Let's talk about what's possible.™

Project Week Presentation

Data Pirates – Junyang Tang, Zongzhen Lin, Jinsong Zhen, Chin-Heng Lin

BEST BUY

Agenda

Data Overview	• Data Summary • Data Visualization
Feature Engineering	• Drop missing value • One-Hot encoded variables • Scale on numeric feature • Feature Elimination • Advanced models to select features
Models	• XG Boost • Prophet
Recommen-dations	• Business Recommendations

Data Summary

Besides the raw data, we added the following features to the dataset for analysis:

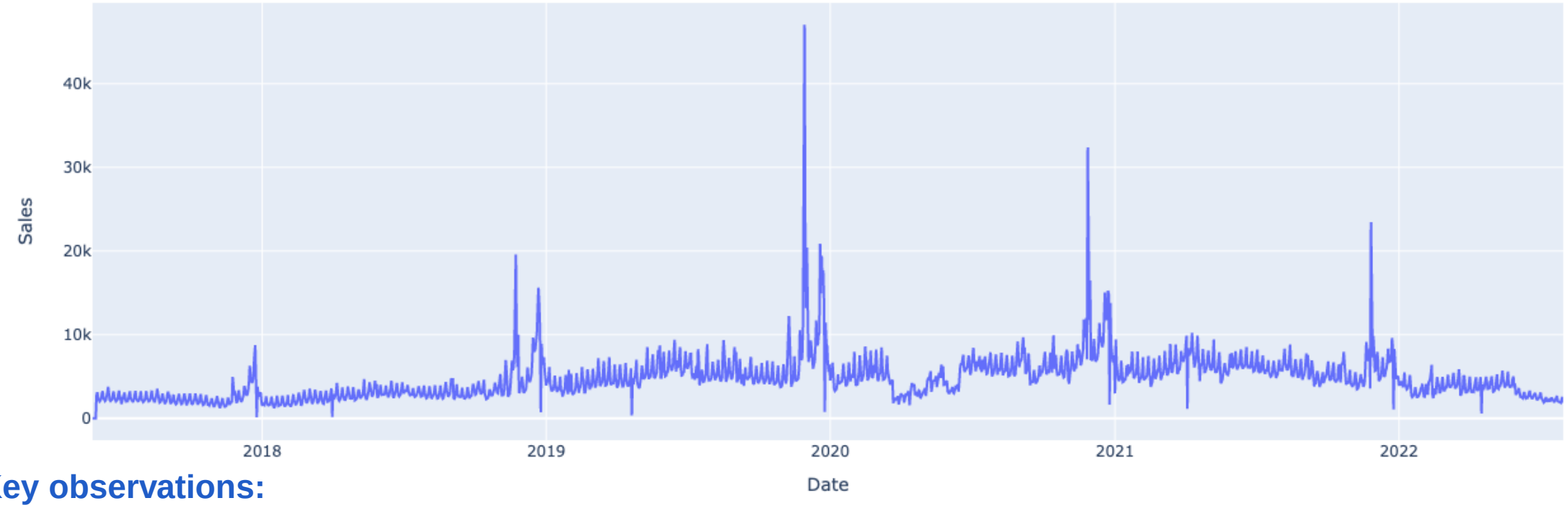
Feature Name	Description	Type
CPIAUCSL	Consumer Price Index for All Urban Consumers	Numerical / Monthly
UNRATE	Unemployment Rate	Numerical / Monthly
CCI	Consumer Confidence Index	Numerical / Monthly
PCE	Personal Consumption Expenditure	Numerical / Monthly
GSCPI	Global Supply Chain Pressure Index	Numerical / Monthly
CSI	Consumer Sentiment Index	Numerical / Monthly
Is_holiday	Import from the “holidays” package to identify important US holidays	Categorical (1: Holiday; 0: Nonholiday)
Weekday	Identify if SALES_DATE is on weekday or weekend	Categorical (1: Weekday; 0: Weekend)
DISCOUNT	Calculate the discount rate by PROMO_PRICE (if any)	Numerical / Daily

Data Visualization

Aggregate sales daily overview

Sum of sales unit by day

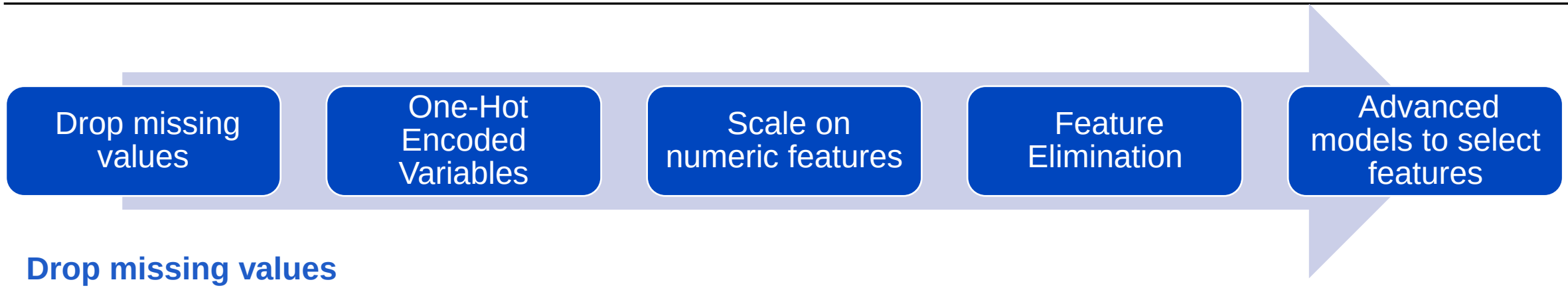
Sales unit



Key observations:

- **Seasonality and variation:** Strong variation exists across years, signalling that there is probably a trend which is different by year
- **Holiday effect:** Strong holiday effect apparently present, particularly obvious on key dates for the retail industry such as Black Friday and Cyber Monday

Feature Engineering



Drop missing values

- Dropped competitor price since 63% of data are missing
- Due to read_csv function errors, there could be erroneous NAN rows, dropped rows with only NAN values

One-hot Encoded variables

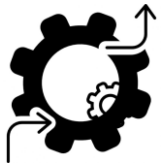
- Used get_dummies function to encode categorical variables
- The purpose is to investigate the effects of categorical variables

Scale on numeric features

Feature Elimination, Advanced models to select features

- Detailed on next slide

Feature Engineering



To investigate the contribution of each data feature, three methods were used:

Feature selection methods	Process	Result
PCA	Checking the number of columns in transformed values	In total three features
Gradient Boosting Regressor	Build a series of decision trees based on reduction of the loss function by features	**Features Chosen:** Discount, Price, Consumer Confidence Index
Random Forest Regressor	Use multiple decision trees to make predictions	**Features Chosen:** Discount, Price, Consumer Confidence Index

Correlation Matrix

Findings

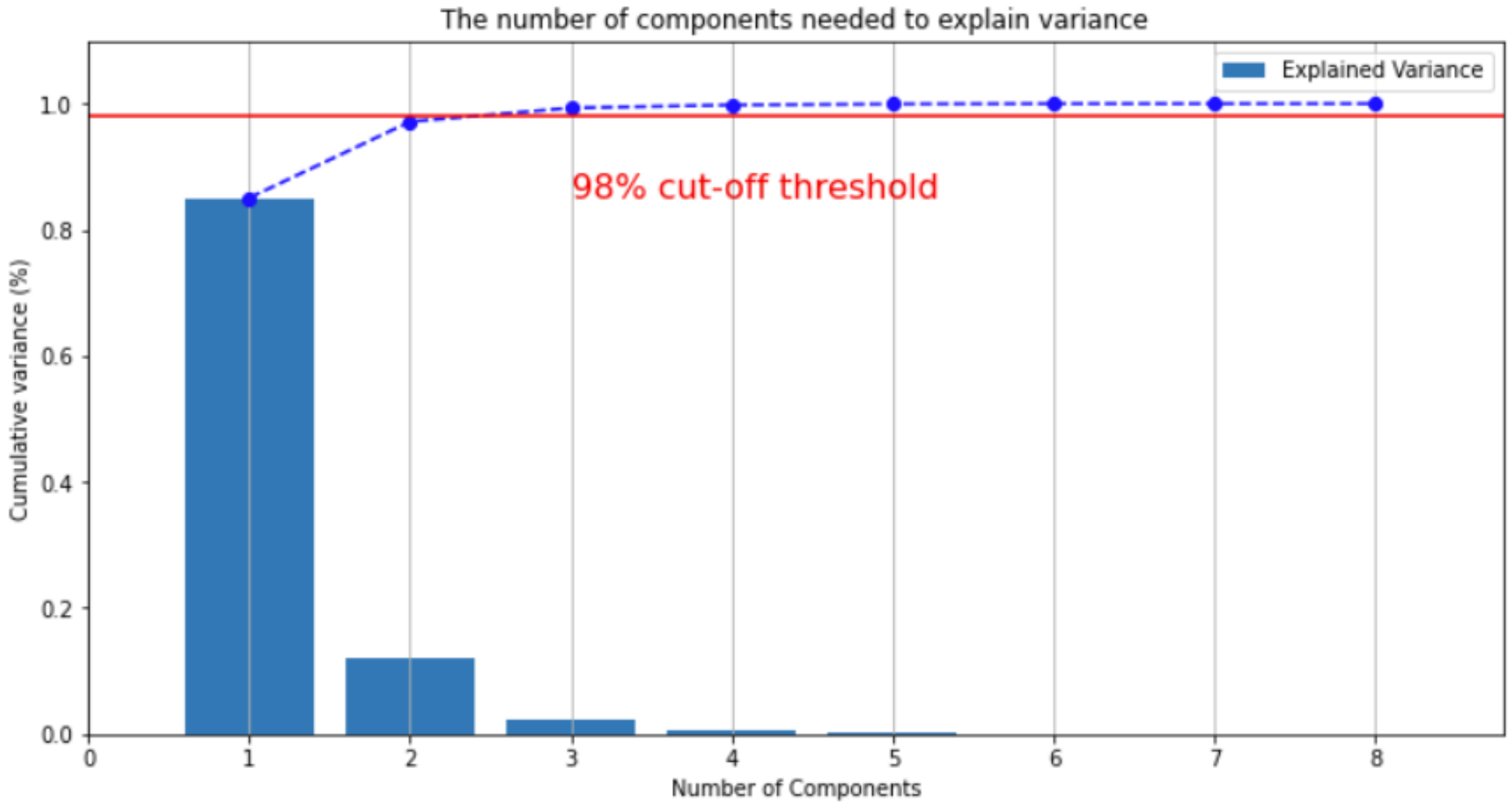
Correlation between Actual sales unit vs. Other Numeric Features

Discount Percentage

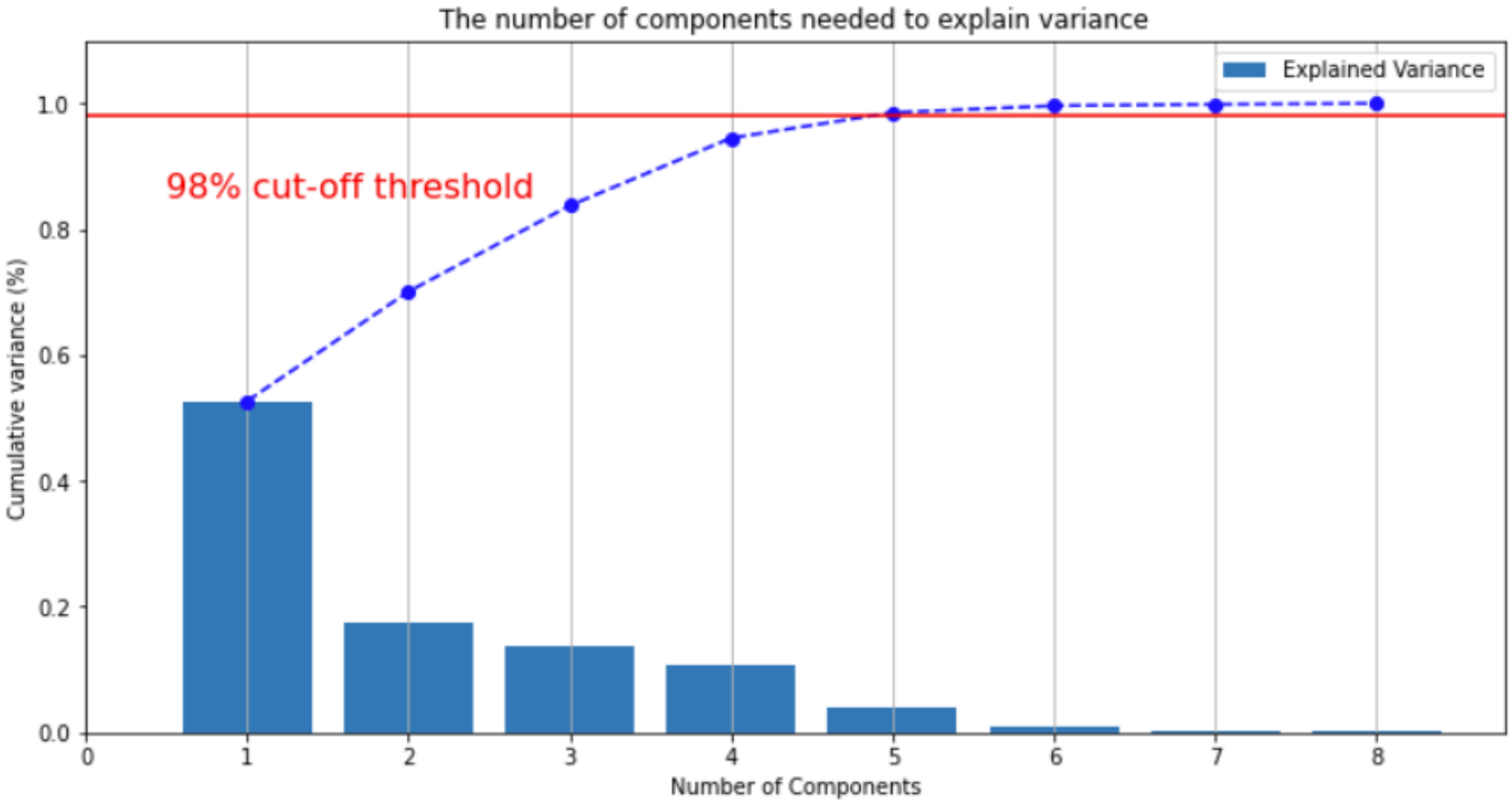
Price

Discount percentage and **price** have stronger correlation with daily units

Feature Selection (After Log Transformation)

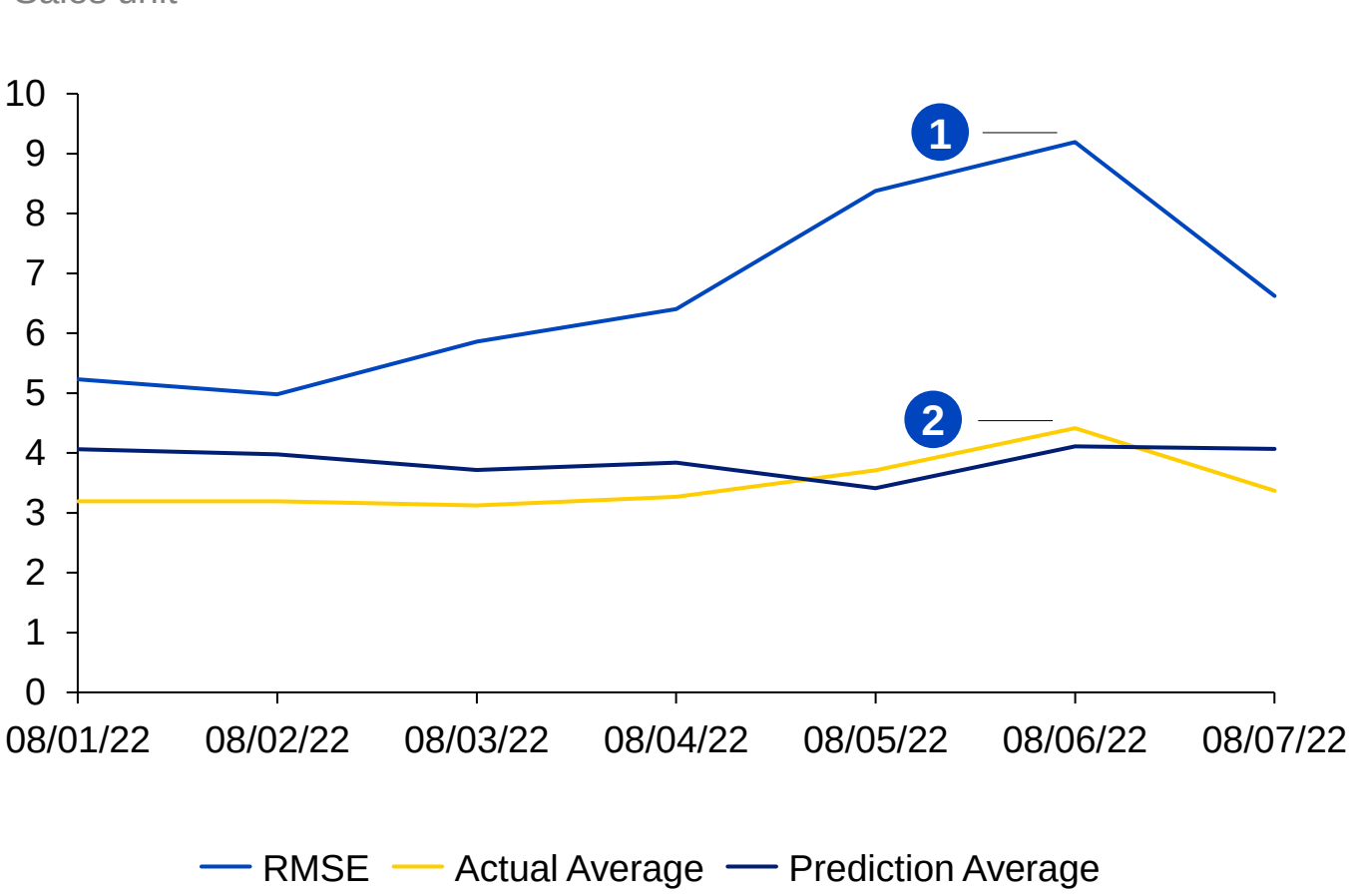


Feature Selection (After Standardization)



XGBoost: Boosting yields an acceptable RMSE results of ~7

Validation Set:
Actual sales unit vs. Predicted sales unit



Observations

1 RMSE heavily influenced by outliers

- SKU #430 & SKU # 469 contribute to **52%** of RMSE
- Effect of discount hard to forecast due to lack of knowledge of **promotion scale/format**

2 Actual sales and sales prediction generally stay close

- Despite the variation of RMSE across individual SKUs, on a per day perspective the XGBoost model did a good job

3 Extraneous data's contribution is limited due to data granularity

- Extraneous data's contribution to prediction is very limited
- since economic data are usually released on a per month basis

Prophet Introduction: Modelling time series with multiple seasonalities

Prophet as a **decomposable** time series model

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t.$$

g(t): Trend function which models the non-periodic changes in the value of the time series

s(t): Represents periodic changes (e.g., weekly and yearly seasonality)

h(t): Represents the effect of holidays

Error term represents any idiosyncratic changes which are not accommodated by the model

Advantages of Prophet

1 Robust to outliers

- Prophet can automatically detect anomaly data…
- And remove outliers that can potentially affect forecast

2 Robust to missing data

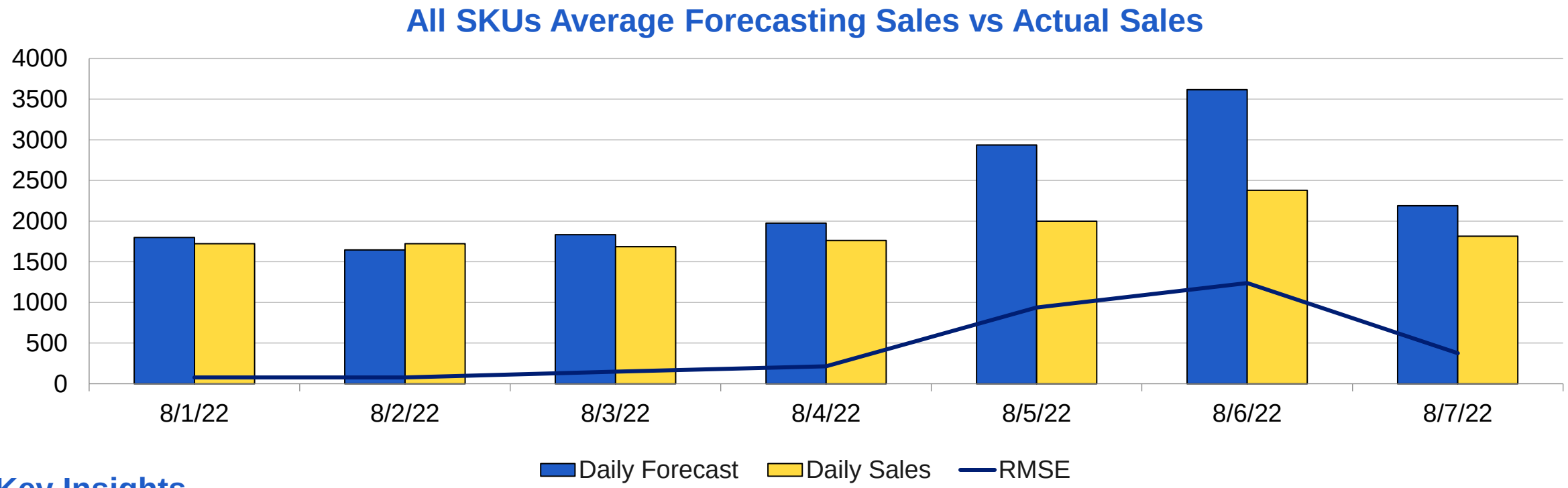
- Prophet can automatically impute data if certain date values are missing

3 Handle seasonal variation well

- Prophet can default fit weekly and yearly seasonality

Since this dataset has strong trend, holidays effects, and several apparent outliers, Prophet will be ideal for time series forecast

Prophet: RMSE reaches 3.32, achieving a ~51% cut



Key Insights

1. Forecast runs with a low RMSE for the first four days, produces higher RMSEs for the next two days, and then decrease back to the average daily level of RMSE. The expected sales are much higher than the actual sales during 8/5/22 - 8/6/22.
2. Prophet is sensitive to the significant outliers from daily sales, leading to less accurate prediction. The model is well designed to take seasonality into consideration, but multiple seasonal patterns can worsen the model performance

Recommendations

- **Price Leadership Strategy**
  - › Why?
    - We find out that the variable "Discount" contribute most to explaining daily sales
  - › Argument
    - Even though such a strategy seems like would lead to pain in short, we believe the overall increase in sales from this strategy and offering more incentives will more than offset the short-term losses.

- **Marketing Strategy**
  - › Why?
    - Personalized marketing is sending right products to the right customer at the right time is critical and making products accessible to expand customer segments.
  - › Solution
    - Focus on marketing products with a price between 25%-50% quantile, invest more on marketing campaigns and stay active on marketing channel like social media, email, video ads, local newspapers, word-of-mouth (viral marketing), etc.

End

Thank You for Your Time

Q & A