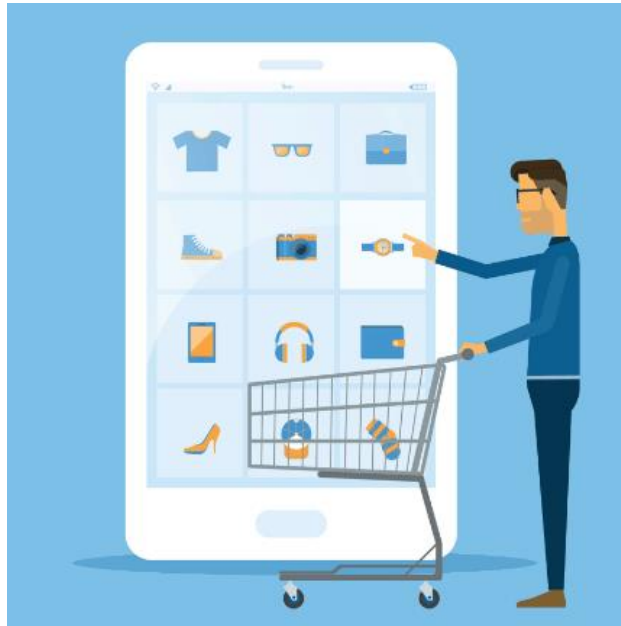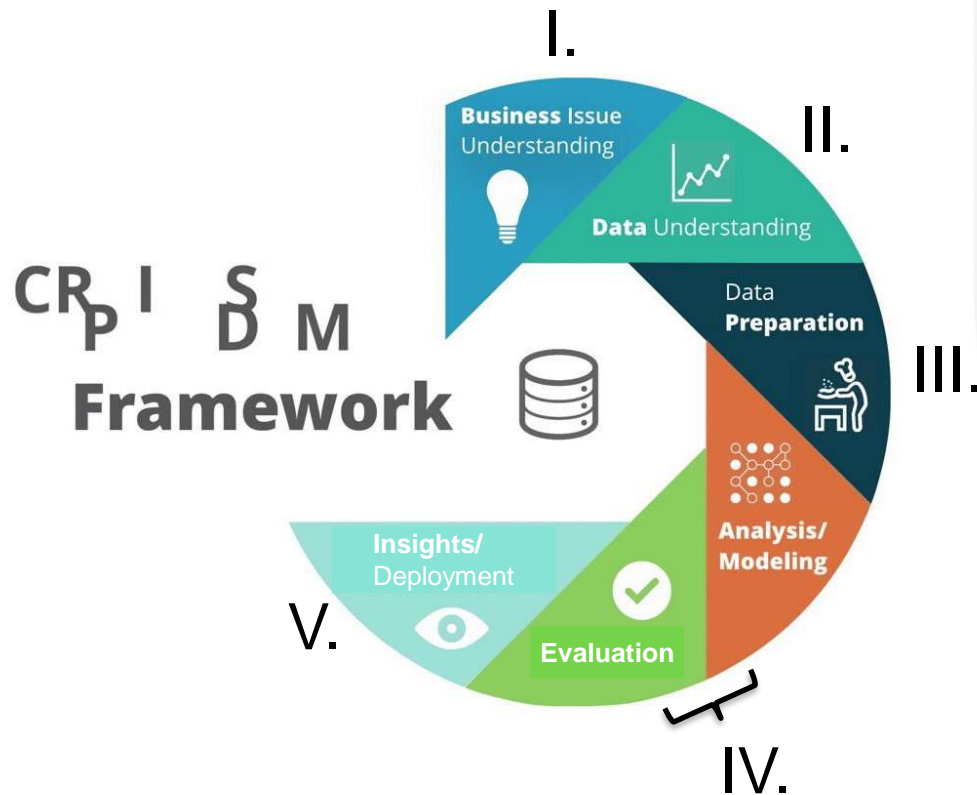# Analysis of
Online Shoppers' 'Conversion'

BA545 Final Project for Machine Team 4

(Michael DiSanto, Dawn Massey & Brian Nicholls)

April 28, 2020

Agenda: How we analyzed 'conversion'

I. Business understanding: Who is the client?

- *Industry/Company/Products:*

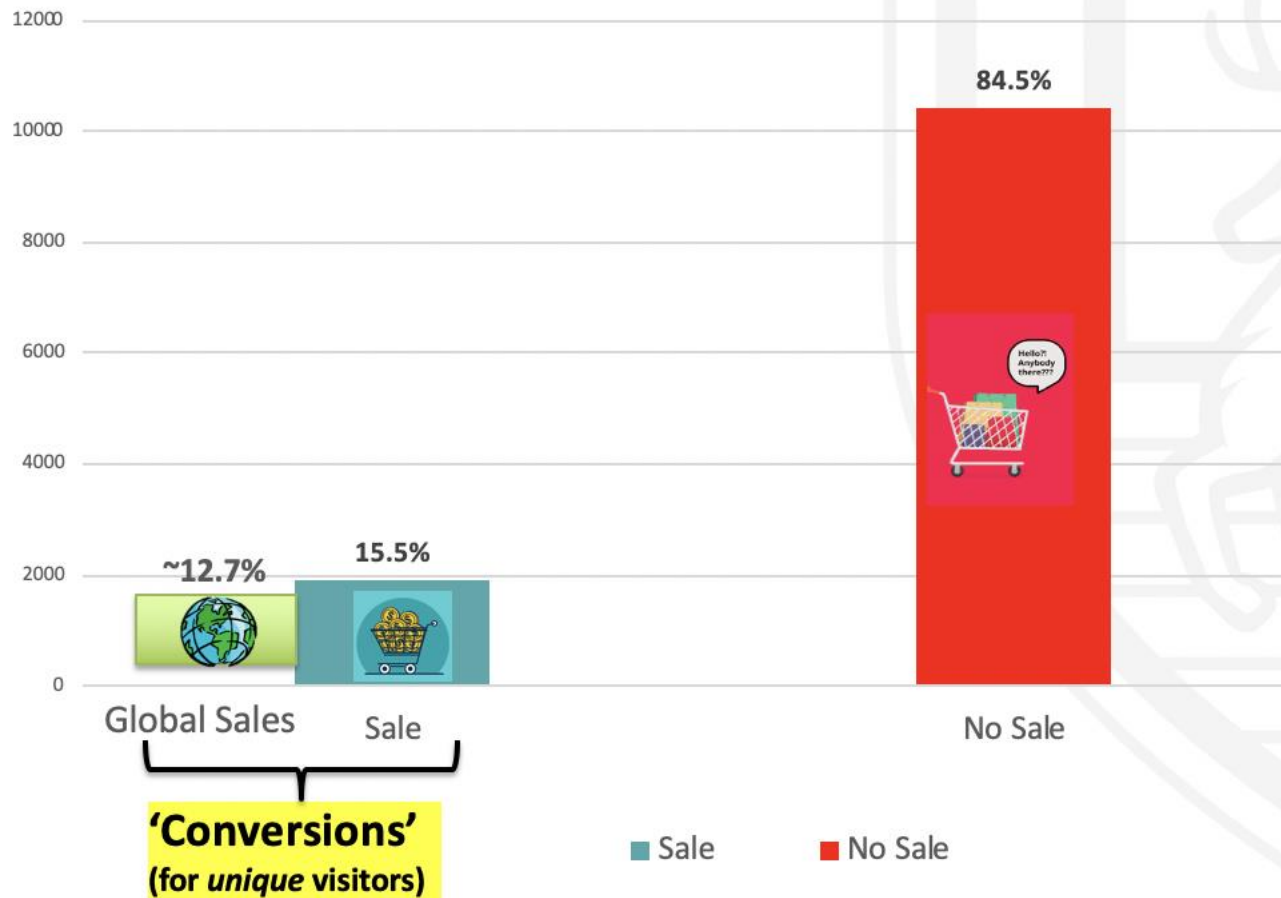# I. Business understanding: What's the issue?

- ***Research Q: What drives 'conversion'?***

Sale/No Sale in the Dataset v. Global Sales

I. Business understanding: Why do we care?

- ***Motivation:***

Conversion increases are impactful!

+10% → +$150,000 sales rev/year*
(Among ***unique*** visitors only)

*12,330 customers in 10 months → ~41 customers/day
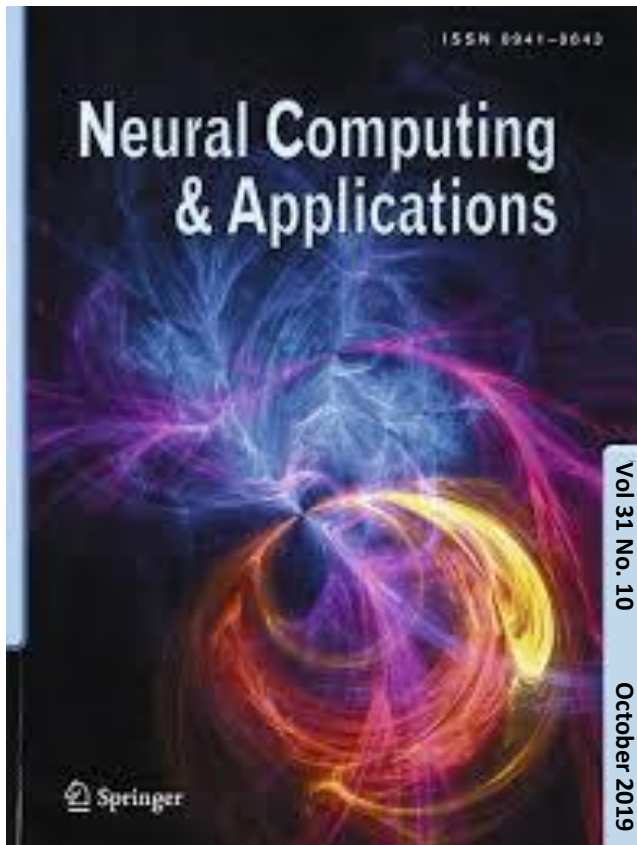
An increase of 10% in conversions → ~4.1 customers/day

At ~365 internet selling days/year → ~1,500 sales/year

At ~$100/sale → $150,000 increase in revenue

I. **Business understanding:** What else do we know?

- *Data source:* Online shopper browsing session data

See: Sakar, C., S. Polat, M. Katircioglu & Y. Kastro. 2019. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*. *31* [10, October]: 6893-6908.

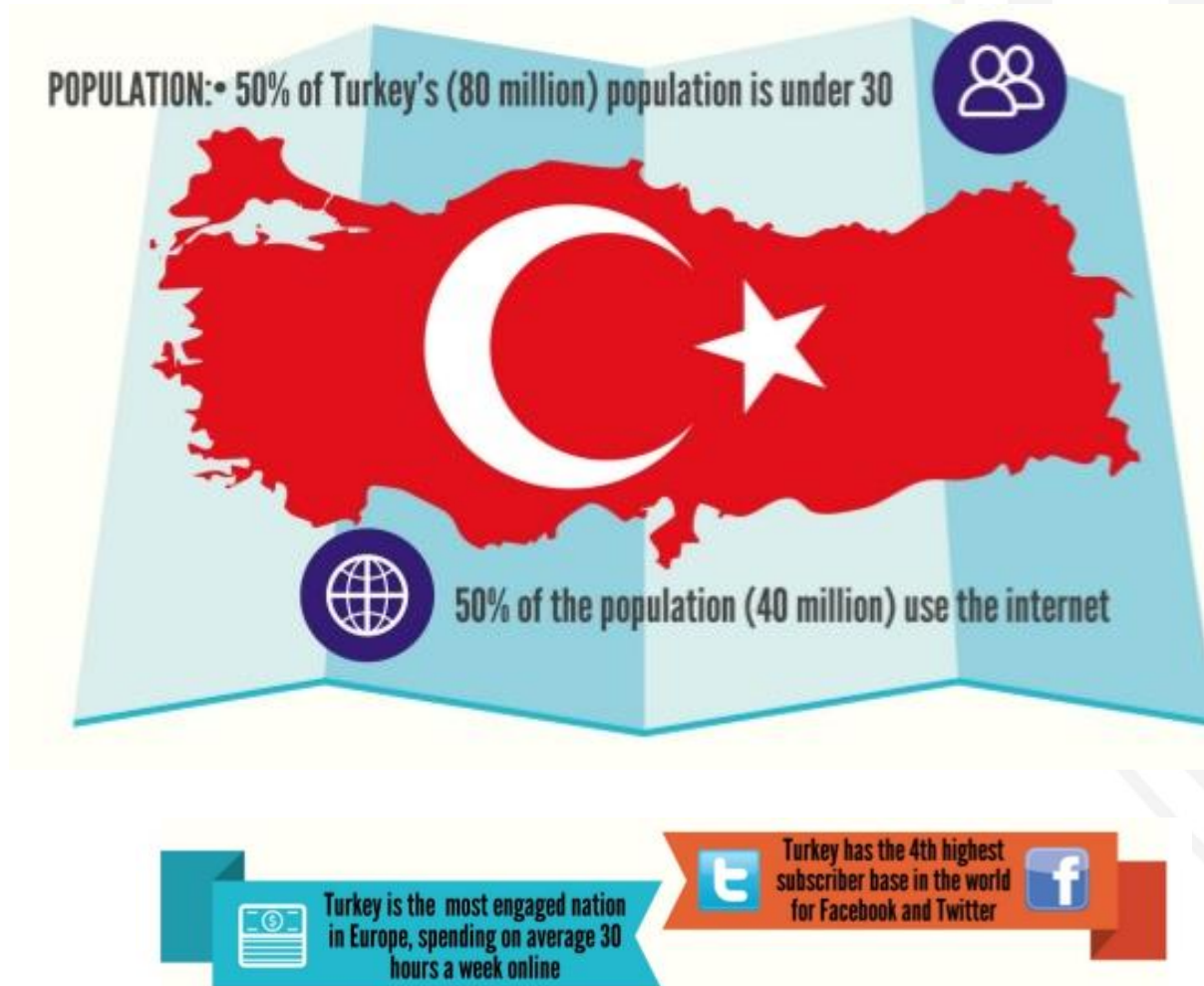# I. Business understanding: What else do we know?

- ## *Company/Industry/Products:*

Online purchases in Turkey (Y/E March 2015)*

I. Business understanding: What else do we know?
- *Geography:*

POPULATION:• 50% of Turkey's (80 million) population is under 30

50% of the population (40 million) use the internet

Turkey is the most engaged nation in Europe, spending on average 30 hours a week online

Turkey has the 4th highest subscriber base in the world for Facebook and Twitter

I. **Business understanding:** What else do we know?

- ***Time Period:***

  – Data collection is between 2004 and July 2017

    - Data from 10 months only → Jan/Apr excluded
    - Just 2 'Special' Days:
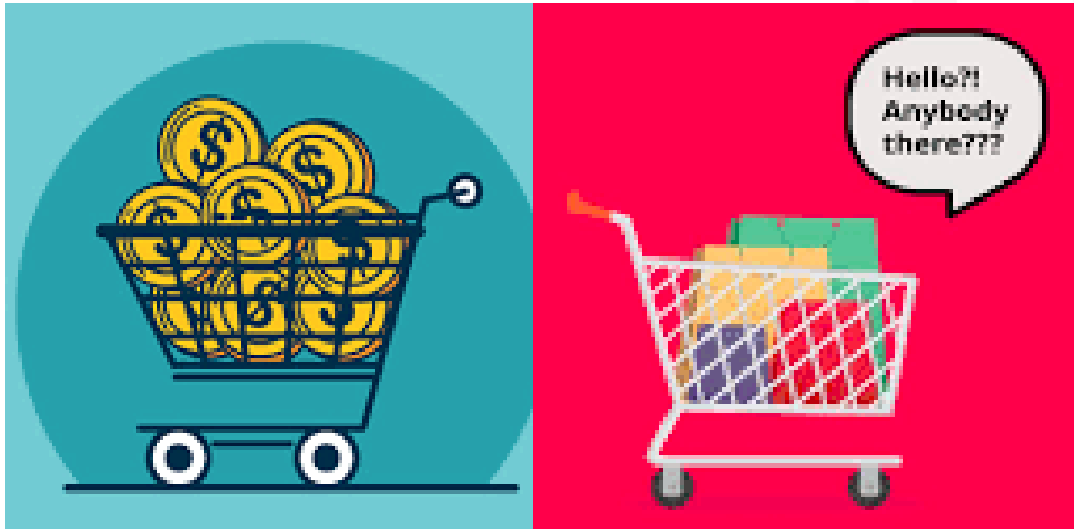   – Valentine's Day
      – Mother's Day

2004

Gozalan/Columbia association

JULY 2017

Journal article submission

II. **Data Understanding:** What's the data like?

- ***Dataset Details:***
  - 12,330 *unique visitors*; 18 features
    - 1 target feature

Conversion (Sale) v. Abandonment (No-sale)

II. Data Understanding: What's the data like?

- ## *Dataset Details (con't):*

  - 10 Numerical Features
    - Behavioral information for web visitors

  - 7 Categorical Features
    - De facto demographics from URL information

II. **Data Understanding:** What does EDA reveal?
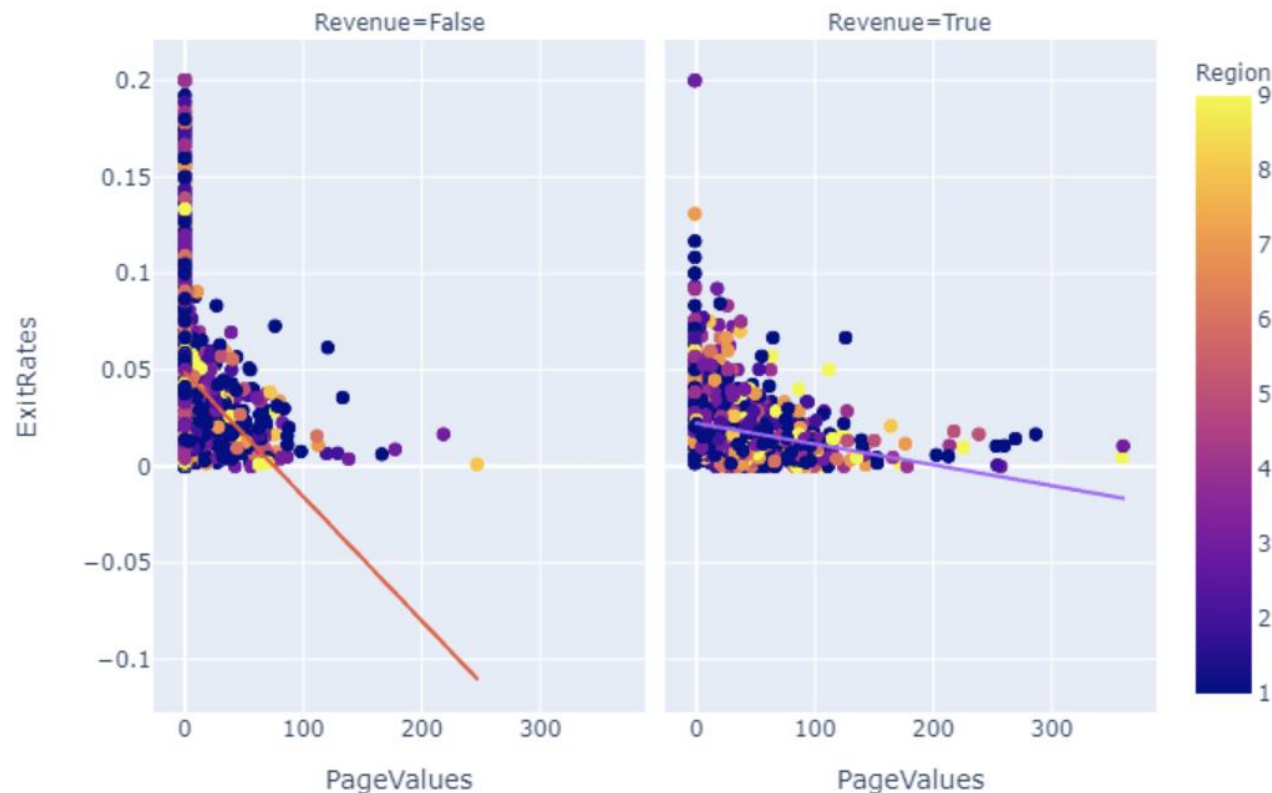
- ***Visualizations/Key Observations:***
  - Do BounceRates v. ExitRates guide us in the predictive models by revealing different trends in Abandonment (Rev=False) v. Conversion (Rev=True)?

II. Data Understanding: What does EDA reveal?

- ## *Visualizations/Key Observations (con't):*
  - Do PageValues v. ExitRates guide us in the predictive models by revealing different trends in Abandonment (Rev=False) v. Conversion (Rev=True)?

II. **Data Understanding:** What does EDA reveal?
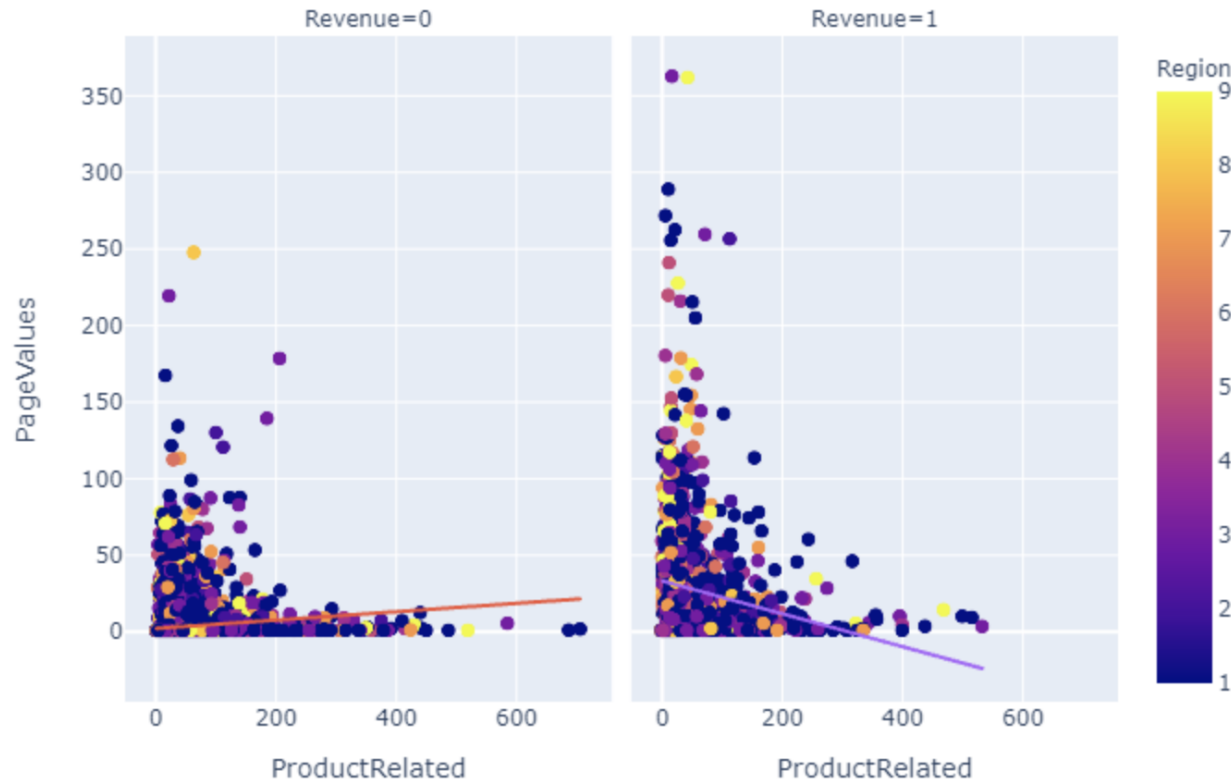
- ## *Visualizations/Key Observations (con't):*
  – Do PageValues v. ProductRelated guide us in the predictive models by revealing different trends in Abandonment (Rev=False) v. Conversion (Rev=True)?
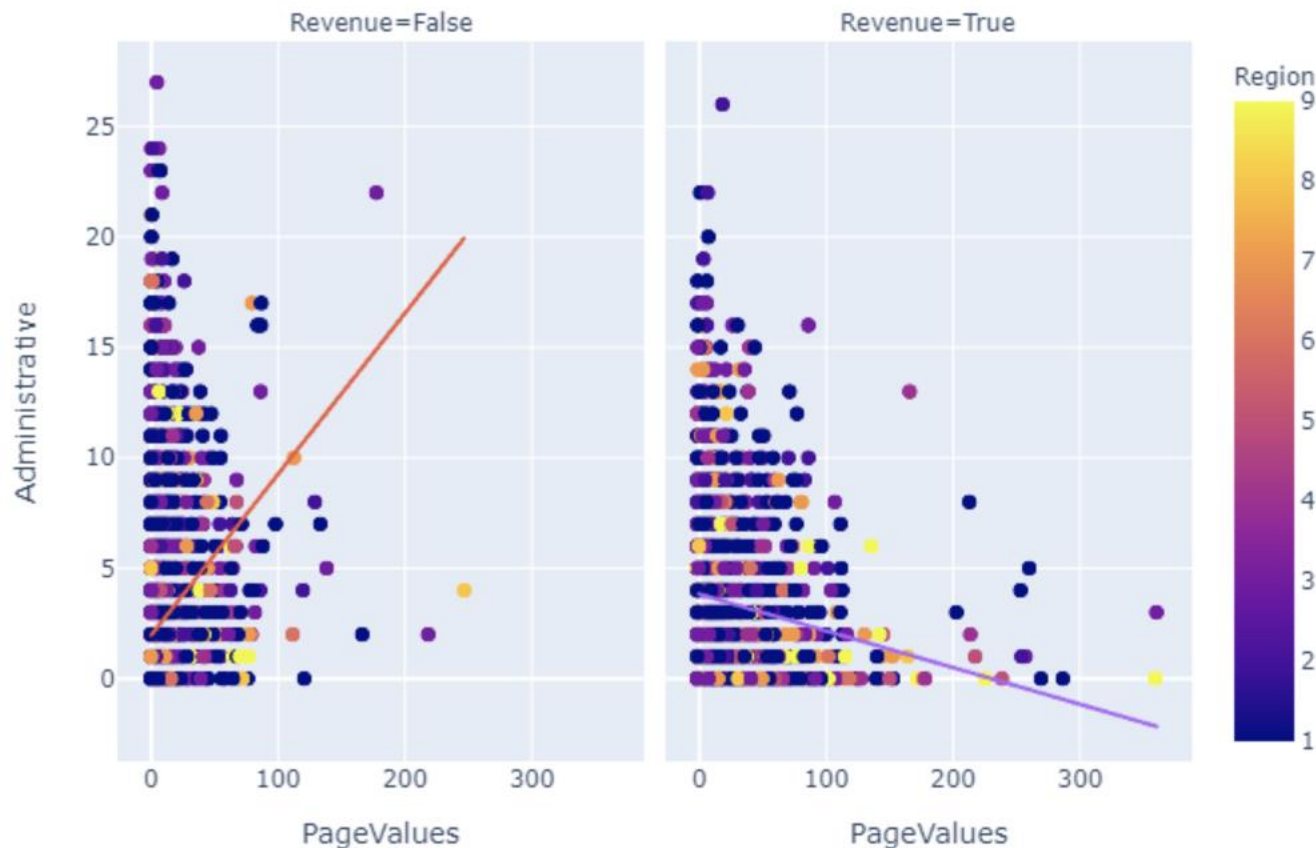
II. **Data Understanding:** What does EDA reveal?

- ## *Visualizations/Key Observations (con't):*
  - Do PageValues v. Administrative guide us in the predictive models by revealing different trends in Abandonment (Rev=False) v. Conversion (Rev=True)?
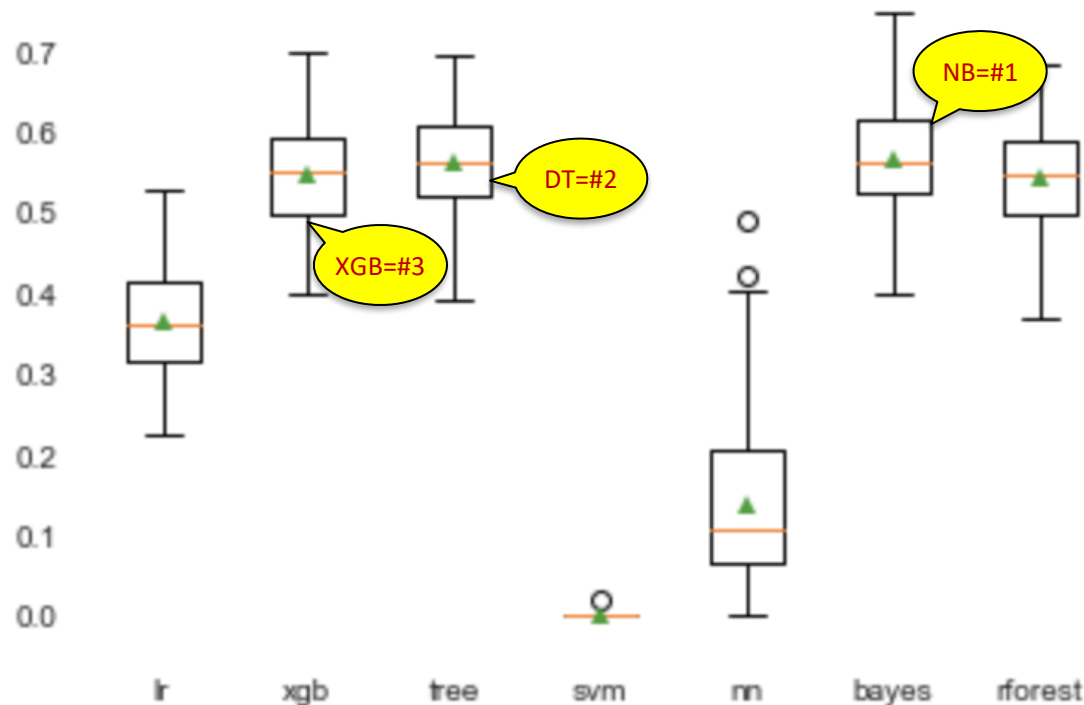
III. Data Preparation: What did we do

- ## *Overview of Steps (after csv imported):*

  ## 1) Preliminary base models:

  - Encode non-numeric (object/boolean) features
  - Run 7 models with all features & rank using repeated K-Fold
  - *Address RQ: What drives **conversion** → **PPV* for 'sale'** (not F1)*
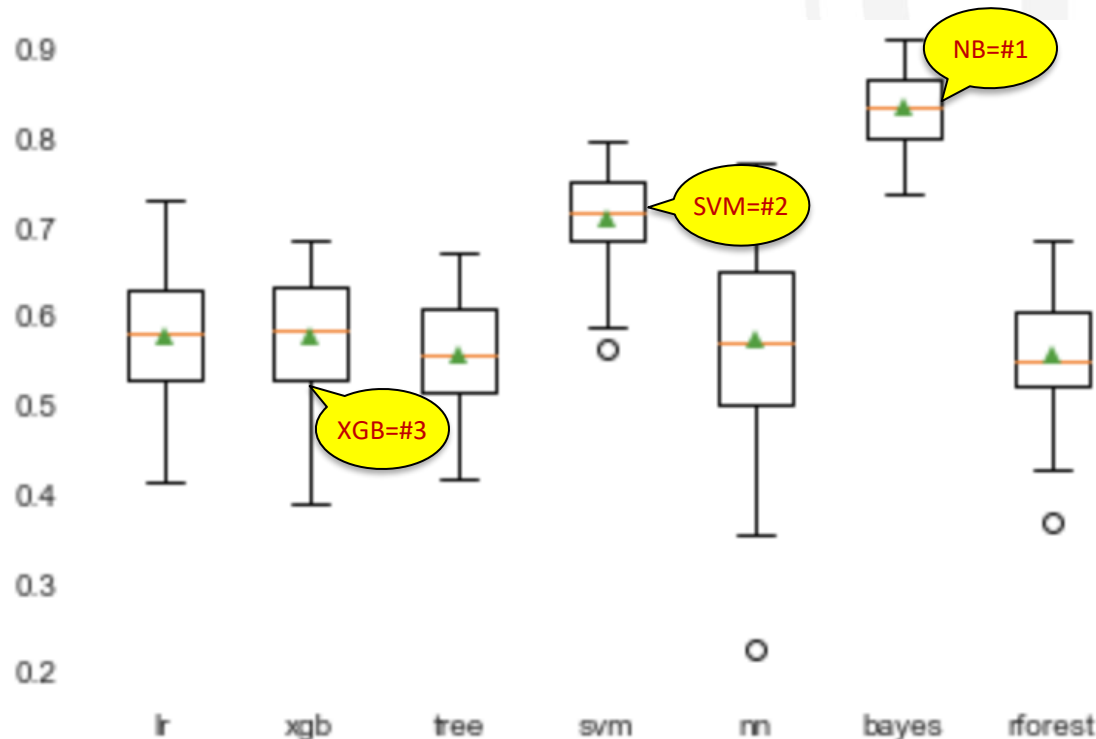
  (*positive predictive value)

III. Data Preparation: What did we do

- ## *Overview of Steps (con't):*

  ## 2) Initial base models/preparation process:

  – Impute, encode, bin (categorical features) – e.g., Bin 9 Regions to 5

  – Transform, engineer features, standardize, normalize (continuous features) – e.g., Engineer PageValues v ExitRates

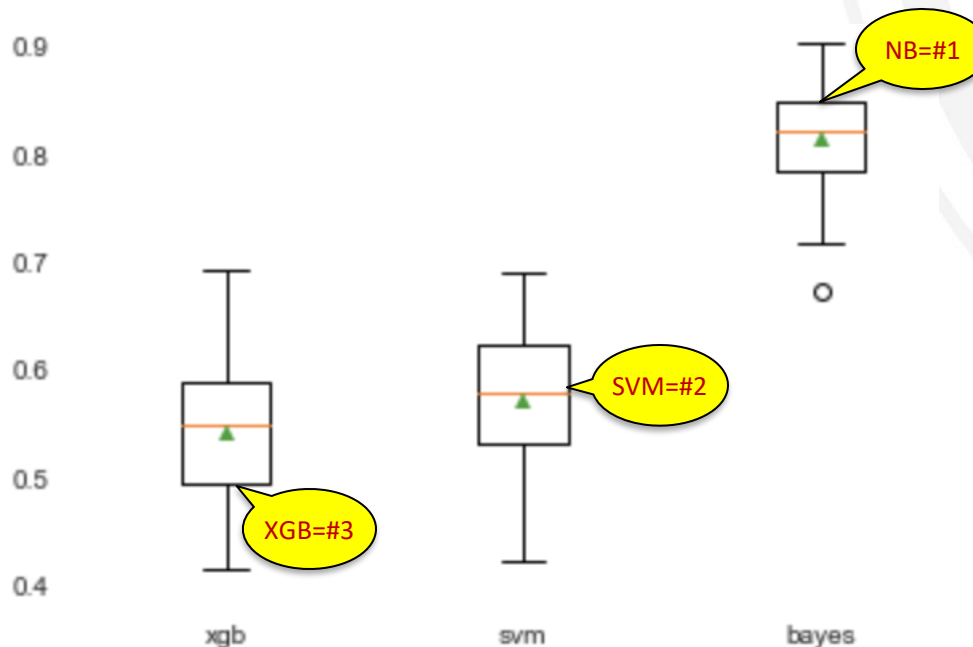 – Run models & rank; evaluate via PPV for 'sale'

IV. Analysis/Modeling/Evaluation: What did we do?

- ***Overview of Steps (after data prep):***

3) Revised base models:

- – Remove highly correlated features
- – Use 3 best models & determine top features for each
- – Run 3 'best features' models (NB, SVM, XGB)
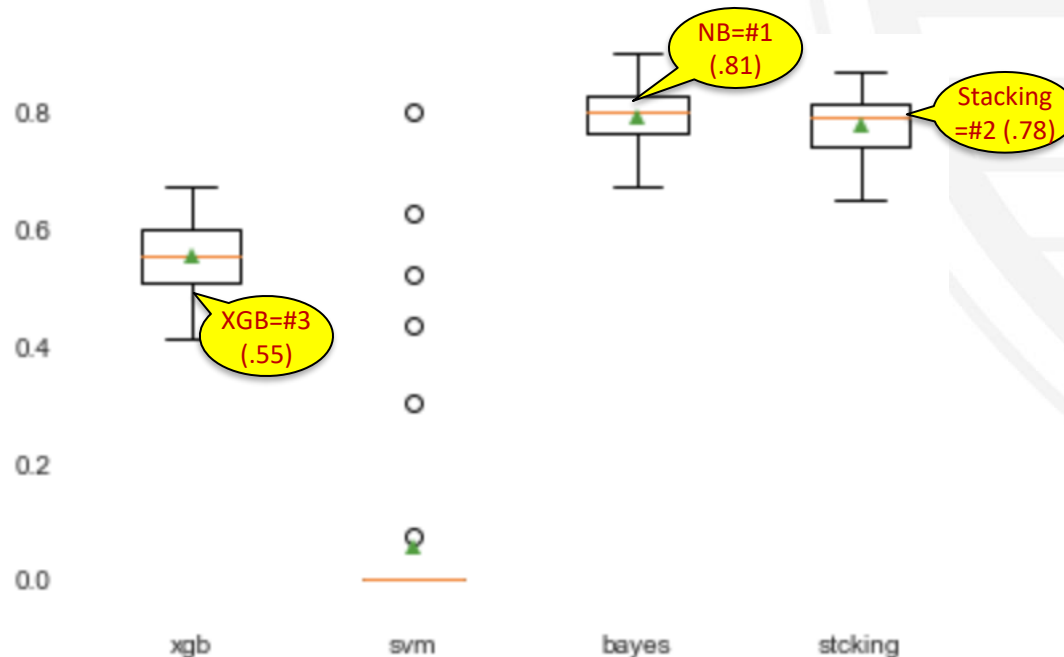- – Evaluate results via PPV for 'sale' & rank models

IV. Analysis/Modeling/Evaluation: What did we do?

- ## *Overview of Steps (after data prep):*

4) Final models (NB, SVM, XGB):

- – Tune HPs* in 2 advanced machine learning models (SVM, XGB)
- – Adjust for imbalanced data via SMOTE
- – Combine optimized models using Stacking Ensemble
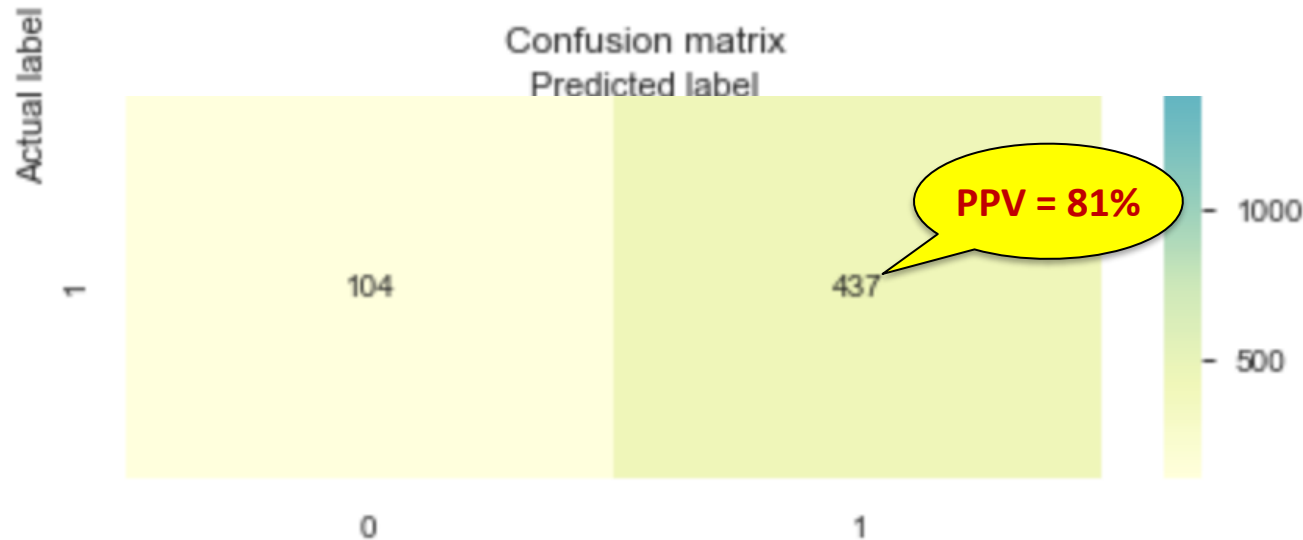- – Evaluate via PPV for 'sale'

*HPs = Hyper-parameters

V. Insights/Deployment: What do we find?

- ***Naïve Bayes** = best model to predict 'conversion'*
  - TPOT confirms Naïve Bayes as best model

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Confusion matrix
Predicted label

	0	1
Actual label 1	104	437

PPV = 81%

- 1000
- 500

Fairfield
UNIVERSITY

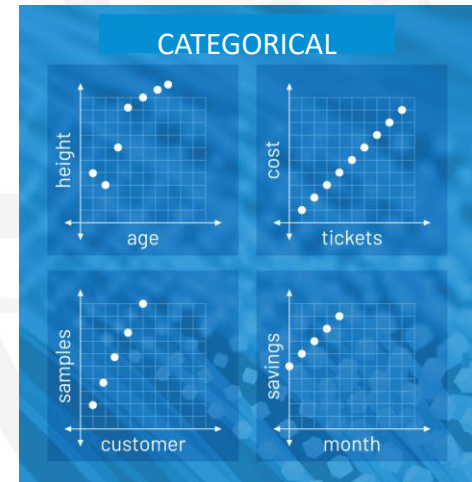# V. Insights/Deployment: What does it mean?

- ***NB*** *model for 'conversion'* → *17 features*
  - Top 3 categorical features:

- **Timing of 'sale'**
  - Later **Quarters**
  - **Weekends**
 - → *Increase yearend/weekend marketing*

- **Customer location**
  - Certain **Regions** (1,2,3,4)
  - → *Target regions in marketing*

CATEGORICAL

height / age
cost / tickets
samples / customer
savings / month

Fairfield UNIVERSITY

V. Insights/Deployment: What does it mean?

- ***NB*** *model for 'conversion'* → *17 features*
  - Top 3 continuous features → behavioral info:
    - **Exit rates**

    - **Page values**
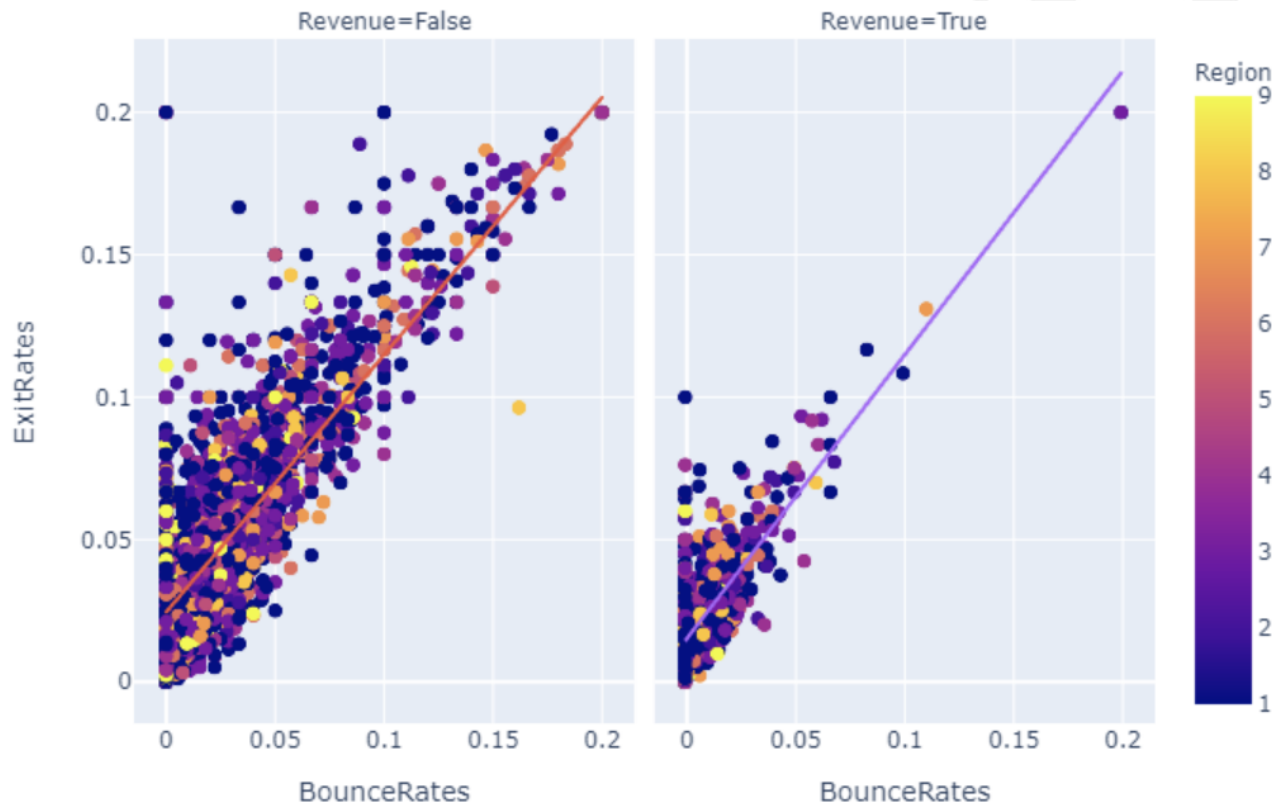
    - **Product-related pages**
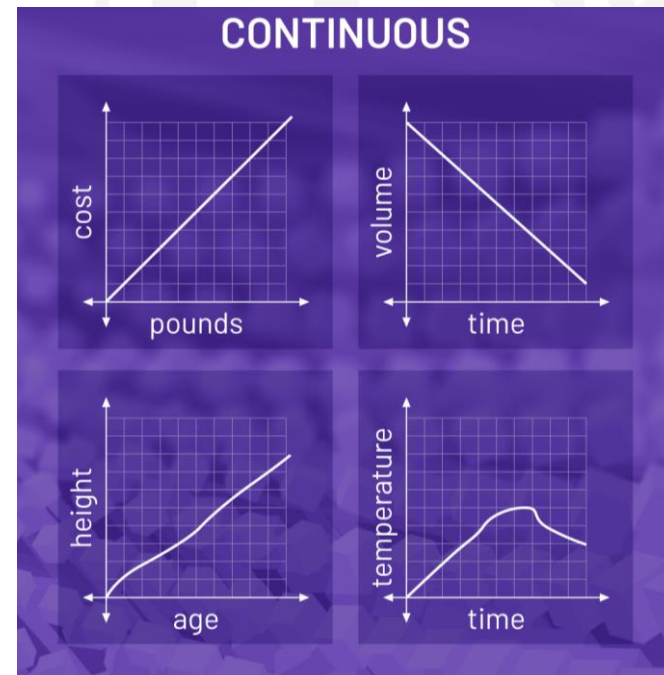
V. Insights/Deployment: What else did we find?

- ## *Exit rates:*
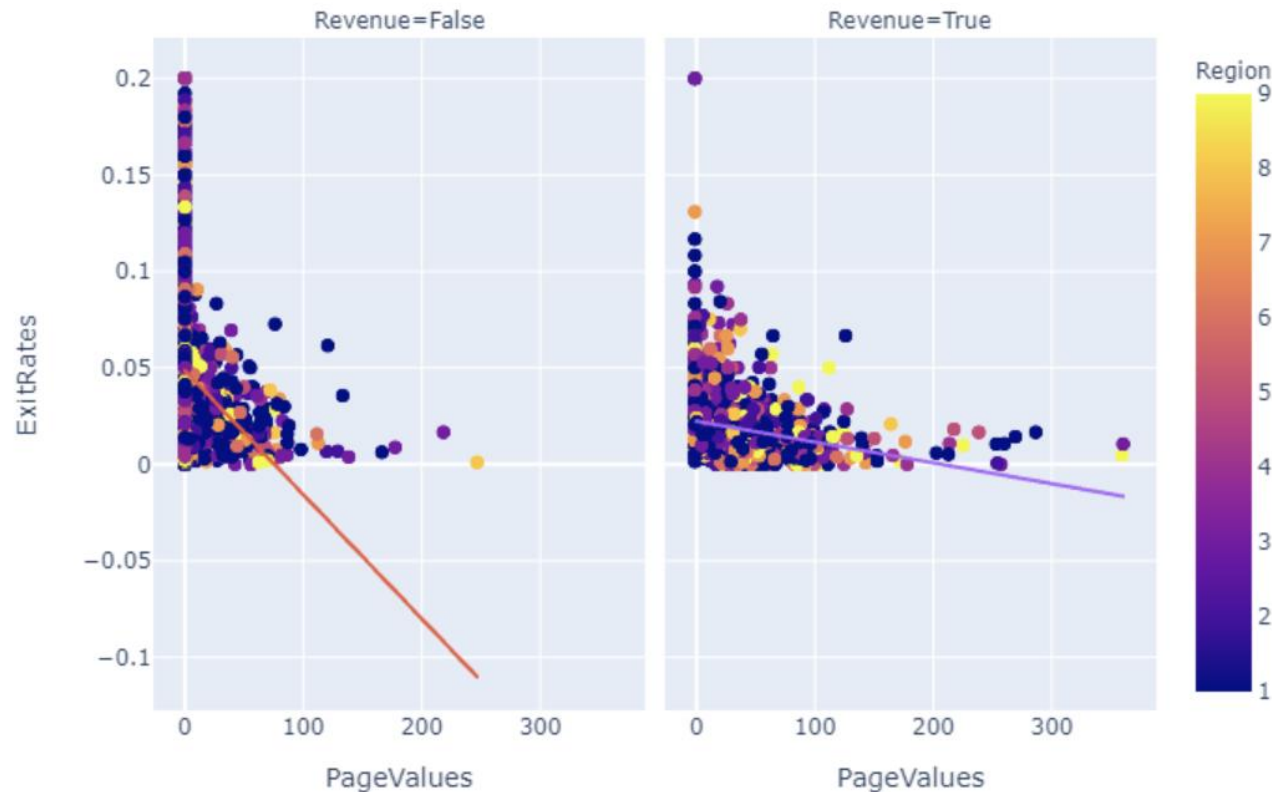 - BounceRates v ExitRates → user satisfaction & conversion*

    (In NB & XGB models)

- **Insight**: Fewer exit-prone pages for conversion
- **Recommendation:** Use exit rates to identify pages to improve

*See cxl.com (n.d.)

V. Insights/Deployment: What else did we find?

- ## *Page values:*
  - PageValues v ExitRates → conversion*

(in all 3 models)

- **Insight**: Higher page values for conversion
- **Recommendation**: Identify what influences high page values

*Source: See Gleason (2019)

V. Insights/Deployment: What else did we find?

- ***Product-related:***
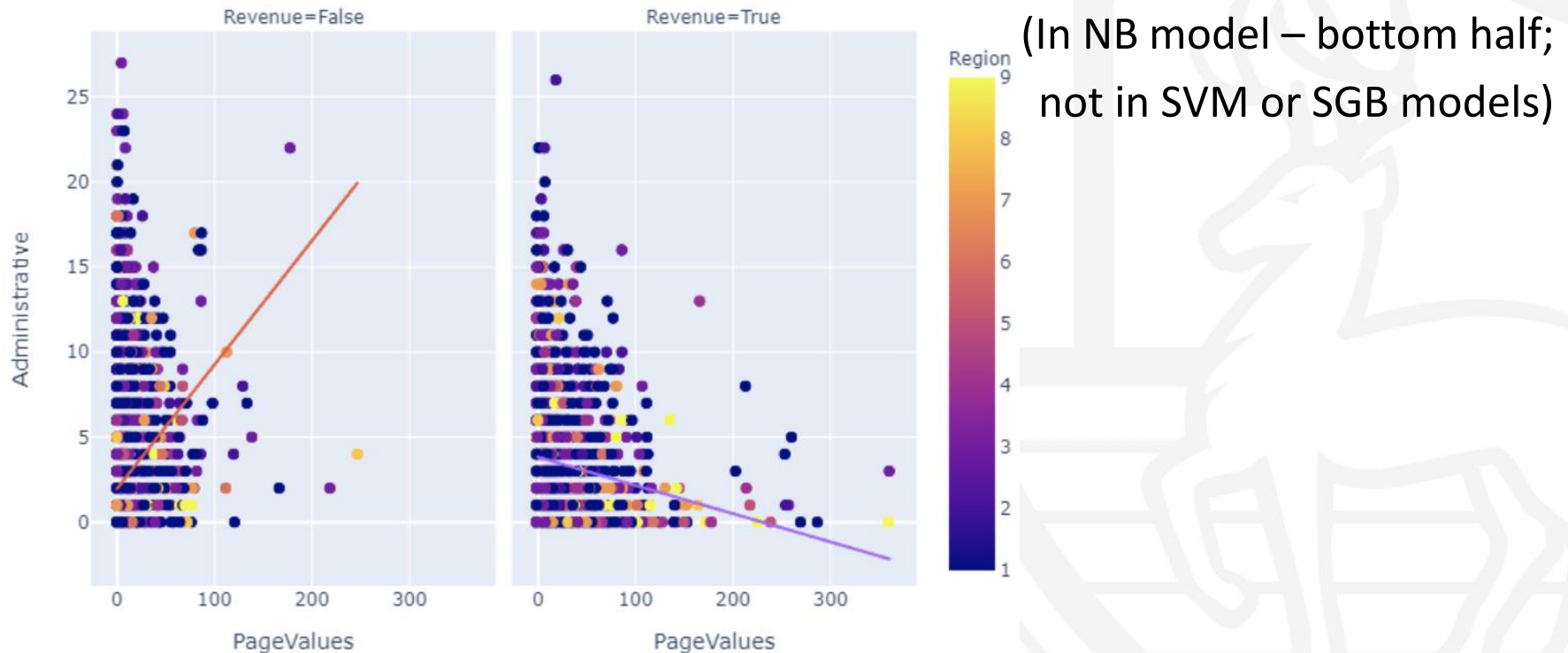 - ProductRelated → product-related web pages
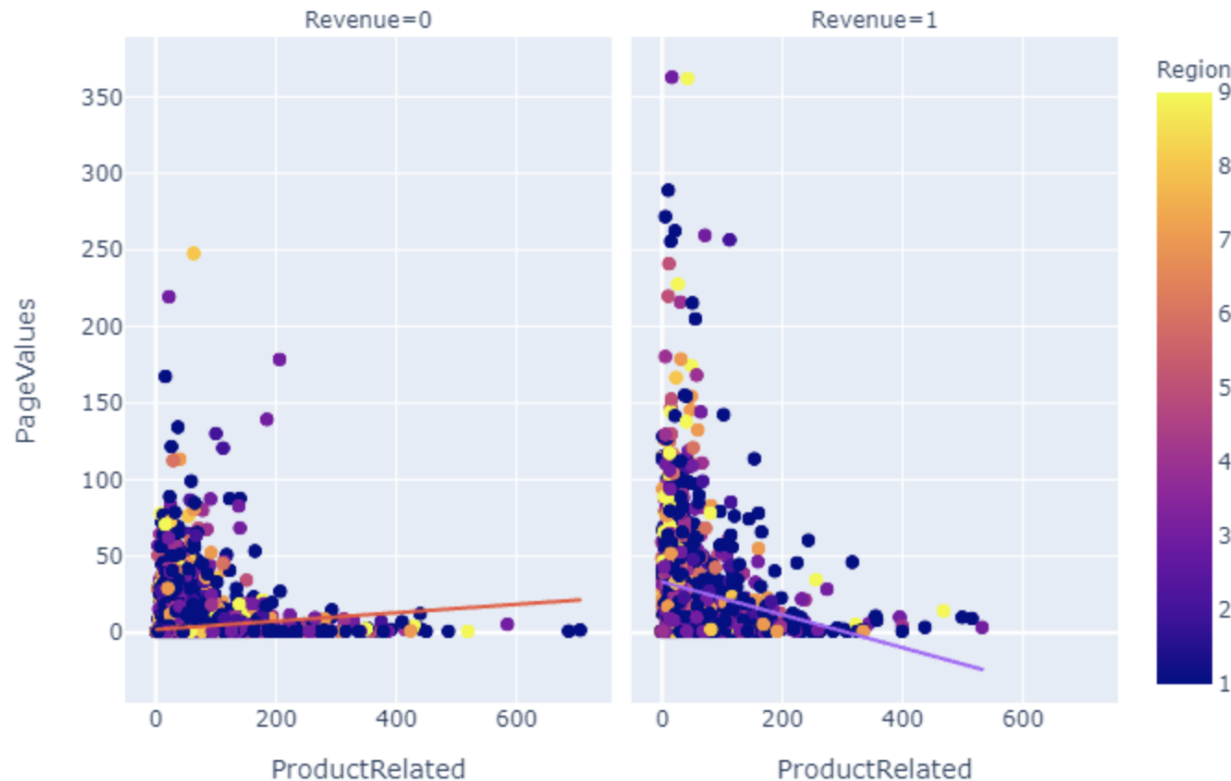
- **Insight**: Fewer product-related page views drive conversion
- **Recommendation**: Identify what reduces product-related page views

V. Insights/Deployment: What else did we find?

- ## *Administrative:*

– Admin v PageValues → conversion...*somewhat*

(In NB model – bottom half;
   not in SVM or SGB models)

– **Insight**: Fewer administrative page views may drive conversion

– **Recommendation**: Minor feature in NB, not in other models;
Be cautious in expending resources to further investigate

V. Insights/Deployment: How can we improve?

- ***Expand domain expertise/business knowledge***
  - Search engine optimization experts
  - Sportswear retail segment
  - Turkish economy/geography/culture/customs
    - Preferred sports? Seasonality?
    - Valentine's Day? Mother's Day?

V. Insights/Deployment: How can we improve?

- ***Obtain additional information/data***

  – Timing of data collection

  – Uncollected data

    - January/April data; additional year(s)

    - Other holidays in Turkey

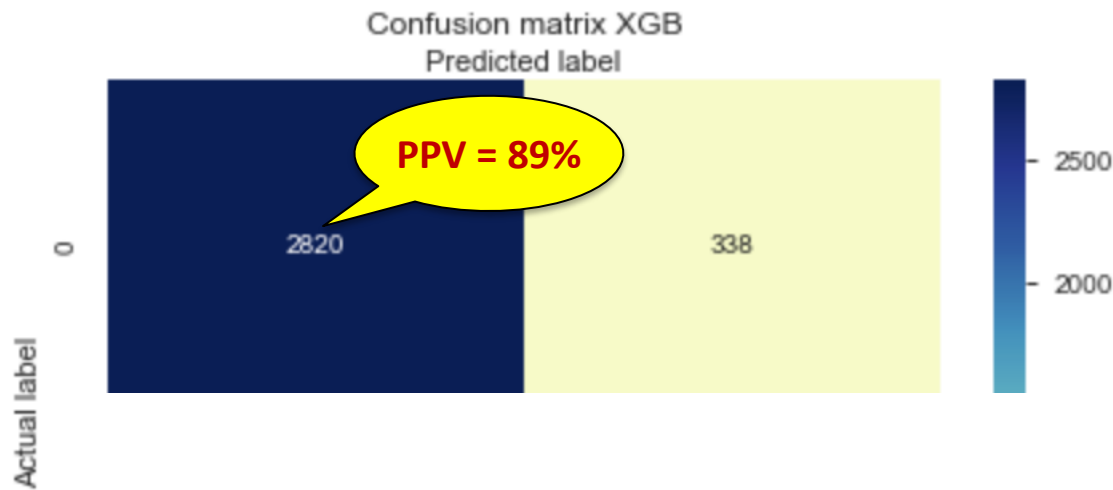   - Other unidentified features → device type?

V. Insights/Deployment: What can we study next?

- What drives online shoppers to abandon:

  - The site? the shopping cart?

  - Admin v PageValues may be important here

- What will reduce abandonment?

- Use XGB to predict just abandonment

V. Insights/Deployment: What do we recommend?

- *Use **XGBoost** model to **uniquely** predict 'abandonment'*

Confusion matrix XGB
Predicted label

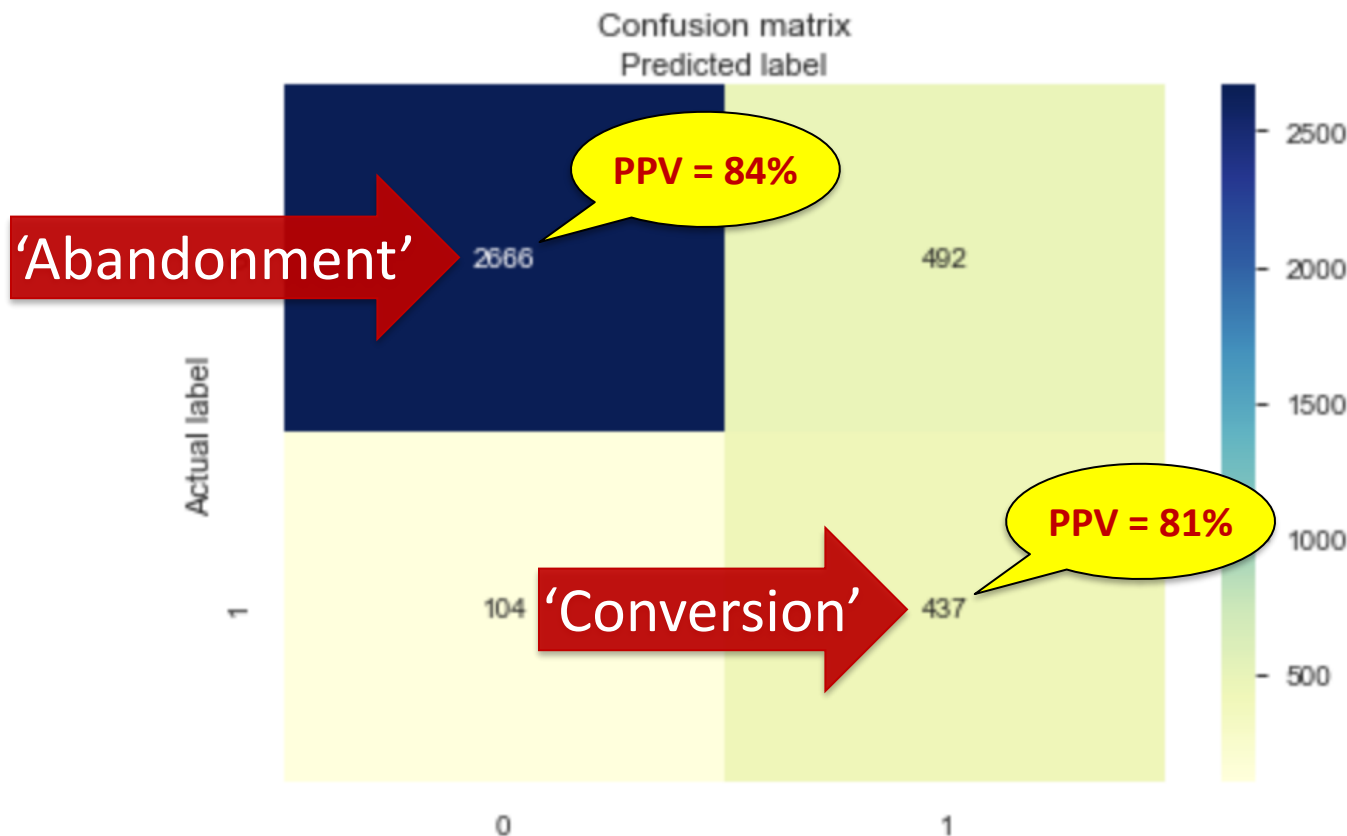PPV = 89%

2820 | 338

Actual label

0

2500

2000

XGBoost

-Best performance for 'abandonment'

V. Insights/Deployment: What do we recommend?

- *Use **Naïve Bayes** model for 'abandonment' and 'conversion'*

-Balanced performance (for 'abandonment' & 'conversion')
-Efficient use of resources

V. Insights/Deployment: What do we find?

- *Side note:*

*'**K**eep **I**t **S**uper **S**imple'*

- – *More sophisticated models*
  - ***Don't add much***
  - ***Require additional resources***
    - – *Data preparation*
    - – *Computing time*

Questions

External Sources:

- Berstein, D. 2014. Ecommerce research chart: Industry benchmark conversion rates for 25 retail categories. Marketingsherpa.com. Sept. 15. 18. Available at: https://www.marketingsherpa.com/article/chart/conversion-rates-retail-categories

- Berstein, D. 2018. Ecommerce marketing chart: Median conversion rates for 25 industries. Marketingsherpa.com. Jun. 18. Available at: https://www.marketingsherpa.com/article/chart/median-conversion-rates-25-industries

- Cxl.com. n.d. Bounce rate vs. exit rate: What's the difference? Blog post. Available at: https://cxl.com/guides/bounce-rate/bounce-rate-vs-exit-rate/

- Gleason, D. 2019. Long sales cycle? Here's how to set the value of a conversion. Cxl.com. Jan 10. Updated Oct 25. Available at: https://cxl.com/blog/analytics-long-sales-cycles/

- Taylor, E. 2014. eTail Turkey - E-Commerce Landscape. LinkedIn Slide Share Presentation. Dec. 17. Available at: https://www.slideshare.net/etailevents/etail-turkey-infographic.

- Ystats.com. 2016. Infographic: Turkey B2C E-Commerce Market. LinkedIn Slide Share Presentation. Apr. 20. Available at: https://www.slideshare.net/yStats/infographic-turkey-b2c-ecommerce-market-2016.