Online Shopper's Intention

BANA 273 Machine Learning

Group 4: Chia-Jo Chen, Dong Wook Kim, Edward Shih-Yu Chung,
Kirti Swapnil Bhalgat, Marielle Dela Cruz

Agenda

1. Introduction

1. Data Preparation

1. Data Visualization

1. Classification

1. Clustering

1. Conclusion

1. Introduction

Objectives

“Black Friday 2020 – Online shopping surges 22% to record $9 billion” (CNBC)

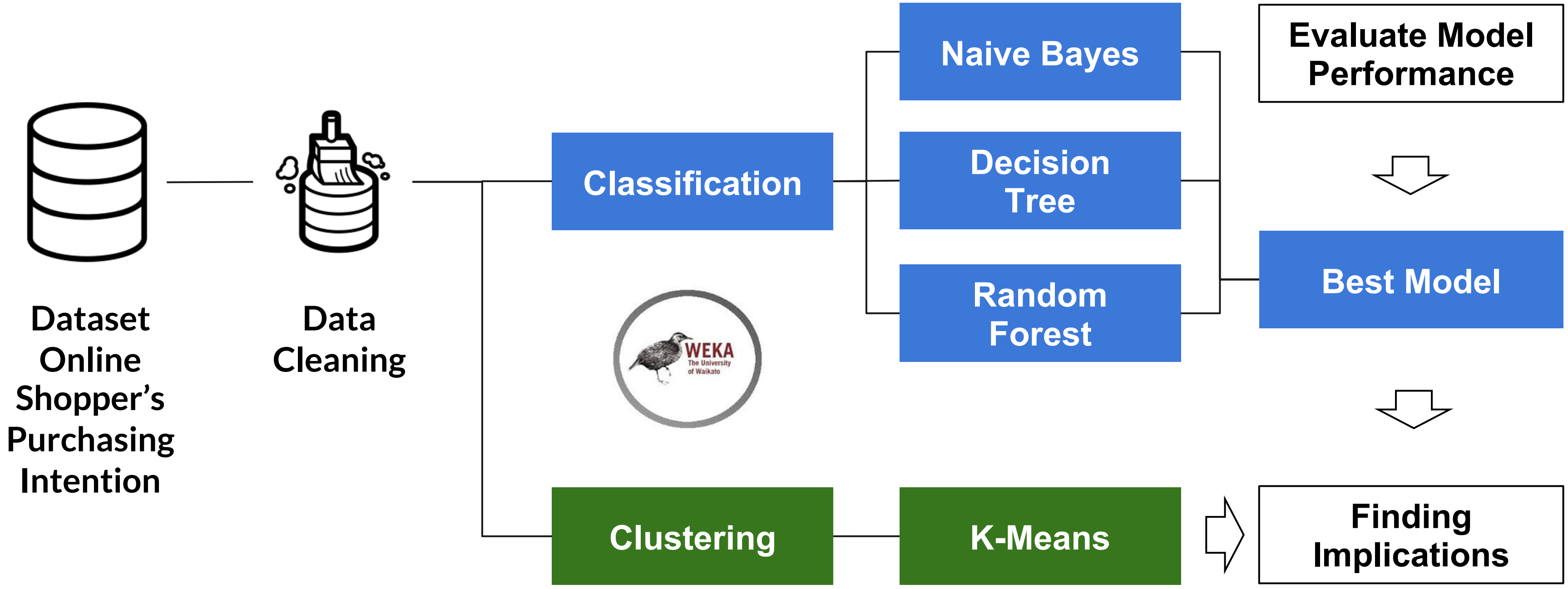
- Online retail consumer behavior has become crucial in analyzing factors that influence online buying behavior to find ways to increase company revenue

- Address the following questions:

- *What factors affect conversion and customer intention?*
- *How to effectively increase company revenue?*

1. Introduction

Project Process



2. Data Preparation

Dataset

- **Data Source: Online Shoppers Purchasing Intention Dataset**

  (UCI Machine Learning Repository by Sakar et al.)

- **Data Structure**

- **12,330 Rows 18 columns**

  (10 numerical and 8 categorical attributes)

- Class label: 'Revenue'

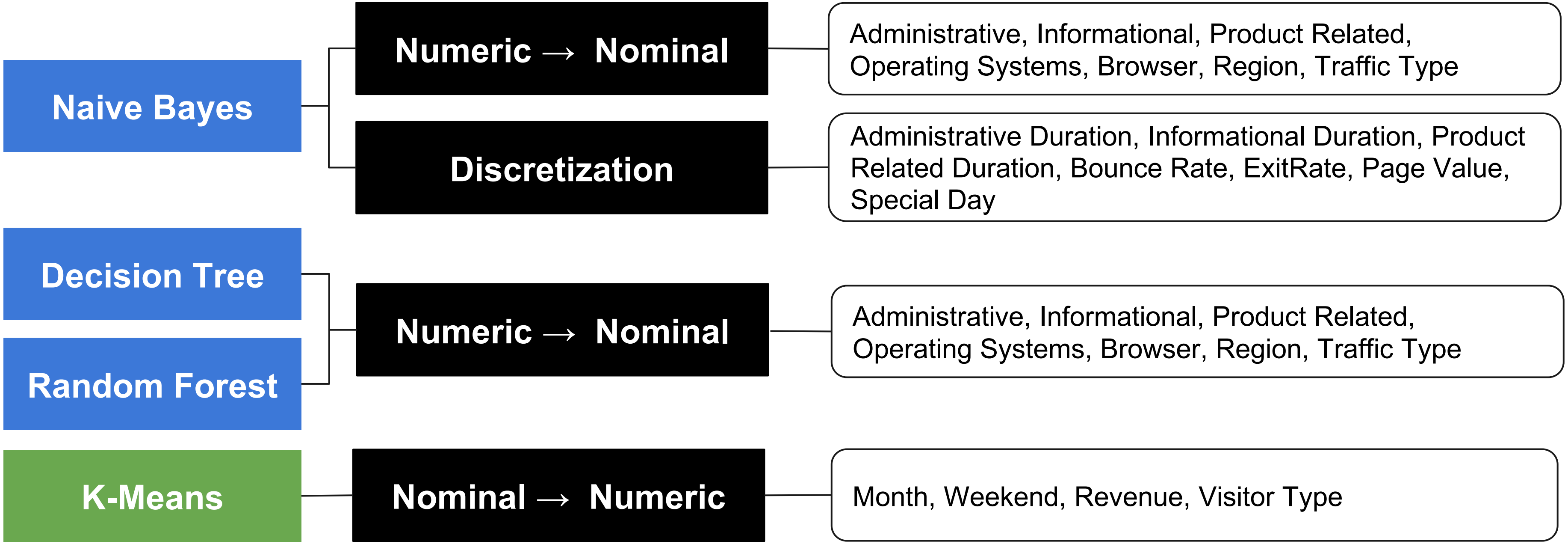
  (Desired target)

Column	**Descriptions**
Bounce Rate	% of visitors who enter the site from that page and then leave without triggering any other requests
Exit Rate	% of visitors that were that the last in the session
Page Value	Average value for a web page that a user visited before completing an e-commerce transaction
Special Day	the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day)
Weekend	A Boolean value indicating whether the date of the visit is weekend
....	....

2. Data Preparation

Data Preprocessing

- Adjust variables to correct data type



Naive Bayes

Numeric → Nominal

Administrative, Informational, Product Related, Operating Systems, Browser, Region, Traffic Type

Discretization

Administrative Duration, Informational Duration, Product Related Duration, Bounce Rate, ExitRate, Page Value, Special Day

Decision Tree

Random Forest

Numeric → Nominal

Administrative, Informational, Product Related, Operating Systems, Browser, Region, Traffic Type

K-Means

Nominal → Numeric

Month, Weekend, Revenue, Visitor Type

3. Data Visualization

Categorical features

Fig. 1a

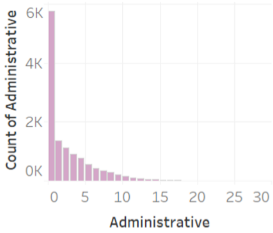


Fig. 1b

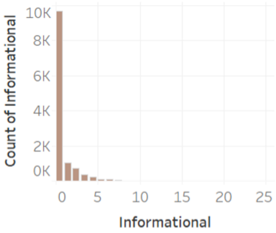
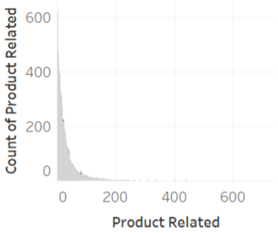


Fig. 1c



- No Null and Missing Values
- Month: only ten months(no Jan. and Apr.)
- Imbalance in Class variable(Revenue)

Fig. 1d

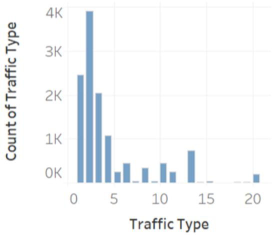


Fig. 1e

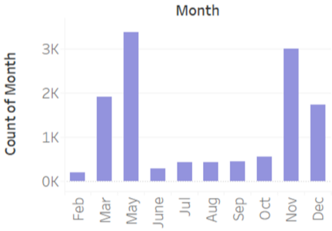
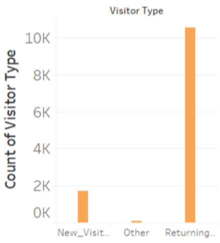


Fig. 1f



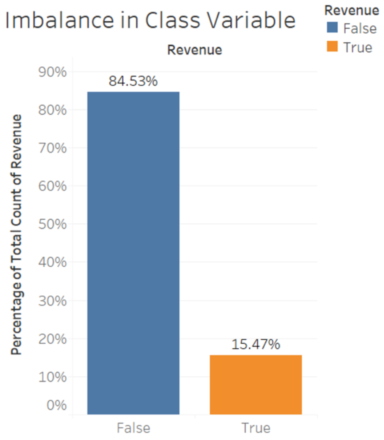


Fig. 1g

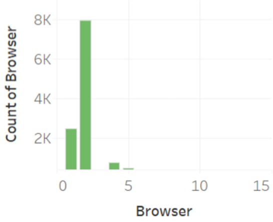


Fig. 1h

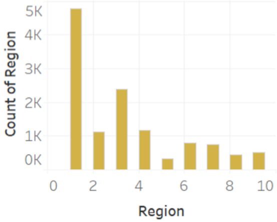
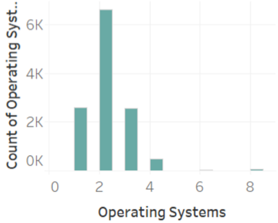


Fig. 1i

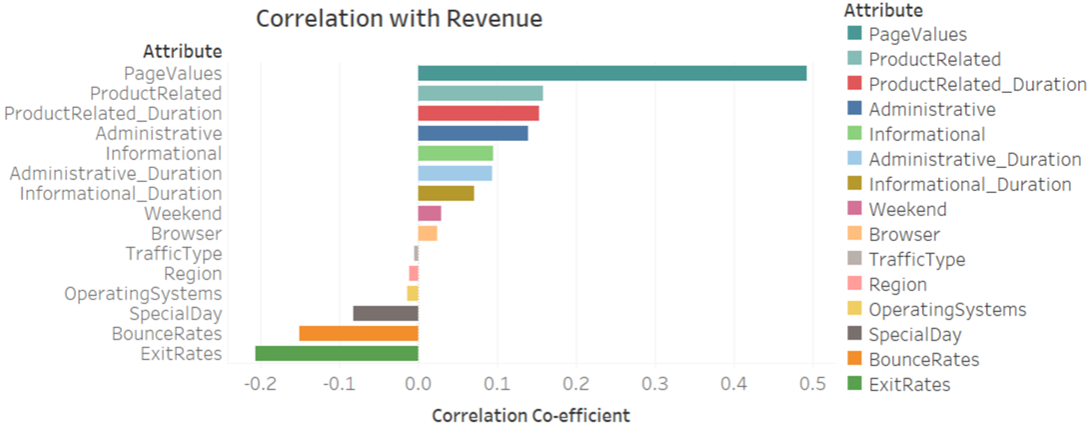


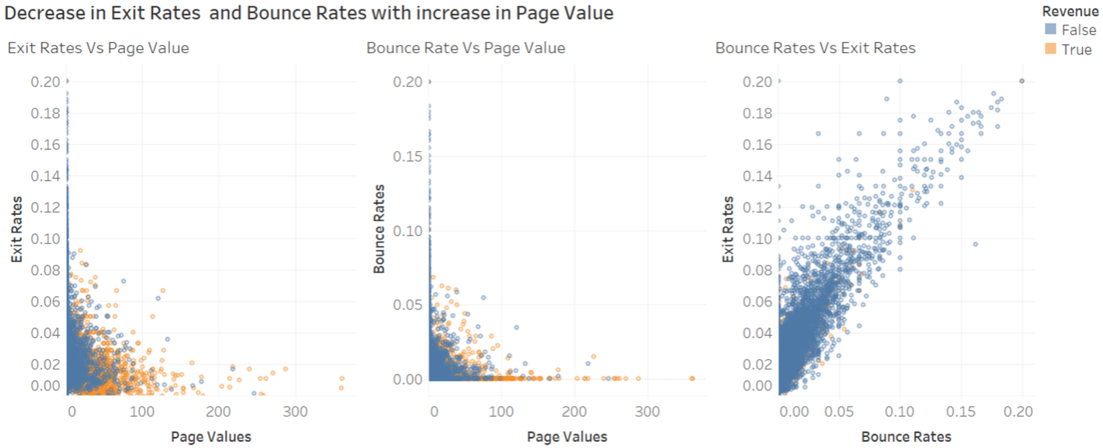
3. Data Visualization

Correlation

- Page values: Highest positive correlation with Revenue

- Bounce Rate/Exit Rate: Highly correlated with each other. Negative correlation with revenue.





4. Classification

Evaluating Performance

1 **Naive Bayes**

-Best Accuracy: 89.6%

2 **Decision Tree**

-Best Accuracy: 91.05%

3 **Random Forest**

-Best Accuracy: 96.14%

4. Classification

Naive Bayes

Without Attribute Selection

Method	**Accuracy**	**ROC area**
Naive Bayes (Non-resampling)	83.97%	0.87
Naive Bayes (Resampling)	79.58%	0.88

With Attribute Selection

Method	**Accuracy**	**ROC area**
Naive Bayes (Wrapper & Non-resampling)	**89.60%**	**0.91**
Naive Bayes (Wrapper & Resampling)	84.74%	0.92

Observation:

- The accuracy decreased significantly after resampling.
- Wrapper provided the better prediction; however, the weighted false positive rate was higher as well.

4. Classification

Decision Tree

Without Attribute Selection

Method	**Accuracy**	**ROC area**
Decision Tree (Non-resampling)	89.11 %	0.85
Decision Tree (Resampling)	89.35 %	0.92

With Attribute Selection

Method	**Accuracy**	**ROC area**
Decision Tree (Resampling and attribute selection)	**91.05%**	**0.93**

Observations:

- Easy to understand and interpret
- Better predictions than Naive Bayes model
- Resampling does not increase accuracy much, but ROC area significantly improved from 0.85 to 0.92

4. Classification

Random Forest

Random Forest

Method	Accuracy	ROC area
Random Forest (Non-resampling)	88.15%	0.908
Random Forest (Resampling)	**96.14%**	**0.996**

Boosting

Method	Accuracy	ROC area
Boosting (Non-resampling)	86.99%	0.868
Boosting (Resampling)	95.68%	0.985

Observations:

- RF achieves diversity by Bootstrap samples and random selection of attributes
- RF gave best results for this dataset with significant improvement over decision trees
- Without resampling the accuracy for boosting is quite low

5. Clustering

K-Means

Without PCA

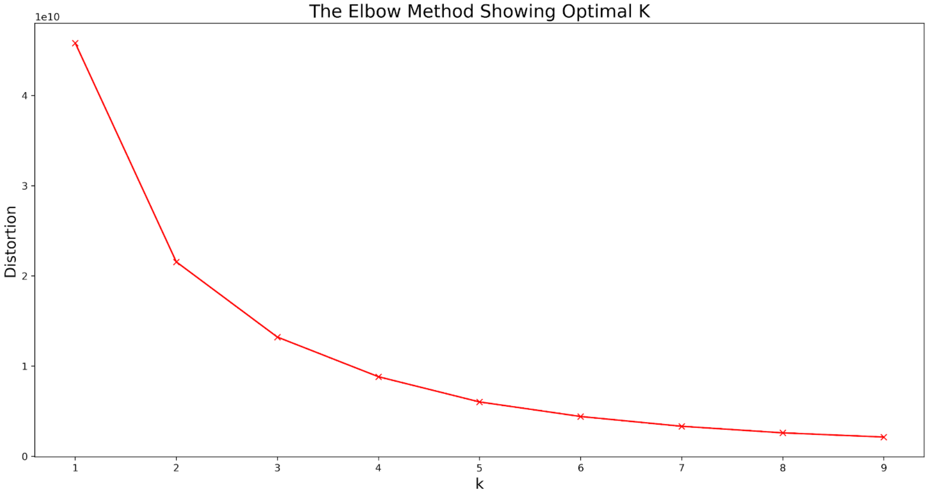
	SSE	Cluster distribution
No ClassBalancer	8089.45	0 = 20%, 1 = 41%, 2 = 39%
With ClassBalancer	1495.95	0 = 19%, 1 = 15%,3 = 65%

With PCA

# of Features	SSE	Cluster distribution
14	943.05	0 = 41%, 1 = 15%, 2 = 44%
10	673.67	0 = 9%, 1 = 15%, 2 = 76%

Characteristics of customers that make a purchase:

- Spends a lot of time on the website
- View more pages
- Have the lowest exit and bounce rates
- More likely to make purchases on the weekends



6. Conclusion

Implications

Machine Learning Models show that

- Classification: Most important attribute is ‘**Page Value**’
- Clustering: **Characteristics of cluster** who make a purchase : spend more time browsing the website, have the lowest bounce rates.

Business Recommendations:

- **Optimization of Landing pages** and creating user friendly interface.
- **Personalized targeted emails and loyalty programs** for returning customers to increase sales and revenue.

Thank You

Your feedback matters