Navigating Markets

Forecasting the Superior Fit for Business Growth

E-commerce vs. Physical Market

MARIJA CVEEVSKA | SULEYMAN ERIM | JOAN ORELLANA RIOS

Outline

1. Preprocessing the Data
2. Exploratory Data Analysis
3. Finding the Best Model
4. Forecasting
5. Conclusion and Possible Improvements

Meet our Company

Meet **Your Marketplace**, a master course project turned online marketplace that operates solely on e-commerce!

E-commerce has taken us this far, but now the company is evaluating whether the leap into the **Physical retail** world can be beneficial.

To make an informed decision, a study will be conducted to assess the viability of the physical market.

As part of our study, we'll also take a closer look at e-commerce, giving us a complete picture of our business.

Data Preprocessing

The Data for this Project was collected from various sources.

The Dataset is for the **USA** and is **Quarterly**, from **2000 - 2022**, with **11 variables.**

All values are obtained from the original sources except for the year 2016 for the **Fashion ecommerce revenue column** **where we did imputation of the value**

Cyber Monday Revenue - values only one quarter per year starting year 2014

Physical Market vs Ecommerce

Definition of **Physical Market** and **E-commerce** sales

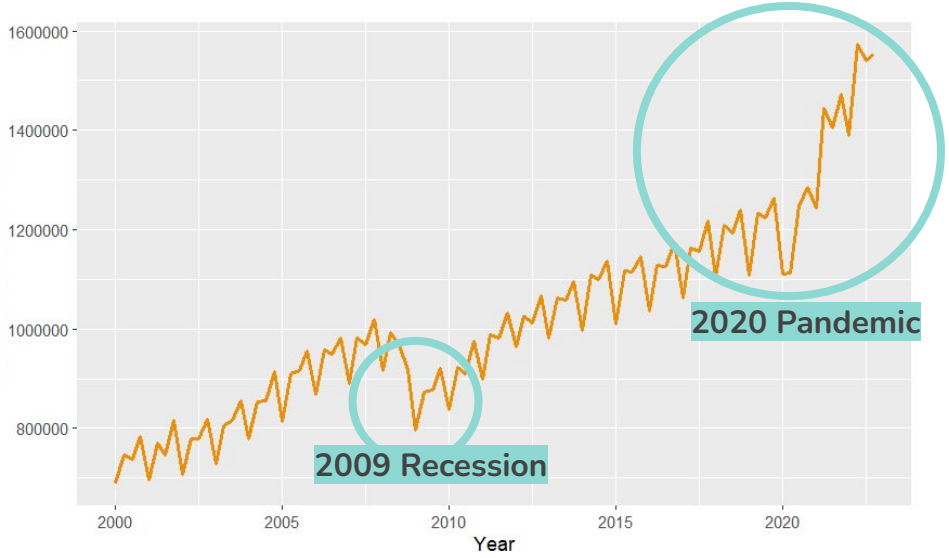
Physical Market

Buying and selling of goods or services directly between buyers and sellers in a physical location

E-commerce

Buying and selling of goods or services over the internet involving digital payments

Physical Market

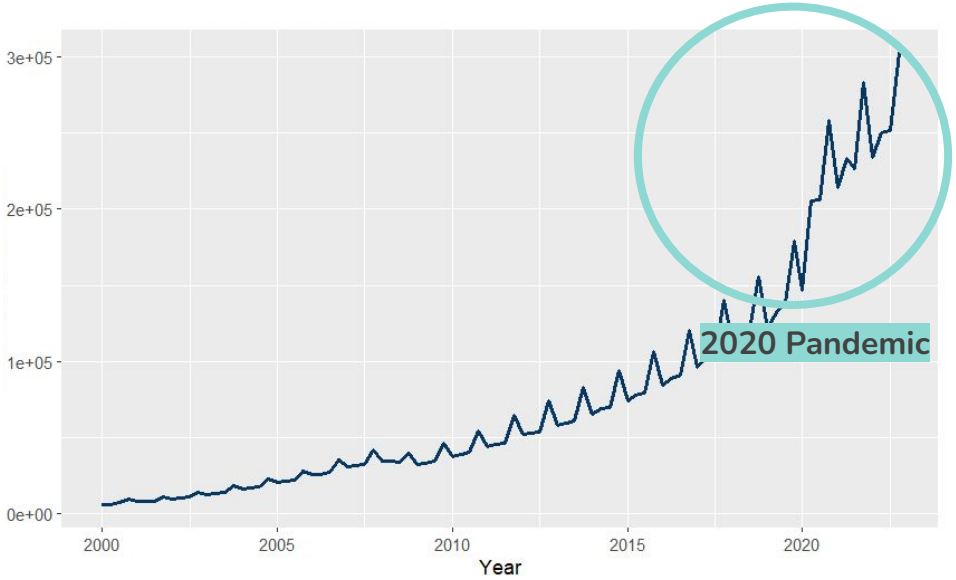


Physical Market Sales Over Time

Why are we witnessing reopening of physical stores post-disruptive events that impact the market?

- In-Person Experience
- Brand Experience and Trust
- Targeting Different Markets
- Multichannel Strategy

E-Commerce



E-Commerce Sales Over Time

Increase in e-commerce shopping can be attributed to several factors:

Convenience

Pandemic Influence

Special Deals and Promotions

Correlation Analysis

E-Commerce

Strong Positive Correlation with:

- **Population**
- **Fashion Revenue**
- **Internet Penetration**

Physical Market

Strong Positive Correlation with:

- **Population**
- **Fashion Stores**
- **Fashion Revenue**

	E-Commerce	Physical Market	Population	GDP	Unemployment Rate	Fashion Stores	Fashion Revenue	Internet Penetration	Covid Cases	Cyber Monday Revenue	Inflation (CPI)
E-Commerce	1	0.93	0.85	0.01	-0.18	0.75	0.98	0.79	0.73	0.55	0.19
Physical Market	0.93	1	0.9	0.09	-0.22	0.89	0.91	0.86	0.68	0.46	0.21
Population	0.85	0.9	1	-0.08	-0.01	0.89	0.87	0.93	0.42	0.35	0.07
GDP	0.01	0.09	-0.08	1	-0.46	0.22	0.04	-0.08	0.15	-0.01	0.23
Unemployment Rate	-0.18	-0.22	-0.01	-0.46	1	-0.23	-0.23	-0.07	-0.25	-0.1	-0.18
Fashion Stores	0.75	0.89	0.89	0.22	-0.23	1	0.77	0.85	0.54	0.29	0.17
Fashion Revenue	0.98	0.91	0.87	0.04	-0.23	0.77	1	0.81	0.69	0.4	0.24
Internet Penetration	0.79	0.86	0.93	-0.08	-0.07	0.85	0.81	1	0.4	0.32	0.1
Covid Cases	0.73	0.68	0.42	0.15	-0.25	0.54	0.69	0.4	1	0.34	0.38
Cyber Monday Revenue	0.55	0.46	0.35	-0.01	-0.1	0.29	0.4	0.32	0.34	1	-0.15
Inflation (CPI)	0.19	0.21	0.07	0.23	-0.18	0.17	0.24	0.1	0.38	-0.15	1

Data Modelling

Physical Market Study

Data Modelling

Type of Model

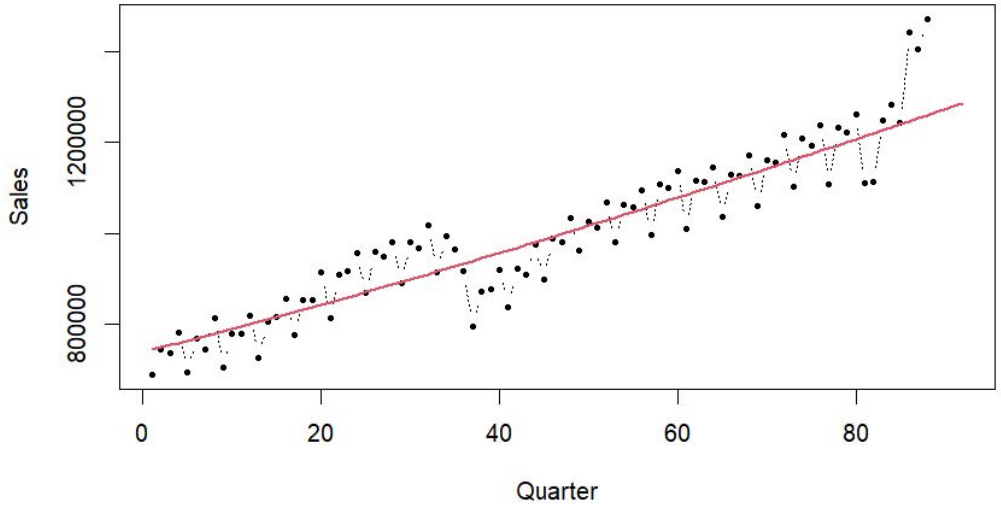
GBM + SARMAX

Data Modelling: Trend

Bass Model

Adjusted R^2 = 0.999981

Parameters	P-values
m	9.041451e+08*
p	8.207887e-04*
q	37.529943e-03***

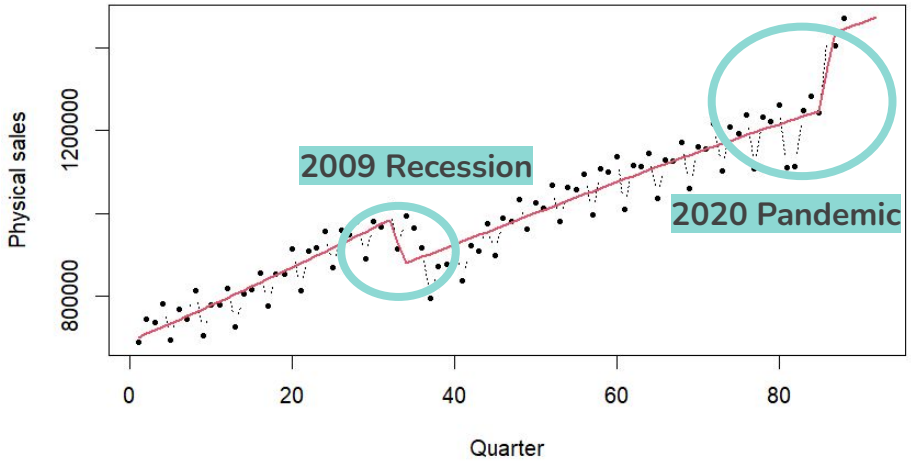


Data Modelling: Trend

GBM with Rectangular Shock

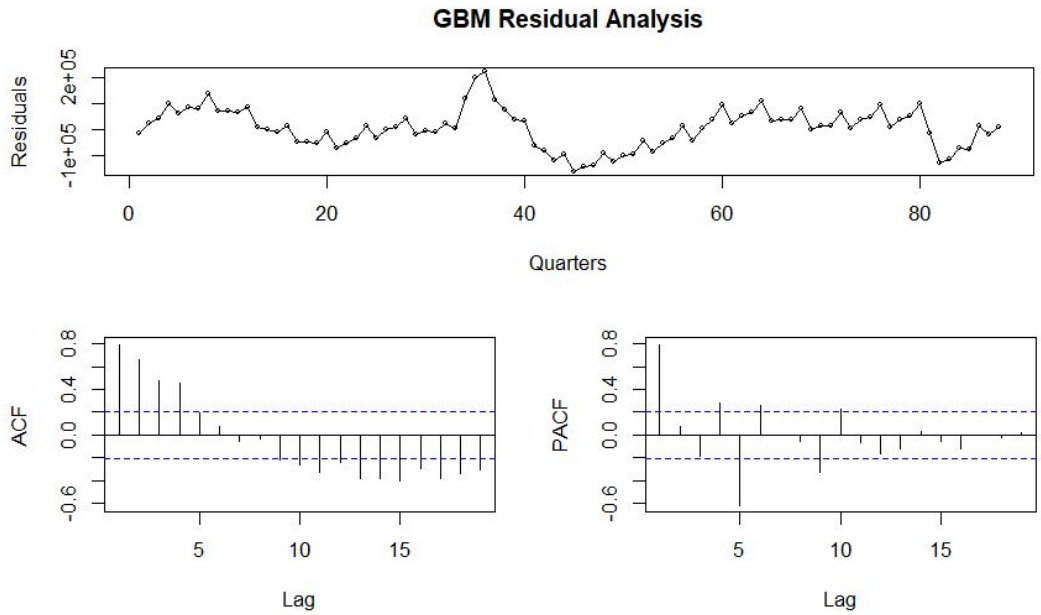
Adjusted R^2 = 0.999991

Parameters	Estimates
m	3.611603e+08***
p	1.930084e-03***
q	1.386370e-02***
a (32)	3.248120e+01***
b (85)	8.541923e+01***
C	1.230221e-01***



Data Modelling: Trend

Analysis of Residuals -> GBM



Ljung-Box test

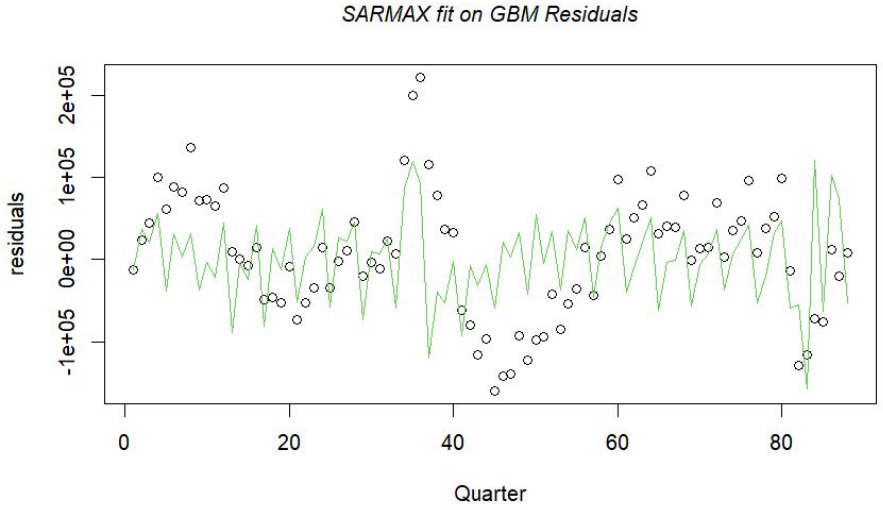
p-value = <2.2e-16***

- Harmonic behaviour of residuals
- Exponential decay in ACF
- Significant spikes in PACF
- Residuals are not WN

Data Modelling: Seasonality

SARMAX(2,1,1)(0,1,1)[4]

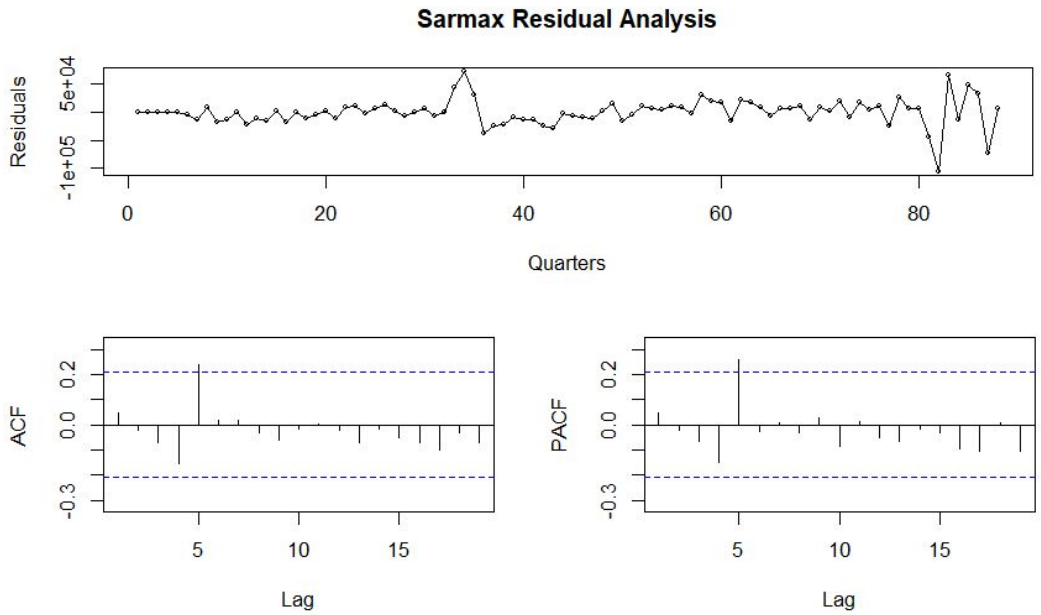
Parameters	Values	Standard Error
ar1	1.4152	0.0903
ar2	-0.5327	0.0896
ma1	-1.0000	0.0685
sma1	-1.0000	0.0909
lambda	1.0014	0.0109



AIC = 1940.39

Data Modelling: Seasonality

Analysis of Residuals -> SARMAX



Ljung-Box test

p-value = 0.5271

- No significant spike
- Insignificant p-value
- Residuals seem WN

Data Modelling

Type of Model

Piecewise Regression

Data Modelling: Time Series Regression

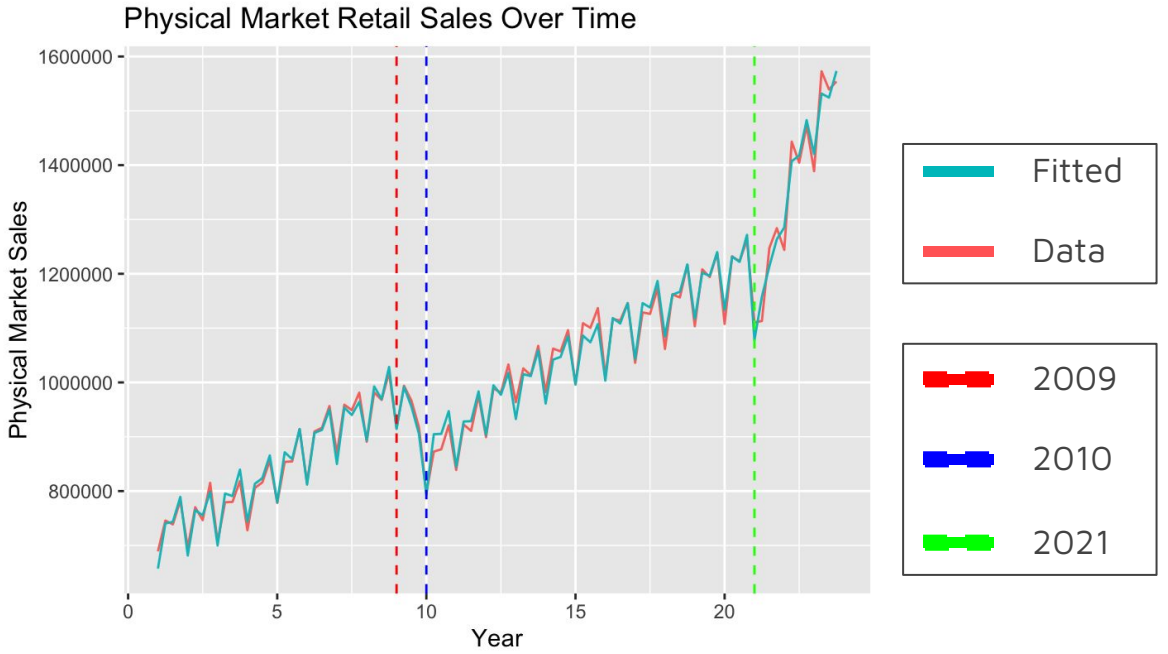
Performance of different Linear and Nonlinear Trends

Model	AIC	Adj R^2	DW Test	BG Test
Linear	1.87e+03	0.985	1.02	8.67e-05
Piecewise	1.81e+03	0.992	**1.8**	**0.07**
Splines	1.88e+03	0.981	0.77	9.12e-06

Data Modelling: Time Series Regression

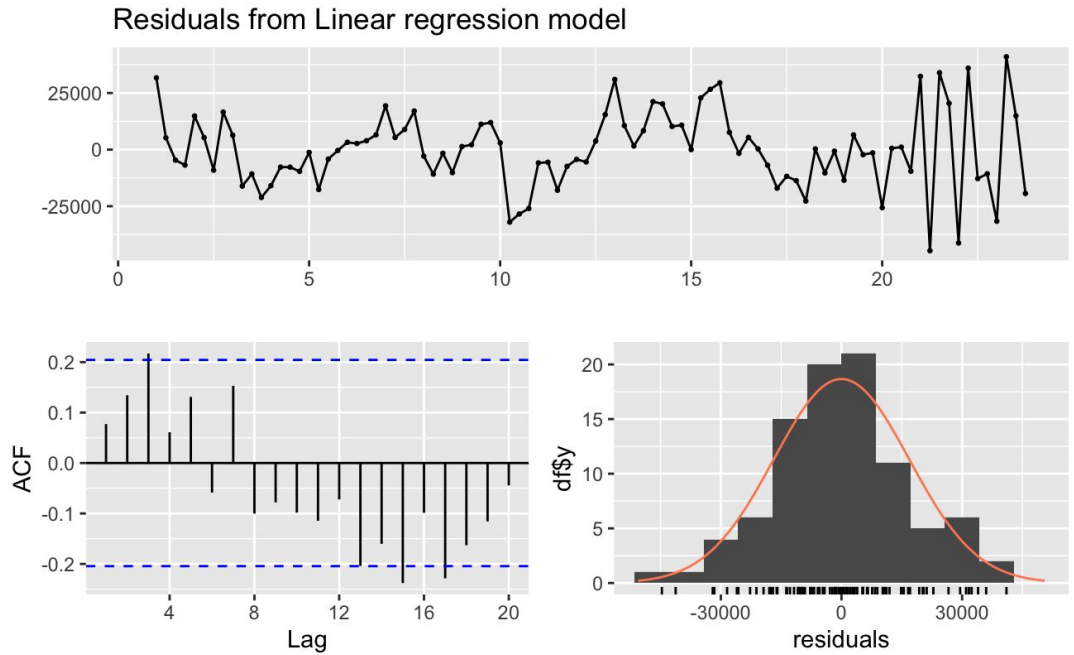
Piecewise Regression: Best Fit

Predictors	P-values
Trend	< 2e-16 ***
Season	< 2e-16 ***
Number of Fashion Stores	< 2e-16 ***
Inflation (CPI)	7.24e-07 ***



Data Modelling: Time Series Regression

Piecewise Regression: Analysis of Residuals



Data Modelling

Other Models Tested (and Discarded)

GAM with Holt-Winters'

AIC = 2.061e+03

DW Stat = 1.98

Data Modelling

E-Commerce Study

Data Modelling

Type of Model

GAM with Holt-Winters’

Data Modelling: GAM with Holt-Winters'

ANOVA for Parametric and Nonparametric Effects

Predictors	P-values
s(trend)	< 2e-16 ***
Season	< 2e-16 ***
s(Covid Cases)	< 2e-16 ***
s(Cyber Monday Revenue)	< 2e-16 ***
Holt-Winters'	< 2e-16 ***

Parametric Effects

Predictors	P-values
s(trend)	1.071e-4 ***
s(Covid Cases)	2.143e-3 **
s(Cyber Monday Revenue)	2.404e-3 *

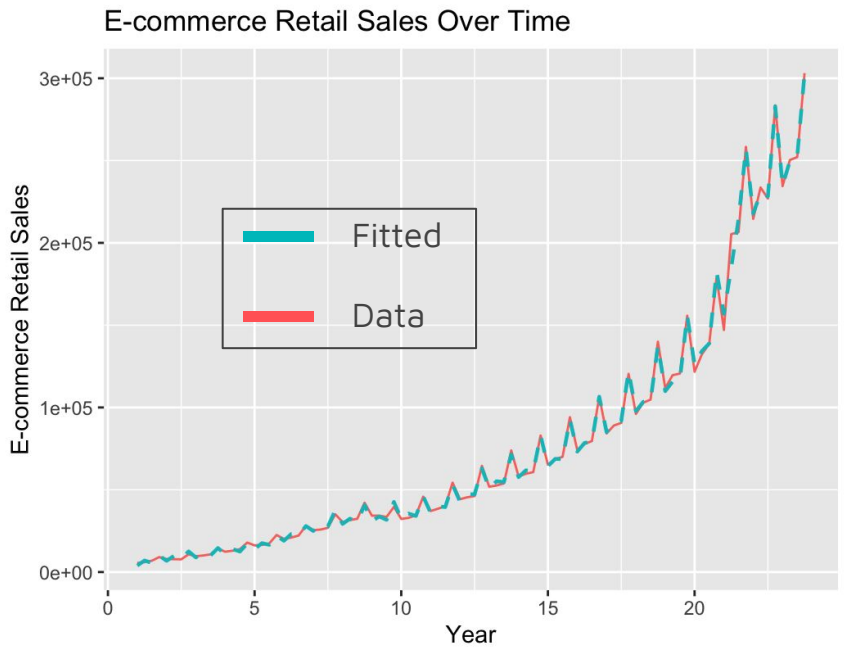
Nonparametric Effects

Data Modelling: GAM with Holt-Winters'

Best Fit and Performance Metrics

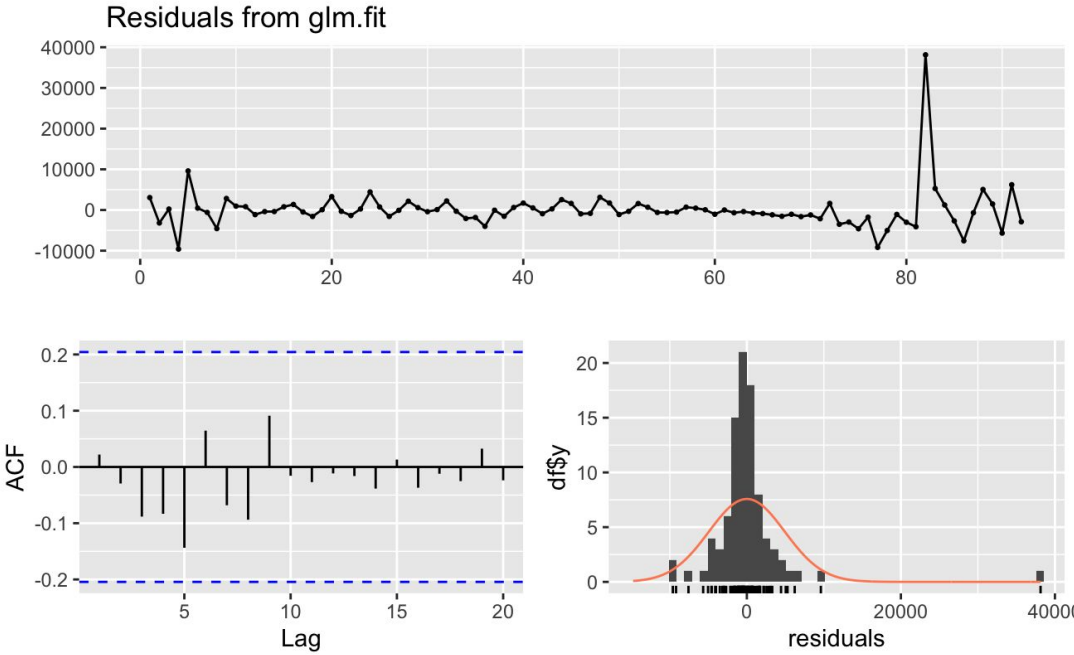
Adjusted R^2 = 0.998

AIC = 1.783e+03



Data Modelling: GAM with Holt-Winters'

Analysis of Residuals



DW Stat = 1.69

Breusch-Godfrey test

p-value = 0.9789

Data Modelling

Other Models Tested (and Discarded)

Model	AIC	Adj R^2	DW Test
Linear	1.66e+03	0.987	1.03
Exponential	-4.22e+02	0.991	0.29
Polynomial	-4.92e+02	0.996	0.56
Splines	1.60e+03	0.993	1.35

Data Modelling

Type of Model

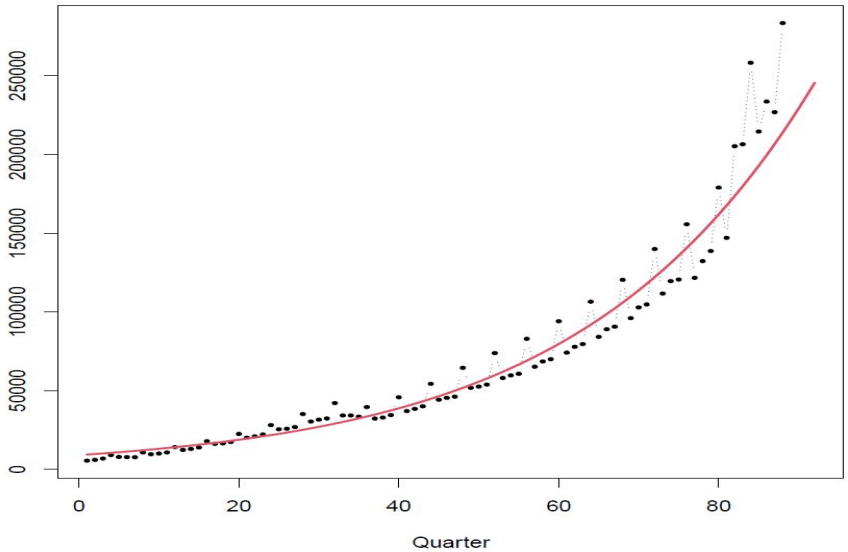
GBM with ARIMA and GB

Data Modelling: Trend

Bass Model

Adjusted R^2 = 0.999656

Parameters	P-values
m	0.495
p	0.506
q	2.49e-58 ***

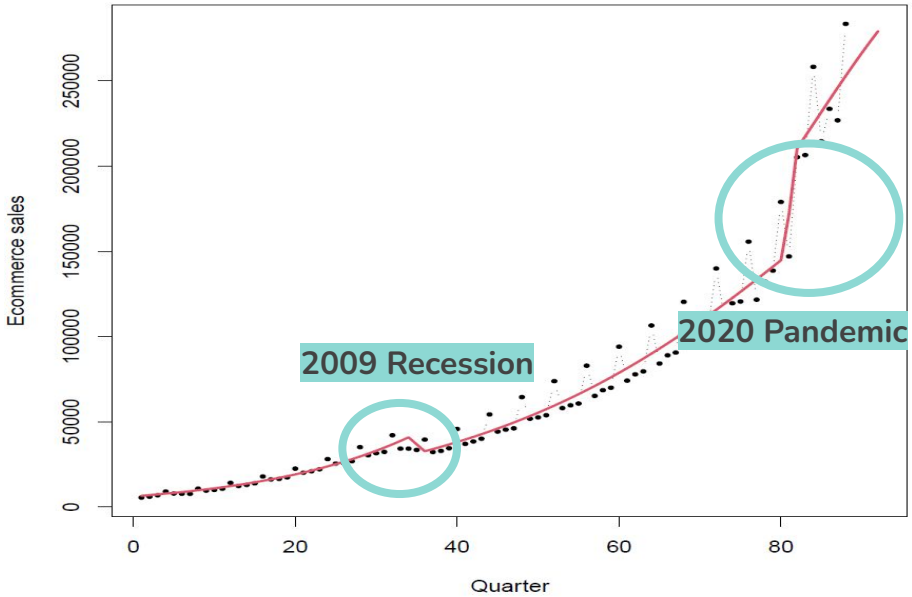


Data Modelling: Trend

GBM with Rectangular Shock

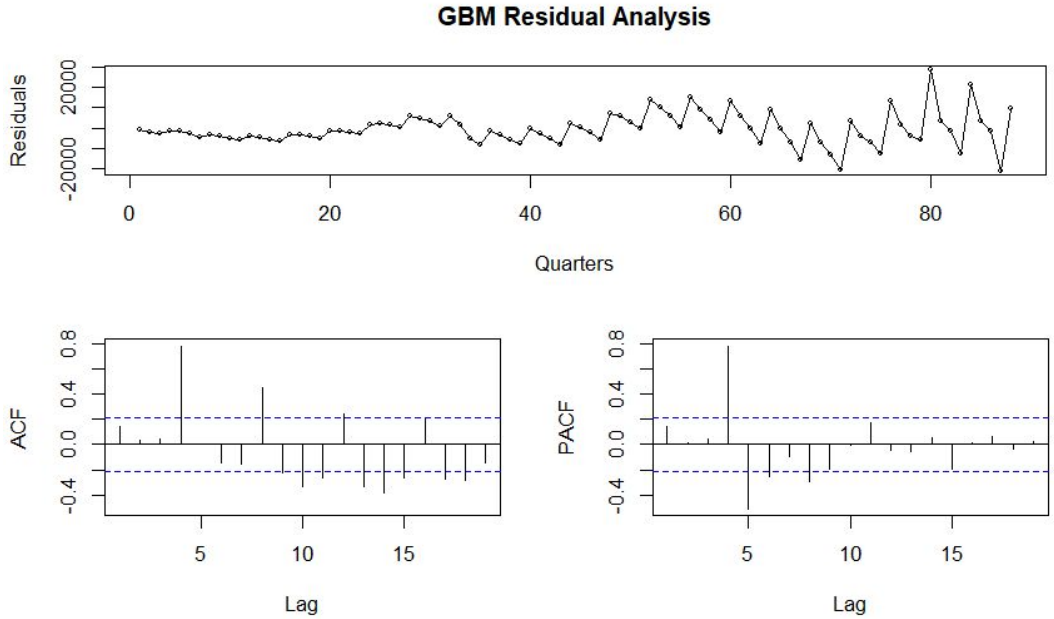
Adjusted R^2 = 0.999977

Parameters	Estimates
m	2.30e+07***
p	2.80e-04***
q	5.64e-02***
a (35)	3.44e+01***
b (85)	8.05e+01***
C	2.69e-01***



Data Modelling: Trend

Analysis of Residuals -> GBM



Ljung-Box test

p-value = <2.2e-16***

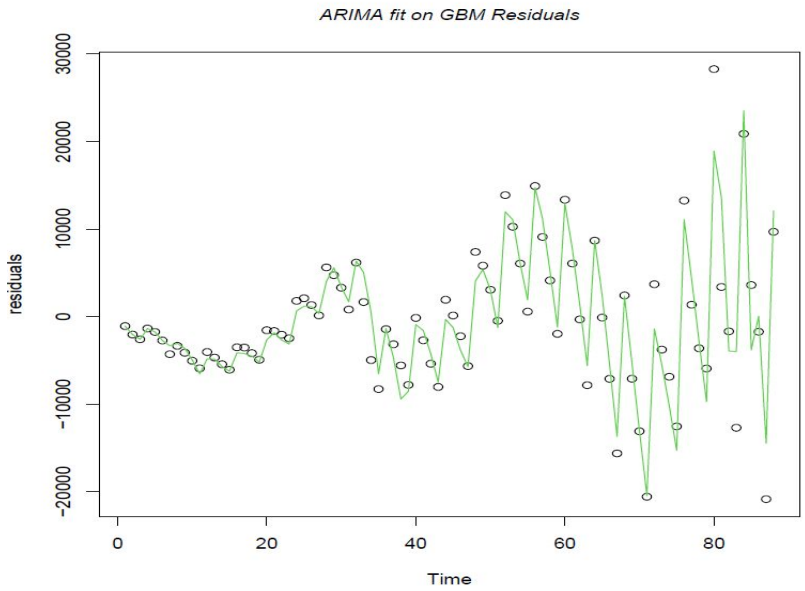
- Increased variance in residuals
- Significant spikes on correlogram
- Significant p-value
- Residuals are not WN

Data Modelling: Seasonality

Seasonal ARIMA(2,1,1)(1,1,1)[4]

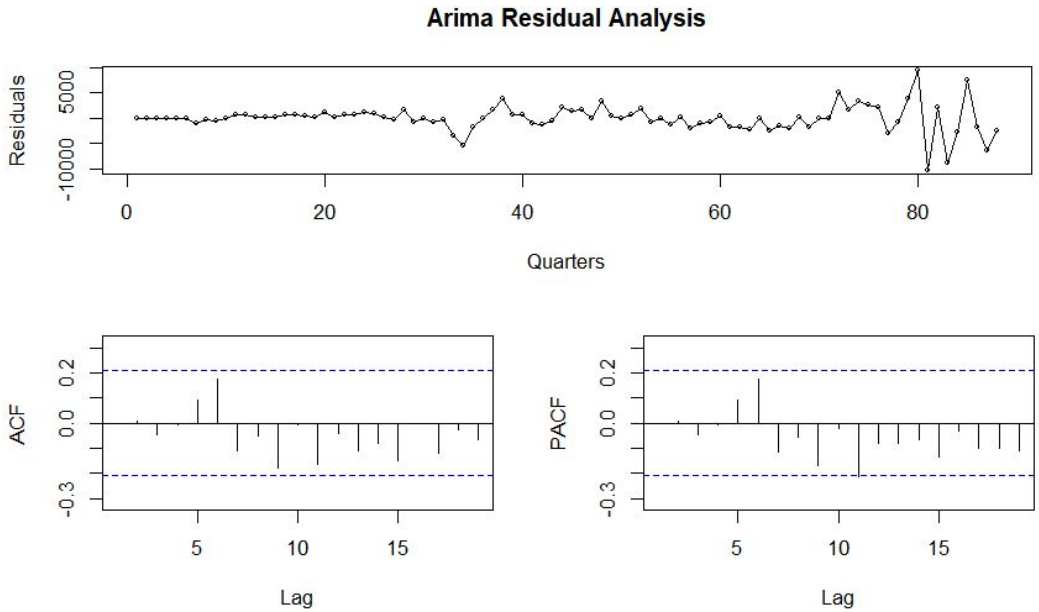
AIC = 1560.7

Parameters	Values	Standard Error
ar1	0.90	0.1122
ar2	-0.12	0.1133
ma1	-1.00	0.0354
sar1	-0.5412	0.5104
sma1	0.7102	0.4442



Data Modelling: Seasonality

Analysis of Residuals -> ARIMA



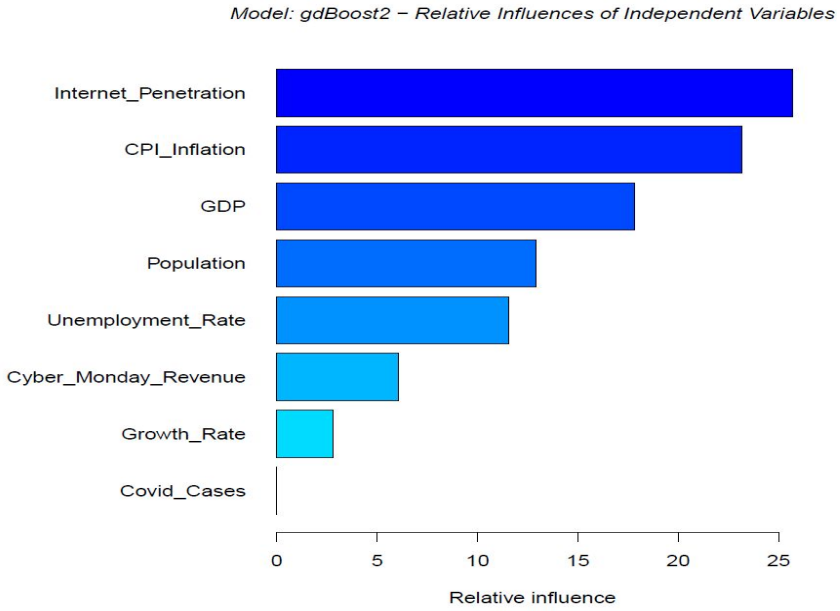
Ljung-Box test

p-value = 0.5633

- Increased variance in residuals
- No significant spike
- Insignificant p-value
- Residuals seem WN

Data Modelling: External Factors

Gradient Boosting



- Fit GB model on residuals of ARIMA
- 7 external factors have impact on residuals

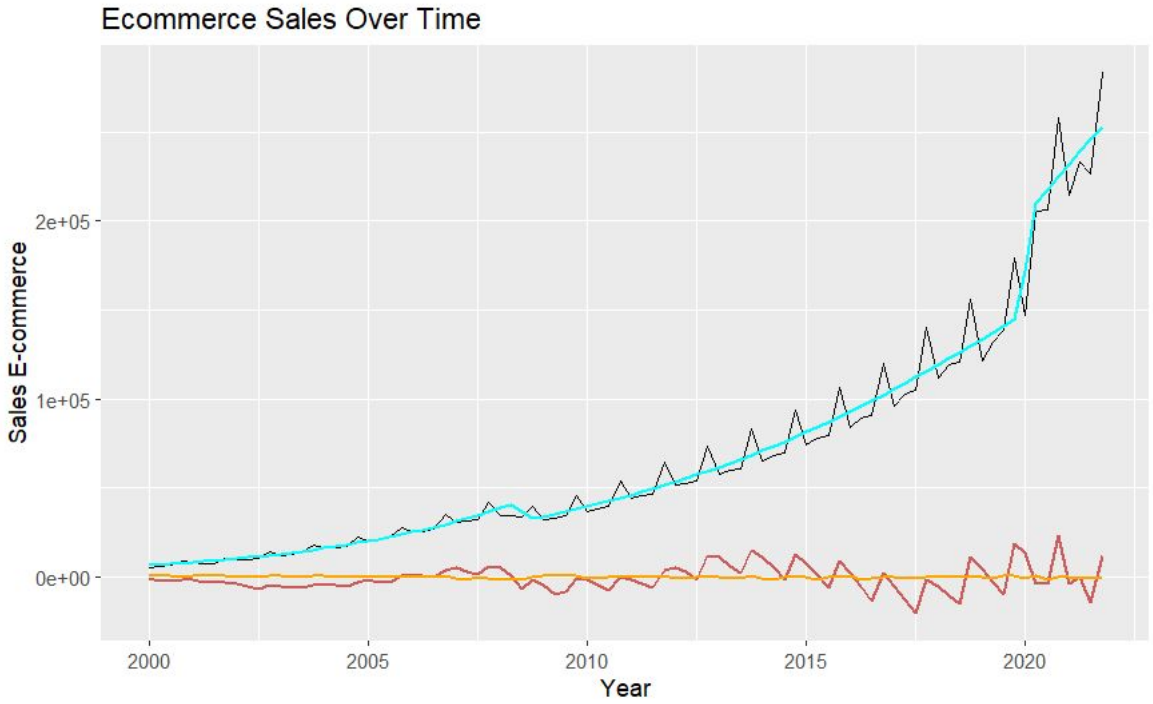
Data Modelling: GBM + ARIMA + GB

Test Set Performance

Model	MAPE
GBM with Rectangular Shock + ARIMA (2,1,1)(1,1,1)[4]	0.0507
GBM with Rectangular Shock + ARIMA (2,1,1)(1,1,1)[4] + 8 Independent Variables	0.0498
GBM with Rectangular Shock + ARIMA (2,1,1)(1,1,1)[4] + Internet Penetration + Inflation	0.0494

External factors influence is time independent, modelled with cross validated Gradient boosting to avoid overfitting.

Data Modelling



Data

Trend - GBM

Seasonality - SARIMA

External Factors - GB

Model Testing

Model Candidates and Testing Method

Physical Market Modelling

1) GBM with SARMAX
2) Piecewise Regression

E-commerce Modelling

1) GAM with Holt-Winters'
2) GBM with ARIMA and GB

Train-Test Split

Training Data = 2000-2021

Test Data = 2022

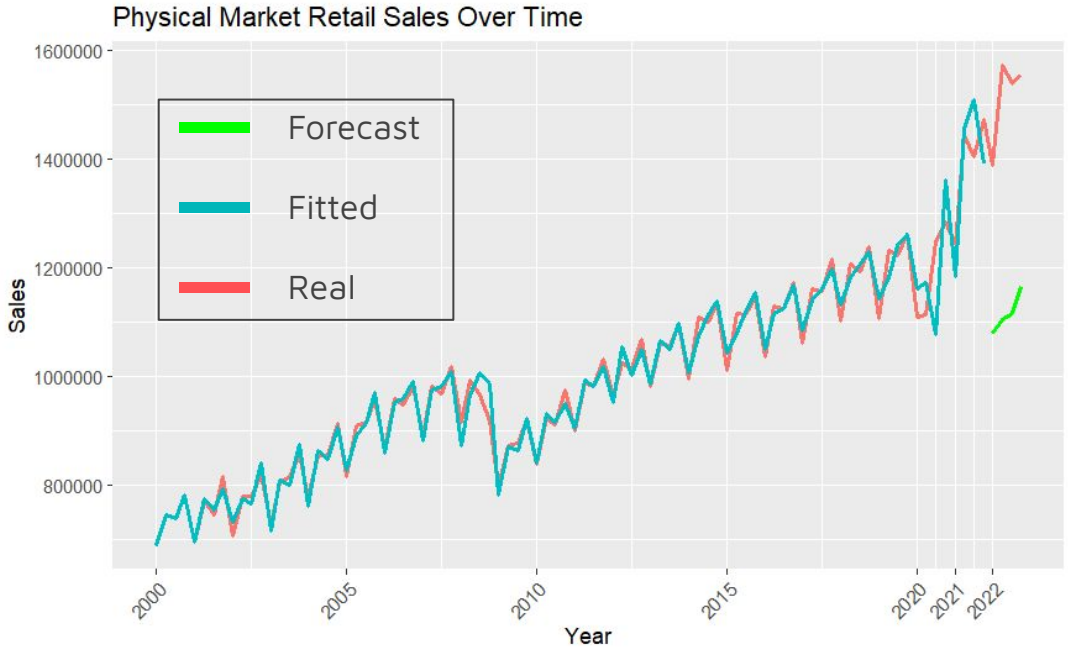
Metric Used

Mean Absolute Percentage Error

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Model Testing

GBM with SARMAX



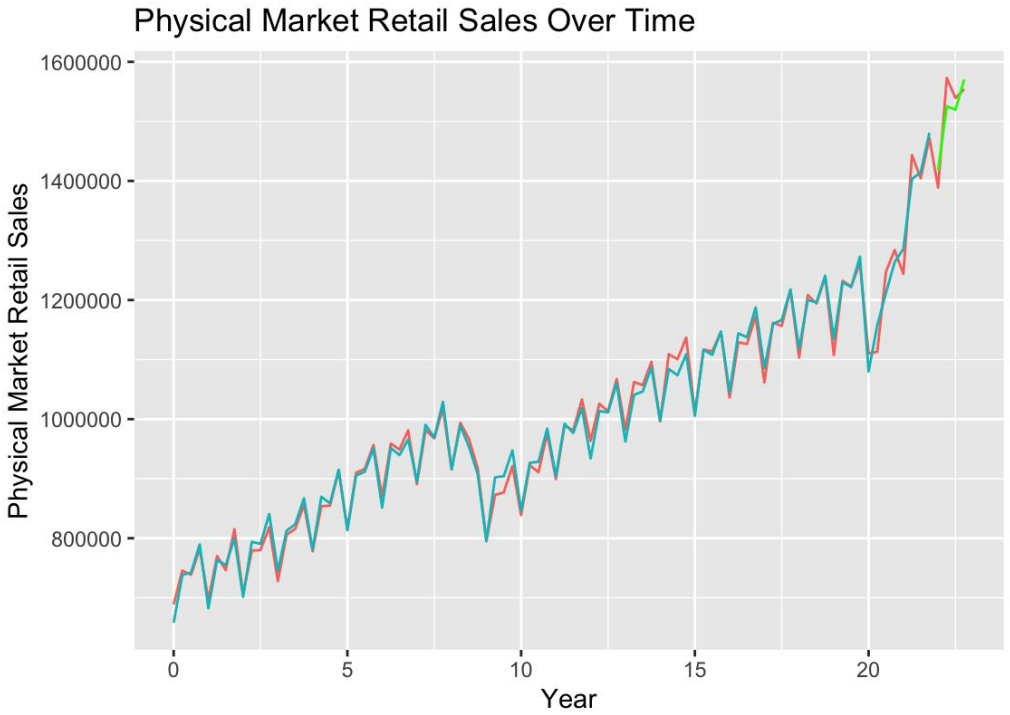
MAPE = 0.26111

	Data	Predictions	Diff
Q1	1388911	1080914	-307996
Q2	1572718	1104198	-468520
Q3	1539296	1116735	-422561
Q4	1554080	1165166	-388913

Model Testing

Piecewise Regression

*When we take as predictors only the **trend**, the **season** and the **breakpoints**, we still reach a MAPE of **0.0228**!*



	Data	Predictions	Diff
Q1	1388911	1417988	29077
Q2	1572718	1525296	-47422
Q3	1539296	1519650	-19646
Q4	1554080	1570233	16153

Forecast

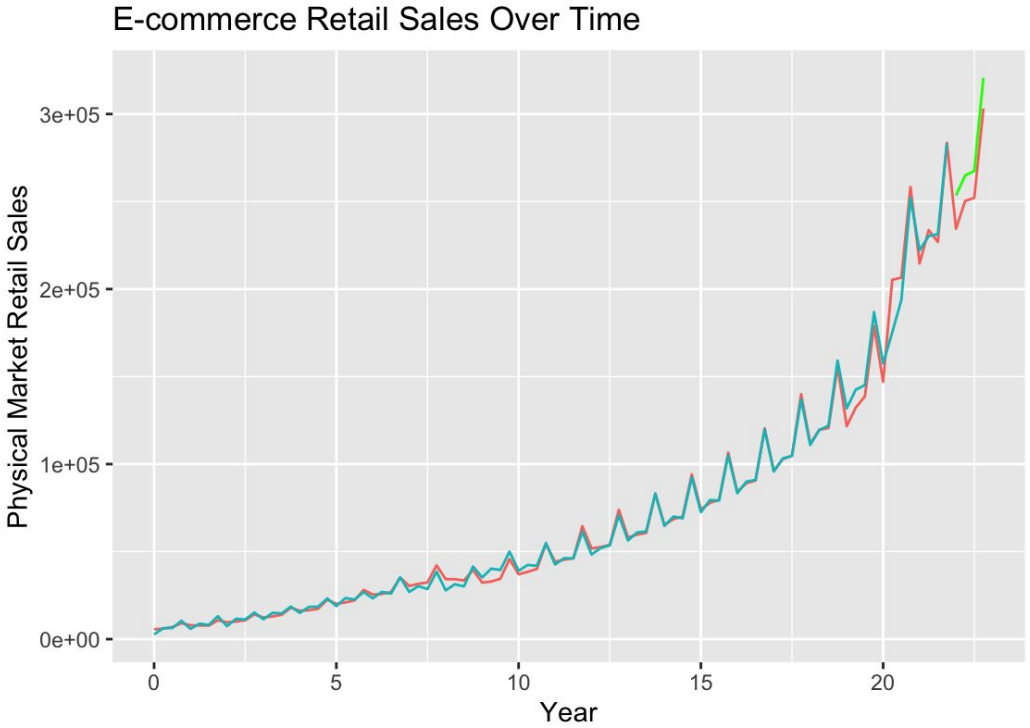
Fitted

Real

MAPE = 0.0185

Model Testing

GAM with Holt-Winters'



	Data	Predictions	Diff
Q1	234454	253510	19056
Q2	250341	264871	14530
Q3	252107	267441	15334
Q4	303120	320633	17513

Forecast

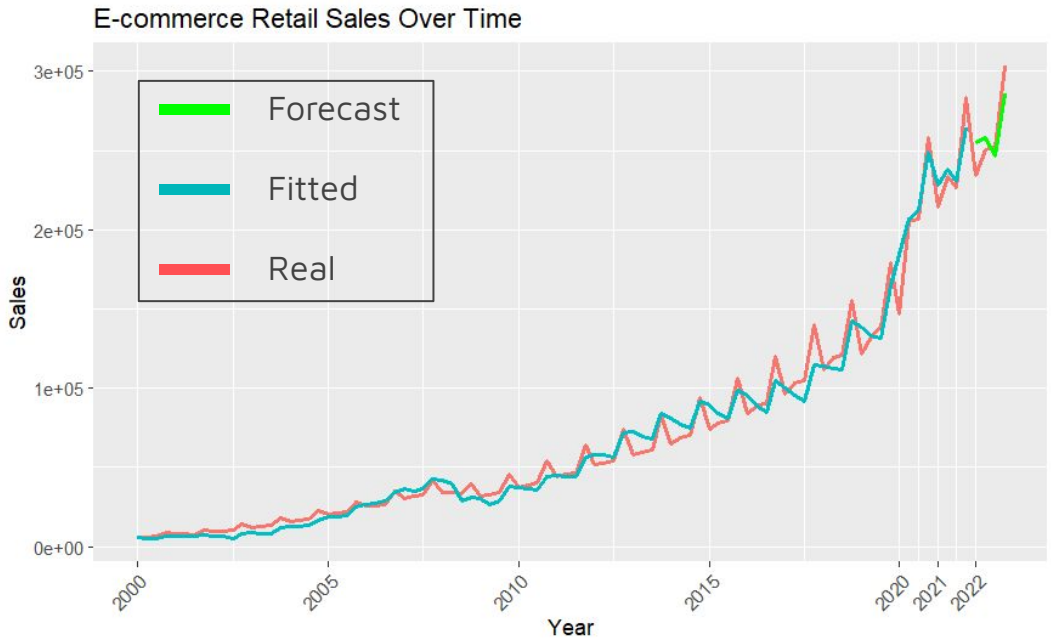
Fitted

Real

MAPE = 0.0645

Model Testing

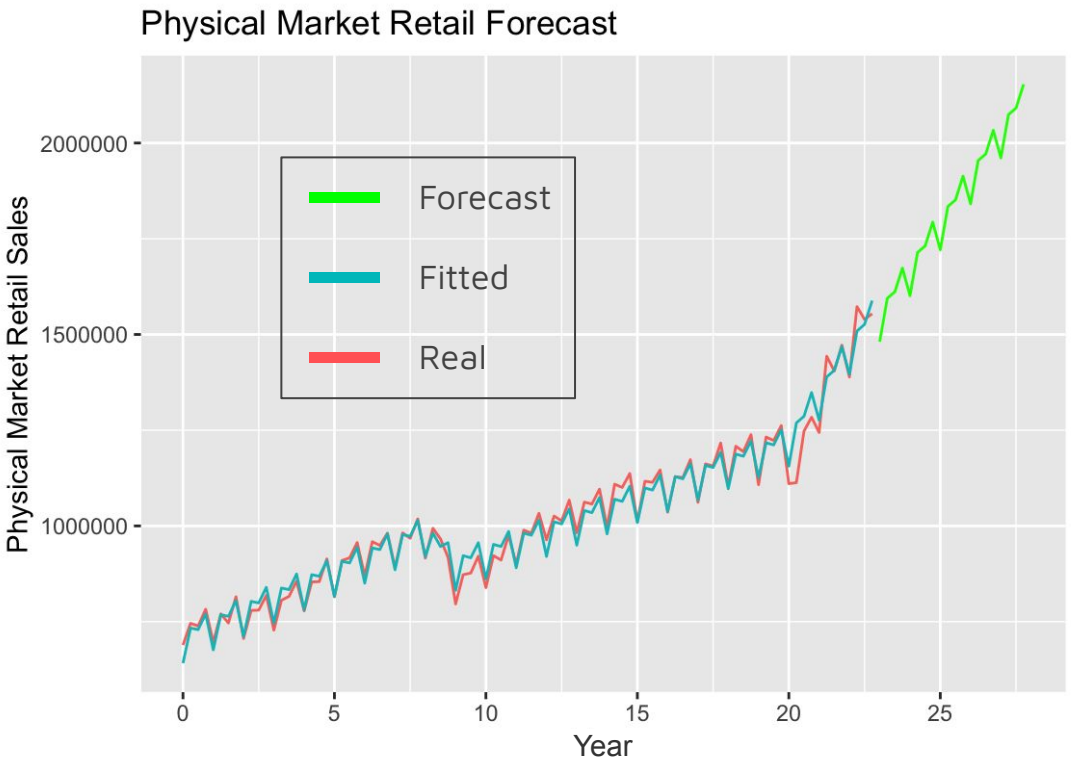
GBM with ARIMA and Gradient Boosting



MAPE = 0.04957

	Data	Predictions	Diff
Q1	234454	255070	20616
Q2	250341	255748	7407
Q3	252107	246300	-5806
Q4	303120	285740	-17379

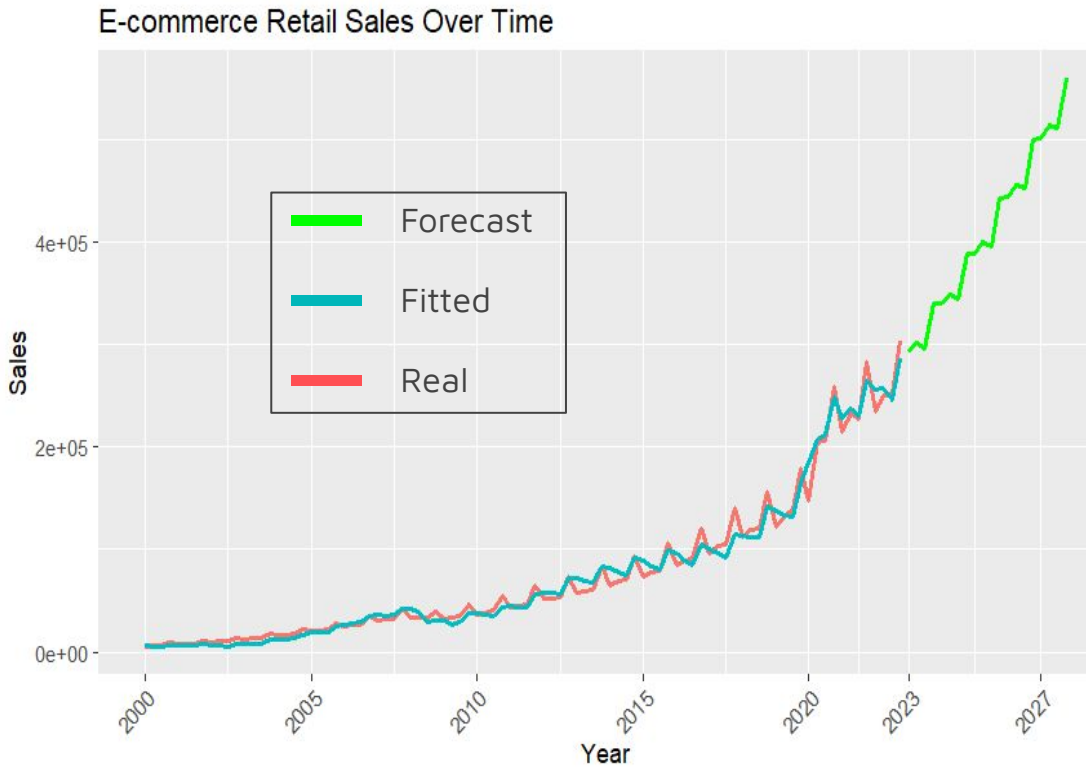
Forecasting



Piecewise Regression

Year (4th Quarter)	Sales Forecast ($)	Growth Rate (%)
2023	1,673,427	7.68
2024	1,793,403	7.18
2025	1,913,378	6.69
2026	2,033,354	6.28
2027	2,153,330	5.90

Forecasting



GBM with ARIMA and Gradient Boosting

Year (4th Quarter)	Sales Forecast ($)	Growth Rate (%)
2023	339,909	12.14
2024	389,060	14.46
2025	442,431	13.71
2026	499,613	12.92
2027	559,325	11.95

Conclusions

Market	Forecasted Sales (2027)	Growth (2022-2027)
Physical	**2,153,330$**	38.6%
E-Commerce	559,325$	**84.5%**

- Positive trends
- To keep in mind: Influence of external factors
- E-Commerce higher growth rate
- Physical market enduring relevance

Conclusions

Market	Forecasted Sales (2027)	Growth (2022-2027)
Physical	**2,153,330$**	38.6%
E-Commerce	559,325$	**84.5%**

It is logical to consider a physical store as a viable business strategy!

Possible Improvements

- Specific focus on area of interest (Fashion, electronics, etc.)
- We can collect the data of 2023 for further analysis and adding to test set
- Even though both markets grow, there are different costs related to different markets (physical store rents, logistics, e-commerce storage area, etc). Detailed analysis for cost should be incorporated for further guidance.

Just before we finish...

DISCLAIMER: This is a fictional company

Data Sources

- **Total Retail Sales** - https://www.census.gov/
- **E-commerce Sales** - https://www.census.gov/
- **Physical Market Sales** - https://www.census.gov/
- **Population** - https://www.statista.com/
- **Growth Rate** - https://www.statista.com/
- **GDP** - https://fred.stlouisfed.org/ - federal reserve economic data
- **Unemployment rate** - https://fred.stlouisfed.org/ - federal reserve economic data
- **Fashion stores** - https://www.census.gov/
- **Fashion ecommerce revenue** - https://www.census.gov/
- **Internet Penetration** - https://fred.stlouisfed.org/ - federal reserve economic data
- **Covid Cases** - https://github.com/CSSEGISandData/
- **Cyber Monday Revenue** - https://en.wikipedia.org/wiki/Cyber_Monday
- **CPI - Inflation** - https://fred.stlouisfed.org/ - federal reserve economic data