Online Customer Intention Prediction

Yash Rajput

IBM Advanced Data Science Capstone

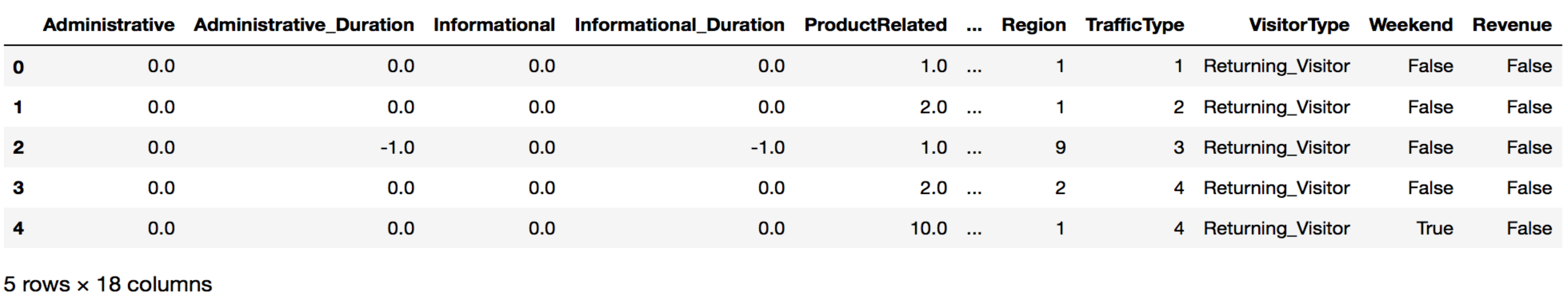
Use Case

- 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



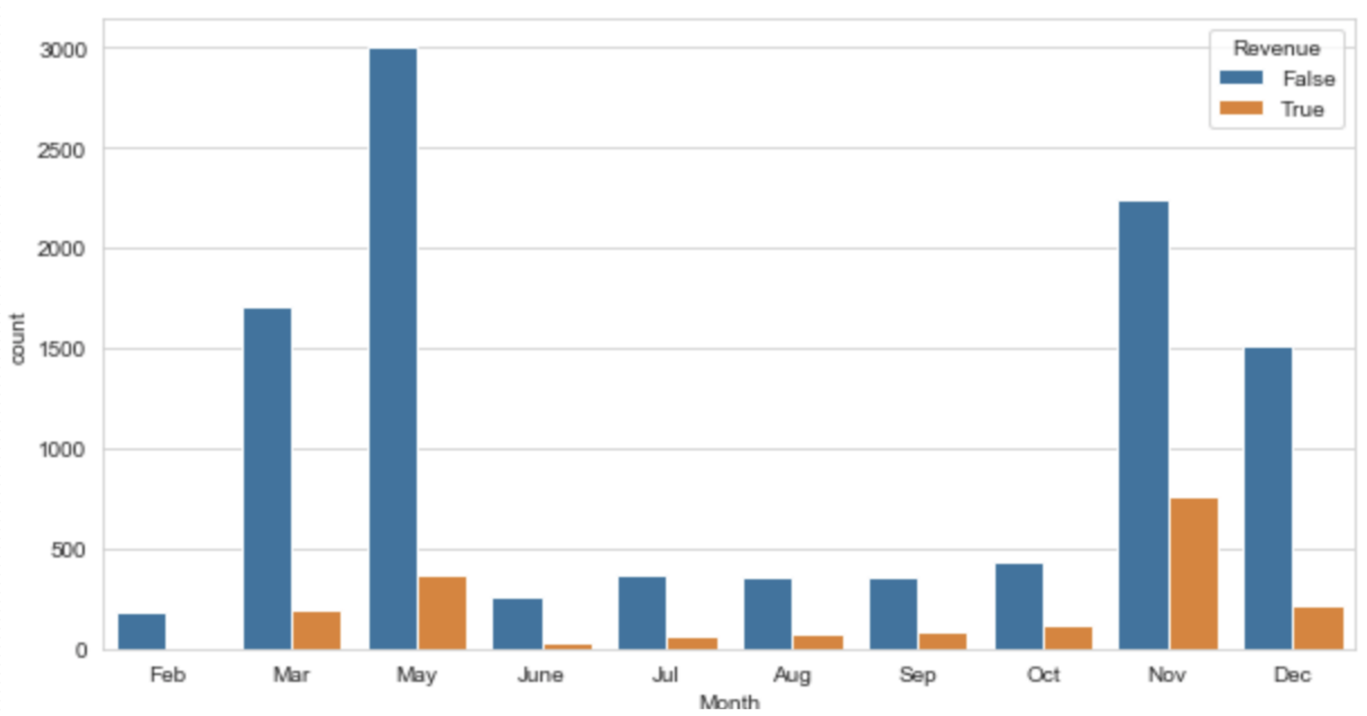
	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	...	Region	TrafficType	VisitorType	Weekend	Revenue
0	0.0	0.0	0.0	0.0	1.0	...	1	1	Returning_Visitor	False	False
1	0.0	0.0	0.0	0.0	2.0	...	1	2	Returning_Visitor	False	False
2	0.0	-1.0	0.0	-1.0	1.0	...	9	3	Returning_Visitor	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 × 18 columns

Data Quality Assessment

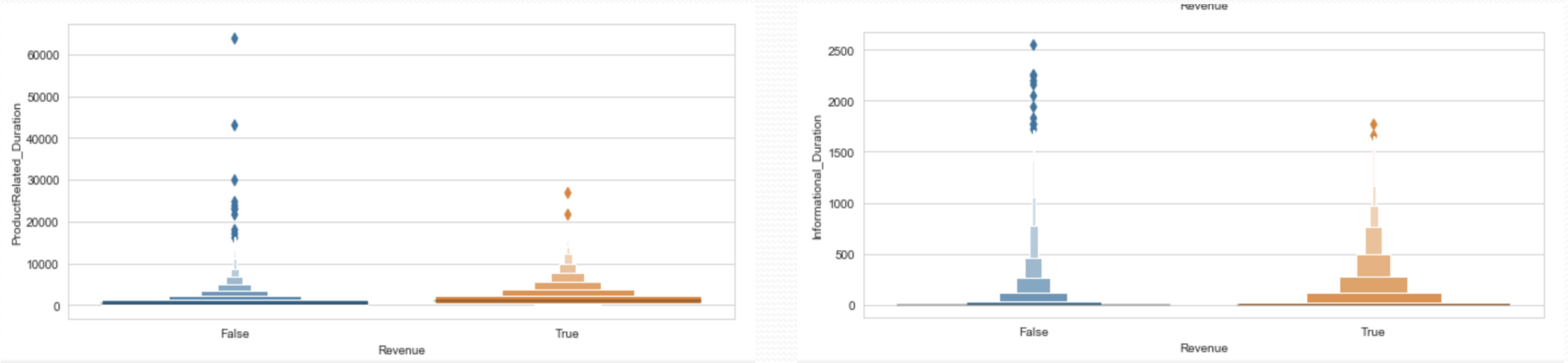
- The Data consists of 0.11 percent Missing values
- Some duration values were negative suggesting outliers or missing values
- The numerical features are highly skewed.
- The Prediction classes are imbalanced with 1908 True and 10422 False classes

Exploration and Visualization



Website visitors and their contribution towards Revenue for different Months

Exploration and Visualization

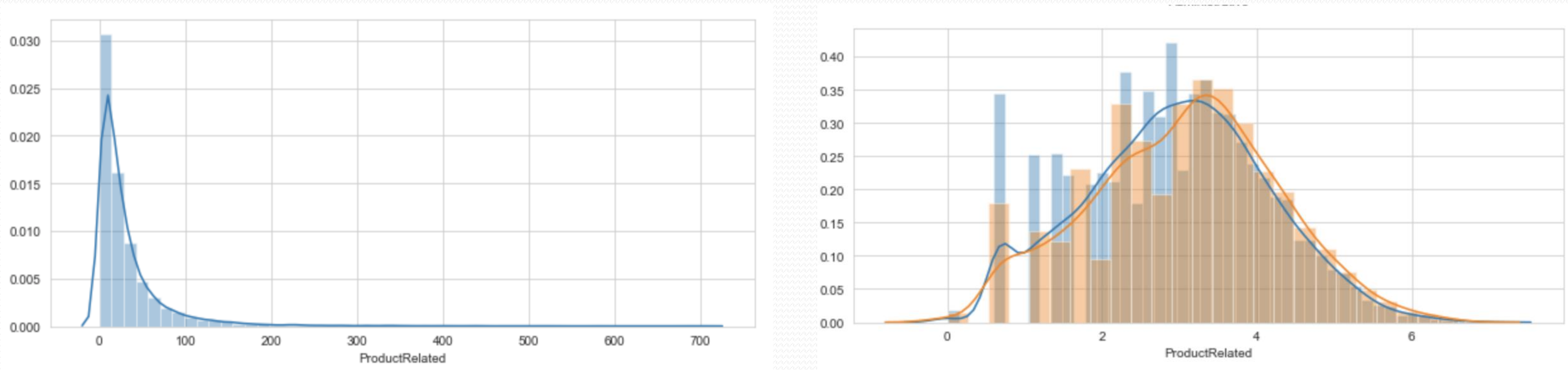


Enhanced Box Plots for outlier Detection

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
  1. No scaling
  2. Scaled using MinMax scaler
  3. Yeo Johnson transformation

Exploration and Visualization



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

Model Performance Indicator

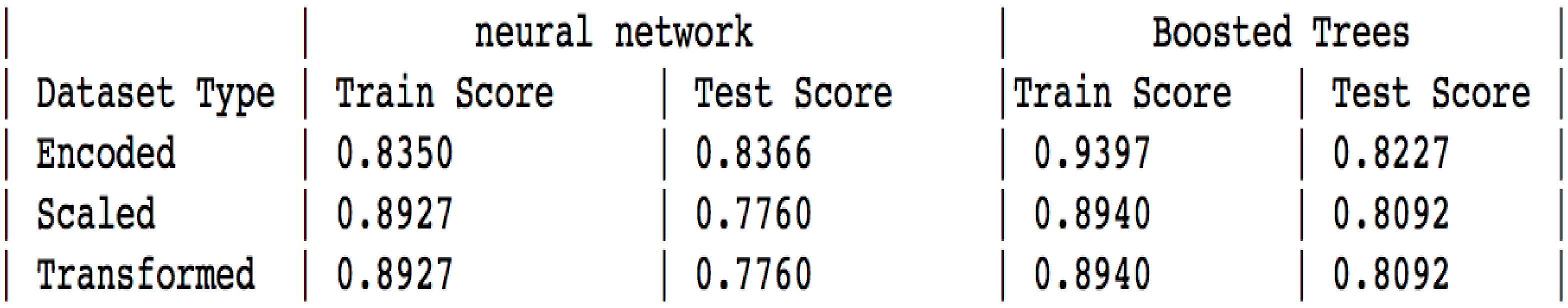
- To handle imbalance of classes Balanced accuracy score was used

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

Model Selection

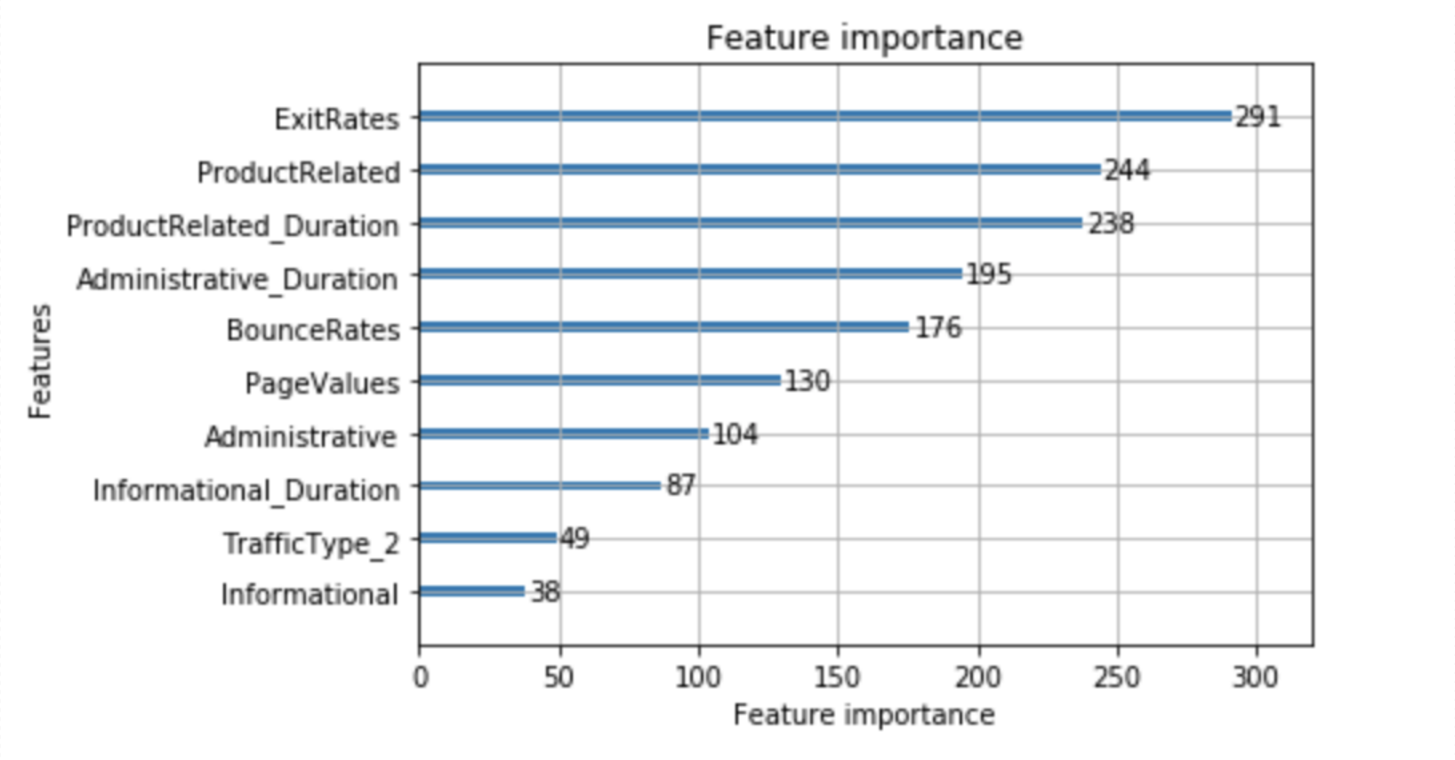
- Two models were uses:-
  1. Gradient boosted tree implemented with LightGBM
  2. Deep Neural Network implemented with Keras
- Hyper-parameters For each models were optimized using Bayesian optimization
- The models were evaluated on all 3 three feature sets.

Results



	neural network		Boosted Trees	
Dataset Type	Train Score	Test Score	Train Score	Test Score
Encoded	0.8350	0.8366	0.9397	0.8227
Scaled	0.8927	0.7760	0.8940	0.8092
Transformed	0.8927	0.7760	0.8940	0.8092

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