Team Project Final Report

Course dec520Q, Section 10A, Team A15

Team Members: Yehudi Baptiste, Hirak Bhayani, Rata Kiewkarnkha, Xinyue Li, Zhipeng Wang

A. The Business Understanding

According to the Commerce Department, the percentage of retail sales done online surpassed general merchandise sales (i.e. sales at department stores, warehouse clubs and super-center) for the first time ever this year.[1] This major milestone for ecommerce highlights the importance that online sales has come to hold for merchandisers and retailers. Retailers can no longer deny the benefits of offering their products online.

But with an online presence comes the need to manage that presence in a way that maximizes conversion i.e. the probability that a visitor will make a purchase. According to Okan Sakar et al, "the fact that … conversion rates have not increased at the same rate of [usage] leads to the need for solutions that present [customization] to online shoppers."[2] Customization requires an understanding of when a visitor is likely to make a purchase, or simply abandon the site empty handed without making any purchases. Visitors who are more likely to abandon can be shown a different version of the site, offered special promotions, or shown different products, to encourage conversion.

Oskar et al. contrast the tools that retailers have online vs offline to motivate conversion. "In physical retailing," according to them, "a salesperson can offer a range of customized alternatives to shoppers based on the experience he or she has gained over time."[3] Bringing such agency online should be a key goal of retailers. We hope to help retailers achieve this goal.

Using session data from an online retailer, we seek to construct a model that will predict the likelihood that a visitor will make a purchase in real time. This prediction will be used to inform customization so that those customers with purchase intent are pushed to convert.

In this supervised binary classification problem, we construct and assess different models (e.g. logistic, Lasso, random forest) to see which model has the most predictive power. This project is relevant to the online retail space. We envision that our model can be used in real time to make predictions that will allow retailers to maximize the likelihood that visitors make a purchase. Since a

---

[1] https://www.cnbc.com/2019/04/02/online-shopping-officially-overtakes-brick-and-mortar-retail-for-the-first-time-ever.html

[2] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). https://doi.org/10.1007/s00521-018-3523-0

[3] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). https://doi.org/10.1007/s00521-018-3523-0

small increase in sales conversion can pay huge dividends, this project will help retailers understand how to maximize their revenues.

B. The Data Understanding

To build our supervised models, we will use session data donated from an online retailer. The data contain session information, click flow data, and purchases for visitors who navigated to the retailer's website over the course of a year. One record reflects one session, in which a user visited the website, perform certain actions, and either purchased a product or not. The dataset consists of 10 numerical variables (e.g average time spent on different pages) along with 8 categorical variables (browser, region etc.).

Our variable of interest (i.e. outcome or target variable) is the attribute "Revenue," which is a binary variable that records if a visitor made a purchase or not. Of the 12,330 sessions in the dataset, 84.5% are false cases in which a visitor did not make a purchase, and the rest are true cases that reflect a purchase. The dataset was obtained from the University of California, Irvine machine learning repository.[4]

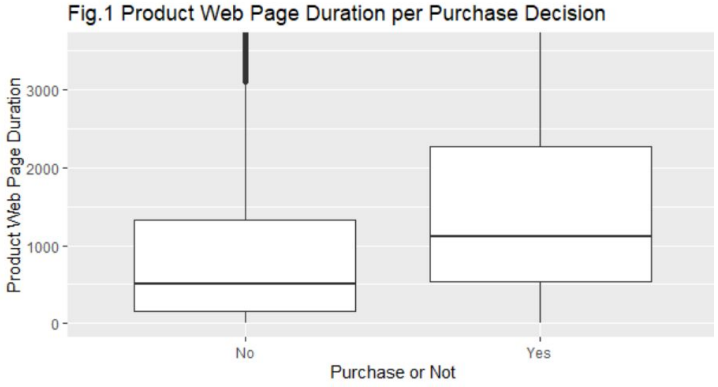
C. Data Preparation

In preprocessing our data, we change some variables into the appropriate format. For example, we change the variable "Revenue" (our outcome variable) into a binary variable from a logical. We also changed the levels of our categorical variables by labeling them with the correct names (e.g. changing Browser 1 to Browser Chrome). Our original dataset includes all complete cases, so there was no need to drop cases.

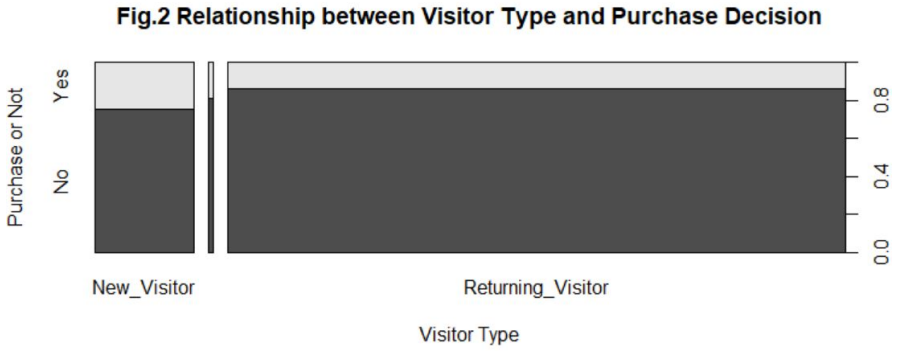
C.1 Exploratory Data Analysis

Before constructing our models, we explore our data using visualizations.

[4] http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset



The box plot above shows the average time spent on product-related pages split by those who made a purchase vs not. Unsurprisingly, we observe that the shoppers who purchased stayed longer on product related pages than those who did not make a purchase.



Next, we look at purchase penetration for new visitors vs. returning visitors. The plot above demonstrates that the majority of the visitors coming to the website are returning visitors. Surprisingly, we observe that the rate of purchase is higher for new visitors.

C.2 Clustering

We conduct a k-Means clustering based on six variables including "Administrative", "Administrative_Duration", "Informational", "Informational_Duration", "ProductRelated", and "ProductRelated_Duration". The Administrative, Informational, and ProductRelated are number of web pages that visitors visit in each category; while, the duration ones are amount of time visitors spend across that type of web pages. HDIC recommends nine clusters; however, we consider that nine is too high to interpret meaningfully. We follow the advice of industry experts such as Pranava Goundan of ZS Associates and settle on five clusters.

Table 1: The Five Clusters

Cluster	1	2	3	4	5
Mean of Revenue (Purchase or Not, the outcome variable)	0.2500000	0.3458333	0.2703180	0.2189239	0.1188061
Size of Cluster	252	240	566	2695	8577

From Table 1, we see that the biggest cluster is cluster 5 with the size of 8577 and mean revenue of about 0.12. This is the group that is the most unlikely to purchase. According to the centers of the clusters (Table 2) in the appendix, cluster 5 has the lowest duration spending on the web pages and they also visit the fewest number of web pages as compared to other clusters. Another interesting cluster is cluster 2 which has the highest chance of buying of about 0.35. According to Table 2, cluster 2 has the highest values for ProductRelated and ProductRelated_Duration, suggesting that product related web pages might lead visitors to make a purchase more than Informational and Administrative web pages, which makes sense.

D. Modeling

D.1: Model Selection with 10-fold cross validation on original dataset.

In the first step of our model framework, we fit several models using 11 of the 17 variables in our dataset. We exclude highly collinear variables, such as "Administrative Duration," which captures similar information to the variable Administrative. (See Table , Appendix). We also dropped variables that likely supply no information for predicting a visitor's purchasing intention, such as operating machine.

We set aside 25 percent of our data for testing and employ 10-fold cross validation to train models on the remaining 75 percent of our data. We run a 1) logistic regression 2) logistic regression with interaction 3) classification tree 4) LASSO with theory rule 5) LASSO with min rule 6) LASSO with 1se rule 7) Post-LASSO with theory rule 8) Post-LASSO with min rule 9) Post-LASSO with 1se rule and 10) a neural network with a sigmoid function at its output layer.

D. 2: Model Evaluation

We show the performance of our models across a number of OOS metrics, including accuracy, True-postive rate (TPR or recall), True-negative rate (TNR or specificity), and F1 score. We show accuracy because it is a familiar metric; although, it is susceptible to class imbalances. Given our business problem, however, we focus on TPR, TNR, and F1 score since we are more concerned with correctly classifying likely purchasers and nonpurchasers to provide appropriate content for each class.

We believe that offering customized content to visitors who are likely to purchase could further motivate them to purchase, increasing conversion. For example, showing simpler web pages--or navigating visitors through a streamlined purchasing funnel--could remove distractions and maximize the chance of conversion for likely purchasers. Failing to predict likely purchasers is highly undesirable if content is shown that would distract from conversion.

Moreover, customized content for non purchasing visits could ensure that we are optimizing on the metrics that we care about for these types of visits. For example, a retailer could seek to maximize engagement and retention for sessions associated with little to no purchasing intent.

We adopt F1 as the final evaluation metric. The F1 score, widely used in the literature, balances precision and recall in addition to accounting for class imbalances.[5] Class imbalances are common in session data since most visits on an online retailer do not end in conversion.

In Table 3, we display the performance of the classifiers on the test sample. The neural network produces the highest F1 score, followed by the classification tree and LASSO with min rule.

Table 3: Test Results

Model	Out of Sample Performance Measurement Matrix			
	Accuracy	TPR	TNR	F1
DNN	0.89	0.61	0.94	0.63
tree	0.89	0.51	0.96	0.58
L.min	0.88	0.35	0.98	0.48
logistic with interaction	0.88	0.36	0.97	0.48
logistic	0.88	0.34	0.98	0.47
L.1se	0.88	0.32	0.98	0.46
L.theory	0.88	0.30	0.98	0.43
average	0.87	0.24	0.99	0.37
PL.min	0.68	0.24	0.76	0.19
PL.1se	0.68	0.22	0.77	0.18
PL.theory	0.69	0.22	0.77	0.18
null	0.85	0.00	1.00	NaN

[5] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). https://doi.org/10.1007/s00521-018-3523-0

D.4: Oversampling and undersampling to mitigate class imbalance

Table 3 shows the neural network as the model with the highest F1 score. This model, however, is an imbalanced classifier. The number of negative class instances (i.e. visits with no purchase) exceeds that number of positive class instance in our original data set, on the order of 5 to 1. The differences between TPR and TNR across the models signify that the classifiers tend to discriminatively label visits in the hold out sample as negative test instances. Class imbalances are common in the online retail environment since most visits do not end with conversion. The needs to improve our true-positive rate and compensate for class imbalances call for a balanced classifier.

We use two methods to create balanced datasets for training. We first create a balanced data set using oversampling. We set aside 30 % of our original data for OOS testing. Using the remaining 70%, we oversampled minority class instances (i.e. visits with purchases) so that our training data is split 50/50 between negative and positive classes.

We also undersample. As with oversampling, we divided our data between test and training. Then, we undersampled majority class instances (i.e. visits with no purchases) so that the data is split 50/50 between negative and positive classes. Performance metrics for models trained with balanced data using undersampling and oversampling are shown in Table 4 and Table 5, respectively. The performance of the models in both cases improve compared to the performance of the models trained with imbalanced data. Undersampling, however, performs better than oversampling.

Table 4: Test Results with Undersampling

Model	Out of Sample Performance Measurement Matrix			
	Accuracy	TPR	TNR	F1
tree	0.88	0.78	0.90	0.66
L.theory	0.89	0.71	0.92	0.66
average	0.89	0.68	0.93	0.65
L.1se	0.88	0.70	0.91	0.64
L.min	0.87	0.76	0.89	0.63
PL.theory	0.88	0.71	0.91	0.63
logistic	0.87	0.73	0.89	0.62
PL.1se	0.87	0.73	0.89	0.62
PL.min	0.83	0.80	0.83	0.57
logistic.interaction	0.81	0.77	0.82	0.55
DNN	0.83	0.58	0.87	0.50
null	0.85	0.00	1.00	NaN

Table 5: Test Results with Oversampling

Model	Out of Sample Performance Measurement Matrix			
	Accuracy	TPR	TNR	F1
tree	0.87	0.79	0.89	0.65
average	0.88	0.70	0.91	0.64
L.theory	0.87	0.74	0.89	0.63
PL.theory	0.86	0.76	0.87	0.62
logistic	0.86	0.76	0.87	0.62
L.1se	0.86	0.75	0.87	0.62
PL.1se	0.84	0.74	0.86	0.59
L.min	0.83	0.75	0.85	0.58
PL.min	0.83	0.74	0.85	0.58
logistic.interaction	0.83	0.75	0.85	0.58
DNN	0.74	0.91	0.71	0.52
null	0.16	1.00	0.00	0.27

D.5: Feature Selection

We seek to further improve performance using feature selection. Although the models are constructed with only 11 variables, less factors improve the scalability of a model in a live environment and make constructing accurate predictions in real time easier. The less information is needed about a visit to accurately predict purchasing intent, the faster the system can respond with customized content.

We use filter-based feature selection to avoid the need to use a learning algorithm that produces features specific to a classifier.[6] We rank the 11 features using mutual information (MI) and minimum redundancy feature selection (mRMR) and train our best performing model from undersampling, the classification tree, using different subsets of the top ranked features according to mRMR. We see no performance improvement using feature selection.

Table 6: Feature Selection

Ranking	Mutual Information	mRMR
1	PageValues	PageValues
2	Administrative_Duration	ExitRates
3	ExitRates	ProductRelated_Duration
4	ProductRelated	BounceRates
5	Browser	ProductRelated
6	Month	Month
7	TrafficType	Administrative
8	OperatingSystems	Administrative_Duration
9	Region	TrafficType
10	VisitorType	Informational

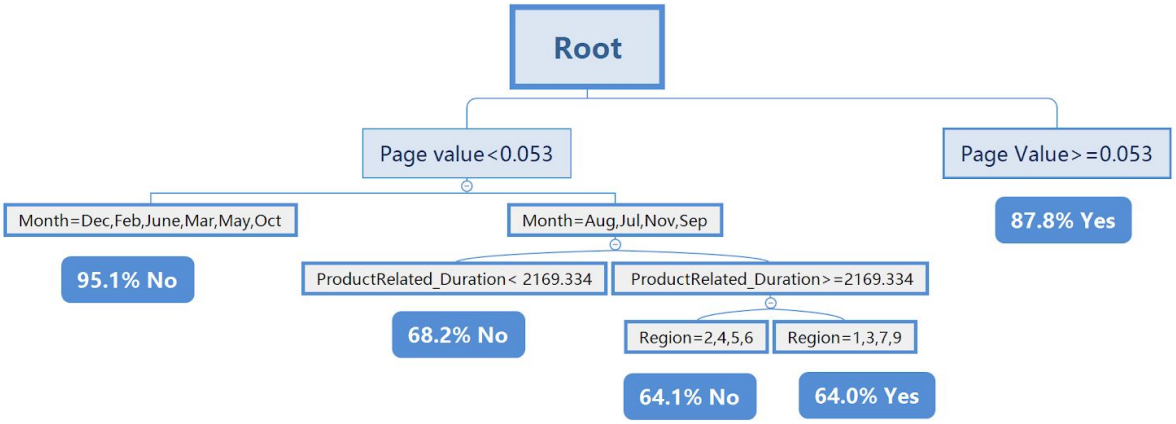
Final Model

We train the tree using undersampling with the whole data set. The final model is displayed in Figure 3. Page value is an important factor in predicting intent. Page value is a measure of how often a page is associated with conversion. It is a way to assign monetary performance to individual pages on a site by attributing a dollar amount to pages that appear in a transaction.[7]

[6] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). https://doi.org/10.1007/s00521-018-3523-0

[7] https://support.google.com/analytics/answer/2695658?hl=en

Figure 3: Final Classification Tree



Deployment

Our model has real world implications for online retailers. As Peretz mentions, discerning intent "should be central to the process of building a website, and particularly its conversion funnel." [8] In the age of hyper customization, retailers should factor purchase intent into how customers are served online. For customers who are inclined to buy in a visit, it is "important to structure the website in a way that facilitates [the] end goal seamlessly and intuitively."[9] This means that the user experience should be constructed in a way that minimizes friction and distractions, maximizing the likelihood of conversion. For visitors with no purchase intent, content tailored to optimize other metrics should be offered.

In our supervised binary classification problem, we evaluate a number of classifiers. We first train our models on an unbalanced data set. We then seek improvement by oversampling and undersampling to construct balanced models, which outputs a decision tree trained with undersampling as the model with the highest F1 score. Different subsets of features, ranked using mRMR, are used to try to improve the tree, without success . Our final model is trained with undersampling using all the data.

Similar models have been shown to be easily deployed online for retailers to make accurate predictions in real time. The models are run on a server instance, which is called whenever new predictions need to be made. The value of the features used for predictions "can be stored in the

[8] https://www.instantsearchplus.com/understanding-shoppers-intent-maximize-ecommerce-conversion/
[9] https://www.instantsearchplus.com/understanding-shoppers-intent-maximize-ecommerce-conversion/

application database for all web pages of the e-commerce site in the developed system and updated automatically at regular intervals."[10]

The challenges of deploying a real time predictive model online include data collection and data processing. Precise code tracking needs to be implemented in order for a model to spit out accurate predictions using correct data. Moreover, real time processing of session data, in which there are likely millions of records, requires a robust and scalable infrastructure.[11] Finally, if a retailer decides to completely overhaul its web site, it might need to retrain the classifier.

We evaluate models using TPR, TNR, and F1. We do not consider False Positives i.e. clients who are predicted to purchase but do not actually purchase and True Negatives i.e. clients who are not predicted to purchase but end up purchasing. Hence one of the potential risks to the model is that it might not be optimal if the cost of false positives are high or true negative are more desirable.

Moreover, we do not go into the exact content that would be optimal to show to a purchaser vs a non purchaser. We merely stress the need to show different content, which is a shortcoming of this study. We believe, however, that the best content to display will depend on the retailer and should be selected using A/B testing.

Cost-Benefit Matrix

For now, we don't have any data about the value of customers and the cost of customizing content for the buyers. Stepping forward, we can consider collecting these data and constructing a cost-benefit matrix to decide whether or not to provide a certain customer customized content. For example, we can construct a business model like this:

E[benefit lift|x,content]=[P(buy|x, content)-P(buy|x, origin content)]*V(buy,x)-C(content)

where x is the features of a certain visit.

If the E[benefit lift|x,content] is larger than 0, we can consider providing the customer the customized content.

[10] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). https://doi.org/10.1007/s00521-018-3523-0

[11] https://conferences.oreilly.com/strata/big-data-conference-sg-2015/public/schedule/detail/45045