Predicting Smartphone Purchase of Multi Category Store

Pei-Chieh (Paget) Hsiao

IOD
Capstone Project

Agenda

Business overview and insights

Question definition

EDA/Featuring engineering

Project delivery

Summary and next steps

Research

- PhD in Photovoltaics
- Researcher at University of New South Wales
- Finite Element Analysis

Data Science

- Data Scientist
- Python, Sklearn, Tensorflow

About Me

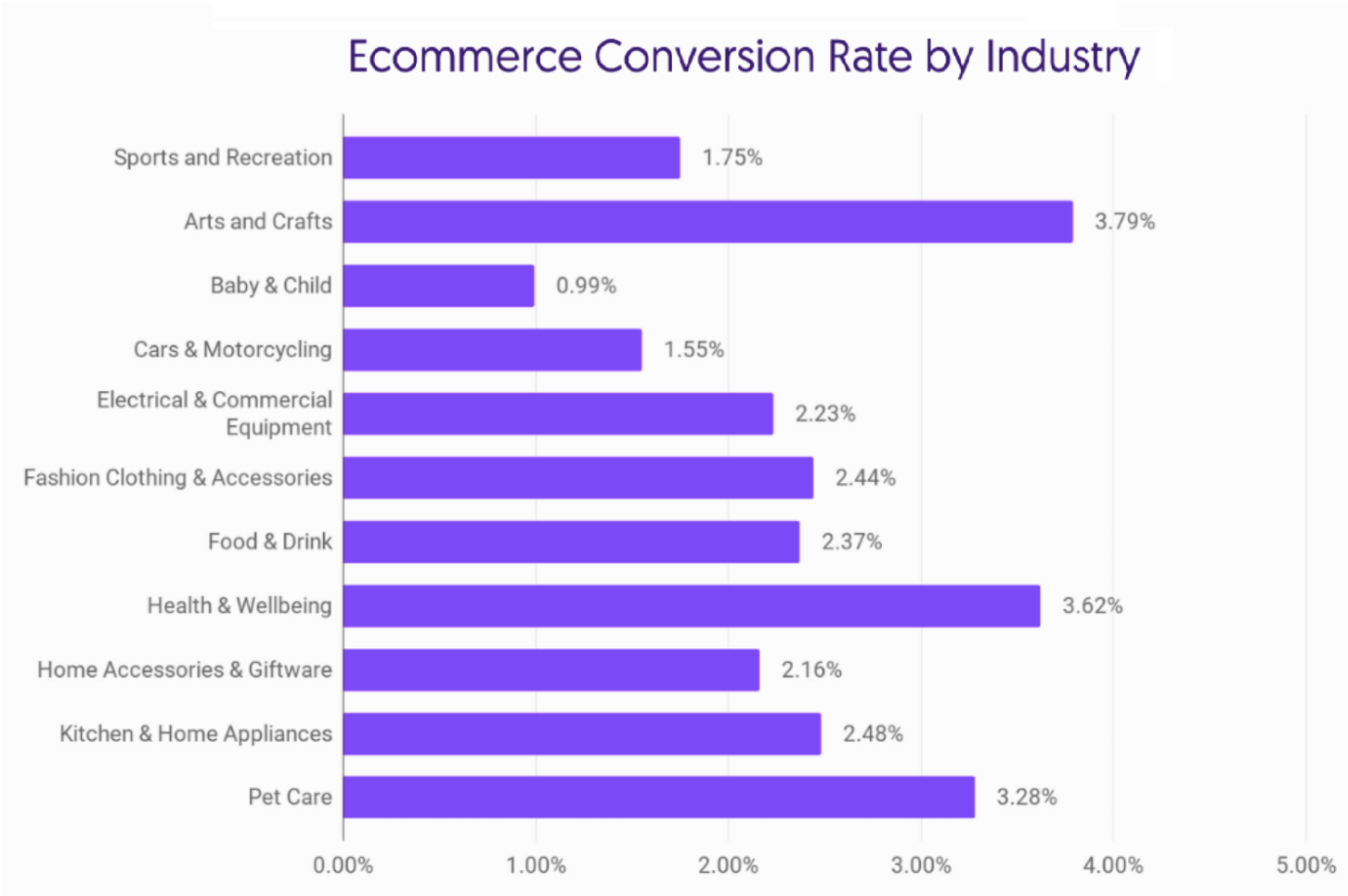
Pei-Chieh (Paget) Hsiao

Background

Conversion rate

- The ratio of transactions to sessions expressed in percentage

- KPI to review the effectiveness of e-commerce sites



https://www.smartinsights.com/ecommerce/ecommerce-analytics/ecommerce-conversion-rates/

Business Overview

B2C e-commerce store

- 13 categories
- 58 products
- $2.67 B revenue*

Key Insights

- **1.87%** overall conversion rate
- **2.51%** conversion rate in “electronics”
- **85%** revenue in “electronics”
- **40%** of customers’ activities on “smartphone” products

- Customers tend to browse through the week and make decision to purchase on weekends.
- There is a surge in purchase coming toward black Friday sale.

**** Estimated from two-month revenue***

Business Question

Can we predict if smartphone products are purchased in customer activity sessions?

Stakeholder:

Michael Kechinov, CEO

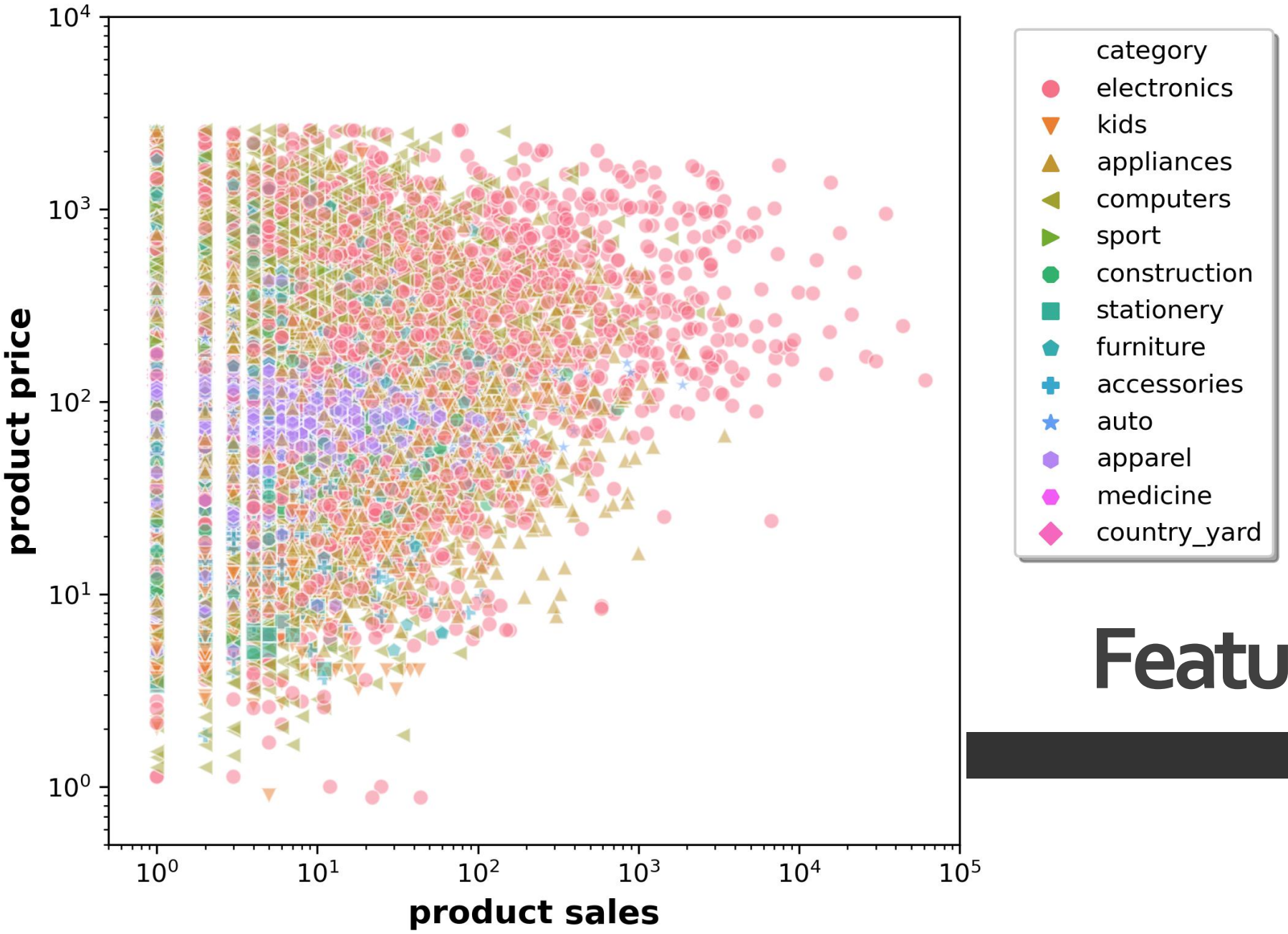
Data Question

What features are important to predict the purchase of smartphone products in each session?

Dataset

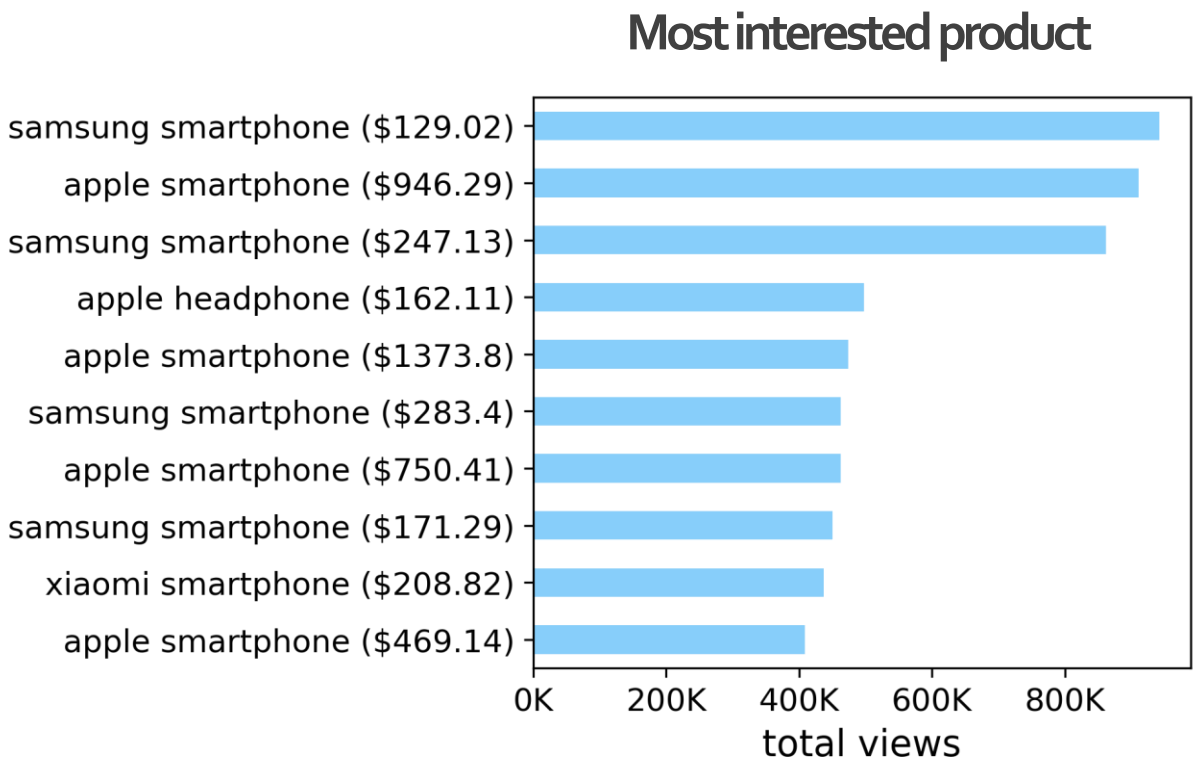
Two months (Oct – Nov 2019)	
	Unique numbers
instances	109,950,743
Session time	Date & time
event type	3 (view, cart, purchase)
product ID	74,532
category code	129
brand	2044
User ID	4,167,673
User session	15,957,931

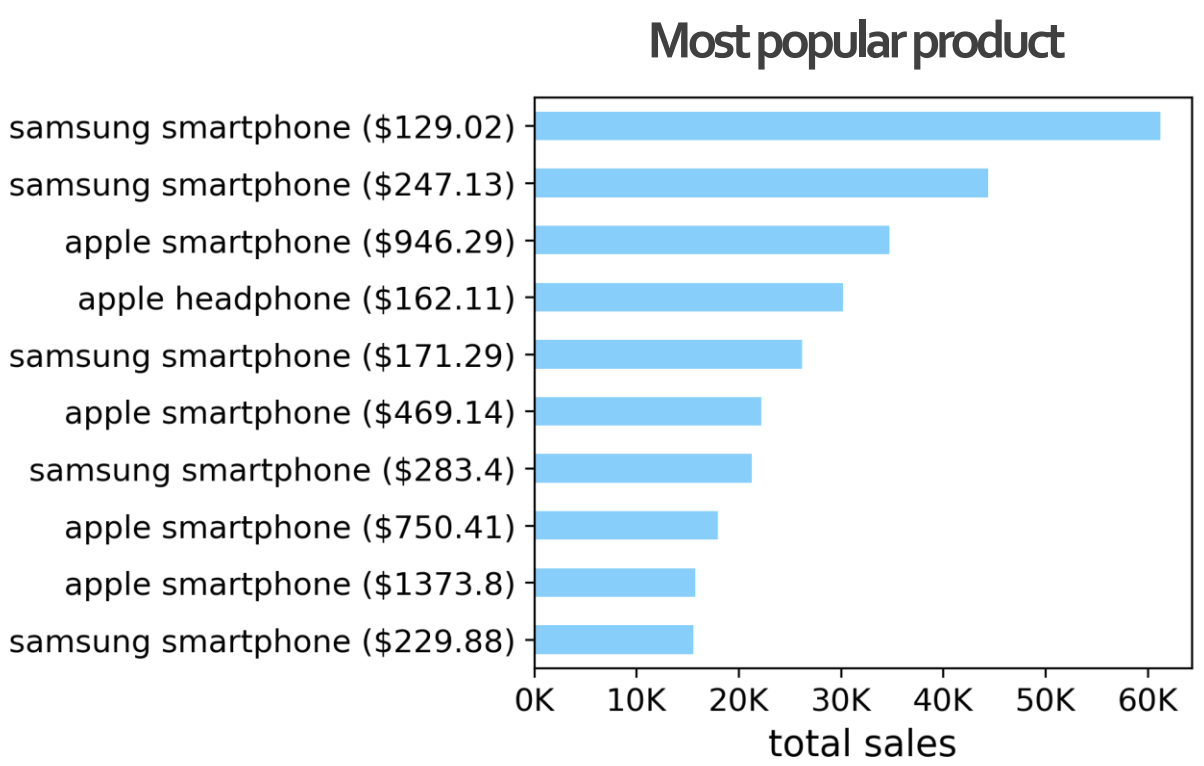
Dataset is rich in volume but lack in features. How to predict?



EDA and Feature Engineering

Product Attributes

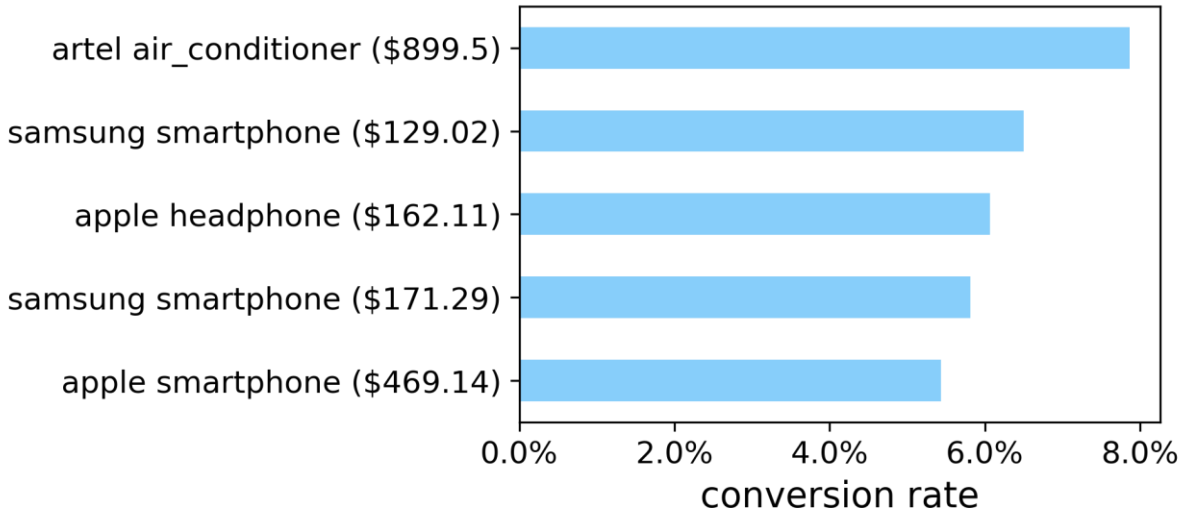




Not all viewed products are proportionally transformed into purchased.

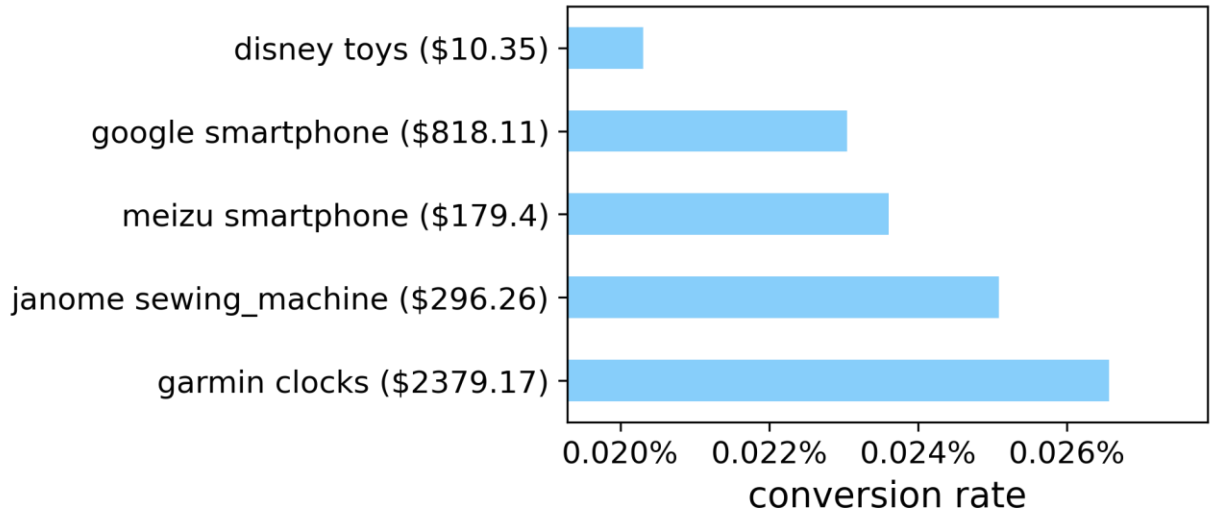
Conversion Rate Attributes

High conversion rate product



NOTE: minimum 200 sales

Low conversion rate product



NOTE: 31,113 products with zero sales

Smartphone Purchase Prediction

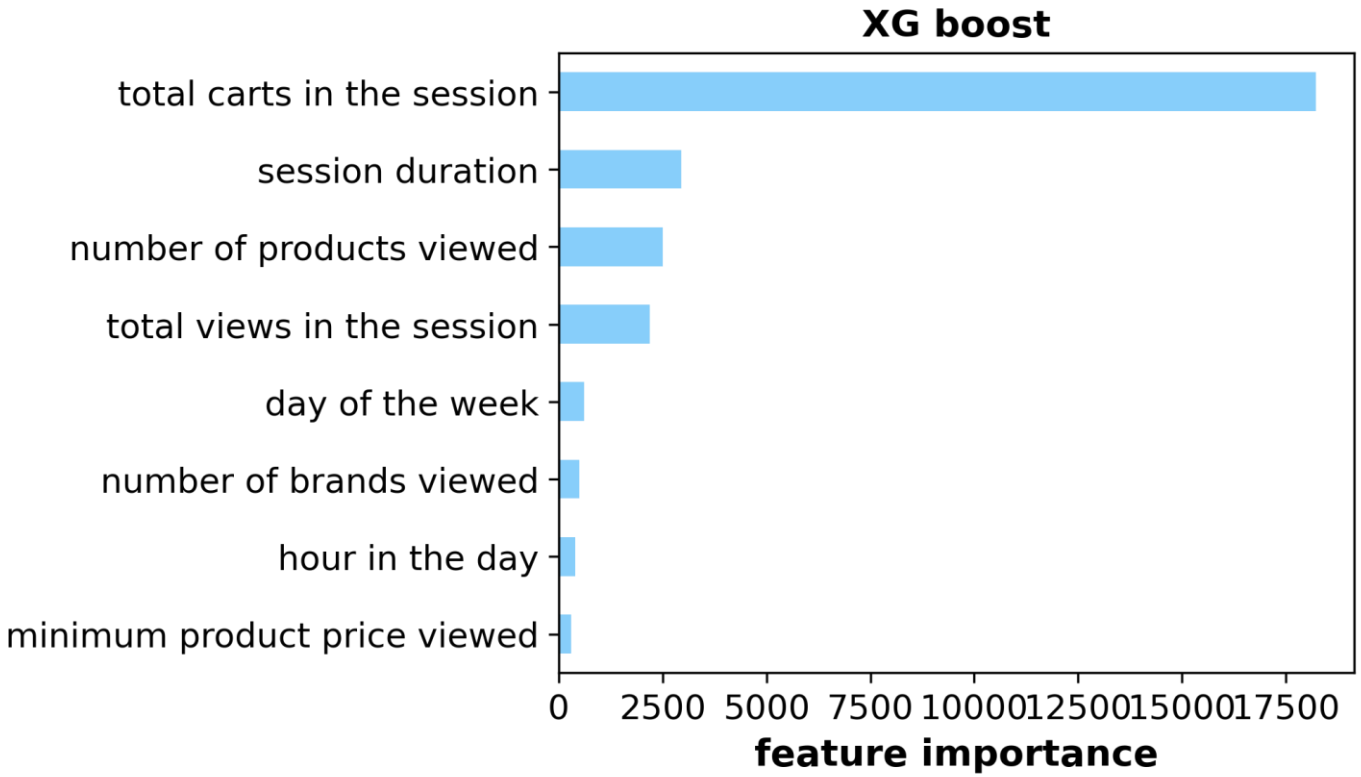
Use historical data to predict if smartphone is purchased in the sessions

Train dataset: week 40~47 , Test dataset: week 48

Smartphone Purchase Prediction

The dataset is highly imbalanced: 8.3% purchased

Model	precision	recall	*f1 score*
Logistic	55.8%	98.4%	71.2%
Ridge	56.3%	98.5%	71.6%
Decision Tree	57.0%	55.1%	56.0%
AdaBoost	69.5%	53.4%	60.4%
Bagging	68.9%	55.9%	61.7%
Gradient Boost	70.7%	61.8%	66.0%
Random Forest	71.8%	58.5%	64.5%
XGBoost	65.9%	83.6%	73.7%



Smartphone Purchase by High-Value Customers

Use historical data to predict if smartphone products are purchased in sessions of high-value customers

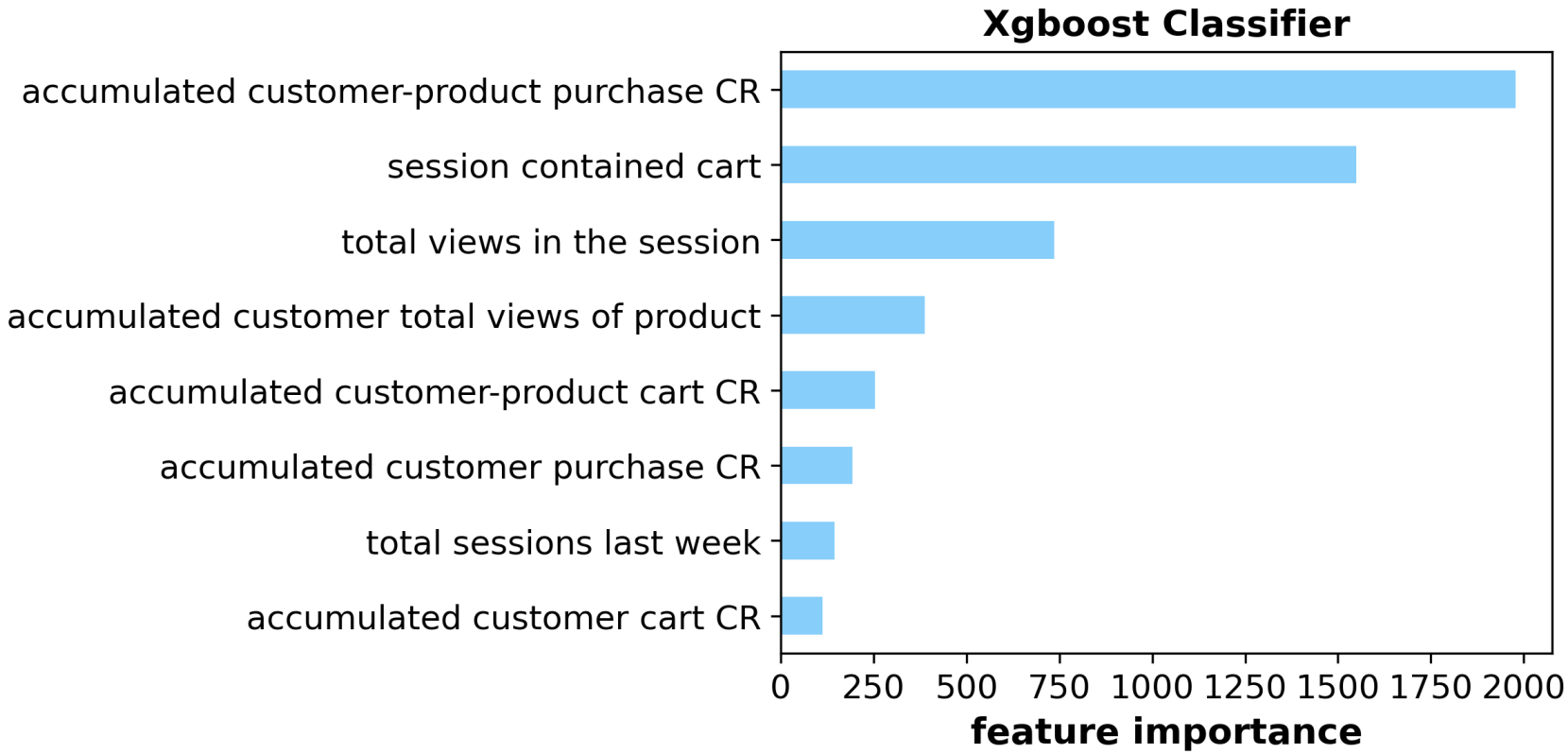
Train dataset: week 40~47 , Test dataset: week 48

Purchase Prediction of High-Value Customers

High-value customers (15% smartphone revenue contribution)

- *Customers: 2096*
- *Products: 896*
- *Sessions: 70689*

Model	accuracy
Logistic	59.4%
Ridge	83.5%
Gradient boost	83.5%
XGBoost	83.6%



CR stands for "conversion rate"

Summary and Next Steps

Dataset
- Preparation
- Cleaning

Data Engineering
- EDA
- Create features

Model Prediction
- Comparison
- Hyperparameter tuning

Implement Solution

Data Answer

Session duration, Customers' product conversion rate (cart and purchase), cart activities, total views, historical session numbers are important features for prediction

Business Answer

Yes, we can predict if smartphone is purchased in each session, with a f1 score of 73.7% for all and an accuracy of 83.6% for high-value costumers.

Next Steps

- Implement the model
- Collect more data and customer information to improve model accuracy
- Apply real-time marketing incentives to increase conversion rate

Thank You

Paget Hsiao

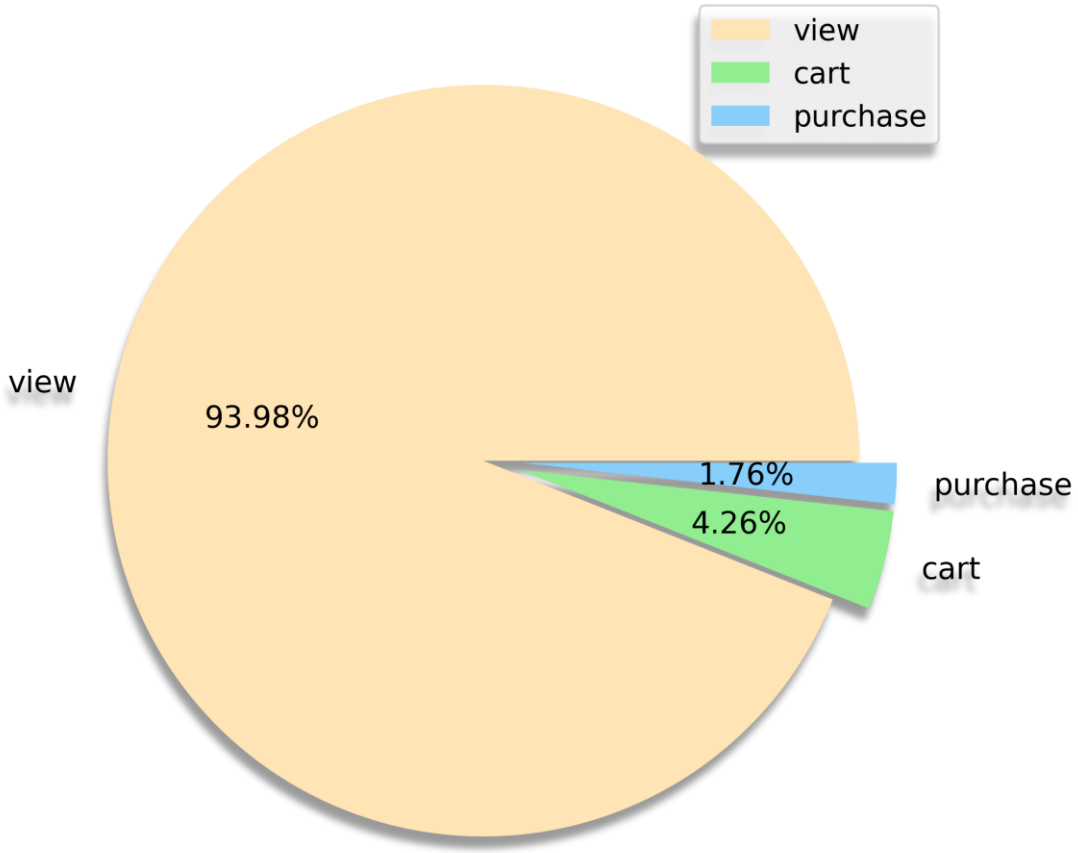
Paget.Hsiao@gmail.com

Github

Linkedin

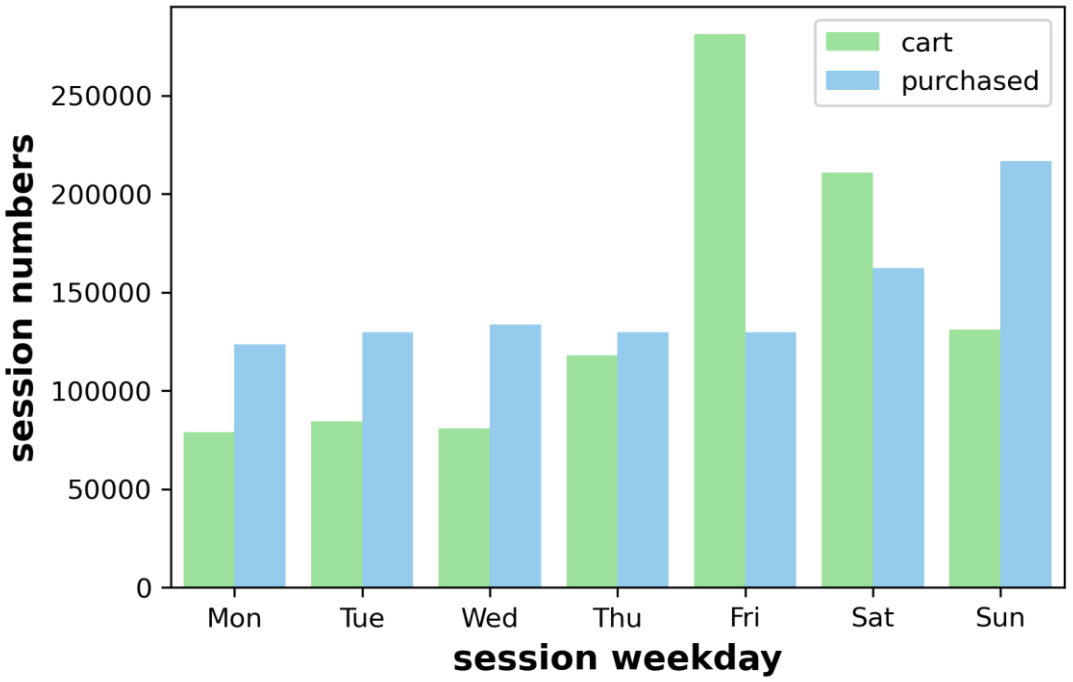
Supplementary Information

Session Statistics

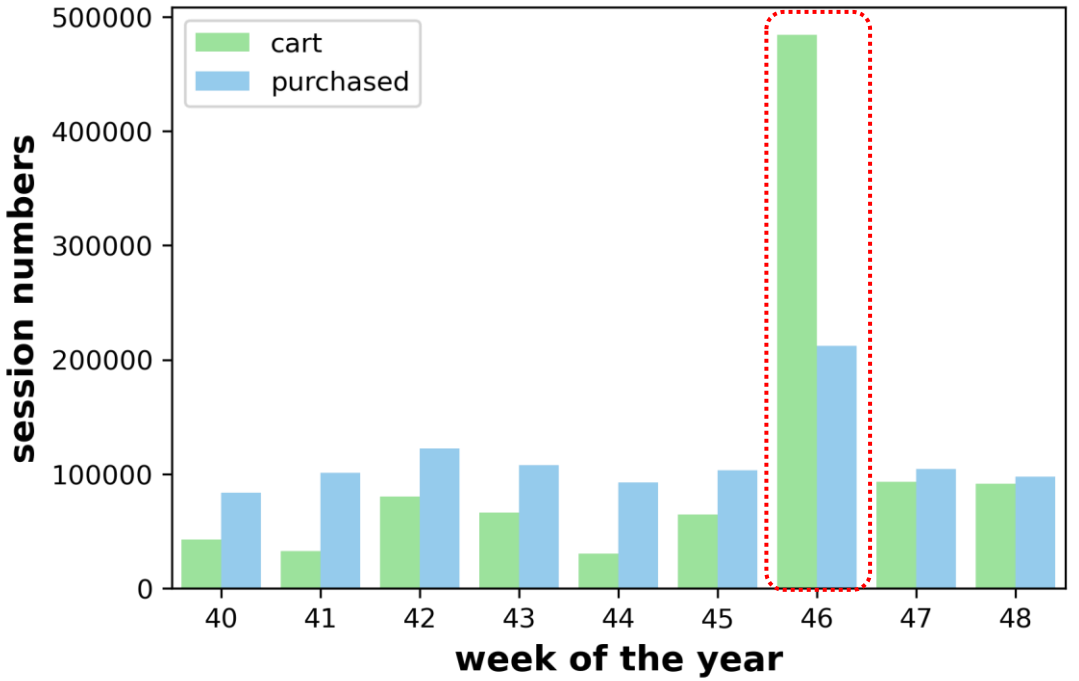


Sessions with cart or purchase	Average
“session” activities	7.9
“view” in each session	5.8
“cart” in each session	1.4
“purchase” in each session	0.6
“session” duration	220 s

Session Analysis



Customers tend to browse through the week and make decision to purchase on weekends.

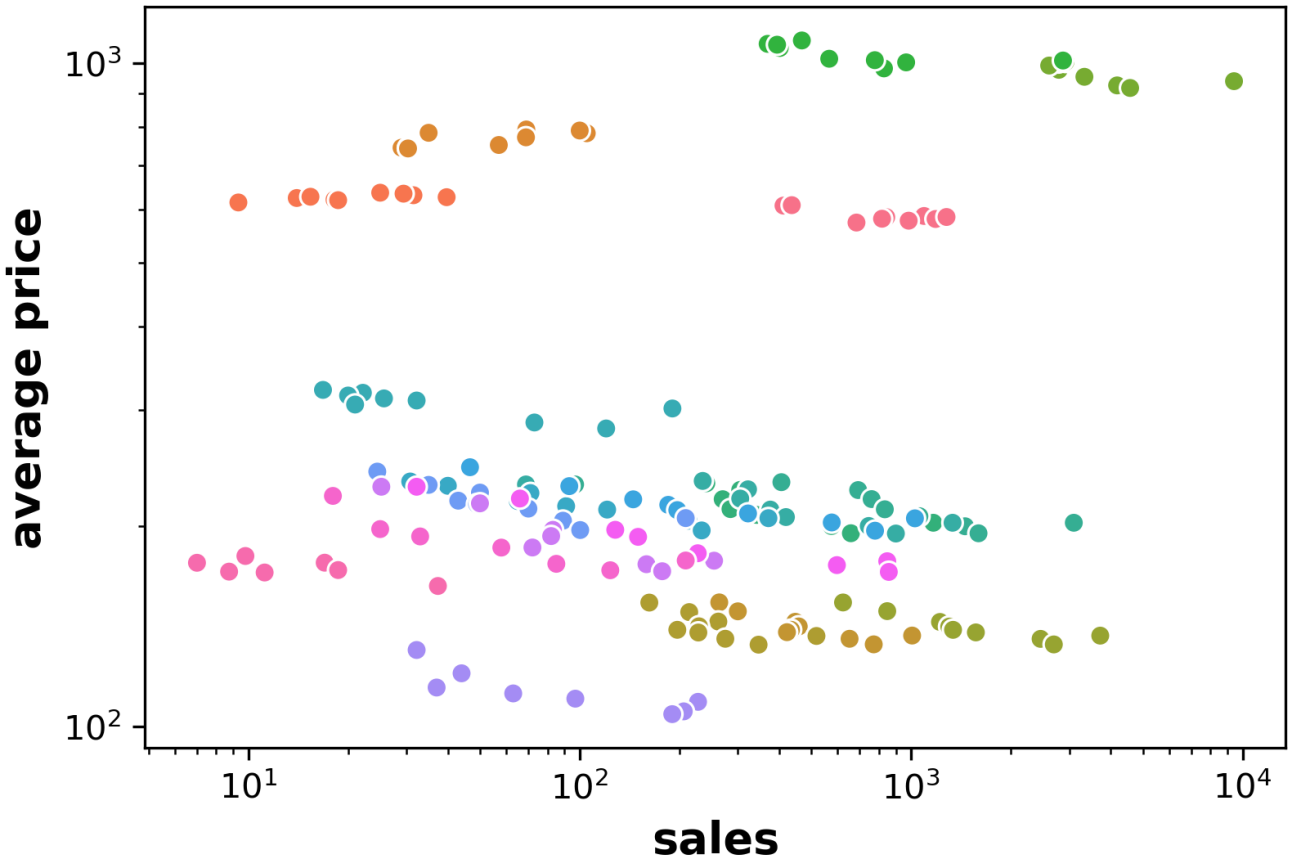


There is a surge in purchase coming toward black Friday sale.

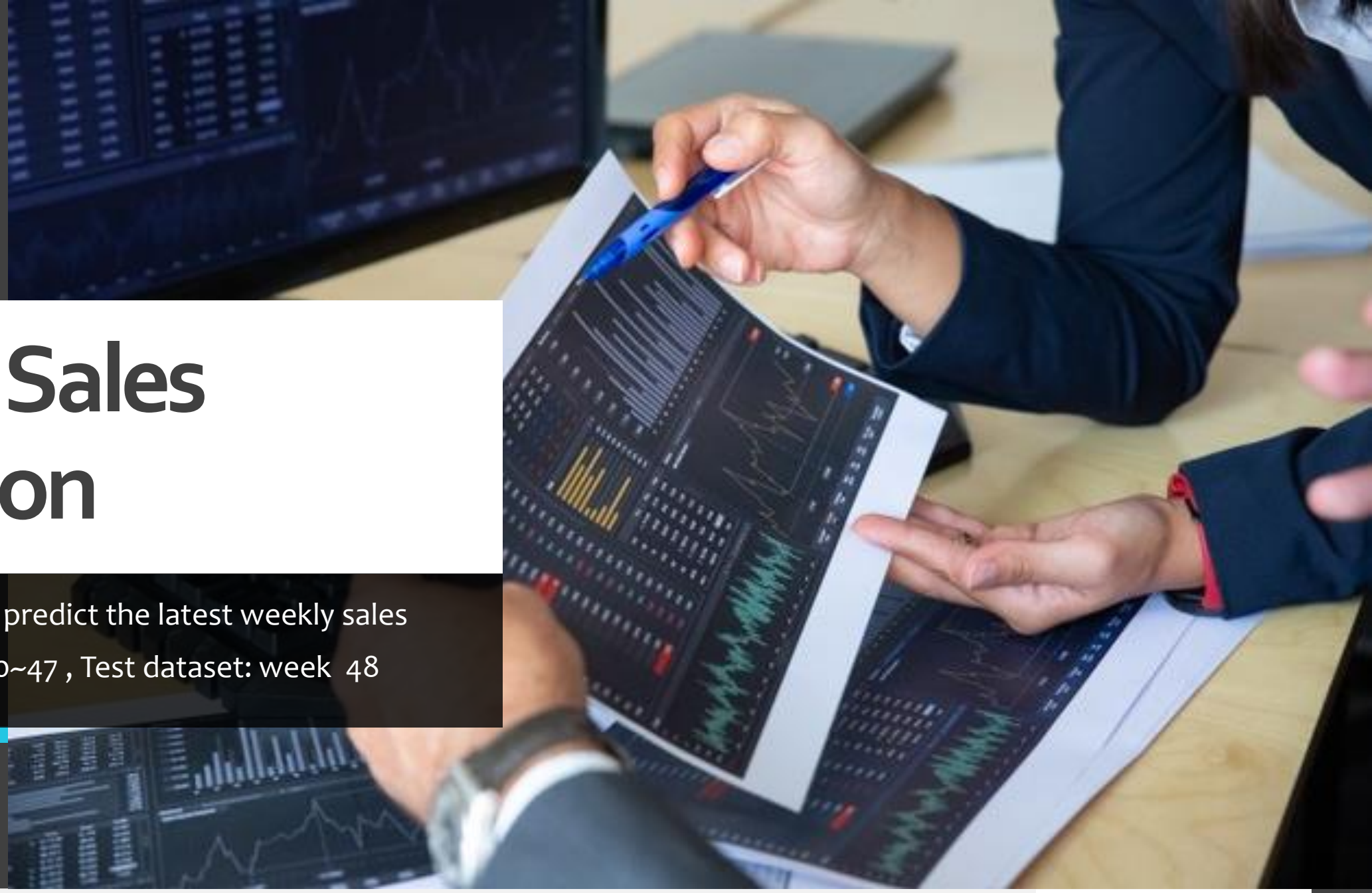
Dynamic Pricing

There are 76 products with more than 100 dynamic prices over two months.

From the fitted price elasticity of demand, only 20 products are truly price elastic (using p-values null hypothesis test)



Weekly Sales Prediction



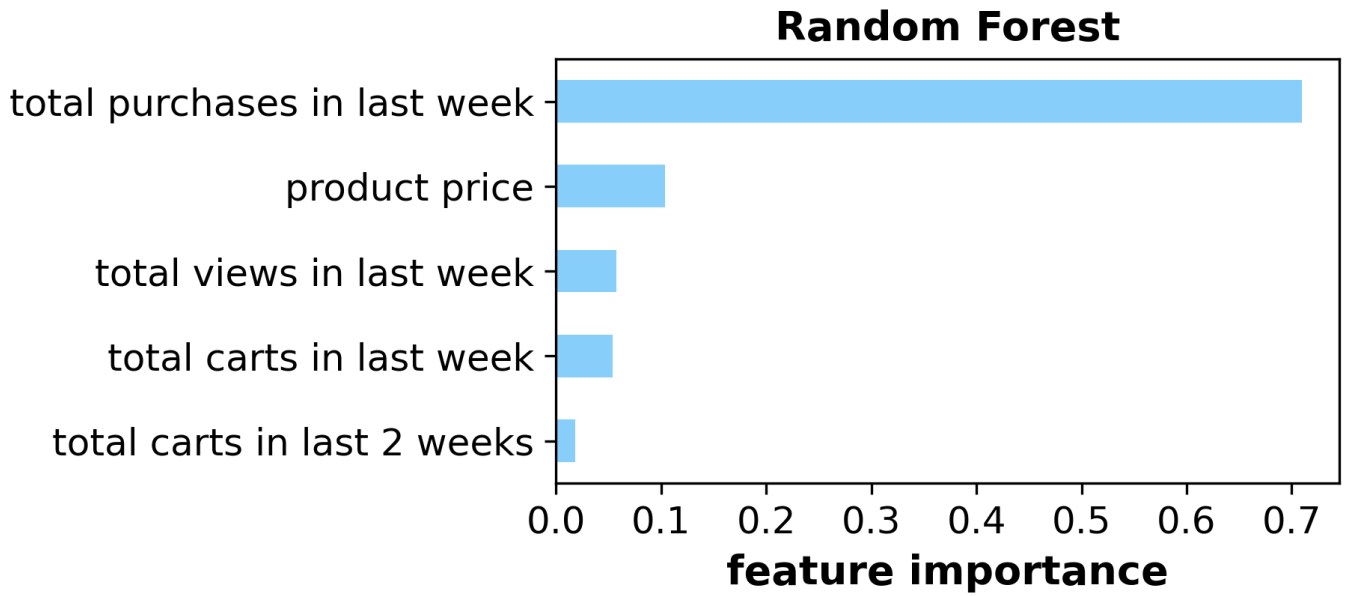
Use historical sales to predict the latest weekly sales

Train dataset: week 40~47 , Test dataset: week 48

Sales Prediction

Model	R^2 score
Linear	0.210
Ridge	0.210
Lasso	0.197
SVM (“rbf” kernel)	0.985
AdaBoost	0.869
Gradient Boost	0.941
Random Forest	0.953
XGBoost	0.940

374 products with > 100 purchases



Feature importance of SVM with ‘rbf’ kernel can not be extracted