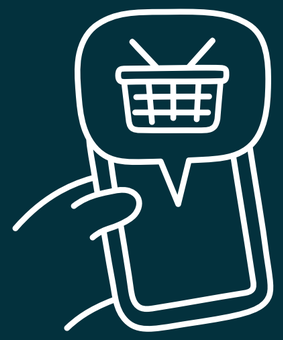
ONLINE SHOPPERS PURCHASING BEHAVIOR

Poojitha Vijjapu

OVERVIEW



DISCUSSION POINTS

About Dataset

Literature Review

Models built and methods used

Results

ABOUT DATASET : ONLINE SHOPPER'S PURCHASING INTENT

- Data is a collection of various predictor variables related to a session of customer’s visit to an e-commerce website
- Each observation tells if the customer generated a revenue or not
- Goal is to identify revenue generating customers and target them.

WHY

- Understand Consumer Behavior and predict the outcomes of actions
- Relatable dataset
- E-commerce is increasing exponentially - able to target consumers efficiently is important.
- This dataset enables application of the classroom concepts of Logistic Regression and Decision Trees.

Features of the Dataset

- Multivariate : This dataset is a multi-variate dataset, with a combination of numerical and categorical features.
- Binary Response Variable : The response variable is a binary variable with True or False values
- Imbalanced Data : The dataset is imbalanced in terms of response variable, with over 81% lesser customers generated revenue compared to those who did not.

18 Columns

12330 Rows

Features of the Dataset

Predictor Variables-Numerical
Administrative
Administrative duration
Informational
Informational duration
Product related
Product related duration
Bounce rate
Exit rate
Page value
Special day

Predictor Variables-Categorical
OperatingSystems
Browser
Region
TrafficType
VisitorType
Weekend
Month

Response Variables-Categorical
Revenue

Dataset Description

Feature name	Feature description
Administrative	Number of pages visited by the visitor about account management
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages
Product related	Number of pages visited by visitor about product related pages
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce rate	Average bounce rate value of the pages visited by the visitor
Exit rate	Average exit rate value of the pages visited by the visitor
Page value	Average page value of the pages visited by the visitor
Special day	Closeness of the site visiting time to a special day

Table 2 Categorical features used in the user behavior analysis model

Feature name	Feature description
OperatingSystems	Operating system of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)
VisitorType	Visitor type as “New Visitor,” “Returning Visitor,” and “Other”
Weekend	Boolean value indicating whether the date of the visit is weekend
Month	Month value of the visit date
Revenue	Class label indicating whether the visit has been finalized with a transaction

LITERATURE REVIEW

- Paper from 2018
- Real-time online shopper behavior analysis system
- It proposes two models which act simultaneously and in real-time: predicts visitor's 1. purchasing intention and 2. likelihood to abandon the site.
- Used filter feature selection to determine the most discriminative factors in predicting purchasing intention
- Did not use knn as it not applicable for real time
- Used long short-term memory (LSTM) recurrent neural network (RNN) (LSTM-RNN) (with sequential clickstream data to predict the probability that the user will leave the site within a certain time)

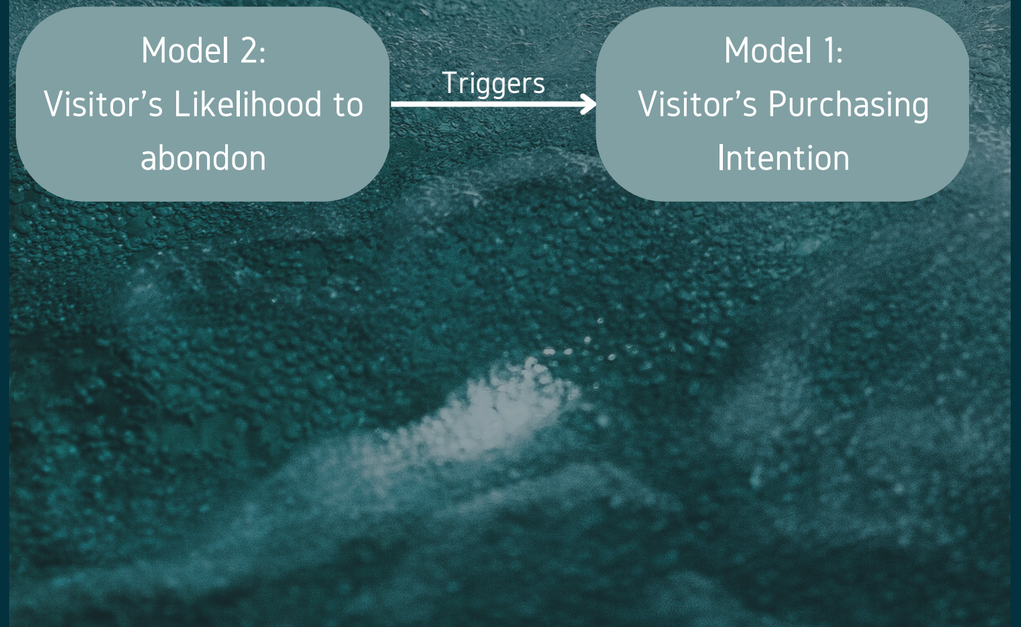
Neural Computing and Applications (2019) 31:6893–6908
https://doi.org/10.1007/s00521-018-3523-0

ORIGINAL ARTICLE

Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks

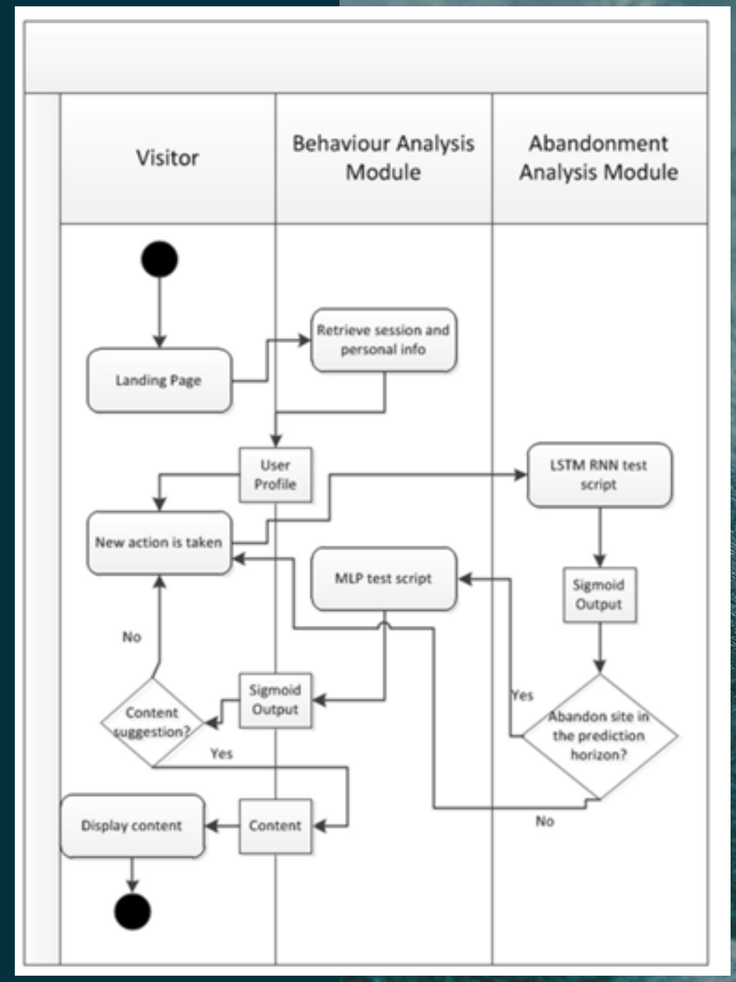
C. Okan Sakar[1] · S. Olcay Polat[2] · Mete Katircioglu[1] · Yomi Kastro[3]

Received: 18 July 2017 / Accepted: 4 May 2018 / Published online: 9 May 2018



LITERATURE REVIEW

- Used MLP, SVM and Random Forest algorithms
- MLP(Multilayer Perceptor) is a feedforward artificial neural network model that is made up of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.
- SVM (Support Vector Machine) is a discriminant-based algorithm which aims to find the optimal separation boundary called hyperplane to discriminate the classes from each other.
- Decision Trees: Choose Random Forest as it proved to be effective for many classification problem.
- Used Feature Selection Technique to improve performance
- Imbalanced data: Used oversampling to get a balanced data. Oversampling on the 70% training data only to avoid data leakage



LITERATURE REVIEW-INSPIRED

Following techniques and methods are inspired from the literature review :

SAMPLING

Using Oversampling to set the unbalanced data

DECISION TREE

Using Ramdom Forest decision tree as it is a proven successful in similar datasets

SVM

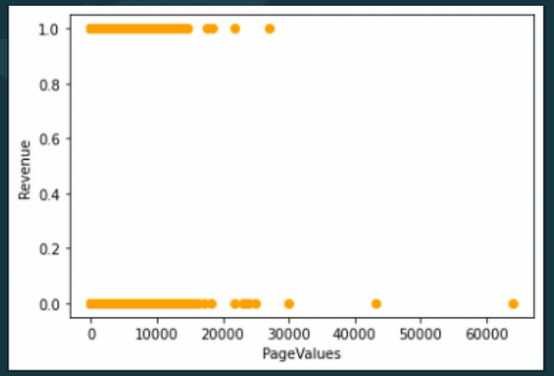
Using Support Vector machine with rbf kernel

DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

EDA

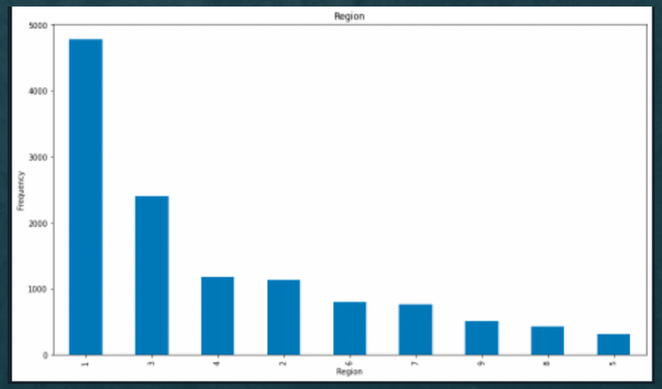
- There are 10 numericals and 8 categorical variables
- The response variable in unbalanced with 85% False and 15% only true data
- There are no missing or null values in the dataset

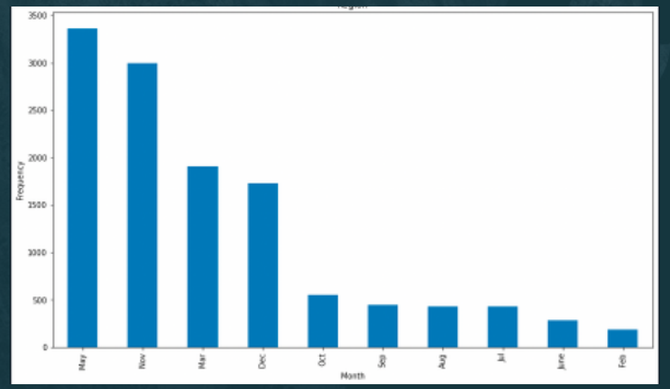
Class	Count
TRUE	1908
FALSE	10422

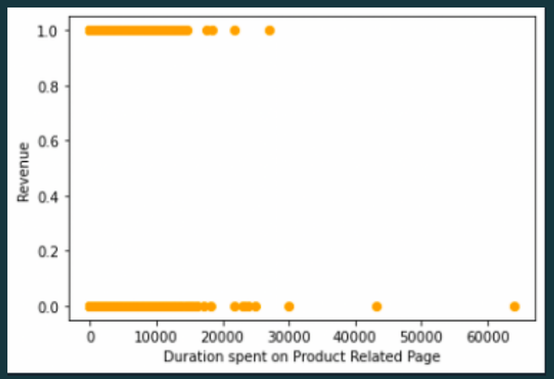


18 Columns

12330 Rows



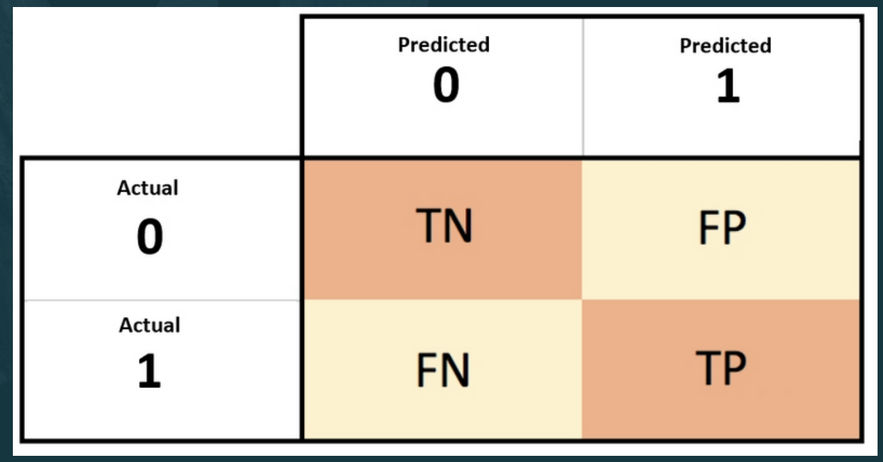




DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

GOALS & EXPLANATIONS

- Need to identify Revenue generating customers
- Want a model which predict high number of True Positives
- In order to check for high number of true positives, we need to check for :
  - -High Positive Predictive Value (TP/(TP+FP))
  - -High Accuracy
  - -High Roc_Auc Score
  - -High Youden Score
  - -Low FPR
- A good model in this scenario should have high sensitivity



DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

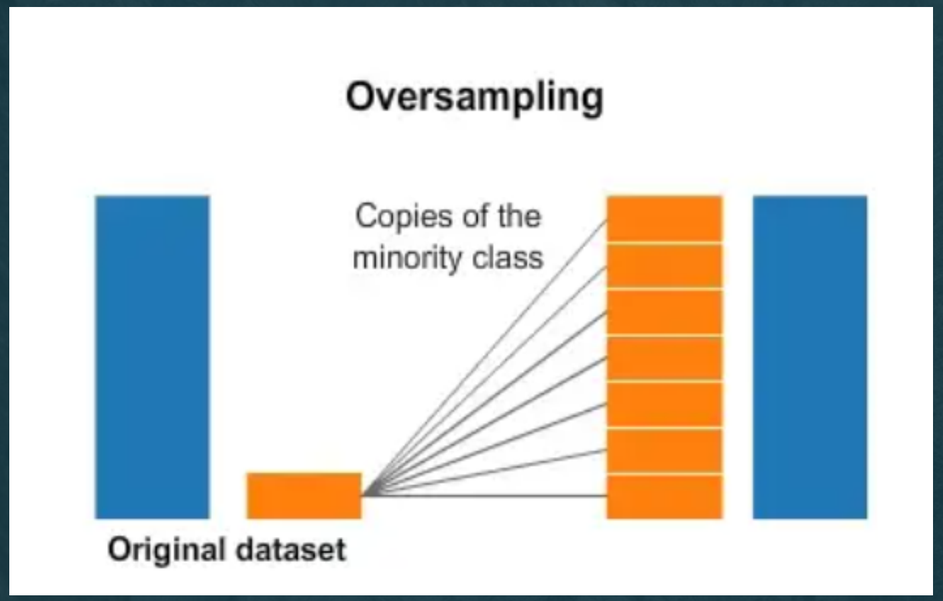
OVERVIEW OF METHODS AND STEPS

- Oversampled Data
- Logistic Regression
- Logistic Regression with Lasso with tunning parameters
- Random Forest
- SVM

DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

WORKING, RESULTS AND INTERPRETATIONS

Oversampling : It means to create multiple copies of the minority class. This allows equal distribution of data.



No Sampling of Data →

```
For threshold  0.15151515151515152  the confusion matrix is
 [[2697  452]
 [ 123  427]]
And other metrics are
Sensitvity(or TPR):  0.7763636363636364
Specificity(or TNR):  0.8564623690060337
FNR:  0.22363636363636363
FPR:  0.14353763099396633
Accuracy:  0.8445525817788592
Youden Index:  0.6328260053696702
Precision:  0.4857792946530148
 Roc_Auc_Score 0.9044048615722162
PositivePredictionValue 0.4857792946530148
```

Oversamplied Data →

```
For threshold  0.3535353535353536  the confusion matrix is
 [[2619  530]
 [  92  458]]
And other metrics are
Sensitvity(or TPR):  0.8327272727272728
Specificity(or TNR):  0.831692600825659
FNR:  0.16727272727272724
FPR:  0.168307399174341
Accuracy:  0.8318464449851312
Youden Index:  0.6644198735529319
Precision:  0.46356275303643724
 Roc_Auc_Score 0.9067952308092035
PositivePredictionValue 0.46356275303643724
```

Logistic Regression without tunning or penalizations

DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

WORKING, RESULTS AND INTERPRETATIONS

Logistic Regression - Tuning and Penalizing

- Logistic Regression with l1 penalization - Lasso Regression
- Tuned : solvers, c_values

```
The results of Lasso Regression with threshold 0.15 are
 the confusion matrix is
 [[2576  573]
 [ 105  445]]
And other metrics are
Sensitvity(or TPR):  0.8090909090909091
Specificity(or TNR):  0.8180374722134011
FNR:  0.19090909090909092
FPR:  0.18196252778659894
Accuracy:  0.8167072181670721
Youden Index:  0.6271283813043103
Precision:  0.43713163064833005
```

No Sampling of Data

```
The results of Lasso Regression with threshold 0.35 are
 the confusion matrix is
 [[2270  879]
 [  64  486]]
And other metrics are
Sensitvity(or TPR):  0.8836363636363637
Specificity(or TNR):  0.7208637662750079
FNR:  0.11636363636363634
FPR:  0.2791362337249921
Accuracy:  0.745066234117329
Youden Index:  0.6045001299113717
Precision:  0.35604395604395606
 Roc_Auc_Score 0.891345881809521
PositivePredictionValue 0.35604395604395606
```

Oversamplied Data

DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

WORKING, RESULTS AND INTERPRETATIONS

Decision Trees- Random Forest

- Choose Random Forest as according the paper it proved to be successful in these kinds of datasets
- Tuned : n_estimators and max_features

```
The results of Random Forest are
 the confusion matrix is
 [[3019  130]
 [ 230  320]]
And other metrics are
Sensitvity(or TPR):  0.5818181818181818
Specificity(or TNR):  0.9587170530327088
FNR:  0.4181818181818182
FPR:  0.04128294696729118
Accuracy:  0.902676399026764
Youden Index:  0.5405352348508905
Precision:  0.7111111111111111
 Roc_Auc_Score 0.8935806460925546
PositivePredictionValue 0.7111111111111111
```

No Sampling of Data

```
The results of Random Forest are
 the confusion matrix is
 [[2793  356]
 [ 115  435]]
And other metrics are
Sensitvity(or TPR):  0.7909090909090909
Specificity(or TNR):  0.8869482375357256
FNR:  0.2090909090909091
FPR:  0.1130517624642744
Accuracy:  0.8726682887266829
Youden Index:  0.6778573284448166
Precision:  0.549936788874842
 Roc_Auc_Score 0.891345881809521
PositivePredictionValue 0.549936788874842
```

Oversamplied Data

	feature	importance_1
8	PageValues	0.608396
7	ExitRates	0.066374
5	ProductRelated_Duration	0.060585
4	ProductRelated	0.050305
6	BounceRates	0.049204
1	Administrative_Duration	0.033424
16	Month_Nov	0.032395
0	Administrative	0.031409
3	Informational_Duration	0.012450
66	VisitorType_Returning_Visitor	0.009467
15	Month_May	0.005958
2	Informational	0.005529
46	TrafficType_2	0.004289
19	OperatingSystems_2	0.003827
67	Weekend_True	0.003543

No Sampling of Data

DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT

WORKING, RESULTS AND INTERPRETATIONS

SVM -

- Used rbf kernel
- Both are not so good.

```
The results of SVM using rbf kernel are
 the confusion matrix is
 [[3019  130]
 [ 230  320]]
And other metrics are
Sensitvity(or TPR):  0.03272727272727273
Specificity(or TNR):  0.9911082883455065
FNR:  0.9672727272727273
FPR:  0.00889171165449354
Accuracy:  0.8486077318194106
Youden Index:  0.023835561072779177
Precision:  0.391304347826087
 Roc_Auc_Score 0.8935806460925546
PositivePredictionValue 0.391304347826087
```

No Sampling of Data

```
The results of SVM using rbf kernel are
 the confusion matrix is
 [[3111   38]
 [ 530   20]]
And other metrics are
Sensitvity(or TPR):  0.03636363636363636
Specificity(or TNR):  0.9879326770403303
FNR:  0.9636363636363636
FPR:  0.012067322959669724
Accuracy:  0.8464449851311165
Youden Index:  0.02429631340396665
Precision:  0.3448275862068966
 Roc_Auc_Score 0.5121481567019833
PositivePredictionValue 0.3448275862068966
```

Oversamplied Data

DATA ANALYSIS ON THE DATASET: ONLINE SHOPPER'S PURCHASING INTENT CONCLUSION

- From the interpretations of different models, it is clear that the Oversampled data is performing better in most cases for the cause of identifying True Positives or Revenue generating customers.
- The best model of all is formed from Logistic Regression with Lasso Penalty with the oversampled data. It has 88.3% Sensitivity, 72 % Specificity, low False Negative Rate and low False Positive rates of 11% and 27%; Accuracy of 74.5% and high Roc_Auc score of 89.1 % .
- The best features discovered from this model are PageValues, ProductRelated, Informational_Duration, ProductRelated_Duration and Administrative_Duration

Thank You