|  |  |
| --- | --- |
| Predicting loyalty in e-commerce | Abstract  Can we predict an e-commerce user’s loyalty? The major objective of this study is to develop a model that can use signals from a user’s actions on a website to predict whether they will return to the website in 15 days. Actions that will be used include what a user is clicking on, their screen size, and the sequence they perform these actions. The research showed that there are signals to predicting “loyalty”, but a more robust dataset would be needed to improve performance.  Baron Curtin  Data 698 – Professor Khansari |

Table of Contents

[Project Proposal 2](#_Toc36059087)

[Introduction 2](#_Toc36059088)

[Data Source 2](#_Toc36059089)

[Background 2](#_Toc36059090)

[Hypothesis 3](#_Toc36059091)

[Literature Review 3](#_Toc36059092)

[Introduction 3](#_Toc36059093)

[Definition of Loyalty 4](#_Toc36059094)

[Predictors 5](#_Toc36059095)

[Previous Methods 6](#_Toc36059096)

[Methodology 8](#_Toc36059097)

[About the Data 8](#_Toc36059098)

[About the Algorithm 9](#_Toc36059099)

[References 10](#_Toc36059100)

# Project Proposal

## Introduction

*Data Science Company* currently uses machine learning to be able to predict consumer behavior in real-time when they visit e-commerce websites. As the company looks to expand and grow, it has come to light that the predictions made currently do not “fit” as well to other industries as it does the industry the model is currently trained to perform in, but the current models perform well enough that they are suitable in the interim. As it is, there are currently ideas to create new predictions that are better suited for scaling into new industries, with the ultimate hope that the new model can be agnostic across industries.

## Data Source

*Data Science Company*, together with its partners, have a wealth of information and user log data that can be used for the purposes of the new prediction. The user level data available for the purpose of this project amounts to over 180 TB (compressed). Information such as what the user purchased, when they purchased, and how much the purchase was are just some of the fields available.

## Background

In this project, I will attempt to use the existing data and create a “loyalty score” for each user. “Loyalty Score” is currently not something that is defined but I have identified a few ways this new metric could be defined:

* How long the user will remain a customer?
  + This could also then be defined as how many purchases per year a user will make?
* How many purchases will a customer make on a visit?
* What is a user’s customer lifetime value?
* Churn rate?
  + Will need to address how much time can pass between visits before a customer is considered a churner?
* Will a customer come back to the site (“Repeat Visit Score”)?
  + Time frame will need to be established, (e.g. within the next 30 days)

One of the methods above will need to be chosen for the purpose of this project. The reason this problem is interesting is because if a more industry-agnostic prediction can be developed, it will be easier for *Data Science Company* to expand and grow, as it will now be able to sell the service to a much larger customer-base. The revenue potential of the product skyrockets if the machine learning platform has wider applications.

## Hypothesis

The new metric “loyalty score”, however it comes to be defined, will be a better performing model in the new industry *Data Science Company* is currently testing, but also perform just as well or have a negligible difference in performance on the current industry the model is trained on. Customers who are searching and looking to make a new purchase, have some inherently common behaviors regardless of the product they are looking for. Those common behaviors amongst customers are what will make the model performant across industries.

# Literature Review

## Introduction

What drives internet users to visit, to buy, and to repeat those actions? More importantly, how to predict which action the users takes and when? Today’s race for clicks, views, mentions, and purchases has created a market for machine-learning based predictions in this space. *Data Science Company* currently predicts whether a user will convert (make a purchase), on any site visit within the travel industry. The need to grow and expand into other industries has raised the questions: “Are we answering the best question, with our prediction, as it pertains to a different industry and is it also possible to make a better prediction that spans multiple industries?” Currently, *Data Science Company* makes a prediction for each user, for each site, for each page, whether the user will purchase or not. To transcend that prediction in the travel industry, this research project will look to examine two things: how loyalty can be defined in an online context and if loyalty is a better predictor of user value than just whether a not a user will generate a sale. This literature review, as a result, will examine how customer loyalty can be defined in an online/cyber context, some potential predictors for user intent, and some of the methods that have been applied to the prediction of a user’s web-browsing behavior.

## Definition of Loyalty

In a traditional environment, loyalty is defined as “a strong feeling of support or allegiance”, but there is reason to believe that the traditional definition is not so easily applied to the cyber-universe. It has previously been hypothesized that *trust* and more importantly, the three precursors to *trust*: comprehensive information, shared value, and communication are the most important variables in defining customer loyalty on the world wide web [1]. *Comprehensive information* is defined as the extent to which a customer has enough information to make a purchase decision, *shared value* is defined as the extent to which customers have beliefs in common about the type of behaviors, goals, and policies that are important and appropriate, and *communication* can be defined as the formal as well as the informal sharing of meaningful and timely information between buyers and sellers [1]. It has been concluded from a study of web users in Internet stores that when customers have more information and more detailed and thorough information about the products in the store, if there is a common thread they share with other patrons, and if they have more diverse means of communication, customers would feel a higher level of trust in the store [1]. The ideas are hard to quantify and hard to capture in data but for *Data Science Company*, some of these ideas could manifest themselves in the amount of search terms a user inputs, the kinds of users that are visiting, and where the users are coming from. If data is available, these will be evaluated as part of the feature set. A different paper tested the ideas of recommendation and retention and concluded that recommendation could be predicted by a patrons preconceived attitude towards a [site], while retention could be predicted more by repeat patronage than attitude [2]. This means that users with repeat visits are more likely to have a higher loyalty. With the data available at *Data Science Company*, loyalty scores would be higher on subsequent visits but there must be other features that can predict loyalty on the first visit. From these papers, loyalty appears to be determined by an idea of trust stemming from comprehensive information, communication of information, shared values across users, and repeat patronage.

## Predictors

The internet contains a wealth of information and sometimes users are not ready to act, or in this case purchase, but rather they want to conduct research. One study that aimed to predict whether a user was performing exploratory research or intending to purchase concluded that users with intent to purchased had a higher ad clickthrough rate than those with research intents [3]. Conclusions to draw and ideas to keep in mind from this study are users that click more will generally have a higher intent to purchase, what a user clicks on could be a key predictor, and what context the user is clicking could also be significant. The authors of the paper also state that they plan to expand the model to consider user interactions and page context in future studies [3]. This could mean where a user comes from, what their journey is, is also important in being able to predict research versus purchase intent. Another study aims at analyzing mobile commerce intent concluded that performance and effort expectancies, social influence, and trust in mobile commerce had a significant predictive impact on purchase intentions [4]. Performance and effort expectancies seems to refer to a user’s belief or preconceived notion that a website will effectively handle or get the task completed relative to the amount of energy for the undertaking. If a user wants to buy a pair of sneakers, will the user be able to easily get what they want? To extend that idea to this project, the idea of a browsing session would need to be created and relevant metrics such as clicks per session, session duration, clicks per 10 seconds or some other duration, and actions per 10 seconds. Final metrics could be determined based on looking at the data to see how long on average a typical session last. In a separate study, researchers concluded that web site investment influenced a user’s purchase intentions and that web site investment was a signal of a firm’s ability [5]. It will be hard to measure the investment a particular company invests in their website and that may continue to be a limitation of this research. The question is still worth answering and if there are features that can be created that can signify investment, it would help to create a better feature set for training purposes. Another study looked to predict online purchase conversion using web path analysis. Results showed that more promotional messages on a page, removing the presence of price information on a page, and reducing the number of hypertext links can positively affect the purchase conversion rate of users who are surfing but those same changes can negatively impact visitors who are purchase oriented [6]. These learnings can be effectively used in a model for training. Finding features that encompass the number of elements on a page, whether the price is on a page, and whether promotions or discounts are on a page can lead to potentially significant predictive power. The other big conclusion reached from this study was that the sequence information of a user’s journey through a website doubled the fit of the model [6]. This demonstrates the importance of having features that handle not only what a user viewed, but when they viewed and the sequence they viewed.

## Previous Methods

While previous studies may have confirmed the significance of a user’s journey through a website, another study trained a classification algorithm using a modified Markov model that took the sequencing of a user’s actions into account. This study utilizing the modified Markov model built upon the work done in a previous study that confirmed that a Hidden Markov Model is able to reliably predict user intent [7]. The researchers further concluded that the poor precision and recall results could be improved upon [7]. With a recall score of 73%, the Hidden Markov Model was correctly classifying approximately 3 out of 4 converters. At that recall rate, sequencing should be taken into consideration for this project as it would definitely be a boon for the business. In the subsequent study, a session (a user’s actions over a set duration of time), was depicted as an n-gram [8]. The researchers in the study also developed a modified Markov Model reduces the complexity of the original Markov model while achieving comparable prediction accuracy [8]. The research were also able to conclude that a smaller number of n-gram models perform better than higher number of n-gram models because of the lack of sessions generated with the larger number of n-grams [8]. This idea can apply to this project even without the use of the Markov model given that it means that certain actions will innately be linked. Creating those links as features should improve prediction accuracy. In another study, researchers looked at the idea that the second most important quality for a successful e-commerce site was “perceived latency” [9]. Their idea was to create a model that could predict what a user’s next HTTP request was, so that the predicted page could be pre-cached, thereby reducing perceived latency. The researchers used a rule extraction algorithm and a rule selection algorithm to integrate purchase patterns and path traversal patterns to improve the accuracy of prediction on an e-commerce site [9]. For improvement to the model and application to an e-commerce site, the researchers had to alter their approach slightly from traditional models because of the purchase aspect of e-commerce sites. For the purposes of this research project, the end result of using the predictions to pre-load content is not as relevant as the methods applied. A separate study further reinforces the idea of capturing purchase patterns such as the sequencing of user actions [10]. Their two-pronged approach tackled modelling by extracting purchase patterns, and then predicting purchase probability [10]. The study is also highly applicable to *Data Science Company* as the userbase was completely anonymous. The researchers had no prior information about the users upon web visit, and all of the information gathered was during the session or extracted from the session. It becomes clear through the various studies that capturing the sequence of events in the model will improve accuracy and predictive power. Another study decided to approach predicting purchase behavior from a “task-completion” perspective [11]. The researchers identified three tasks that must be completed for a purchase to take place: completion of product configuration, input of complete personal information, and order confirmation with provision of credit card data [11]. With each task being conditioning on the preceding one, the model wasn’t forced to predict rare events [11]. This may be applicable to this project. Currently, *Data Science Company* predicts the rare events but because of the size of the training data, *Data Science Company* is able to attain respectable accuracy rates. Using a task-completion approach should yield better measures of predictive accuracy based on the study. An interesting conclusion also made in the study stayed that the researchers did not find a relationship between repeat site visitation, a measure of loyalty defined previously, and a successful purchase [11]. The study was not conducted in an e-commerce space, but the lack of relationship is noted. It would still be best to include a measure of loyalty and repeat visitation for the purposes of this project.

# Methodology

## About the Data

The data to be used in this study will be the data collected by *Data Science Company*. *Data Science Company* currently collects user log data when users visit a site that is configured with its own proprietary JavaScript tags. The information collected on each user includers, but is not limited to, what site they visiting, when they visited, what they clicked on, and whether they purchased an item. The information sources are contained in separate tables within a single database. To create the training set, data will have to be sourced from the separate tables and joined together. The user’s journey through a website will also require a lot of manipulation to be put in a format usable for the training set. Currently, a user’s journey is logged as various events where each event is a separate row of data.

## About the Algorithm

The algorithm to be used for this project will be logistic regression regularized with elastic net. *Data Science Company* currently produces predictions via logistic regression and there is a current edict to maintain the use of logistic regression because it is cost-effective and produces reliably accurate results. The validation set could be used to test other algorithms but the primary focus will be the success of the logistic regression. The algorithm itself outputs a probability. In terms in this project, the goal is to be able to predict whether a user will be loyal to the site. For the purposes of this study, it may be best to try and create multiple predictions and then see which one provides the most value and makes sense in the business context. Loyalty will be defined as the probability a user will return to the site within 30 days, the probability a user will return to the site within 60 days, the probability a user will return to the site within 90 days, the probability a user will spend longer than the average user on the site for a session, or the probability that a user is doing research as opposed to having a purchase intention.

# Process

Before continuing, I must caveat the proceeding with the fact that due to the coronavirus, I was not able to get access to the data that I had originally planned on using. The original data would have used a company that was more directly associated with retail and e-commerce. As a substitute, the data that I am performing the methodology on was from a travel website that primarily sells flights to students. Aside from those differences, the methodology will remain the same and II will be using a user’s actions on the website, along with other features to predict if a user will revisit the same site within 15, 30, 45, and 60-day intervals. That will be how loyalty is determined in this analysis. Without the resources available to me that I would have had prior to the pandemic, I was forced to limit the data to users utilizing Google Chrome as their web browser, users on desktops only, and users using Windows operating system. The final resulting dataset was about 2 million rows, down from around 200 million rows. It is hypothesized that this reduction in sample size may affect model performance.

## Exploratory Analysis

### Numerical EDA

Exploratory analysis began with the numerical features of the data set. These features include:

* documentHeight/documentWidth: the height and width of the website in pixels
* deviceVentileGroup: proprietary prediction made by *Data Science Company* on a user’s likelihood to make a purchase
* cookieSize: how large a user’s cookie is or how much tracking data is available on the user
* screenPixelDepth: how many pixels a user’s screen can display at a time
* screenHeight/screenWidth: the height and width of a user’s screen
* daysFromTravel: how many days prior to the users travel date the flight is being purchased
* travelDuration: how long the user is travelling for
* travelers: how many people are accounted for in the flight purchase
* viewportHeight/viewportWidth: the height and width of the window the user is viewing the website from

The information on conversions or flight purchases was very sparse. Less than 1% of the users made a flight purchase. As a result, all of the conversion information may have to be dropped from the dataset. Under normal circumstances, the conversion rate would not be an issue with a more robust dataset as there would be enough signals drawn from the larger sample of converters. With only 2 million rows in this dataset, it would be best to remove the conversion information. To that effect, I will be removing all of the conversion information outside of a Boolean value that indicates whether or not a purchase was made. The hope is the feature would provide some signal information to the model as well as that it makes business sense to include this feature. Numerical EDA also revealed very weak correlations across the numerical features. The strongest correlated features were *documentHeight*, *deviceVentileGroup*, *cookieSize*, and *screenPixelDepth* ranging from ~.05 - .06. *cookieSize* does suffer from a lot of missing values and will have to be removed.

### Categorical EDA

The categorical features include:

* campaignReferrer: where a user came from if coming from a link from another site
* pageId: what page a user was viewing
* productPath: what product the user was viewing on the website. Users only viewed Flights and Hotels
* publisherUserId: the unique identifier for a user
* segmentationType: is a categorical label based on deviceVentileGroup. Will be removed because there was only one label and wouldn’t provide any value to the model
* customDimensions.label: represents the element a user interacted with on the website. Will be dropped in favor of a bigram feature of the elements a user interacted with
* tripType: specifies round trip or one-way flight
* nextLabel: the next element a user interacted with in a session
* bigramLabel: a concatenation of customDimensions.label and nextLabel to create a pseudo-sequencing feature. This is the feature that will handle sequencing as opposed to using a markov model

For exploratory purposes, four labels were created for prediction. The categorical target labels include:

* returnIn15Days
* returnIn30Days
* returnIn45Days
* returnIn60Days

As a result of exploration and using business logic, the final feature set includes: browserRequest.screenHeight, browserRequest.screenPixelDepth, browserRequest.screenWidth, campaignReferrer, deviceVentileGroup, pageId, productPath, bigramLabel, converted. It isn’t a very robust dataset but there will be some manipulations that will be done prior to training the model. It should be noted that some of these features were already engineered prior to this point such as bigramLabel.

## Data Processing

Since the main feature is the *bigramLabel* feature and a vast majority of the column contained missing values, I decided to remove all those rows where *bigramLabel* was NA or missing. All of the other missing values were imputed using the KNNImputer class of scikit-learn. The categorical features of *campaignReferrer*, *pageId*, and *productPath* were imputed with the most frequently occurring values and then one-hot encoded. *bigramLabel* was also one-hot encoded for input into the model.Numerical values were standardized.

## Model Performance & Analysis

I ended up using a stochastic gradient descent logistic regression model. I was not able to use the standard logistic regression model because it failed to converge. Below are the highlights of the model’s performance:

* Training Set
  + 5-fold cross-validation ROC/AUC Score: .58
* Test Set
  + Metrics
    - ROC/AUC Score: .52
    - Precision: .51
    - Recall: .14
    - F1 Score: .22
  + Confusion Matrix
    - True Positives: 115,215
    - False Positives: 12,832
    - False Negatives: 81,722
    - True Negatives: 13,277

Although the initial ROC/AUC score is .52 on the test set, which indicates there is some signal or loyalty, this is barely better than a coin flip. Deploying this model into production probably would not be worth the cost of operation. When taking a closer look at the precision and recall figures, the model is only able to classify 51% of the positive predictions it made and was only able to correctly classify 14% of all the total positive predictions. This is a really egregious indictment against the model. The F1 score further confirms the underlying precision and recall figures, but it must still be noted that the AUC is still indicative of some predictive power of the signals in the model.

# Conclusions

Although a deeper analysis of the results was discouraging, there is some value in the model and the signals that went into the model. As a result of the current pandemic, computing resources were not as available as they would have been prior. To that effect, a lot of data had to be discarded without real prejudice, especially rows with missing values. Columns that were used in the production model also had to be removed because of the lack of computing power. The model at *Data Science Company* is able to generate results with an AUC of ~.68 with a more robust dataset. I suspect that with more computing power, a training dataset with millions more rows could have been used. The dataset used in this study was about 900,000 rows.

# Future Work

The work done here could be improved by increasing the computing power, allowing for a much larger dataset with both more rows and columns. This study took a naïve approach and simply concatenated two actions together to create a feature. It may be beneficial to look into some of the algorithms more specialized in sequencing. This study also lacked a true refinement of hyperparameters as it was again limited by compute power. Taking the time to refine the hyperparameters could have yielded better performance and should be done in future work. Lastly, this study used a training set where every row represented a specific action in a user’s browser session. In a future study, it may be beneficial to aggregate all of a session’s actions into one row and approach the data from a session level instead of an action level.

# References

|  |  |
| --- | --- |
| [1] | J. Lee, J. Kim and J. Y. Moon, "What makes Internet users visit cyber stores again? Key design factors for customer loyalty," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2000. |
| [2] | R. East, P. Gendall, K. Hammond and W. Lomax, "Consumer loyalty: singular, additive or interactive?," *Australasian Marketing Journal (AMJ),* vol. 13, pp. 10-26, 2005. |
| [3] | Q. Guo and E. Agichtein, "Ready to buy or just browsing? Detecting web searcher goals from interaction data," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010. |
| [4] | R. Blaise, M. Halloran and M. Muchnick, "Mobile commerce competitive advantage: A quantitative study of variables that predict m-commerce purchase intentions," *Journal of Internet Commerce,* vol. 17, pp. 96-114, 2018. |
| [5] | A. E. Schlossers, T. B. White and S. M. Lloyd, "Converting web site visitors into buyers: how web site investment increases consumer trusting beliefs and online purchase intentions," *Journal of marketing,* vol. 70, no. 2, pp. 133-148, 2006. |
| [6] | A. L. Montgomery, S. Li, K. Srinivasan and J. C. Liechty, "Predicting online purchase conversion using web path analysis," *Marketing Science,* vol. 23, pp. 579-595, 2004. |
| [7] | C.-J. Lin, F. Wu and I.-H. Chiu, "Using hidden markov model to predict the surfing users intention of cyber purchase on the Web," *J. Global Business Management (JGBM),* vol. 5, 2009. |
| [8] | M. A. Awad and I. Khalil, "Prediction of users web-browsing behavior: Application of markov model," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 42, pp. 1131-1142, 2012. |
| [9] | S. Vallamkondu and L. Gruenwald, "Integrating purchase patterns and traversal patterns to predict http requests in e-commerce sites," in *EEE International Conference on E-Commerce, 2003. CEC 2003.*, 2003. |
| [10] | E. Suh, S. Lim, H. Hwang and S. Kim, "A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study," *Expert Systems with Applications,* vol. 27, pp. 245-255, 2004. |
| [11] | C. Sismeiro and R. E. Bucklin, "Modeling purchase behavior at an e-commerce web site: A task-completion approach," *Journal of marketing research,* vol. 41, pp. 306-323, 2004. |