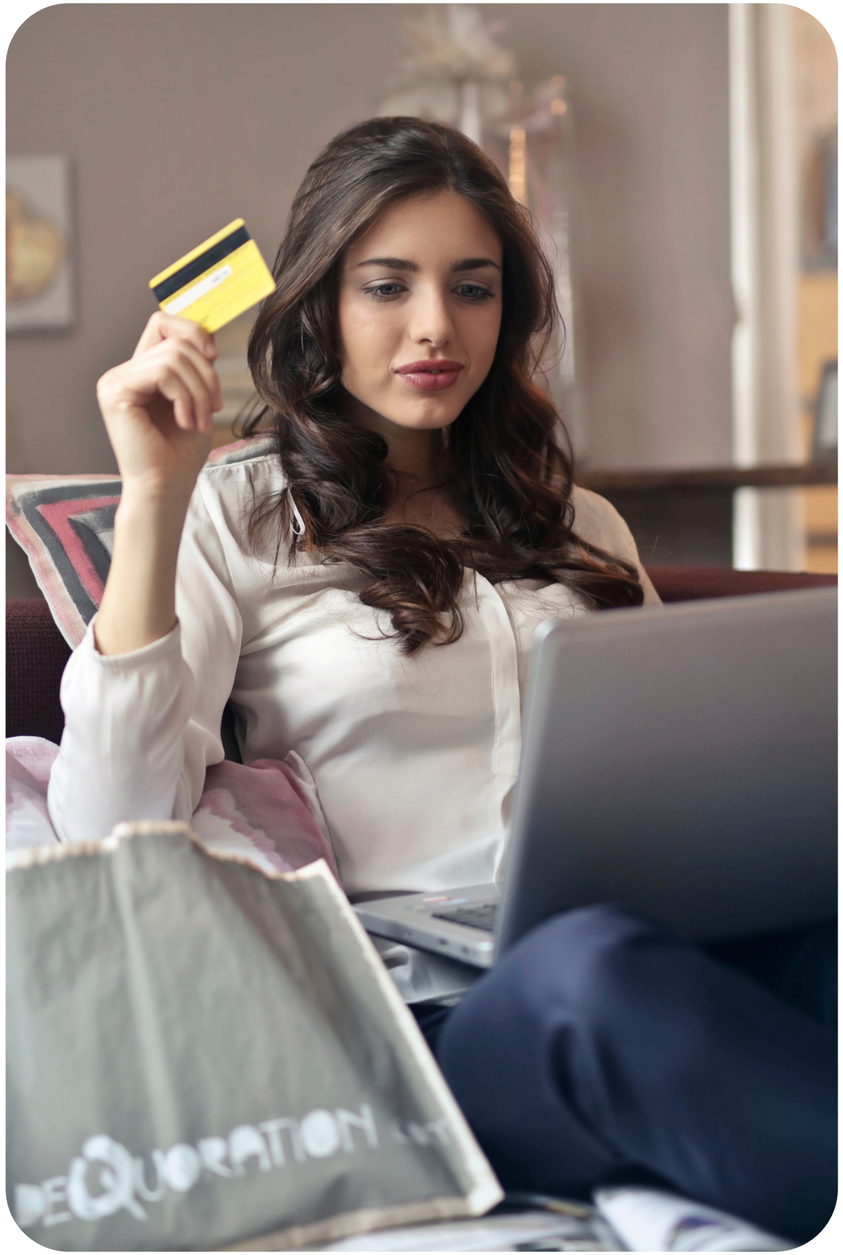
Predicting the likelihood of a retail purchase

Lisa Patel

THANKS TO MENTOR:
DIPANJAN SARKER

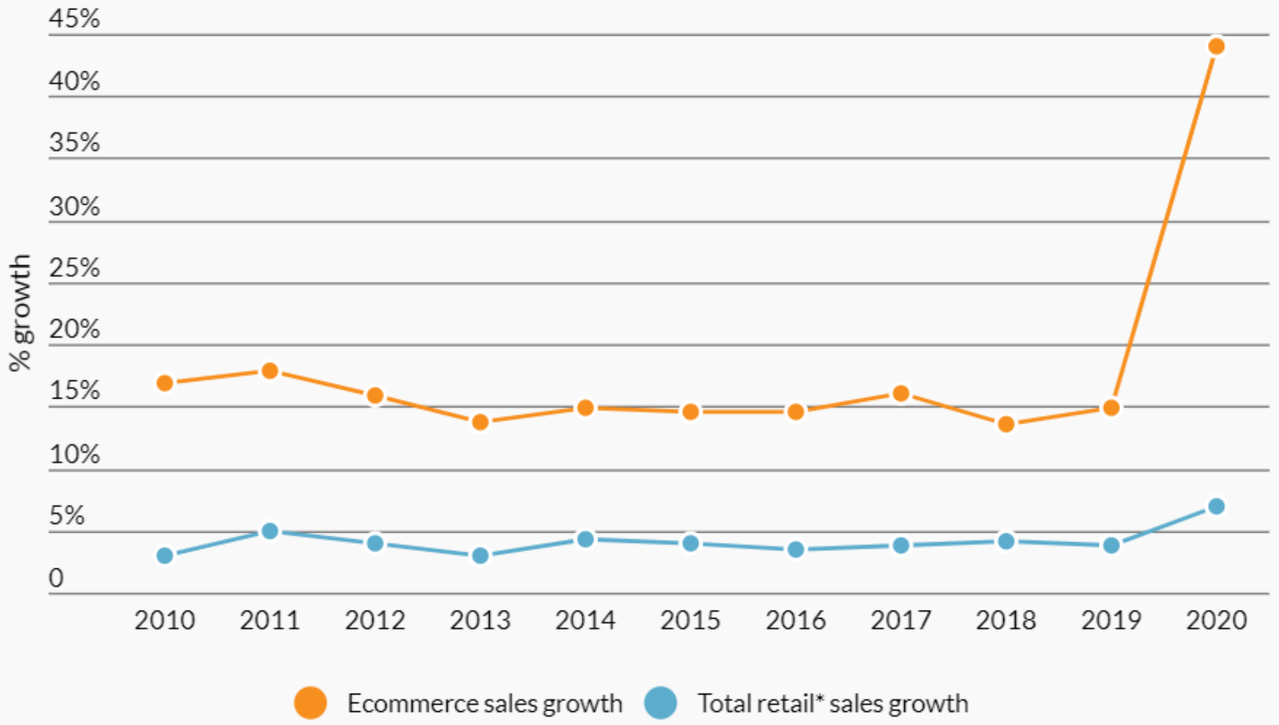


US ecommerce grows 44.0% in 2020.

The top 100 retailers (minus Amazon) had a striking 74.1% share of ecommerce growth in 2020, up significantly from a 49.4% share in 2019.

Comparing growth: US ecommerce vs. total retail* sales

Year-over-year growth, 2010-2020



Source: Digital Commerce 360, U.S. Department of Commerce; updated January 2021

*Total retail figures exclude sales of items not normally purchased online such as spending at restaurants, bars, automobile dealers, gas stations and fuel dealers

1 **Data Wrangling**

Online Shoppers Intention

UCI Machine Learning

12,330 INDIVIDUAL SESSIONS
18 FEATURES

84.5% (10422) DID NOT FINALIZE TRANSACTION

IMPORTANCE

- Web analytics - Answers how many users visit, how long they stay, how many pages they visit, what region they're from andy much more.
- Provides insights for better UX experience and potentially more revenue.
- Optimize conversions rates and measure marketing campaign effectiveness

OBJECTIVE

- Examine the real-time clickstream data to improve business/marketing decisions
- Create a machine learning model that forecasts the visitor’s likelihood of making a purchase.

Exploratory Data Analysis

```
Administrative             int64
Administrative_Duration  float64
Informational              int64
Informational_Duration   float64
ProductRelated             int64
ProductRelated_Duration  float64
BounceRates              float64
ExitRates                float64
PageValues               float64
SpecialDay               float64
Month                     object
OperatingSystems           int64
Browser                    int64
Region                     int64
TrafficType                int64
VisitorType               object
Weekend                     bool
Revenue                     bool
```

BOUNCE RATES:

The percentage of visitors who arrive at a landing page but bounce off without navigating to a second pages.

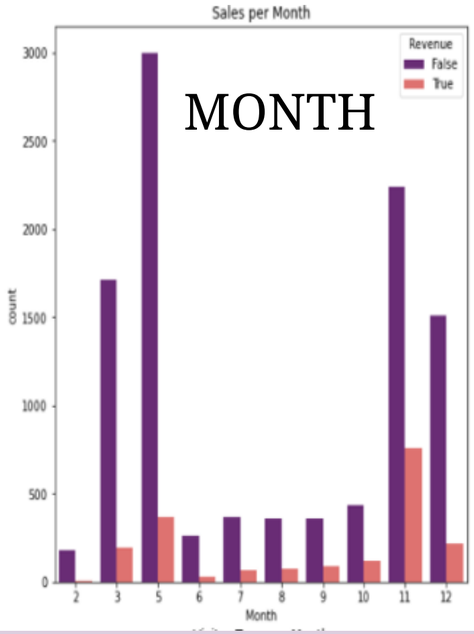
EXIT RATES

The percentage of visitors who exited the site from a particular pages, after visiting any number of pages on that site.

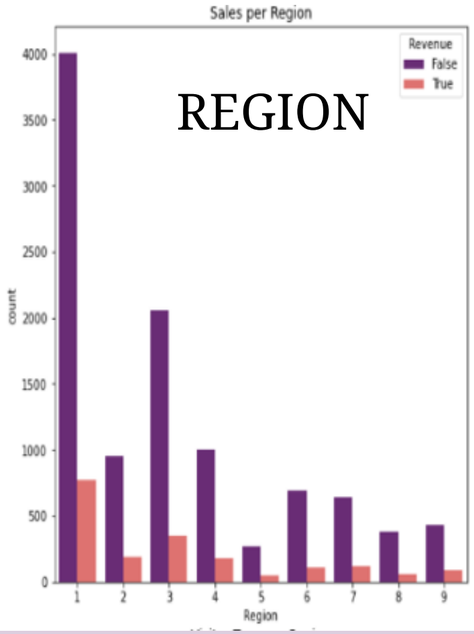
PAGE VALUES

The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction.

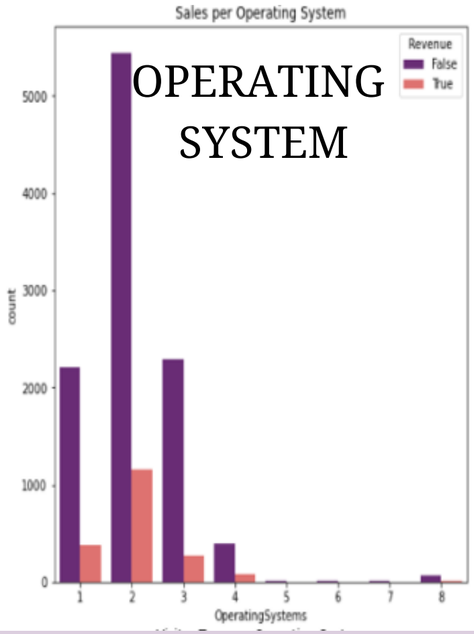
- Answers which page in your site contributed more to your site's revenue.



MONTH



REGION



OPERATING SYSTEM

SALES

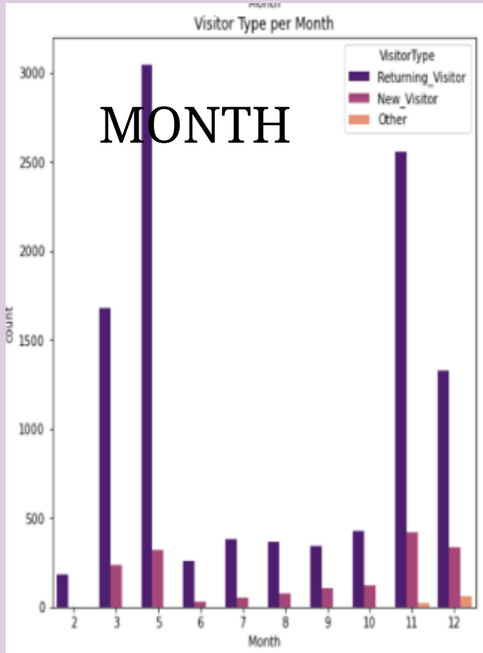
Highest transaction volume in Mar, May, Nov, Dec.
Zero sales in Feb & missing months.

Regions 1 and 3 had more transactions, but need broader regional reach.

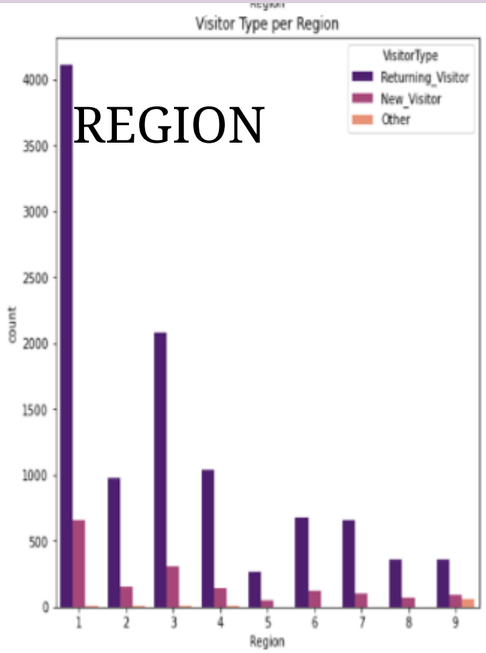
OS 2 more popular

VISITORS

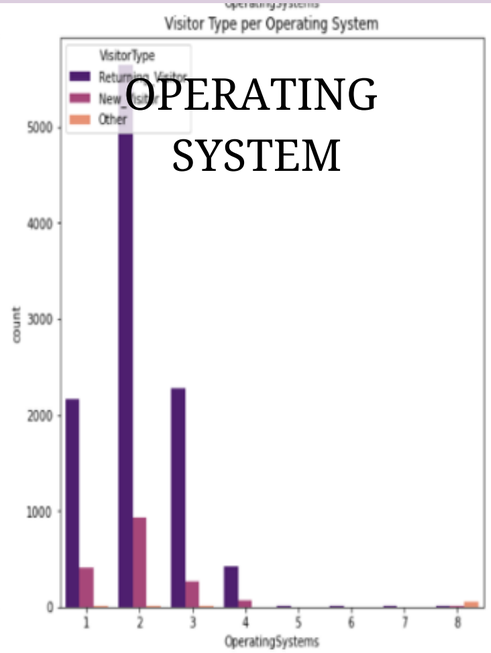
Similar to sales, but new visitors are more likely during Nov and Dec



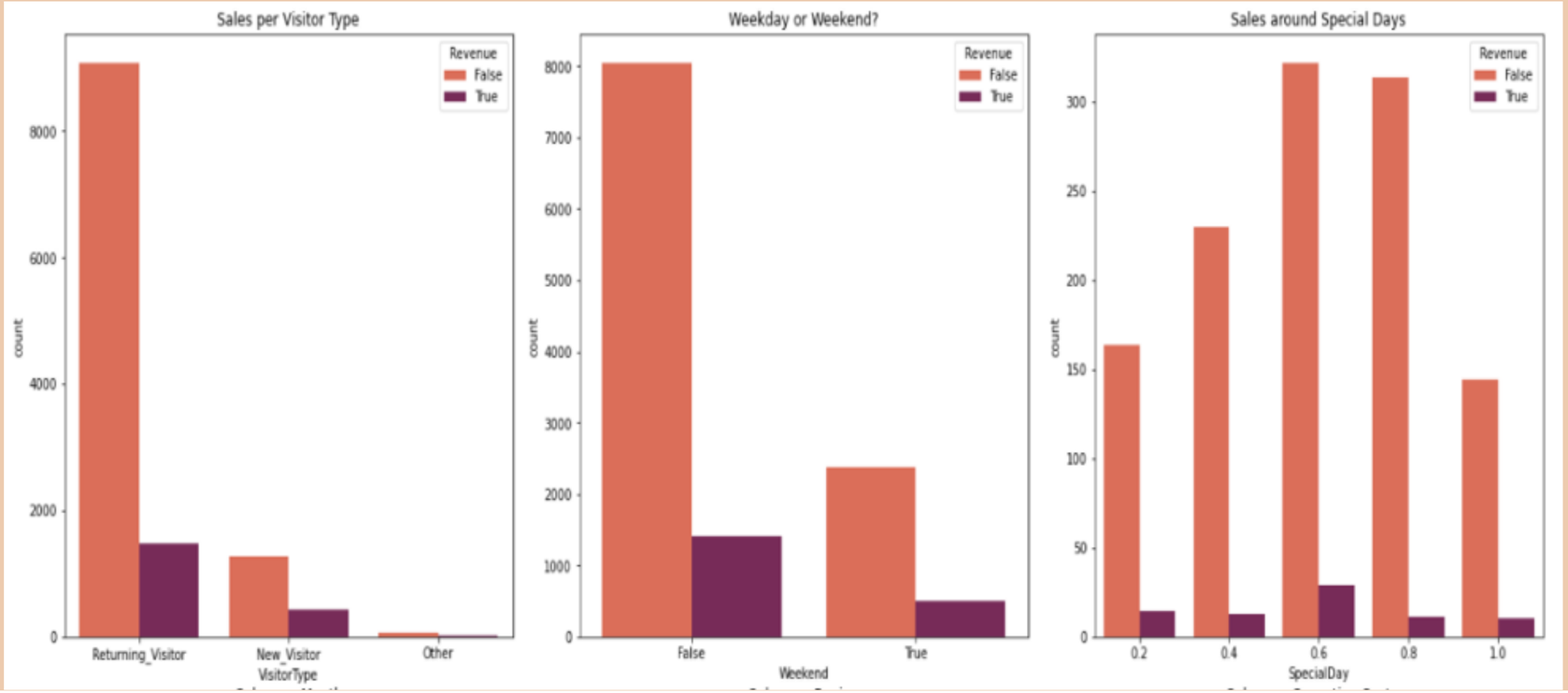
MONTH



REGION



OPERATING SYSTEM



RETURNING VISITORS MAKE MORE PURCHASES

Offer discounts to visitor's friends/family or to new visitors

WEEKDAYS ARE MORE ACTIVE

Promote a weekend sale

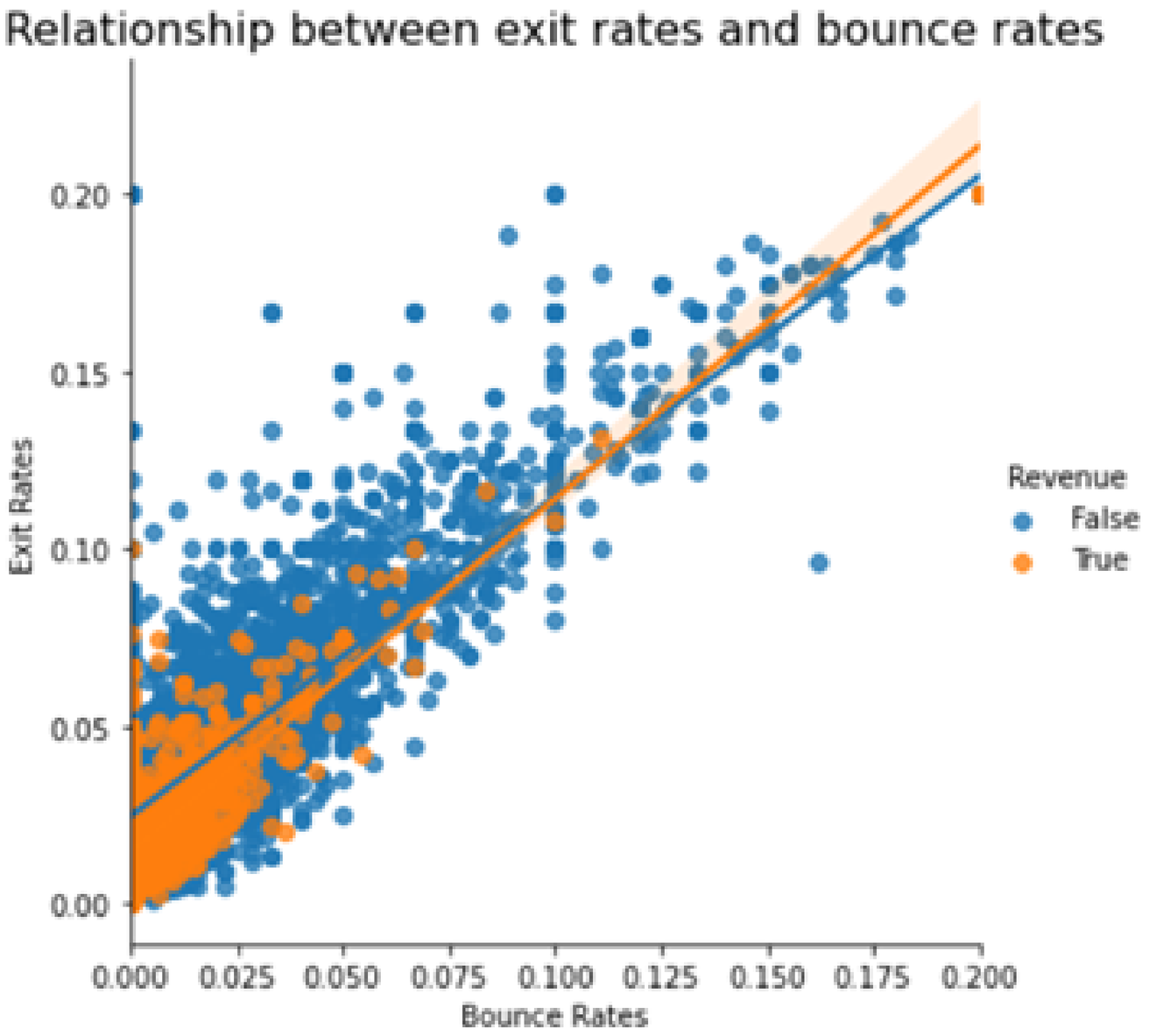
SPECIAL DAYS NEED MORE INCENTIVES

Send special gift emails around birthdays or holidays


Administrative 1 0.6 0.38 0.26 0.43 0.37 -0.22 -0.32 0.099 -0.095 -0.0063 -0.025 -0.0055 -0.034 0.026 0.14
Administrative_Duration 0.6 1 0.3 0.24 0.29 0.36 -0.14 -0.21 0.068 -0.073 -0.0073 -0.015 -0.0056 -0.014 0.015 0.094
Informational 0.38 0.3 1 0.62 0.37 0.39 -0.12 -0.16 0.049 -0.048 -0.0095 -0.038 -0.029 -0.034 0.036 0.095
Informational_Duration 0.26 0.24 0.62 1 0.28 0.35 -0.074 -0.11 0.031 -0.031 -0.0096 -0.019 -0.027 -0.025 0.024 0.07
ProductRelated 0.43 0.29 0.37 0.28 1 0.86 -0.2 -0.29 0.056 -0.024 0.0043 -0.013 -0.038 -0.043 0.016 0.16
ProductRelated_Duration 0.37 0.36 0.39 0.35 0.86 1 -0.18 -0.25 0.053 -0.036 0.003 -0.0074 -0.033 -0.036 0.0073 0.15
BounceRates -0.22 -0.14 -0.12 -0.074 -0.2 -0.18 1 0.91 -0.12 0.073 0.024 -0.016 -0.0065 0.078 -0.047 -0.15
ExitRates -0.32 -0.21 -0.16 -0.11 -0.29 -0.25 0.91 1 -0.17 0.1 0.015 -0.0044 -0.0089 0.079 -0.063 -0.21
PageValues 0.099 0.068 0.049 0.031 0.056 0.053 -0.12 -0.17 1 -0.064 0.019 0.046 0.011 0.013 0.012 0.49
SpecialDay -0.095 -0.073 -0.048 -0.031 -0.024 -0.036 0.073 0.1 -0.064 1 0.013 0.0035 -0.016 0.052 -0.017 -0.082
OperatingSystems -0.0063 -0.0073 -0.0095 -0.0096 0.0043 0.003 0.024 0.015 0.019 0.013 1 0.22 0.077 0.19 0.00028 -0.015
Browser -0.025 -0.015 -0.038 -0.019 -0.013 -0.0074 -0.016 -0.0044 0.046 0.0035 0.22 1 0.097 0.11 -0.04 0.024
Region -0.0055 -0.0056 -0.029 -0.027 -0.038 -0.033 -0.0065 -0.0089 0.011 -0.016 0.077 0.097 1 0.048 -0.00069 -0.012
TrafficType -0.034 -0.014 -0.034 -0.025 -0.043 -0.036 0.078 0.079 0.013 0.052 0.19 0.11 0.048 1 -0.0022 -0.0051
Weekend 0.026 0.015 0.036 0.024 0.016 0.0073 -0.047 -0.063 0.012 -0.017 0.00028 -0.04 -0.00069 -0.0022 1 0.029
Revenue 0.14 0.094 0.095 0.07 0.16 0.15 -0.15 -0.21 0.49 -0.082 -0.015 0.024 -0.012 -0.0051 0.029 1
Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
OperatingSystems
Browser
Region
TrafficType
Weekend
Revenue
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

For ecommerce and retail websites, a bounce rate between 30-50% is acceptable. [2]

Less than 5% indicates a google analytics code inserted twice. [3]



Modeling Overview

TYPE

Supervised Learning | Binary Classification

IMBALANCED DATA

84.5% (10422) did not make a purchase. Will affect results

TOOLS

Python's Scikit learn and imblearn

PRE-PROCESSING

One-hot Encoding: Change necessary categorical to numerical

Data Splitting into 75% training/ 25% testing sets

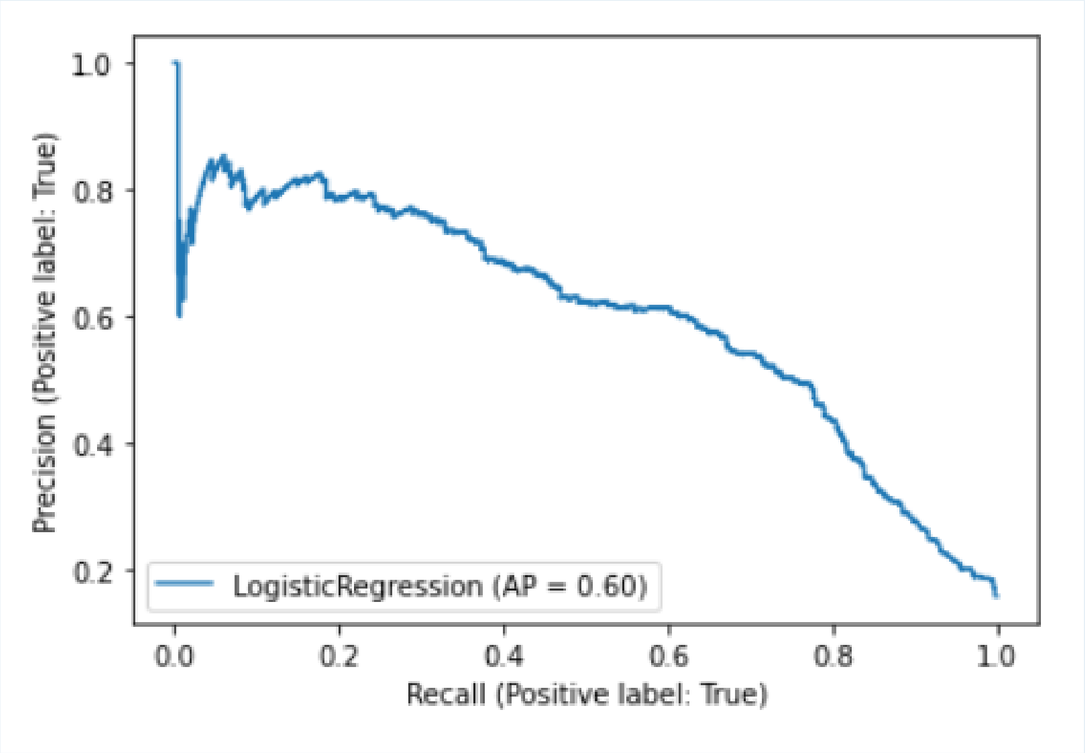
Standard Scaling

Model Evaluation

Accuracy is not reliable for imbalanced data

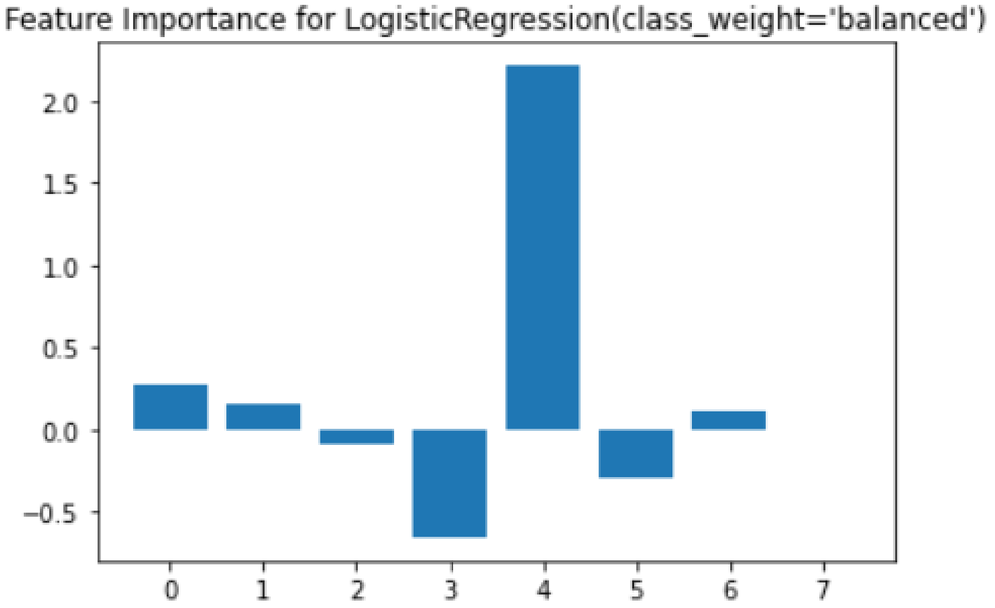
Use F1 for a balance between Precision and Recall

	Logistic Regression	Balanced Logistic Regression	Gaussian Naïve Bayes	Optimized Random Forest
Precision	0.73	0.54	0.37	0.73
Recall	0.36	0.68	0.69	0.51
F1	0.48	0.61	0.48	0.60
TP/TN	170/2544	326/2332	327/2053	244/2514
FP/FN	307/62	151/274	150/553	233/92

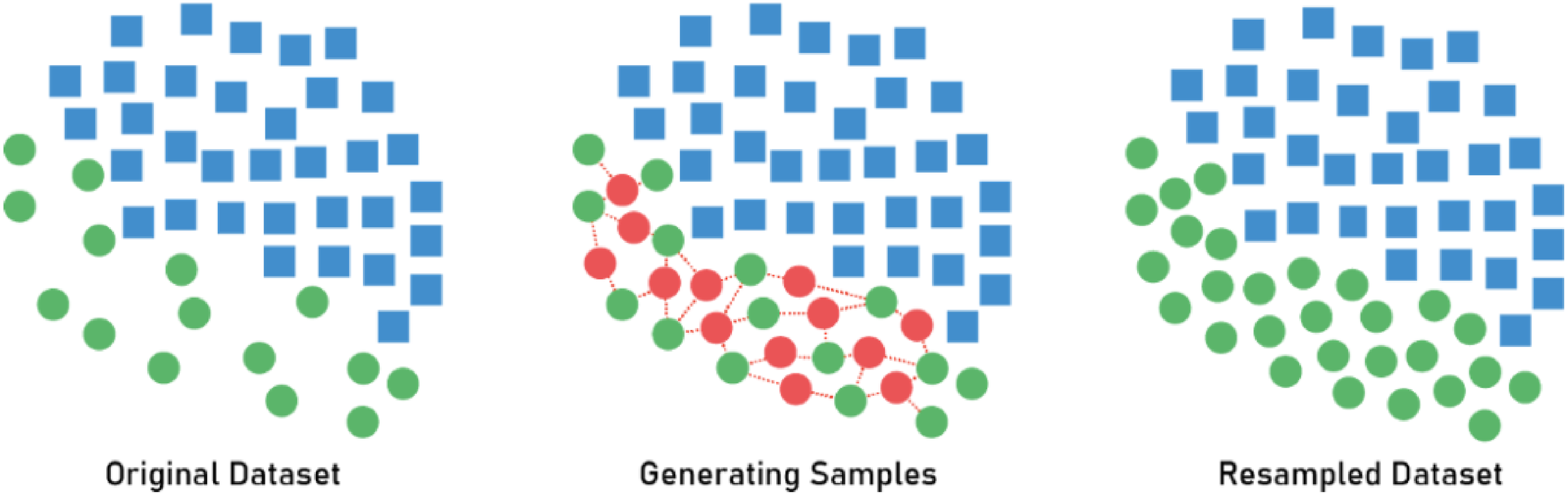


AP summarizes a precision-recall curve as the weighted mean of precision achieved across all threshold. Can be used to calculate AUC-PR

```
Feature: Administrative Score: 0.2728776005044198
Feature: Administrative_Duration Score: 0.14947177359687086
Feature: Informational Score: -0.09041231871566069
Feature: Informational_Duration Score: -0.6550851112489039
Feature: ProductRelated Score: 2.21229590893485
Feature: ProductRelated_Duration Score: -0.28649483723607594
Feature: BounceRates Score: 0.11649145058676816
Feature: ExitRates Score: -0.006448377088902264
```



Synthetic Minority Oversampling Technique



One way to solve this problem of too few examples of the minority class is to oversample the minority examples using synthetic, duplicated samples. These new examples will be close to existing examples in the feature space, but different in small but random ways.

Improvements | Future Work

FIX BOUNCE RATE

Too low bounce rate is a cause for concern and affects interpretation of data

OPTIMIZE FURTHER W/ MORE HYPERPARAMETER TUNING

Find out if the models improve with a more rigorous tuning on all models through random grid search

TRY OTHER RESAMPLING METHODS

Undersampling or combining it with oversampling might yield interesting results

EXTRACT MORE SPECIFIC INFO ON FEATURES

Useful to make additional detailed business-related recommendations.

CONCLUSION

- There is a room for improvement in attracting and retaining visitors, which can be accomplished through various discounts or online marketing strategies.

- Out of 5 supervised classification mode, Logistic Regression with a class balance parameters gave the best results.

- Precision = 0.54 | Recall = 0.68 | F1 = 0.61 | AP = 0.60

Thank you!

Do you have any questions?

References

- https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
- https://www.bigcommerce.com/blog/bounce-rates/
- https://upsidebusiness.com/blog/my-website-bounce-rate-is-5-is-that-too-low-is-it-too-good-to-be-true/#:~:text=A%20very%20low%20Bounce%20Rate,%2C%20don't%20start%20celebrating.

Lisa Patel

Email: Patel_Lisa @ou.edu

Github: lisavisa14

LinkedIn: www.linkedin.com/in/lisa-patel14