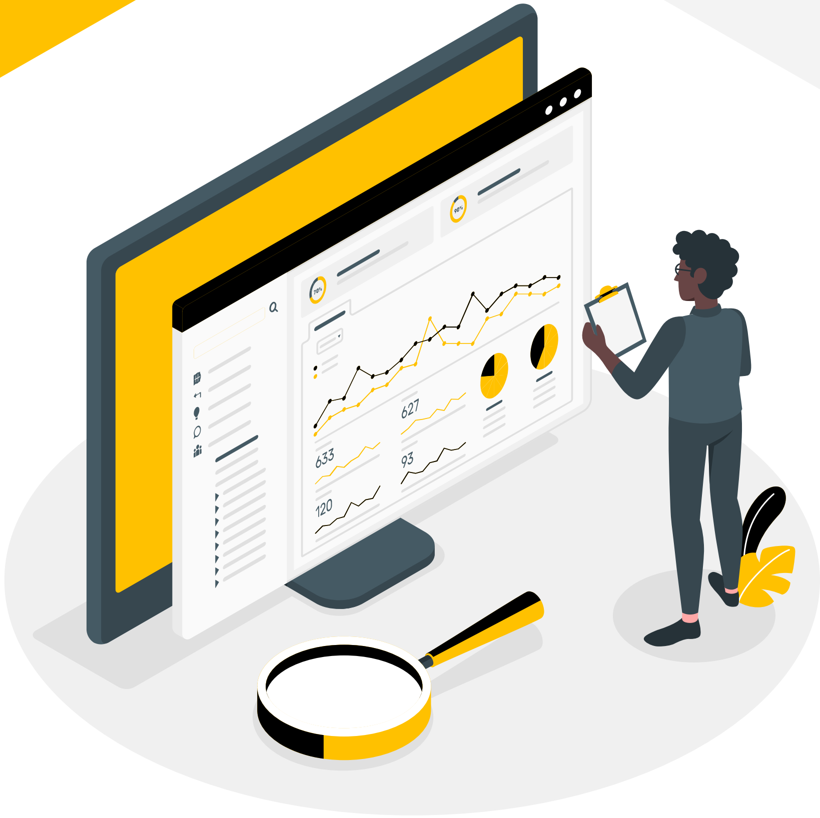
Online Shopper Intention



By Thompson Pham

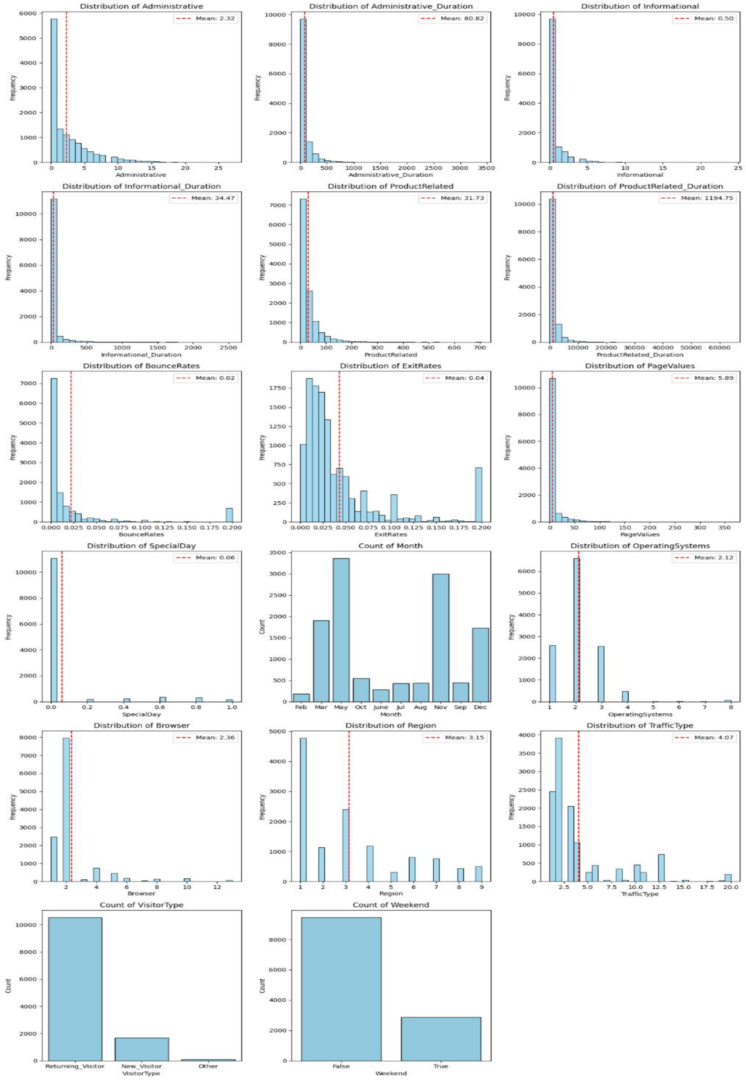
Introduction & Background

01

- E-commerce - buying and selling of goods and services over the internet.
- Rapid growth of online shopping
- Understanding consumer behavior and predicting purchasing patterns importance

Distribution

- Predominate left skew of most attributes
- Reflects the majority of classification of negative instance



Distribution of Administrative

Distribution of Administrative_Duration

Distribution of Informational

Distribution of Informational_Duration

Distribution of ProductRelated

Distribution of ProductRelated_Duration

Distribution of BounceRates

Distribution of ExitRates

Distribution of PageValues

Distribution of SpecialDay

Count of Month

Distribution of OperatingSystems

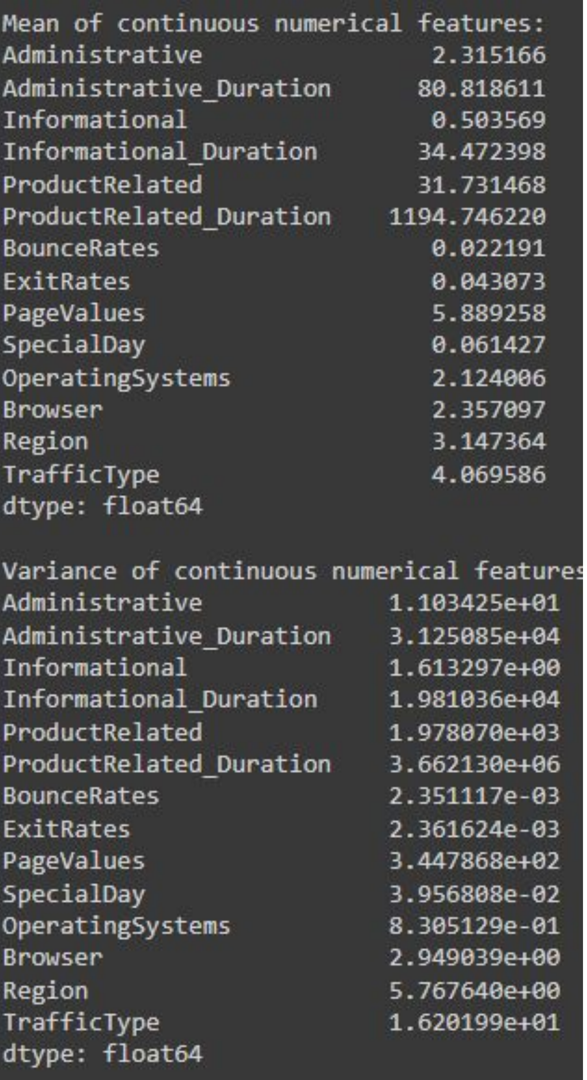
Distribution of Browser

Distribution of Region

Distribution of TrafficType

Count of VisitorType

Count of Weekend



```
Mean of continuous numerical features:
Administrative                2.315166
Administrative_Duration      80.818611
Informational                 0.503569
Informational_Duration       34.472398
ProductRelated               31.731468
ProductRelated_Duration    1194.746220
BounceRates                   0.022191
ExitRates                     0.043073
PageValues                    5.889258
SpecialDay                    0.061427
OperatingSystems              2.124006
Browser                       2.357097
Region                        3.147364
TrafficType                   4.069586
dtype: float64

Variance of continuous numerical features
Administrative             1.103425e+01
Administrative_Duration    3.125085e+04
Informational              1.613297e+00
Informational_Duration     1.981036e+04
ProductRelated             1.978070e+03
ProductRelated_Duration    3.662130e+06
BounceRates                2.351117e-03
ExitRates                  2.361624e-03
PageValues                 3.447868e+02
SpecialDay                 3.956808e-02
OperatingSystems           8.305129e-01
Browser                    2.949039e+00
Region                     5.767640e+00
TrafficType                1.620199e+01
dtype: float64
```

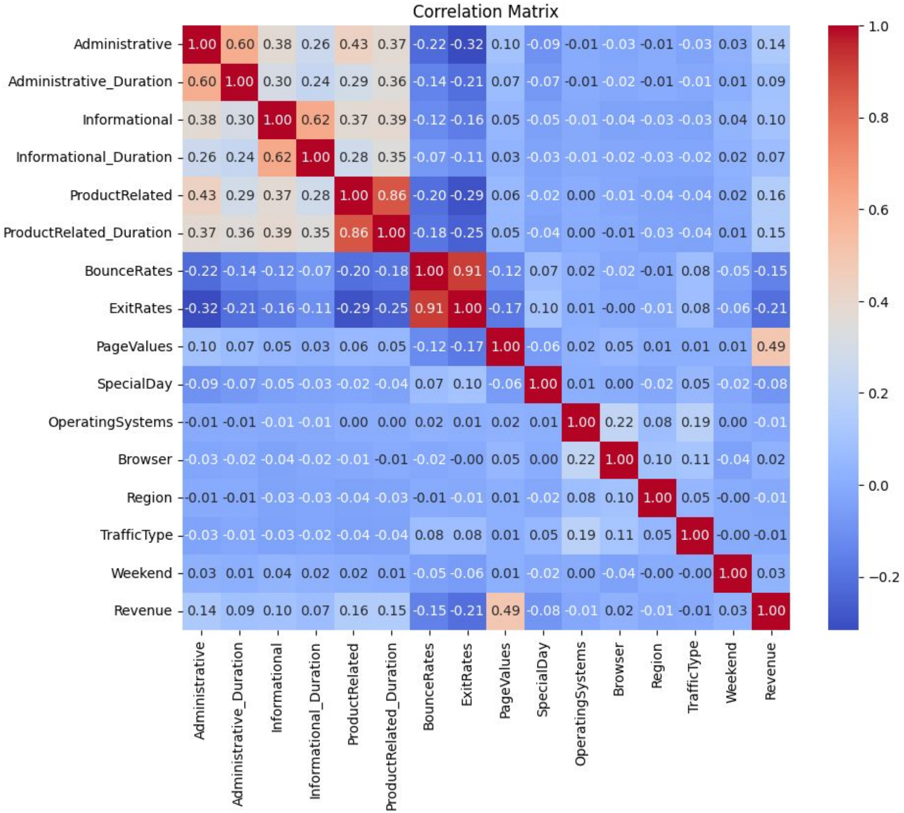
Mean and Variance

- Mean collaborate with distribution, excluding ProductRealated_Duration
- High variance → users completing transaction
- Concern: Noise, overfitting, complexity

Correlation

Variance

Interesting: Attributes which had high variance has high correlation



Correlation Matrix

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType	Weekend	Revenue
Administrative	1.00	0.60	0.38	0.26	0.43	0.37	-0.22	-0.32	0.10	-0.09	-0.01	-0.03	-0.01	-0.03	0.03	0.14
Administrative_Duration	0.60	1.00	0.30	0.24	0.29	0.36	-0.14	-0.21	0.07	-0.07	-0.01	-0.02	-0.01	-0.01	0.01	0.09
Informational	0.38	0.30	1.00	0.62	0.37	0.39	-0.12	-0.16	0.05	-0.05	-0.01	-0.04	-0.03	-0.03	0.04	0.10
Informational_Duration	0.26	0.24	0.62	1.00	0.28	0.35	-0.07	-0.11	0.03	-0.03	-0.01	-0.02	-0.03	-0.02	0.02	0.07
ProductRelated	0.43	0.29	0.37	0.28	1.00	0.86	-0.20	-0.29	0.06	-0.02	0.00	-0.01	-0.04	-0.04	0.02	0.16
ProductRelated_Duration	0.37	0.36	0.39	0.35	0.86	1.00	-0.18	-0.25	0.05	-0.04	0.00	-0.01	-0.03	-0.04	0.01	0.15
BounceRates	-0.22	-0.14	-0.12	-0.07	-0.20	-0.18	1.00	0.91	-0.12	0.07	0.02	-0.02	-0.01	0.08	-0.05	-0.15
ExitRates	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1.00	-0.17	0.10	0.01	-0.00	-0.01	0.08	-0.06	-0.21
PageValues	0.10	0.07	0.05	0.03	0.06	0.05	-0.12	-0.17	1.00	-0.06	0.02	0.05	0.01	0.01	0.01	0.49
SpecialDay	-0.09	-0.07	-0.05	-0.03	-0.02	-0.04	0.07	0.10	-0.06	1.00	0.01	0.00	-0.02	0.05	-0.02	-0.08
OperatingSystems	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.02	0.01	0.02	0.01	1.00	0.22	0.08	0.19	0.00	-0.01
Browser	-0.03	-0.02	-0.04	-0.02	-0.01	-0.01	-0.02	-0.00	0.05	0.00	0.22	1.00	0.10	0.11	-0.04	0.02
Region	-0.01	-0.01	-0.03	-0.03	-0.04	-0.03	-0.01	-0.01	0.01	-0.02	0.08	0.10	1.00	0.05	-0.00	-0.01
TrafficType	-0.03	-0.01	-0.03	-0.02	-0.04	-0.04	0.08	0.08	0.01	0.05	0.19	0.11	0.05	1.00	-0.00	-0.01
Weekend	0.03	0.01	0.04	0.02	0.02	0.01	-0.05	-0.06	0.01	-0.02	0.00	-0.04	-0.00	-0.00	1.00	0.03
Revenue	0.14	0.09	0.10	0.07	0.16	0.15	-0.15	-0.21	0.49	-0.08	-0.01	0.02	-0.01	-0.01	0.03	1.00

Highlight

Good:
Pagevalues, ExitRates, and BounceRates

Bad:
Multicollinearity & less interpretable attributes

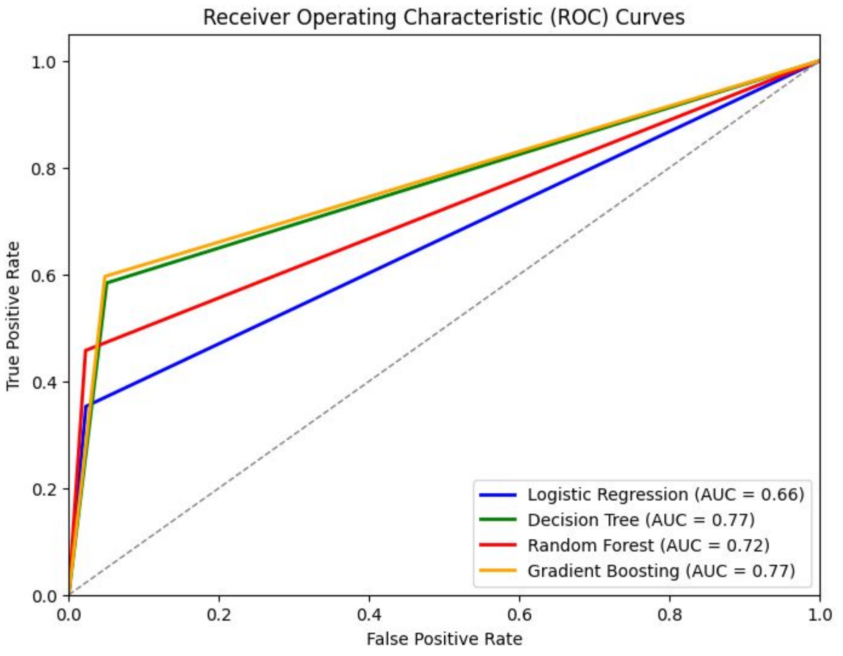
Models

Logistic Regression
Accuracy: 87.31%

Random Forest
Accuracy: 88.93%

Decision Tree
Accuracy: 88.85%

Gradient Boost
Accuracy: 89.13%



DT & GB

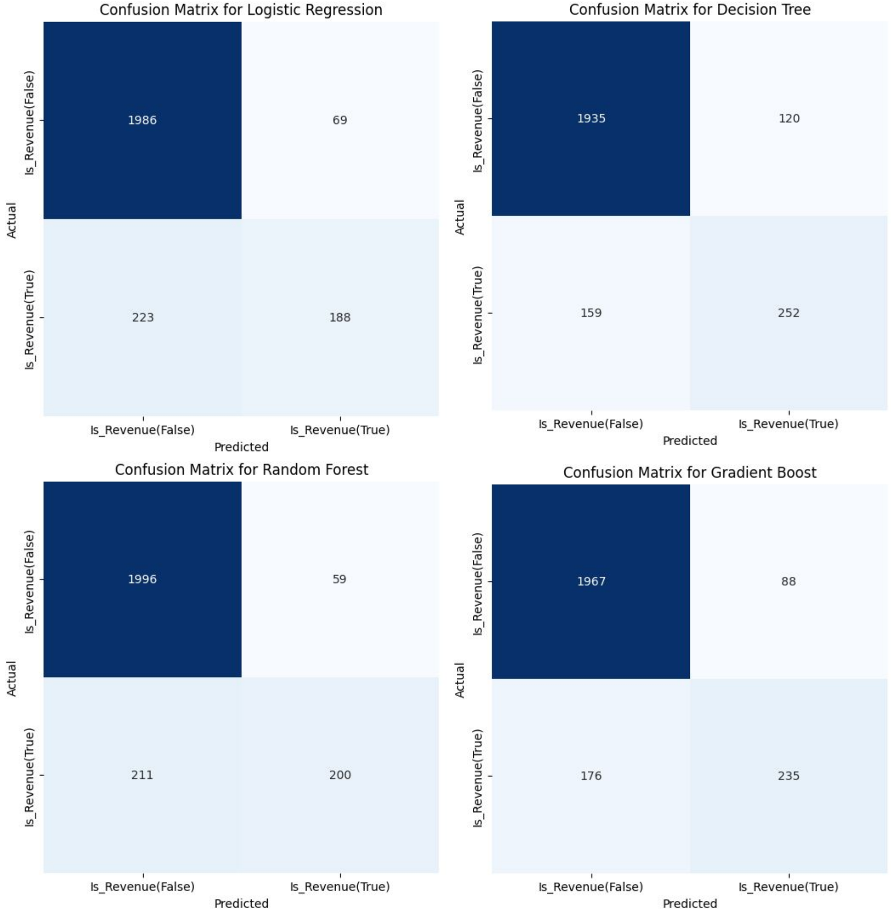
Highest AUC (0.77)
Moderately well at seperating class 1 and class 2

Logistic Regression

Lowest AUC (0.66)
More prone to misclassify class 1 as class 0

Confusion Matrix

- DT edges over RF(misclassification less pronounced)
- Number of misclassifications (top right & bottom left), reflect accuracy



Confusion Matrix for Logistic Regression

Actual \ Predicted	Is_Revenue(False)	Is_Revenue(True)
Is_Revenue(False)	1986	69
Is_Revenue(True)	223	188

Confusion Matrix for Decision Tree

Actual \ Predicted	Is_Revenue(False)	Is_Revenue(True)
Is_Revenue(False)	1935	120
Is_Revenue(True)	159	252

Confusion Matrix for Random Forest

Actual \ Predicted	Is_Revenue(False)	Is_Revenue(True)
Is_Revenue(False)	1996	59
Is_Revenue(True)	211	200

Confusion Matrix for Gradient Boost

Actual \ Predicted	Is_Revenue(False)	Is_Revenue(True)
Is_Revenue(False)	1967	88
Is_Revenue(True)	176	235

Conclusion

Best:
Gradient Boosting - Balance of accuracy and predictive performance
Alternative:
Decision Tree - Marginally less accurate, but scalable

Citations

Slide Template:
Slidesgo

Sources

- “Online Shoppers Purchasing Intention Dataset.” *UCI Machine Learning Repository*, archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset. Accessed 8 May 2024.