ONLINE CUSTOMER INTENTION RECOMMENDATION USING ML AND DEEP LEARNING

Md. Anas Mondol

IBM Advanced Data Science Capstone

Github Link: Click Here

USE CASE

Model user behavior based on their interactions with an e-commerce website

Measure which website actions correlate with revenue and sales

Identify seasonality and trends in buying behaviors

DATASET

The dataset was obtained from Kaggle (https://www.kaggle.com/roshansharma/online-shopper-s-intention)

Dataset consists of 12316 samples and 18 features of which 8 are categorical and 10 are numeric

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	...	Region	TrafficType	VisitorType	Weekend	Revenue
0	0.0	0.0	0.0	0.0	1.0	...	1	1	Returning_Visitor	False	False
1	0.0	0.0	0.0	0.0	2.0	...	1	2	Returning_Visitor	False	False
2	0.0	-1.0	0.0	-1.0	1.0	...	9	3	Returning_Visitor	False	False
3	0.0	0.0	0.0	0.0	2.0	...	2	4	Returning_Visitor	False	False
4	0.0	0.0	0.0	0.0	10.0	...	1	4	Returning_Visitor	True	False

5 rows x 18 columns

DATA QUALITY ASSESSMENT

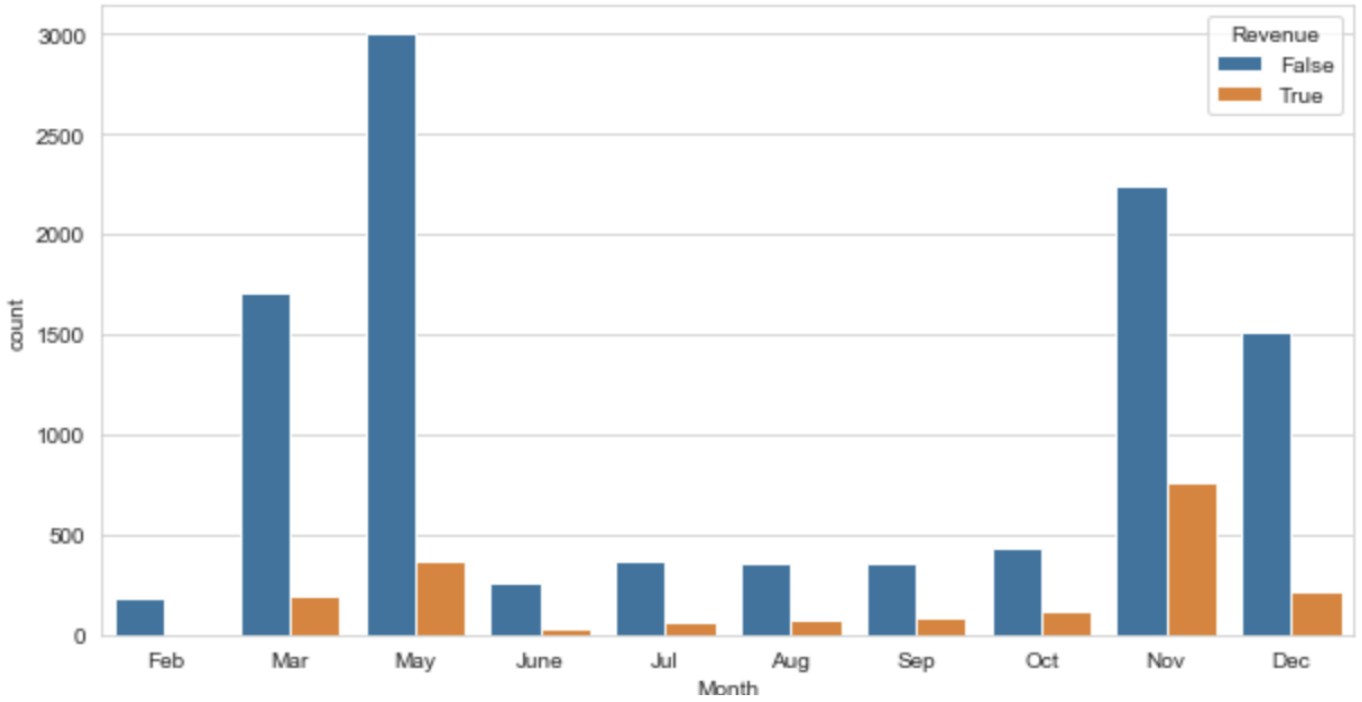
The Data consists of 0.11 percent Missing values

Some duration values were negative suggesting outliers or missing values

The numerical features are highly skewed.

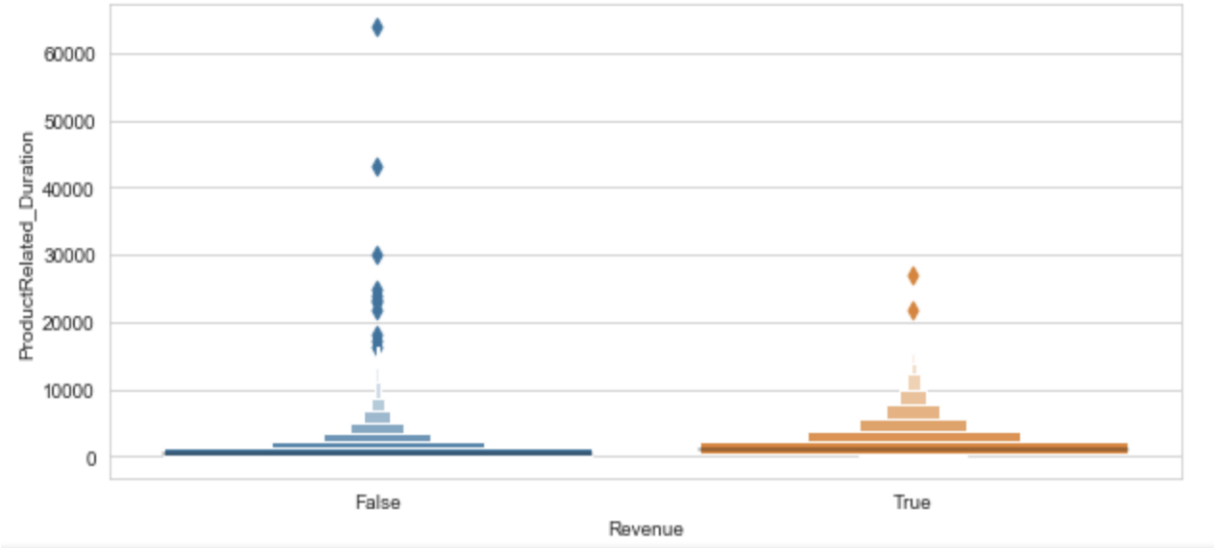
The Prediction classes are imbalanced with 1908 True and 10422 False classes

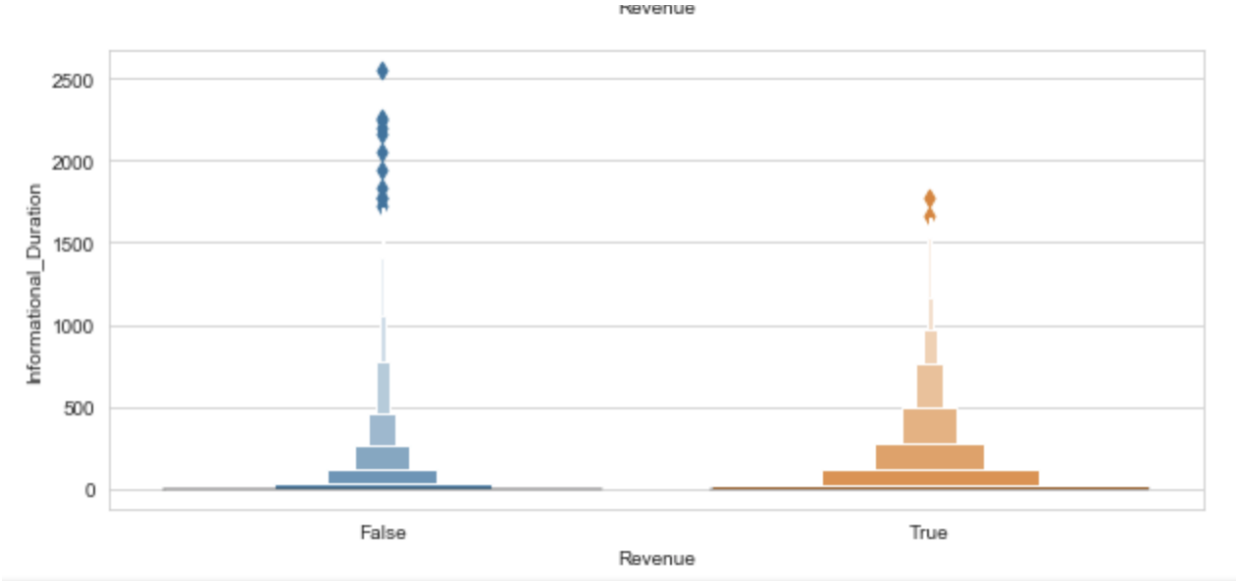
EXPLORATION AND VISUALIZATION



Website visitors and their contribution towards Revenue for different Months

EXPLORATION AND VISUALIZATION



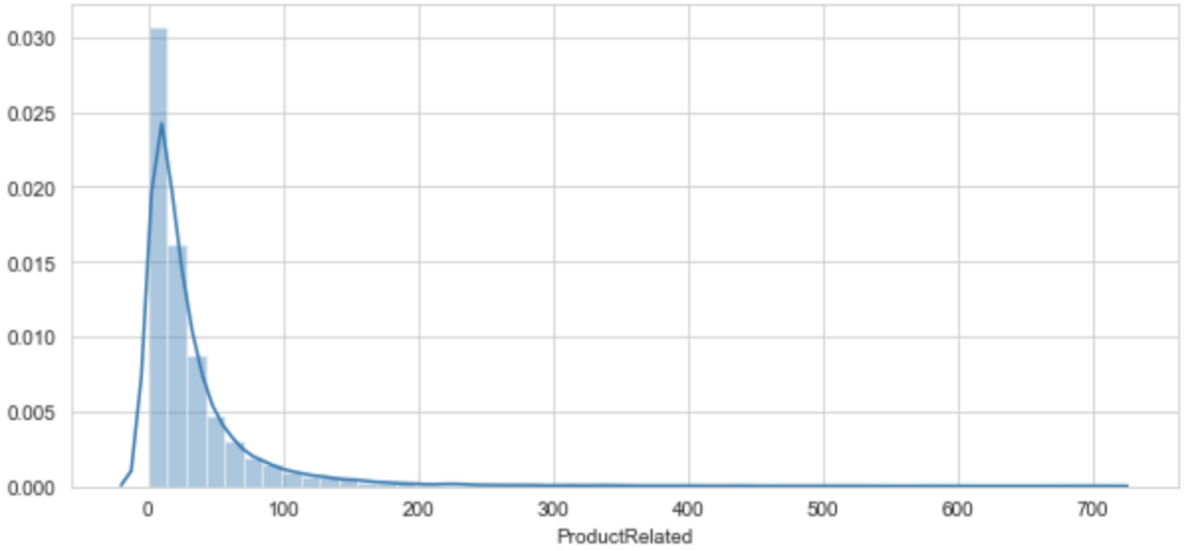


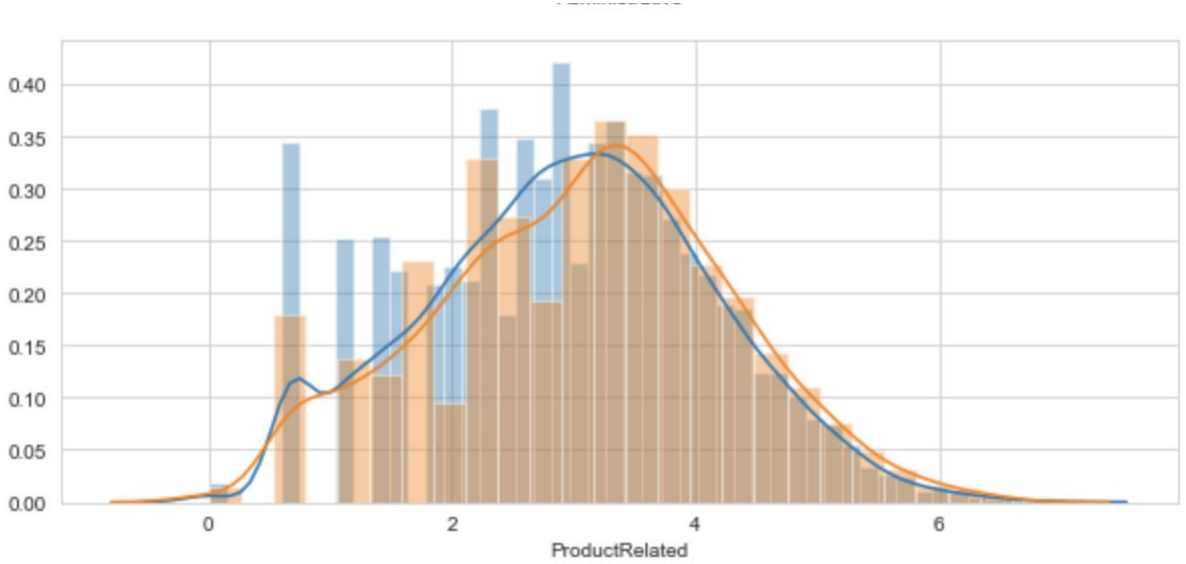
Enhanced Box Plots for outlier Detection

FEATURE ENGINEERING

- Missing values were dropped
- Each numerical values was clipped to remove outliers
- Categorical variables were One Hot encoded
- Three different feature sets were generated
  - No scaling
  - Scaled using MinMax scaler
  - Yeo Johnson transformation

EXPLORATION AND VISUALIZATION

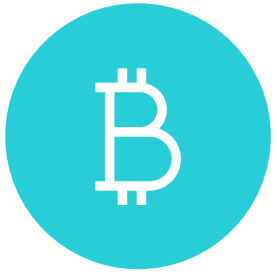




Before (left) and after applying Yeo Johnson Transformation on one feature

MODEL PERFORMANCE INDICATOR

To handle imbalance of classes Balanced accuracy score was used



Balanced accuracy is calculated as the average of the proportion corrects of each class individually.

MODEL SELECTION

Two models were uses:-

- Gradient boosted tree implemented with LightGBM
- Deep Neural Network implemented with Keras

Hyper-parameters For each models were optimized using Bayesian optimization

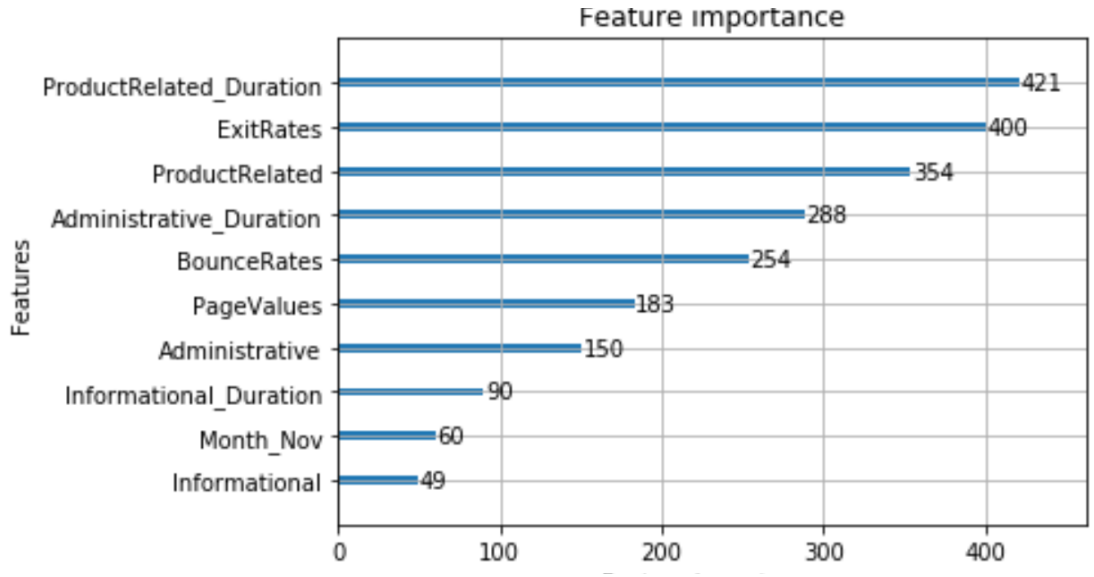
The models were evaluated on all 3 three feature sets.

RESULTS

	neural network		Boosted Trees	
Dataset Type	Train Score	Test Score	Train Score	Test Score
Encoded	0.7531	0.7172	0.8919	0.8118
Scaled	0.8956	0.7804	0.9455	0.8191
Transformed	0.8956	0.7804	0.9455	0.8191

RESULTS

TOP 10 most important features



CONCLUSION

- Dataset was cleaned , explored and visualized
- Three transformations were tested and applied
- Two models were trained
- Top 10 correlating features were isolated

THANK YOU