: 0.71%
Covid_Flag: 0.12%
```

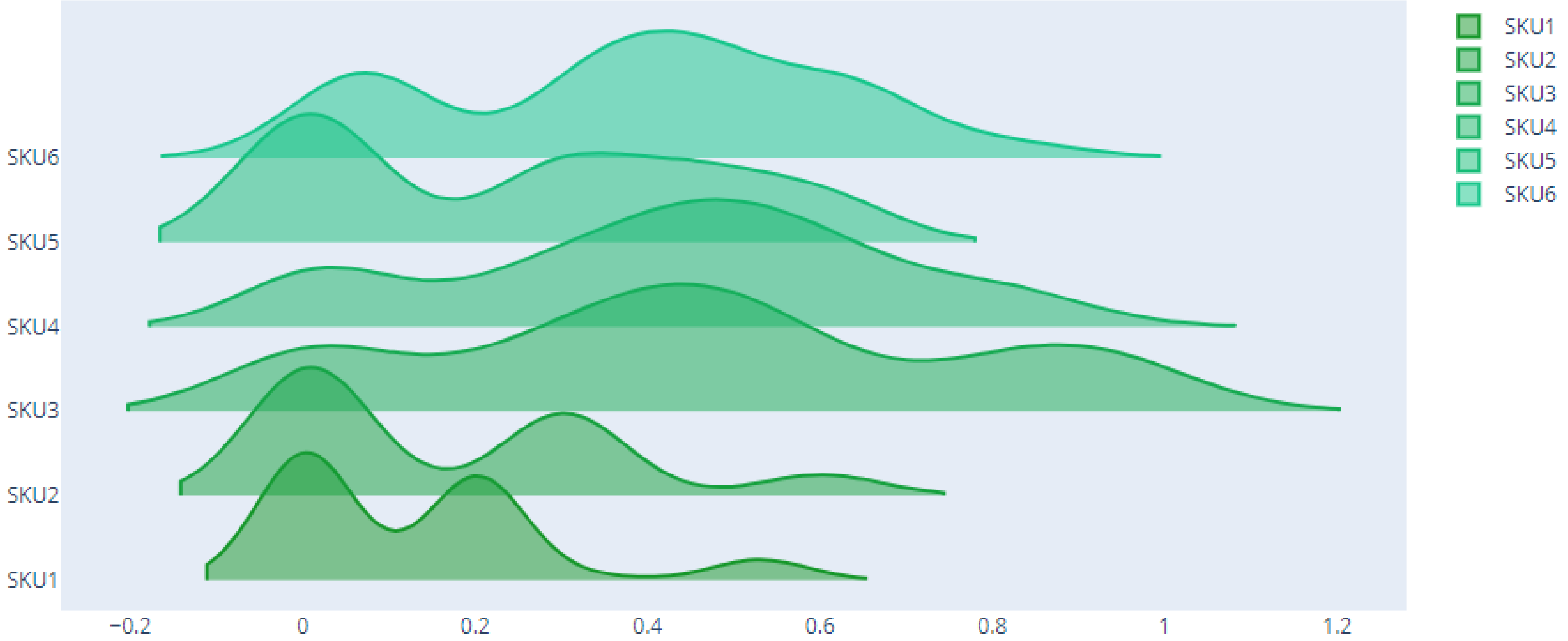
Result 9

EDA

The following discounts ([0,1] <=> [0%, 100%]) are optimal for product sales increase and company's profit balance:

SKU1: [0.11, 0.15]
SKU2: [0.14, 0.19]
SKU3: [0.42, 0.50]
SKU4: [0.39, 0.46]
SKU5: [0.22, 0.28]
SKU6: [0.36, 0.42]

Data Glacier
Your Deep Learning Partner



EDA Summary

This forecast project has yielded valuable insights and recommendations. The analysis has shown that offering price discounts, especially during sales drops, can significantly impact sales. Prioritizing investments in top-selling products (SKU1, 3, 6) is crucial. SKU1 performs exceptionally well during Christmas and Easter, with over 10% increased sales. In-store promotions have an immediate effect on SKU1-5 sales, while Catalogue promotions consistently increase sales the next day for all products. Store-End promotions notably enhance SKU4 sales the day after implementation. Each SKU's daily sales range (95%) aids budget planning. Google Mobility has no sales impact. Optimal discount ranges to maximize sales and profit balance vary for each SKU (1-6). These findings offer actionable insights for effective sales strategies and resource allocation.

RECCOMENDATIONS

Holiday Recommendations:

Leverage SKU1's exceptional performance during Christmas and Easter by strategically aligning marketing efforts and product availability to capture the increased demand during these holiday periods.

Promotional Recommendations:

Utilize end-store promotions to stimulate immediate sales for SKU4. These promotions have shown to yield instant results and can be employed strategically to enhance sales performance.

Utilize in-store promotions to stimulate immediate sales for SKU1-5. These promotions have shown to yield instant results and can be employed strategically to enhance sales performance.

Utilize catalogue promotions to generate increased sales for all products. These promotions consistently lead to higher sales figures on the day following the promotion.

RECCOMENDATIONS

General Recommendations:

Focus investment efforts on SKU1, 3, and 6 due to their high sales volume. Allocating resources to these products can yield substantial returns and sustain the company's growth.

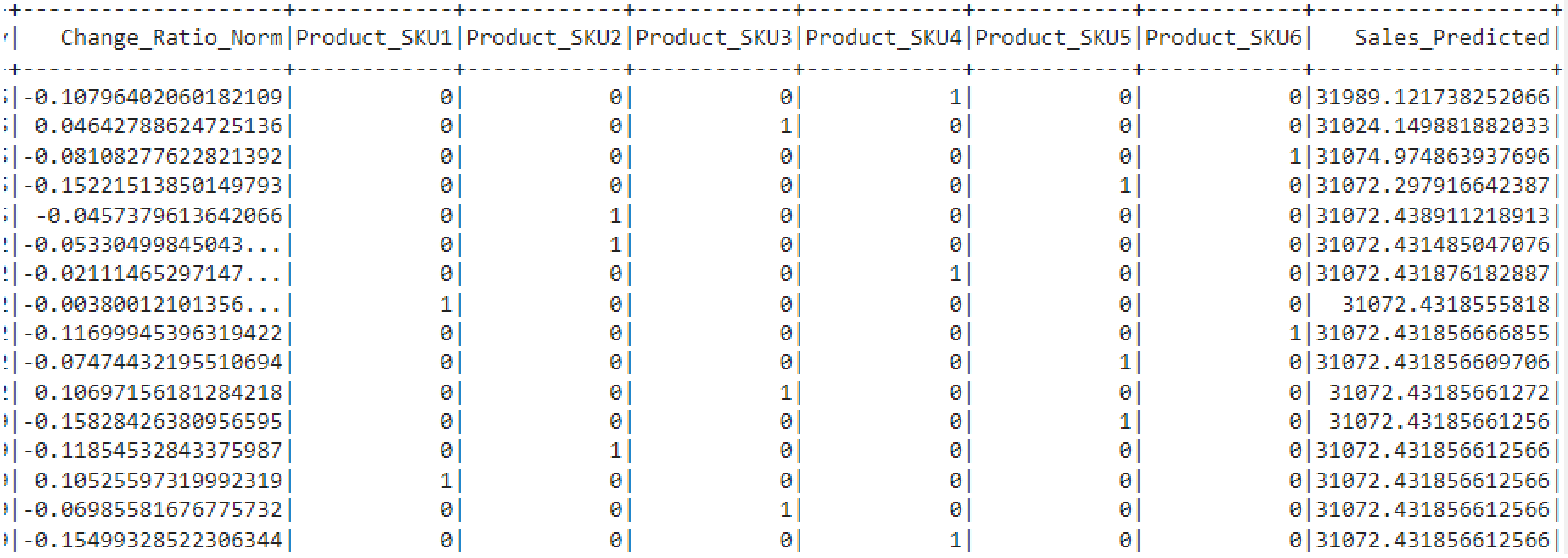
Utilize the provided daily sales probability ranges (95%) for each SKU (1-6) to facilitate budget planning and resource allocation, ensuring readiness for various sales scenarios.

Given the lack of sales impact from Google Mobility, there is no need to make drastic changes to sales strategies based on this factor. The focus can remain on other proven strategies.

Implement the recommended optimal discount ranges for each SKU (1-6) to strike a balance between boosting sales and maintaining a healthy profit margin.

AIRMA

Model Results

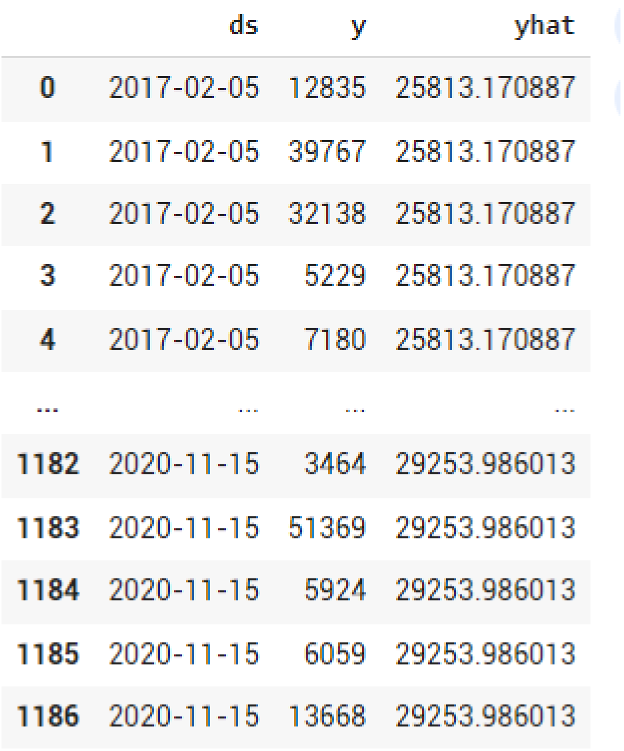


```
+--------------------+------------+------------+------------+------------+------------+------------+------------------+
|   Change_Ratio_Norm|Product_SKU1|Product_SKU2|Product_SKU3|Product_SKU4|Product_SKU5|Product_SKU6|   Sales_Predicted|
+--------------------+------------+------------+------------+------------+------------+------------+------------------+
|-0.10796402060182109|           0|           0|           0|           1|           0|           0|31989.121738252066|
| 0.04642788624725136|           0|           0|           1|           0|           0|           0|31024.149881882033|
|-0.08108277622821392|           0|           0|           0|           0|           0|           1|31074.974863937696|
|-0.15221513850149793|           0|           0|           0|           0|           1|           0|31072.297916642387|
| -0.0457379613642066|           0|           1|           0|           0|           0|           0|31072.438911218913|
|-0.05330499845043...|           0|           1|           0|           0|           0|           0|31072.431485047076|
|-0.02111465297147...|           0|           0|           0|           1|           0|           0|31072.431876182887|
|-0.00380012101356...|           1|           0|           0|           0|           0|           0|  31072.4318555818|
|-0.11699945396319422|           0|           0|           0|           0|           0|           1|31072.431856666855|
|-0.07474432195510694|           0|           0|           0|           0|           1|           0|31072.431856609706|
| 0.10697156181284218|           0|           0|           1|           0|           0|           0| 31072.43185661272|
|-0.15828426380956595|           0|           0|           0|           0|           1|           0| 31072.43185661256|
|-0.11854532843375987|           0|           1|           0|           0|           0|           0|31072.431856612566|
| 0.10525597319992319|           1|           0|           0|           0|           0|           0|31072.431856612566|
|-0.06985581676775732|           0|           0|           1|           0|           0|           0|31072.431856612566|
|-0.15499328522306344|           0|           0|           0|           1|           0|           0|31072.431856612566|
```

AIRMA can't handle complex non-linear relationships between variables, nor binary columns, nor multivariable training, leading to a mean prediction which doesn't supply the expected output for the company, so I'll be moving to the next option for forecast non-multivariable model.

Model Results

Prophet

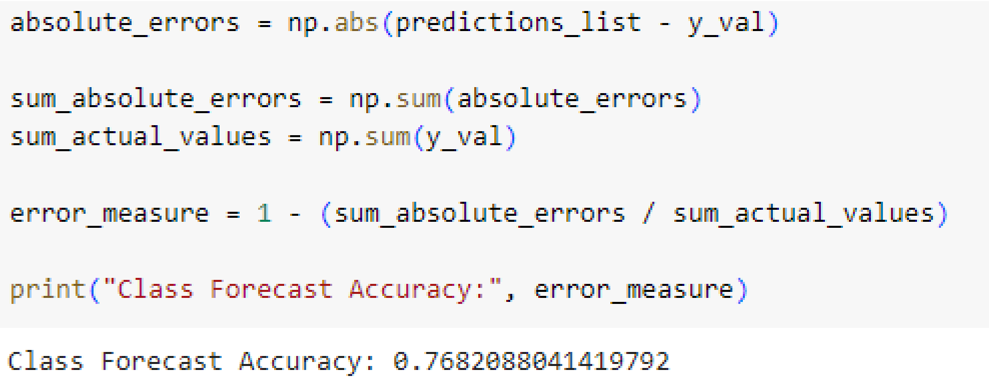


	ds	y	yhat
0	2017-02-05	12835	25813.170887
1	2017-02-05	39767	25813.170887
2	2017-02-05	32138	25813.170887
3	2017-02-05	5229	25813.170887
4	2017-02-05	7180	25813.170887
...	...	...	...
1182	2020-11-15	3464	29253.986013
1183	2020-11-15	51369	29253.986013
1184	2020-11-15	5924	29253.986013
1185	2020-11-15	6059	29253.986013
1186	2020-11-15	13668	29253.986013

Prophet fails because it's a model for date only predictions, and by results from EDA, this dataset doesn't have sales patterns alone for it to predict, so moving to the next model

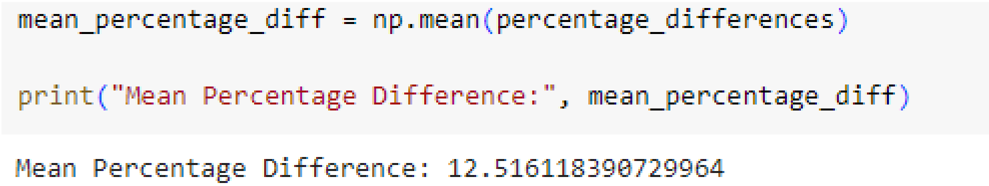
Model Results

LSTM



```
absolute_errors = np.abs(predictions_list - y_val)

sum_absolute_errors = np.sum(absolute_errors)
sum_actual_values = np.sum(y_val)

error_measure = 1 - (sum_absolute_errors / sum_actual_values)

print("Class Forecast Accuracy:", error_measure)
```

```
Class Forecast Accuracy: 0.7682088041419792
```



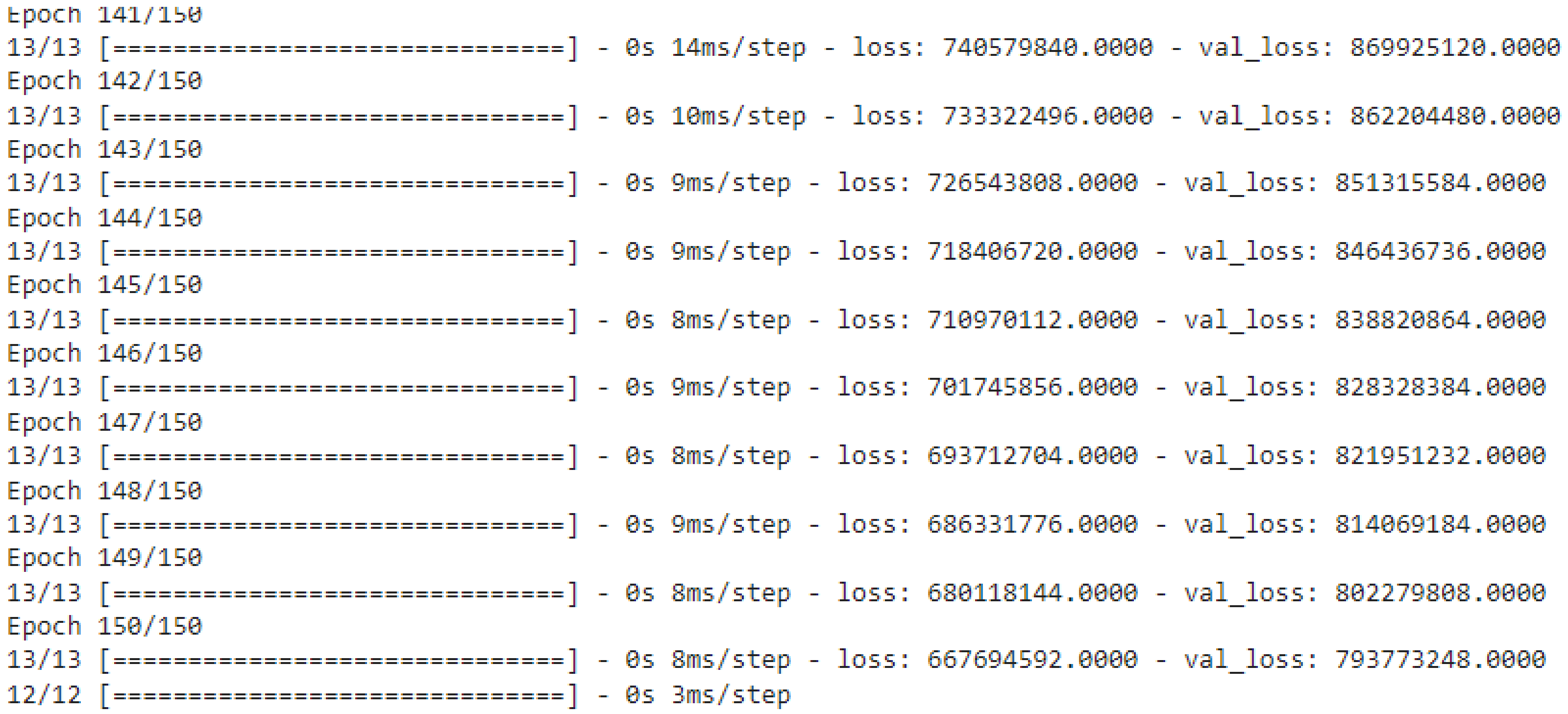
```
mean_percentage_diff = np.mean(percentage_differences)

print("Mean Percentage Difference:", mean_percentage_diff)
```

```
Mean Percentage Difference: 12.516118390729964
```

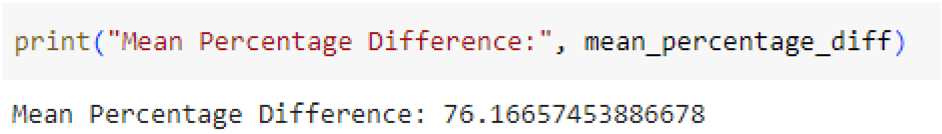
LSTM is by this point the best model for predictions based on all feature columns provided, it handles perfectly complex correlations and multivariable dependecies, making it's Class Forecast Accuracy being 76% (tested with 100 samples, a mean difference of 11% predictions vs actual values).

Gated Recurrent Unit

Model Results



```
Epoch 141/150
13/13 [==============================] - 0s 14ms/step - loss: 740579840.0000 - val_loss: 869925120.0000
Epoch 142/150
13/13 [==============================] - 0s 10ms/step - loss: 733322496.0000 - val_loss: 862204480.0000
Epoch 143/150
13/13 [==============================] - 0s 9ms/step - loss: 726543808.0000 - val_loss: 851315584.0000
Epoch 144/150
13/13 [==============================] - 0s 9ms/step - loss: 718406720.0000 - val_loss: 846436736.0000
Epoch 145/150
13/13 [==============================] - 0s 8ms/step - loss: 710970112.0000 - val_loss: 838820864.0000
Epoch 146/150
13/13 [==============================] - 0s 9ms/step - loss: 701745856.0000 - val_loss: 828328384.0000
Epoch 147/150
13/13 [==============================] - 0s 8ms/step - loss: 693712704.0000 - val_loss: 821951232.0000
Epoch 148/150
13/13 [==============================] - 0s 9ms/step - loss: 686331776.0000 - val_loss: 814069184.0000
Epoch 149/150
13/13 [==============================] - 0s 8ms/step - loss: 680118144.0000 - val_loss: 802279808.0000
Epoch 150/150
13/13 [==============================] - 0s 8ms/step - loss: 667694592.0000 - val_loss: 793773248.0000
12/12 [==============================] - 0s 3ms/step
```



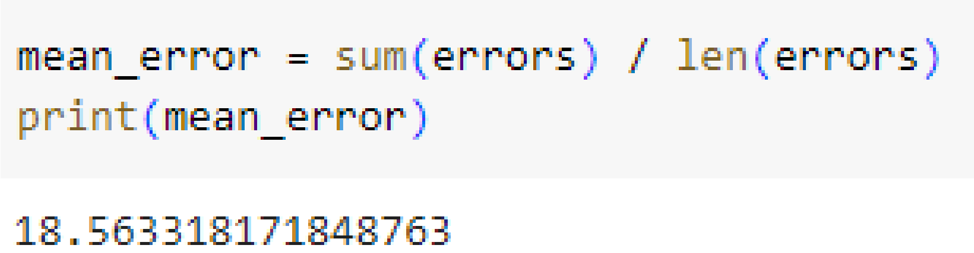
```
print("Mean Percentage Difference:", mean_percentage_diff)
```

```
Mean Percentage Difference: 76.16657453886678
```

Even if it was capable of multivariable training, it's performance wasn't better than LSTM, so it's discarded.

Gradient Boosting Regression

Model Results



```
mean_error = sum(errors) / len(errors)
print(mean_error)
```

```
18.563318171848763
```

This is the second best model so far with margin error of 18%.

Final Results

LSTM was the best model for predictions from the 5 model list proposed in week 11 + Gradient Boosting Regression, it has a best case accuracy of 76% with 11% error margin per prediction, it's input data in order of relevance goes as (based on EDA and prediction testing):

1. Price Discount (%) (Contributes at least 40%)
2. ProductX (The product to sale) (Contributes at least 25%)
3. Day of month (Contributes at least 15%)
4. Month of year (Contributes at least 10%)
5. In-Store Promo (Contributes at least 5%)
6. Store End Promo (Contributes at least 5%)

and undefined/lowest relevance input features are the rest of the original features provided in the dataset along with a daily change ratio (this can be calculated with value 0 starting at day 0, and by applying difference subtraction daily from there).

Final Results

However, holiday and promotions effects are also described in EDA results section, which are important and should not be discarded based only on the model's weights and biases.

So, any new day the company could use this model to predict for example how much payback would product SKU1 produce at the end of the day 15 of October of any year, and use EDA results to add promotions and holiday awareness to the model and so get an approximate of the expected profit that day.

Forecasting in weekly buckets would not only be possible by creating 7 predictions in order, but also it would be customizable as each parameter could be different from each day of the forecasted week, and for the best, it wouldn't be misguided by past week's unlikely events.

Forecast Project – Final Results

Data Glacier

Your Deep Learning Partner

Thank You