Online shoppers purchasing intention prediction

Serena Wong, **P**atryk Kozak, **E**mma Lacourarie, **R**uta Czaplinska, **M**ikolaj Kacki

December 2019

Abstract

This project aims to predict purchasing intention of a visitor while scrolling a website of e-commerce company.

Key words: e-commerce, AdaBoost, support vector machines, feature engineering

CONTENTS

1 AIMS OF PROJECT AND ANALYSIS

1.1 AIMS

The role of online shopping in everyday life and the global market is growing rapidly from one year to another. Due to the increasing availability of the internet worldwide, e-commerce businesses can gather large amounts of data. Consequently, it creates an increasing need to effectively analyze it and make relevant conclusions to help businesses reach more clients and maximize their revenue. This project aims to create a model that would enable those companies to track whether a client is likely to end up making a transaction, so that companies may tailor their marketing approaches to a particular client. The strategies that could be implemented by companies using this model include customised suggested products, and would not be limited to promotions targeted at certain shoppers. Information about the browser used by the shopper, how they got to the website and whether they ended up making a purchase can make the website owners more informed [1] about, and then target the shoppers' overall experience. According to Rajamma et al [2]. the perceived transaction inconvenience is the major predictor of shopping cart abandonment, followed by risk and waiting time. Therefore, looking at the product more holistically, rather focusing solely on tailor price cuts should allow for much better results in terms of providing the best online shopping experience.

1.2 ANALYSIS

This data set came from the machine learning repository of University of California, Irvine and consists of 12 330 observations (therefore 12 330 sessions of internet surfing) followed by 17 categorical and numerical features. The response variable is categorical and obtains True if surfing the site by customer was finalized with a transaction, and False otherwise. By analyzing the data, it was found that there were no missing variables, meaning the data set did not have to be cleaned. We have also gained an overview of the data set by producing some descriptive statistics [3].

Table 1.1: Numerical features derived from the URL information.

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

Table 1.2: Categorical features.

Operating System	Operating system of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
Traffic Type	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)
Visitor Type	Visitor type as $NewVisitor$, $ReturningVisitor$ 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

Table 1.3: Numerical features measured by "Google Analytics".

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

For clarification purposes, in website traffic analysis, bounce rate relates to the percentage of visitors to a particular website who navigate away from the site after viewing only one page and hence not triggering any other requests to the analytics server during that session. Exit rate is the percentage of visitors to a page on the website from which they exit the website to a different website.

	Revenue	PageValues	ProductRelated	ProductRelated_Duration	Administrative	Informational	Administrative_Duration	Informational_Duration	Weekend	Browser
Revenue	1	0.49	0.16	0.15	0.14	0.095	0.094	0.07	0.029	0.024
PageValues	0.49	1	0.056	0.053	0.099	0.049	0.068	0.031	0.012	0.046
ProductRelated	0.16	0.056	1	0.86	0.43	0.37	0.29	0.28	0.016	-0.013
ProductRelated_Duration	0.15	0.053	0.86	1	0.37	0.39	0.36	0.35	0.0073	-0.0074
Administrative	0.14	0.099	0.43	0.37	1	0.38	0.6	0.26	0.026	-0.025
Informational	0.095	0.049	0.37	0.39	0.38	1	0.3	0.62	0.036	-0.038
Administrative_Duration	0.094	0.068	0.29	0.36	0.6	0.3	1	0.24	0.015	-0.015
Informational_Duration	0.07	0.031	0.28	0.35	0.26	0.62	0.24	1	0.024	-0.019
Weekend	0.029	0.012	0.016	0.0073	0.026	0.036	0.015	0.024	1	-0.04
Browser	0.024	0.046	-0.013	-0.0074	-0.025	-0.038	-0.015	-0.019	-0.04	1

1

Figure 1.1: Correlation between URL numerical features.

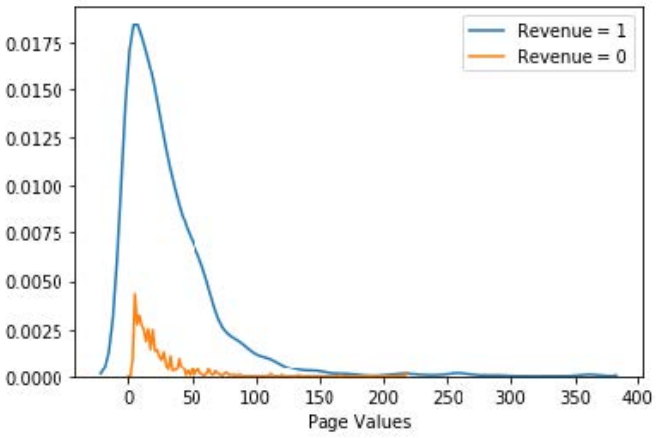

Figure 1.2: Page values.

Figure 1.3: Bounce rates.

1.3 Feature Engineering

To improve the model, new variables were created from pre-existing ones. Three variables representing average time spent per visit of respectively administration, informational and product-related page were created. Through manual analysis, 'SpecialDay' was not labeled as an important variable. However, the Darwin software indicated that it was important [4]. Therefore, the 'SpecialDay' feature was combined with the 'Weekend' to create a new feature called 'SpecialWeekend', which was labeled 1 when special day happened to be during weekend and 0 otherwise.

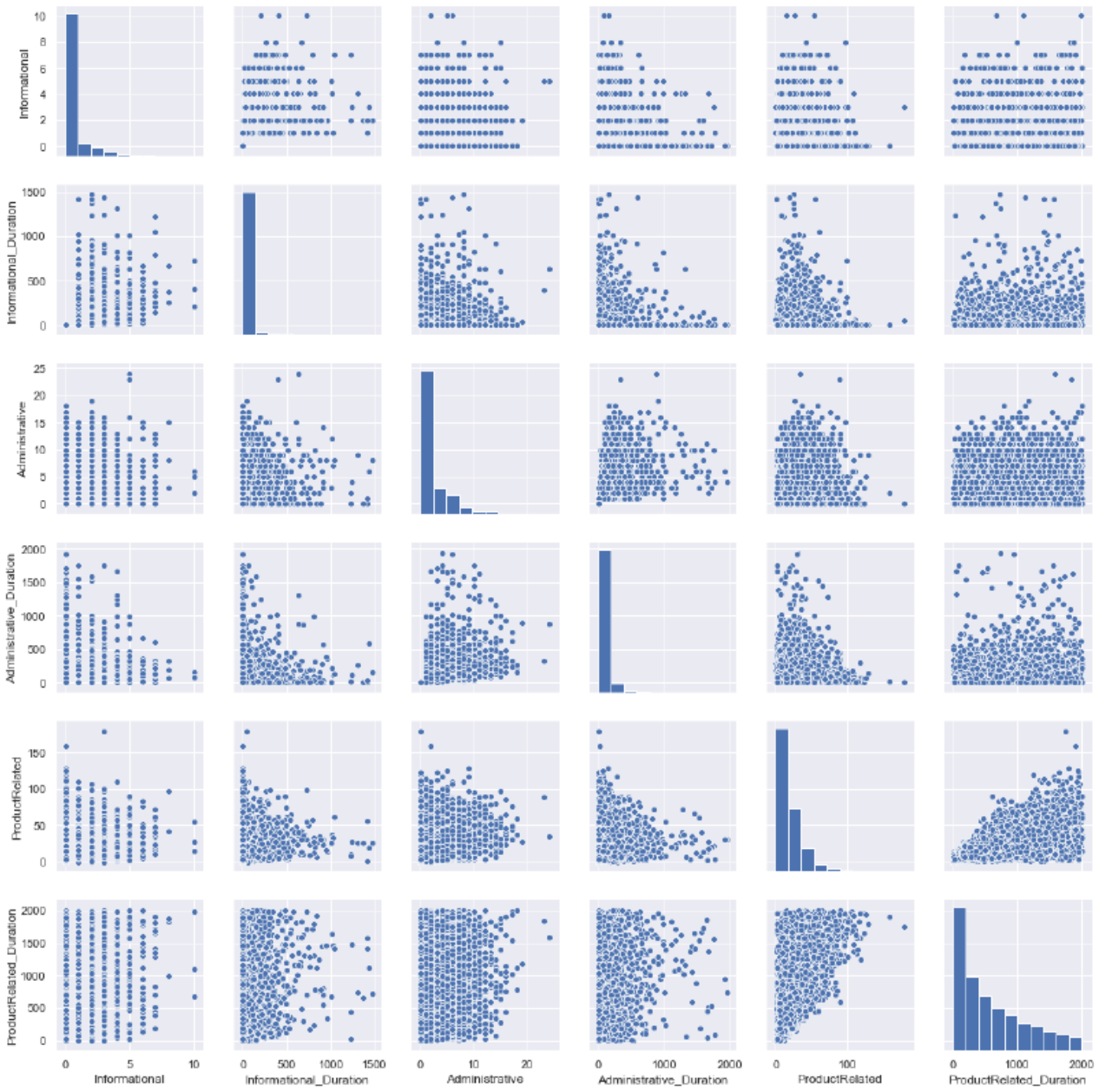

Figure 1.4: Correlation between URL numerical features.

2 Modelling

2.1 Modelling in Python

In Python, five classification algorithms were trained. All five of these gave different accuracy for the problem. The selection of algorithm was based partially on paper [3], and the feedforward neural network proved to be a viable solution for classification problems. The neural network used has two hidden layers of size 16 with a rectified linear unit activation function. The output layer of size 1 uses a sigmoid activation function. AdaBoost turned out to be the best, after training significant outliers were deleted because the algorithm performs worse with noisy data [5].

Table 2.1: Numerical features derived from the URL information.

Learning method	AdaBoost	Support Vector Machine	Random Forest	Multilayer Perceptron	Deep neural network
Mean accuracy	0.916	0.869	0.51	0.802	0.901
Standard deviation	0.0093	0.000312	0.06	0.065	

2.2 Modelling with Darwin

The data was uploaded into the Darwin user interface to create a model [4]. Numeric and categoric types are assigned to the features and engineered features accordingly, and the problem was set to a classification problem. The option for time series set to "no", while the parameters are set to an F1 fitness function and a k = 10 value for the K-Fold Cross-Validation. This model was then trained for 1 hour. The resulting outcome is accuracy of 92.09%, with the choice of the Deep Net algorithm.

3 Results and Discussion

All in all, we managed to develop a model having almost the same accuracy as the one derived from Darwin. With this model, e-commerce websites would be able to predict if the user is likely to leave the website in real-time. Based on the likelihood of whether a transaction will be finalised, website owners can deploy different sales technique that may potentially interest users in making a purchase.

An interesting characteristic of Darwin is that it breaks categorical features into its respected categories when ranking feature importance. For instance, the 'Browser' feature is broken into 'Browser = 1', 'Browser = 2' and so on when ranked. This produces insights for businesses as to which browser used by customers is more likely to result in them finalising a purchase. Businesses can then use this insight to draw conclusions on the cause of this and improve their e-commerce services. For example, their website might be more compatible with a certain type of browser, resulting in more purchases by customers using that particular browser. Businesses could then tailor their online stores for the other browsers to improve user accessibility and navigation.

References

[1] C. Carmona, S. RamÃŋrez-Gallego, F. Torres, E. Bernal, M. del Jesus, and S. GarcÃŋa, "Web usage mining to improve the design of an e-commerce website: Orolivesur.com," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 243 – 11 249, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412005696

[2] R. K. Rajamma, A. K. Paswan, and M. M. Hossain, "Engineering management design," *Journal of Product and Brand Management*.

[3] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks," *Neural Computing and Applications*, pp. 1–16, 2018.

[4] A. M.Shafiee, "DARWIN AI," AI Software (2019).

[5] S. Joglekar. (2016) *A (small) Introduction to Boosting*. [Online]. Available: https://codesachin.wordpress.com/tag/adaboost/?fbclid=IwAR1u2Dn76PzF3tx5E0zWvKkApWTa5z4Oy2GiK1Fntr0sNZLEweRt{_}EiiIbI