Online Buyer Intention

Module 5 Project

Akshay Ghalsasi

The data

- Data set of user sessions on a shopping website
- 10 continuous variables

Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues
0.0	0.0	0.0	0.0	1.0	0.000000	0.20	0.20	0.0
0.0	0.0	0.0	0.0	2.0	64.000000	0.00	0.10	0.0
0.0	-1.0	0.0	-1.0	1.0	-1.000000	0.20	0.20	0.0
0.0	0.0	0.0	0.0	2.0	2.666667	0.05	0.14	0.0
0.0	0.0	0.0	0.0	10.0	627.500000	0.02	0.05	0.0

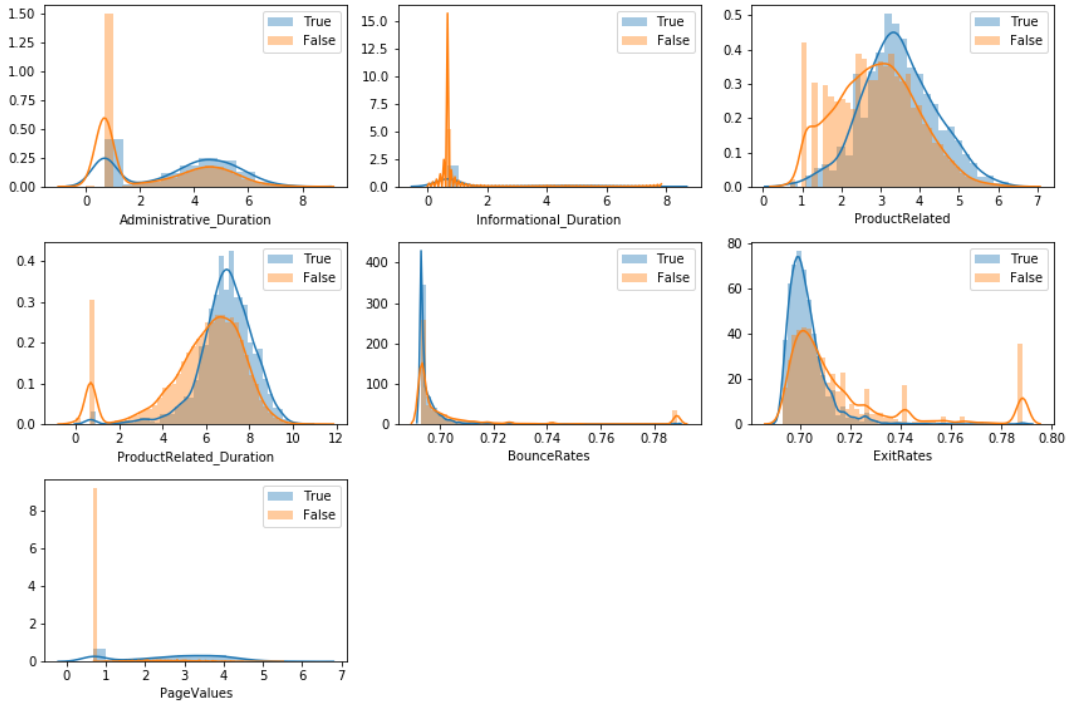
- 7 Categorical variables

SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend
0.0	Feb	1	1	1	1	Returning_Visitor	False
0.0	Feb	2	2	1	2	Returning_Visitor	False
0.0	Feb	4	1	9	3	Returning_Visitor	False
0.0	Feb	3	2	2	4	Returning_Visitor	False
0.0	Feb	3	3	1	4	Returning_Visitor	True

The problem

- One target feature “Revenue” with values [True, False]
- Binary classification problem. Need insight into what features possibly influence the probability of purchase
- Class imbalance 85% False and 15% True

Visualization



Classification Methods

- Used 3 classifiers
  - Naïve Bayes
  - Logistic Regression (GridSearchCV)
  - Random Forests (GridSearchCV)
- Used stacked classifier by stacking the three classifiers

model	precision	recall	accuracy	f1
NB	0.587922	0.676892	0.894452	0.629278
LogisticRegression	0.607460	0.686747	0.897970	0.644675
RandomForest	0.545293	0.834239	0.914208	0.659506
Stacked	0.603908	0.695297	0.899323	0.646388

Conclusion and Future Work

- RandomForest provides best accuracy (91.4%)
- Stacking doesn't seem to improve results
- Try more classifiers (SVM, XGBoost)
- Use Naïve Bayes on PCA to make variables iid.