eCommerce Purchase Decisions

Alex Qaddourah, Alex McLaughlin, Teddy Li, Junji Wiener

AGENDA

Business Understanding

What are the underlying factors that drives an eCommerce decision.

Data Understanding & Limitations

Collection, Familiarity, Problems, Discovery, Planning

Data Preparation

- Raw data transformation
- Preprocessing
- Subsetting
- Feature Engineering
- Label Encoding

Modeling

What models did we choose?

- Decision Tree
- Random Forest
- Linear Regression

Evaluation

What do our classifiers tell us about the data?

Deployment

What can be used for model deployment?

1

Business Understanding

VIEW → CART → PURCHASE

Addressing the Business Problem

Data-Driven Question

What actions within, and circumstances surrounding, a user session on an eCommerce website can help identify the eventual purchase decision?

- Product View
- Cart/Wishlist
- Day or Time

Problem Definition

The average cart abandonment rate for online retailers is 67.91%. To optimize the checkout process, marketers & analysts must be familiar with actions and timelines of user sessions.

- Remarketing
- Sale/Promotion Timelines
- Price Analysis

Motivation & Potential Benefit

A data analytics solution can illuminate characteristics of consumer groups, opportunities to change behavior, and ultimately optimize conversion.

- Motivate (would-be) non-Purchasers
- Retain Purchasers

2

Data Understanding & Limitations

Collection - Familiarity - Problems - Discovery - Planning

Data Understanding

Collection

Where did the data set come from?

Familiarity

What did our team do to familiarize ourselves with the data?

Problems

What problems did we run into during this exploratory phase?

- 42 million rows
- Data integrity gap

Discovery

- User_session
- User_id
- Product_id
- Brand
- Various Events

Planning

Is there opportunity to help answer our business problem?

Data Limitations

- Limited features
- Imbalanced
- Large (42M rows)
- Highly correlated features (view -> cart)
- Data structure

3

Data Preparation

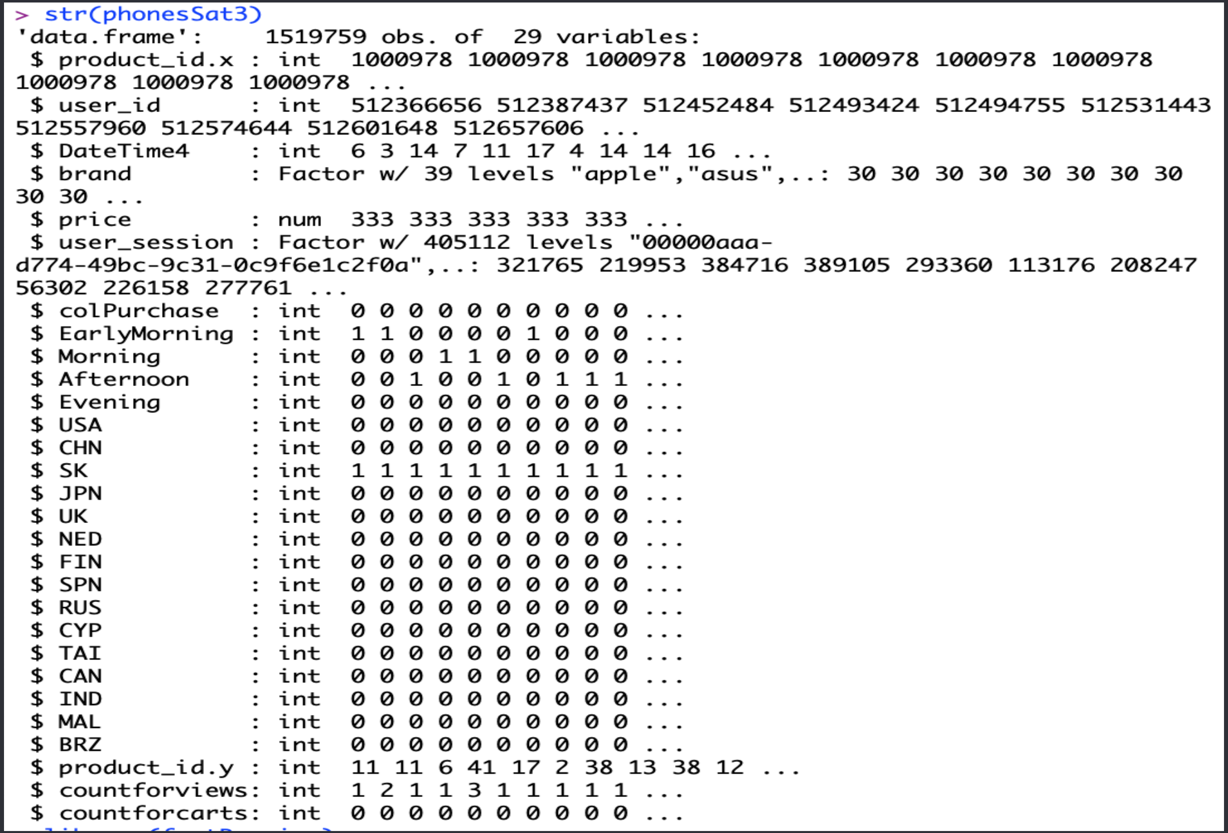
Making the data useable

Preprocessing

- Null Value Handling
- Feature Engineering
- Multicollinearity

- Subsetting
- One-Hot Encoding
- Scaling

Cleaned Data Structure



```
> str(phonesSat3)
'data.frame':	1519759 obs. of  29 variables:
 $ product_id.x : int  1000978 1000978 1000978 1000978 1000978 1000978 1000978
1000978 1000978 1000978 ...
 $ user_id      : int  512366656 512387437 512452484 512493424 512494755 512531443
512557960 512574644 512601648 512657606 ...
 $ DateTime4    : int  6 3 14 7 11 17 4 14 14 16 ...
 $ brand        : Factor w/ 39 levels "apple","asus",..: 30 30 30 30 30 30 30 30
30 30 ...
 $ price        : num  333 333 333 333 333 ...
 $ user_session : Factor w/ 405112 levels "00000aaa-
d774-49bc-9c31-0c9f6e1c2f0a",..: 321765 219953 384716 389105 293360 113176 208247
56302 226158 277761 ...
 $ colPurchase  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ EarlyMorning : int  1 1 0 0 0 0 1 0 0 0 ...
 $ Morning      : int  0 0 0 1 1 0 0 0 0 0 ...
 $ Afternoon    : int  0 0 1 0 0 1 0 1 1 1 ...
 $ Evening      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ USA          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CHN          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SK           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ JPN          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ UK           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NED          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FIN          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SPN          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RUS          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CYP          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TAI          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CAN          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ IND          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MAL          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ BRZ          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ product_id.y : int  11 11 6 41 17 2 38 13 38 12 ...
 $ countforviews: int  1 2 1 1 3 1 1 1 1 1 ...
 $ countforcarts: int  0 0 0 0 0 0 0 0 0 0 ...
```

4

Modeling

Classifiers

- Logistic Regression
- Random Forest
- Neural Network
- XGBoost

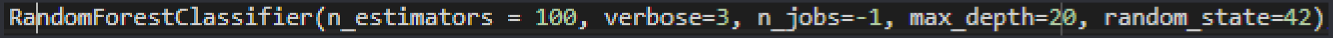
Performance Metrics

- Confusion Matrix
  - Accuracy
  - TPR
  - FPR
  - F1 score
  - Precision
  - ROC AUC

Logistic Regression

- Score - 56%
- Accuracy - 46%
- TPR - 52%
- FPR - 44%
- F1 score - 54%
- Precision - 57%
- ROC AUC - 56%

Random Forest



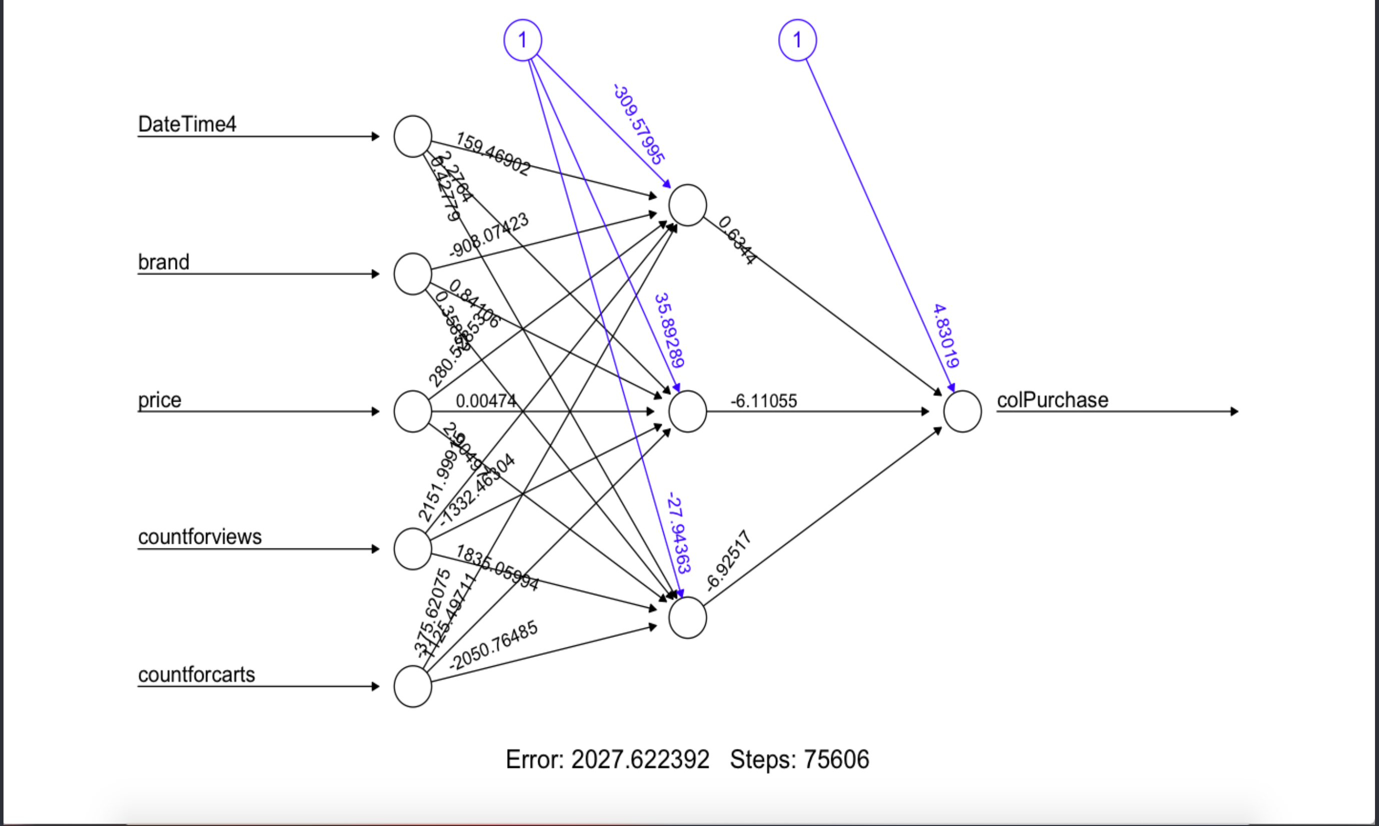
```
RandomForestClassifier(n_estimators = 100, verbose=3, n_jobs=-1, max_depth=20, random_state=42)
```

- Score - 64%
- Accuracy - 38%
- TPR - 53%
- FPR - 62%
- F1 score - 60%
- Precision - 66%
- ROC AUC - 70%

Neural Network

- Accuracy - 91%
- TPR - 0.12%
- FPR - 0.04%
- F1 score - 0.23%
- Precision - 21%

Neural Network



XGBoost

- Score - 57%
- Accuracy - 40%
- TPR - 57%
- FPR - 41%
- F1 score - 58%
- Precision - 59%
- ROC AUC - 62%

5

Evaluation

What did our classifiers find?

Confusion Matrix

Random Forest Classifier

	1	0
1	148497	128850
0	69668	207211

Logistic Regression (upsampled)

	1	0
1	144237	133110
0	109328	167551

Neural Network

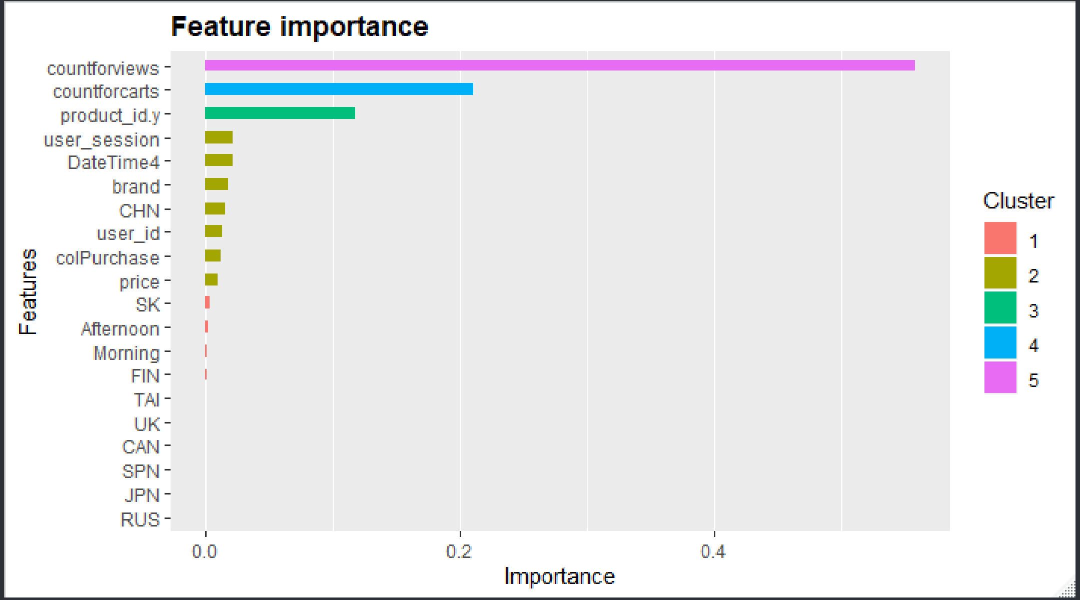
	0	1
0	1384991	572
1	134040	156

XGBoost (upsampled)

	1	0
1	164157	113190
0	123354	153525

Feature Ranking

- Product View Counts (total)
- Cart View Counts (total)
- Unique Products Viewed
- User Sessions (total)
- Hour of Day (24hr)
- Brand
  - Samsung
  - Xiaomi
  - Apple
  - Huawei
  - Oppo
  - Meizu
  - Nokia
  - Vivo
  - TP-Link



Interpreting Models

- Which model had the highest Accuracy?
  - Neural Network (unbalanced)

- Which model had the best TPR?
  - XGBoost

- Which model had the best F1 Score?
  - Random Forest

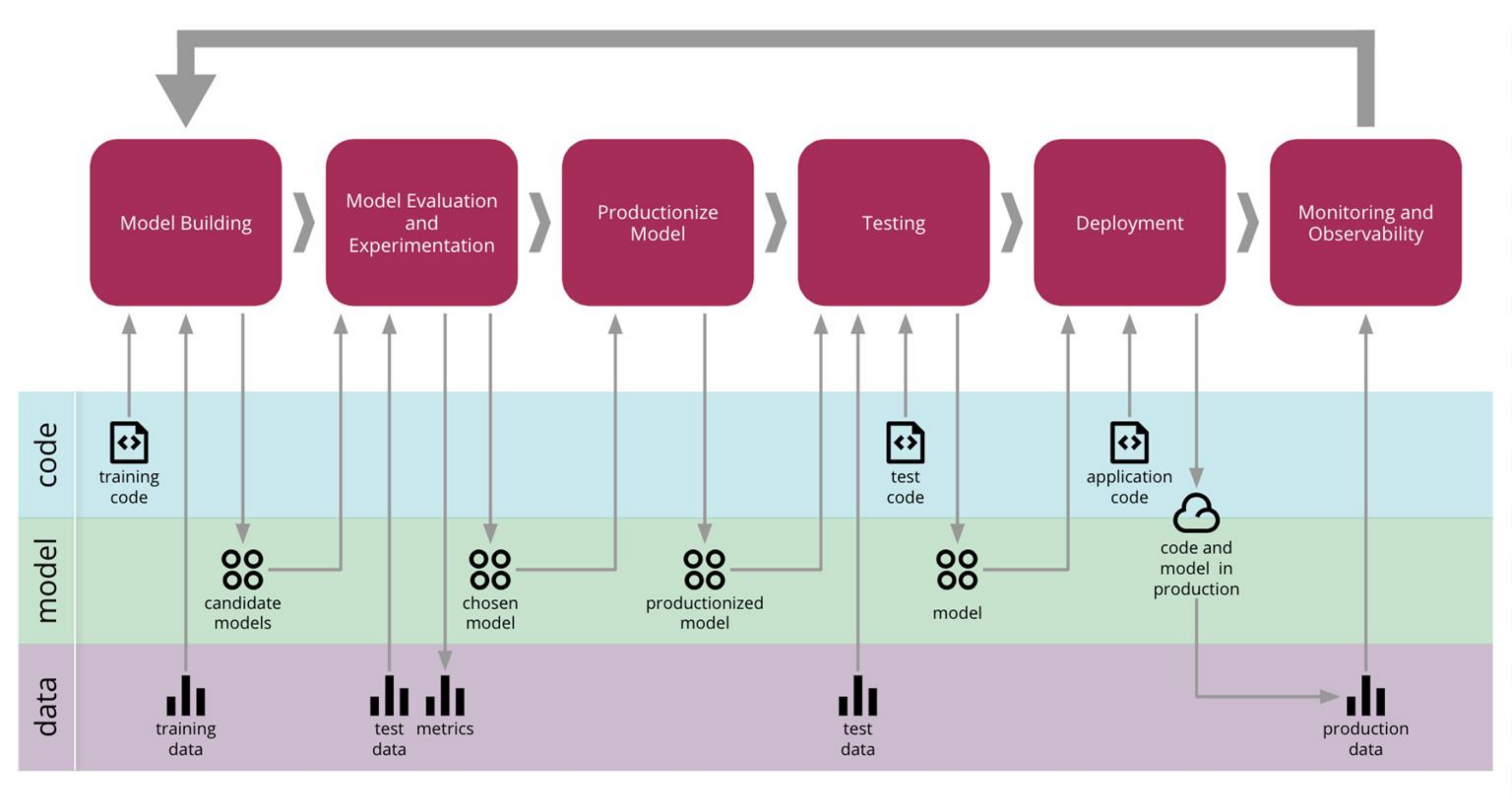
- Which model had the highest Precision?
  - Random Forest

6

Deployment

Continuous Delivery for Machine Learning (CD4ML)

End to End Continuous Delivery for Machine Learning



Thanks!

ANY QUESTIONS?

GitHub’s:

- 8Jun
- alexqaddourah
- almc6742
- TeddyCU

Citations

- Continuous Delivery for Machine Learning
  - https://martinfowler.com/articles/cd4ml.html#initial-ml-process.png
    - Danilo Sato, Arif Wider, Christoph Windheuser
- Kaggle - https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store