International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

The Role of Data Analytics in Predicting Customer Churn in E-Commerce

Karan Salot (1), *Juhi Khare* (2)

Dwarkadas J. Sanghvi College of Engineering
Email: salotkaran@gmail.com, juhi87khare@gmail.com

ABSTRACT:

In the fiercely competitive geography of e-commerce, understanding and mitigating client churn has come consummate for businesses seeking for sustained growth and profitability. This paper explores the vital part of data analytics in prognosticating and managing client churn in thee-commerce sector.

Client churn, the miracle where guests cease their engagement with an e-commerce platform, presents significant challenges and openings. Using data analytics ways offers a comprehensive approach to attack this issue. This paper reviews colorful data sources; including sale history, click stream data, client feedback, and demographic information, which can be exercised to make prophetic models for relating implicit churners.

The study delves into the methodologies and algorithms generally employed in prognosticating client churn. Machine literacy ways similar as logistic retrogression, decision trees, arbitrary timbers, and neural networks are bandied in the environment of e-commerce data. Likewise, the paper explores the use of advanced ways like natural language processing and deep literacy to prize perceptivity from unshaped data sources like client reviews and social media relations.

Also, this exploration discusses the challenges and ethical considerations associated with data analytics ine-commerce, including data sequestration enterprises and algorithmic impulses. It emphasizes the significance of transparent and responsible data practices in erecting client trust.

The benefits of effective client churn vaticination are illustrated; including reduced marketing costs, increased client retention, and enhanced client experience. Real- world case studies are examined to illustrate successful executions of data analytics- driven churn vaticination strategies.

Keywords: Customer Churn, E-commerce, Data Analytics, Predictive Analytics, Customer Retention, Machine Learning, Data Mining, Customer Behavior, Customer Segmentation, Big Data, Customer Engagement, Customer Satisfaction, Customer Lifetime Value (CLV)

1. INTRODUCTION

In the fiercely competitive geography of E-commerce, retaining guests and minimizing client churn is of consummate significance. Client churn, or client waste, refers to the miracle where guests stop copping from a business or switch to a contender. It's a significant challenge for E-commerce businesses as acquiring new guests is frequently more precious than retaining being bones To address this challenge effectively, data analytics plays a vital part in prognosticating and mollifying client churn.

Data analytics involves the methodical examination and interpretation of data to uncover meaningful perceptivity. When applied to E-commerce, it can give inestimable information about client geste, preferences, and the factors that impact churn. In this environment, understanding the part of data analytics in prognosticating client churn is pivotal for E-commerce businesses to enhance client retention and boost profitability.

In order to get favorable product, there's a chance of switching to other online spots from classic, which leads to loss of client. If this continues for a certain period of time it leads to client churn. Merchandising gives an approach to products to find a good pace to client. It endures snags like investment on labour, gratifying the logrolling client. e-retail give muddle free shopping to buyers, as having colorful preferences. To stay in super competitive online request, retailers should suffer proceedings like vindicating the identity of client, pious and transparent to guests, follow return and refund programs, securing client data.

A client likely to break the relationship or reduce the purchase rate is known as churn. Customer churn occurs when a client stops employing a retailer’s product, stops visiting a specific place of business, shift to lower- league experience or shift to the contender’s products. Retailers need an abiding strategy to manage client churn. Measuring the churn rate is kind of pivotal for retail businesses because the metric reflects client response towards the wares, service, price and competition.

Churn vaticination fantasize the liability of client to churn. It pares the investment on gaining new client and helps to retain the being client. The marketing sweats and quantum spends on attracting a new client is high and delicate than adhering to being client. guests who are doubtful to make a purchase or willing to shift the shopping point because of guardedness with plutocrat, awaiting standard and multifariousness in products can be induced and gripped.

Then is an overview of how data analytics is necessary in prognosticating client churn in E-commerce

1. Data Collection and Integration: E-commerce platforms induce vast quantities of data every day, including sale history, client demographics, website relations, and more.

2. point Engineering Data judges and data scientists identify applicable features or variables within the dataset that can be reflective of churn. These may include purchase frequence, order history, client demographics, website engagement criteria , and client feedback.

3. Prophetic Modeling Data analytics ways similar as machine literacy and statistical analysis are employed to make prophetic models. These models can identify patterns and correlations within the data to prognosticate the liability of client churn.

4. Client Segmentation Data analytics can member guests grounded on their characteristics and actions. Segmentation helps in acclimatizing marketing strategies and retention sweats to specific client groups.

5. Sentiment Analysis Textual data from client reviews and feedback can be anatomized using natural language processing (NLP) ways. Sentiment analysis helps in gauging client satisfaction and relating negative sentiments that may indicate an increased threat of churn.

6. Real- time Monitoring Data analytics can be applied in real- time to cover client geste as it happens. Anomalies or unforeseen changes in geste can be detected instantly, allowing businesses to take immediate action to help churn.

7. Personalization and Recommendations By assaying literal client data, E-commerce platforms can give individualized recommendations and offers to individual guests. This not only enhances the client experience but also increases the liability of retaining guests.

8. Retention Strategies Data- driven perceptivity guide the expression of effective retention strategies. Businesses can apply targeted marketing juggernauts, fidelity programs, and visionary client service interventions grounded on the prognostications made by analytics models.

2. PROPOSED SYSTEM

We used gusto and machine literacy models to produce a single- runner web operation for client churn analysis. We conducted exploratory data analysis to identify missing values, categorical and numerical variables, and columns that have a high impact on client churn in recent times. Our dataset includes 5630 unique client IDs, and all columns with n = 5630 have no missing values. We also resolve the data into a 90 training dataset and a 10 test dataset. We trained four base learners Decision Trees, Random timbers, Support Vector Machines, and KNN classifiers. These models' labors were fed into the meta- classifier of the mounding Classifier, which used logistic retrogression.

We compared the vaticination performance of LR and SVM using three generally used performance pointers Accuracy, Recall, and Precision. still, we believe that client data in e-commerce enterprises is unique and requires individualized approaches. These enterprises frequently modernize product information or upload colorful evaluation information for client retention, which is different from fiscal and telecommunication client information. Therefore, we also considered the functional effectiveness of the models, especially when prognosticating guests in real- time. We set up that SVM's data training time is significantly shorter than that of LR when training client data. We used four base learners in our analysis. Decision Trees are unsupervised machine learning algorithms that can be used for bracket or retrogression. Logistic retrogression models the probability of a double outgrowth and is generally used in bracket problems. Random Forest is a machine literacy algorithm that combines the labors of multiple decision trees to overcome over befitting and bias issues. Eventually, Support Vector Machines are supervised machine learning algorithms that are generally used in bracket problems and have been shown to have good performance in client churn analysis.

3. SYSTEM ARCHITECHTURE

The main end of this exploration was to distinguish between guests who churned and those who stayed, and to determine the factors that contribute to churn. The findings of this study indicate that single manly guests are slightly more likely to churn. In addition, guests who prefer the Mobile order were set up to be further prone to churn. Also, churned guests showed a slightly advanced preference for using a phone or mobile device to log in, which could be due to the client experience handed by the E-commerce platform's phone interpretation. Also, the study linked that churned guests have a advanced mean for complaints, megacity league, number of addresses, and number of registered bias. still, unexpectedly, churned guests had a advanced satisfaction score compared to the retained guests. On the other hand, the term and the count of the number of orders were set up to be lower for churned guests, which is anticipated.

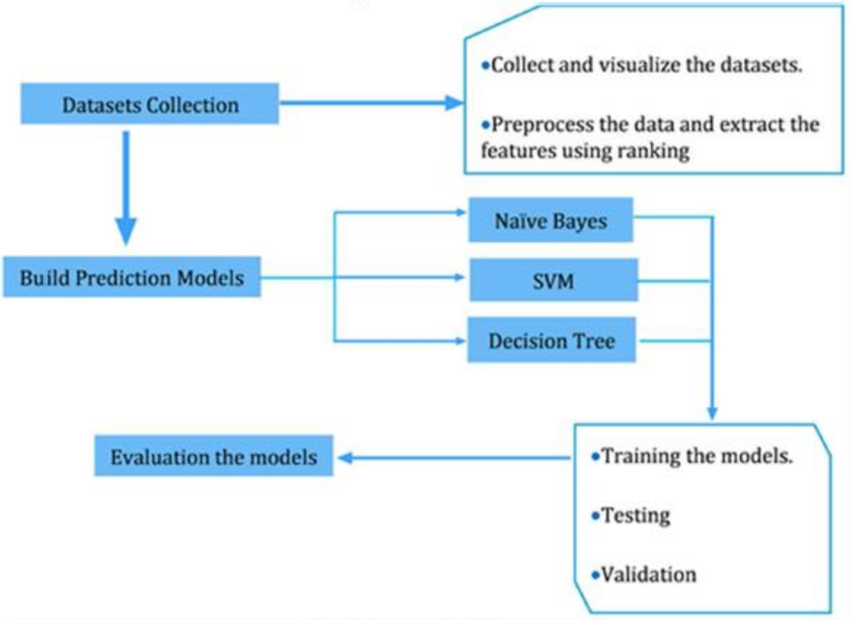
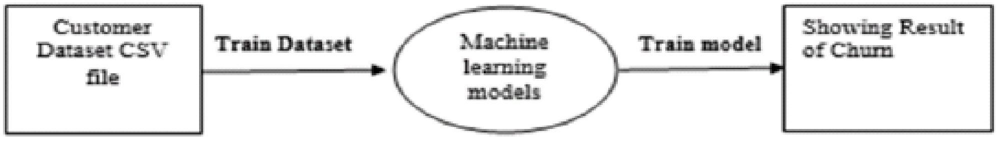


Fig -1: System Architecture

4. DATA FLOW DIAGRAM

The system illustration depicts an E-commerce platform as a blockish shape, which includes colorful factors similar as client Data, Data Pre-processing, Model Building, Trained Model, and Churn vaticination. The input data for the system is represented by the client Data element. The Data Pre-processing element includes four sub-components, which are Data Cleaning, Data Integration, Feature Engineering, and point Selection. The Model Building element includes two sub-components, which are Algorithm Selection and Hyper parameter Tuning. The Trained Model element represents the affair of the Model Building process. Eventually, the Churn Prediction element uses the Trained Model to make prognostications on the input data.



5. METHODS USED IN THE ANALYSIS

We tested the impact of colorful predictor variables on client churn. We applied machine literacy modeling, which requires the following way

1. . Pre-processing the variables present in the dataset so that they can be included in the model.
2. Defining the machine learning modeling styles to be used, in particular choice of the metric to be optimized and the type of model.
3. Training the model using colorful sets of variables, and the selection of independent variables which maximize the performance of the proposed model.
4. Running the prognostications from the named models.
5. The methodology used in this study can be divided into four broad orders
6. Styles used in pre-processing applied to the variables present in the dataset.
7. Styles used for variable selection.
8. Machine literacy modeling styles choice of model, cross-validation, up- slice, etc.
9. Styles used to check the strength of the variable’s influence.

Below we describe the details of these styles. Figure 1 presents an overview of the whole analysis.

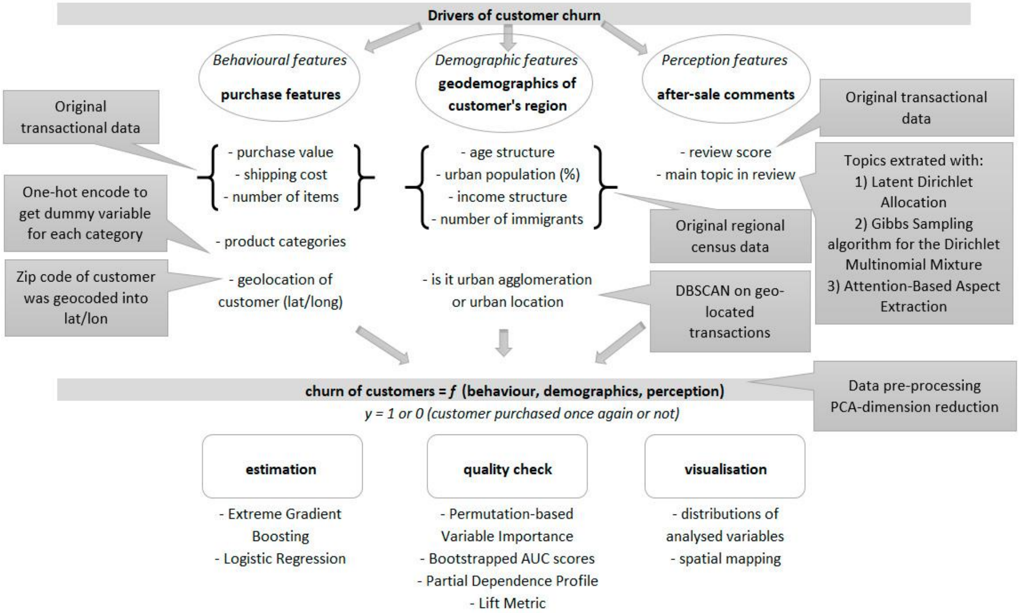
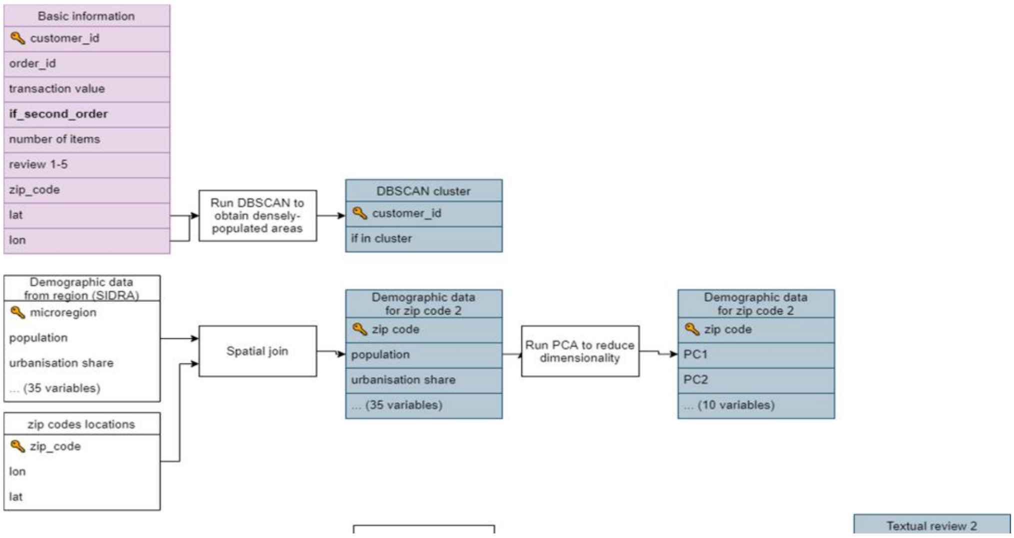

Drivers of customer churn
Behavioural features
purchase features
Demographic features
geodemographics of customer's region
Perception features
after-sale comments
Original transactional data
- purchase value
- shipping cost
- number of items
- age structure
- urban population (%)
- income structure
- number of immigrants
- review score
- main topic in review
Original transactional data
Topics extrated with:
1) Latent Dirichlet Allocation
2) Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture
3) Attention-Based Aspect Extraction
One-hot encode to get dummy variable for each category
- product categories
Original regional census data
Zip code of customer was geocoded into lat/lon
- geolocation of customer (lat/long)
- is it urban agglomeration or urban location
DBSCAN on geo-located transactions
churn of customers = ƒ (behaviour, demographics, perception)
y = 1 or 0 (customer purchased once again or not)
Data pre-processing
PCA-dimension reduction
estimation
quality check
visualisation
- Extreme Gradient Boosting
- Logistic Regression
- Permutation-based Variable Importance
- Bootstrapped AUC scores
- Partial Dependence Profile
- Lift Metric
- distributions of analysed variables
- spatial mapping

Figure 1. Flowchart of analytic steps.

5.1. DATA PRE-PROCESSING

Pre-processing, applied to the entire dataset corridor, can be structured as a flowchart (Figure 2). All of the tables on the left- hand side draw closer directly beginning list (4 tables) and Statistical Office sources (1 table — demographic data). The grandiloquent table is the primary one, and the features from this table were combined with all the remaining sets of variables. The final tables created after pre-processing each of the corridors of the dataset are shown in slate. In the modeling phase, we performed a simple join of the introductory table and the remaining tables (e.g., introductory information order particulars; introductory information DBSCAN clustered.). In this study, we pre-processed three separate groups of variables behavioral (first sale) features, position features and perception features, described in detail below.

Behavioral features of guests were deduced from the financial value of the first purchase, delivery cost, number of particulars bought and the orders of the bought particulars and were included in the model expression. The value of the purchase, as well as the product order, was of central interest, as they've been shown to be significant in other studies on client churn. The purchase value was directly fitted into the model as it didn't bear any pre-processing. In the case of the product order, two way were taken. First, some of the products were veritably rare in the dataset, and therefore were binned into one order because of implicit problems with generalization and slower model training. Secondly, this variable demanded to be converted to a numeric format. therefore, all product orders except the top 15 most popular bones (responsible for 80 of purchases) were binned as a new order “ other ”, also, one-hot encoding was utilised to produce a numeric representation, with the “ other ” order set as a base position. One should flash back that there can be multiple product orders in one order, so it wasn't guaranteed that there would be only one “ 1 ” entry per row, as in the classical one-hot- encoding system.



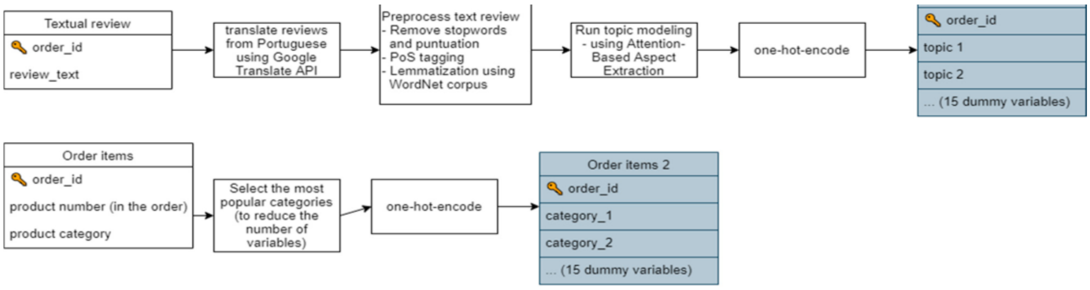

Textual review
order_id
review_text
translate reviews from Portuguese using Google Translate API
Preprocess text review
- Remove stopwords and puntuation
- PoS tagging
- Lemmatization using WordNet corpus
Run topic modeling - using Attention-Based Aspect Extraction
one-hot-encode
order_id
topic 1
topic 2
... (15 dummy variables)
Order items
order_id
product number (in the order)
product category
Select the most popular categories (to reduce the number of variables)
one-hot-encode
Order items 2
order_id
category_1
category_2
... (15 dummy variables)

Figure 2. Data pre-processing—transformations of source tables to obtain final input data.

5.2. METHODS FOR TOPIC MODELLING

Guests' reviews are generally short textbooks, and for CRM analysis, one needs to prize the content (aspect). The most popular model for inferring the content of a textbook is idle Dirichlet Allocation. The system is grounded on the supposition that each document is a admixture of a small number of motifs. At the same time, each content can be characterized by a distribution of word frequency. Still, short textbooks (similar as client reviews) comprise a veritably small number of motifs, generally only one, and because of this, LDA shouldn't be used in similar settings as its hypothetical's are violated. This was verified by an empirical study of short textbooks from Twitter, in which LDA failed to find instructional motifs. An indispensable and advanced system over typical LDA is the Dirichlet Multinomial Admixture model, applied concertedly with the Gibbs Sampling algorithm. The main difference lies in the supposition that each textbook comprises only one content, making DMM superior to LDA in short textbooks. Another innovative volition to LDA and which is a corner in the whole NLP field is word2vec, which is an effective way to bed words in a vector space while conserving their meaning. This system came a base for the Attention- grounded Aspect birth model. In empirical exploration, algorithms grounded on word embeddings outperform LDA in the task of short textbook content modelling. In this study, three algorithms for content modelling were tested and estimated, idle Dirichlet Allocation — because it's a go- to standard for content recognition., Gibbs Sampling algorithm for the Dirichlet Multinomial Admixture— this system is an enhancement over LDA, intended primarily for short textbooks. This is applicable for this case, where utmost of the reviews were just a couple of words long.

Each document can be modelled as a admixture of motifs. Document D can be characterized by the distribution of motifs, D that come from the Dirichlet family of probability distributions. Each content has a distribution of words', k which come from the Dirichlet family. also, a generative process aimed at carrying document D of the length of N words w,(1,..., N) is as follows, To induce a word at position i in the document Sample from the distribution of motifs, D, and gain an assignment of word w_i to one of the motifs k = 1,..., K. This is to gain information from which of the motifs the word should be tried.

Sample from the distribution of words in content', k, and gain the word to be fitted at position.

The parameters of, D for each document D, as well as', k for each of the motifs, should be learned using some system of statistical conclusion. Utmost of the practical executions of the algorithm are grounded on the Anticipation- Maximization system. This iterative approach aims to find the original outside of the liability function for the analyzed dataset.

Gibbs Sampling algorithm for the Dirichlet Multinomial Admixture is veritably analogous to the LDA approach with the difference that each document includes words from only one content. This supposition is grounded on the authors ' claim that generally, in the case of short textbooks, only one content is present. This leads to the following generative process. To induce a document D

Sample from the distribution of motifs, D and gain an assignment of the document to one of the motifs k = 1,...,K.

Sample all words from the content distribution',k.

Attention- Grounded Aspect birth takes a veritably different approach to content modeling compared to the styles over. It isn't grounded on a statistical model but rather on neural network modelling. The ensuing way describes the model armature presented visually by Figure 3. For each document from the corpus

1. Calculate the word embeddings e,(w_1), e,(w_2), e,(w_3),... with dimensionality d for each of the words from the vocabulary grounded on the whole corpus. From this point, one obtains an assignment of the word w to the point vectore _w in the point space Rˆd.
2. Gain document embeddingz_s. This is done by comprising the embeddings of all the words from the document. The normal is ladened by attention weightsa_1,a_2,a_3 given to each of the words. These weights are estimated during the model training and can be allowed
3. of as a probability that the particular word is the right word to concentrate on to infer the document's main content rightly. It's worth noting that the document embedding shares the same point space as the word embeddings.
4. Also, calculate p_t using soft max non-linearity and direct metamorphosis W. This vector p_t is of the same dimensionality as the number of aspects to be learned and can be allowed of as a representation of the probability that the judgment is from the particular aspect. By taking the biggest probability of this vector, one can gain the assignment to the particular content.

5. Increase the dimensionality of the vector p_t to the original dimensionality d by transubstantiating it with aspect matrix T. Vector r_s is attained.
6. The training of the model is grounded on minimizing the reconstruction error between the vectors z_s and r_s.

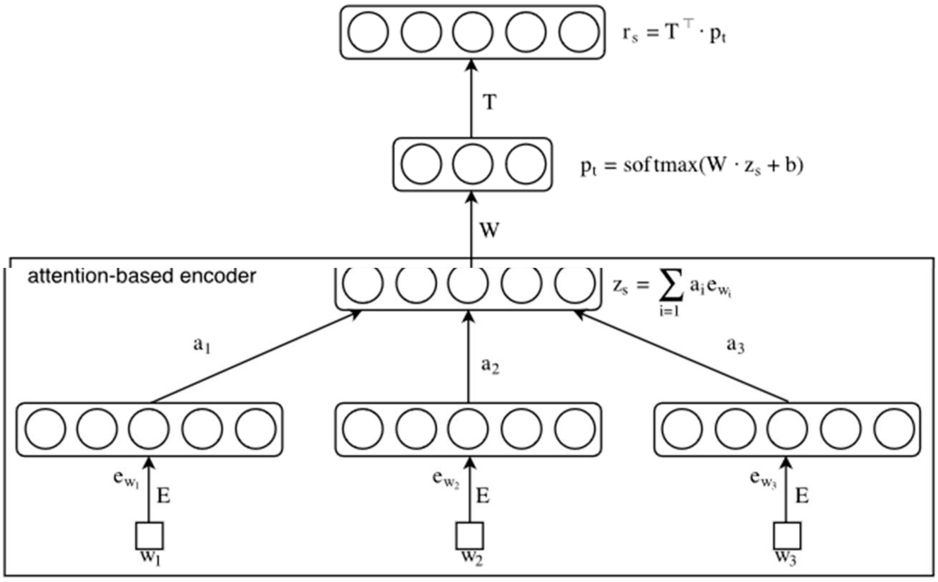


Figure 3. Attention-Based Aspect Extraction model architecture.

This system can capture better, more coherent motifs than LDA and its coequals. The main specified reason is that in LDA, all of the words are assumed to be independent; therefore, information about the “closeness” of the meanings of words is lost and has to be learned by assigning analogous words to the same motifs. This is a grueling task that LDA isn't optimized for. On the negative, using the word embedding approach, the models’ connections are known-priori by the model and can be erected upon. For illustration, indeed without knowing the motifs present in the corpus, one would anticipate that the words “cow” and “milk” should indicate the same content with high probability. Similar information is present in the word embeddings that this system uses. All three algorithms bear colorful types of data pre-processing. There are a many common conduct for all algorithms

6. RESEARCH AND METHODOLOGY

1) Google Collab is a free pall grounded Jupyter tablet terrain that's able of running numerous popular machine learning libraries, which can be fluently imported into the tablet for use.
2) Python is a programming language that has a simple and clear programming style and offers important features through colorful classes. It's also able of fluently integrating with other programming languages like C or C.
3) NumPy is an open- source library used for analysing and calculating data in Python and is essential for enforcing the array data type in Python. It's substantially used for matrix computations.
4) Pandas and Matplotlib are Python libraries that are freely available and generally used for analysing and imaging data.
5) Their main thing is to give druggies with effective tools to perform quick duplications of data analysis, visualization, and debugging.
6) Still, more complex workflows may bear more advanced intertwined development surroundings(IDEs), similar as Visual Studio IDE.

7. PROJECT DESCRIPTION

This project will go through several steps to build a customer churn prediction model. First, the dataset will go through preprocessing where the dataset is cleaned and to get best performance during modeling. After that, we will go through data visualization to get some insights about the data set in addition to the common aspects of churned customers. Finally, machine learning algorithms will be utilized to build customer churn prediction model.

7.1. DATA COLLECTION

The data is collected for this project from Kaggle for an e-commerce website. The empirical study starts in June 2023, and the observation ends in November 2023. The consumption data of customers who purchased goods on the website is selected for analysis and prediction. The dataset contains customer's consumption records in addition to historical behavioral interactions while using the platform.

7.2. DATA DESCRIPTION

Data is collected form a leading E-commerce platform, it is a historical data containing customer details and experience and its outcome is customer churn flag (churn= 1, no churn = 0). The dataset shows more than 5000 customers and their interaction and p references in the platform. The effectiveness of this data is that it contains some specific detailed attributes which will help in customer segmentation such as: preferred login device, Satisfaction store and other attributes. These attributes will help us in studying the causes of churn in each customer's segment to identify the triggers leading to customer churn. (Table 2 includes the name of the attributes, description, and type).

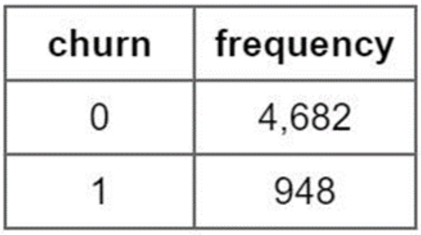
#	**Attribute**	**Description**	**Type**
1	CustomerID	Unique customer ID	Numeric
2	Churn	Churn flag	Numeric
3	Tenure	Tenure of customer in organization	Numeric
4	PreferrdLoginDevice	Preferred login device of customer	Character
5	CityTier	City tier	Numeric
6	WarehouseToHome	Distance in between warehouse to home of customer	Numeric
7	PreferedPaymentMode	Preferred payment method of customer	Character
8	Gender	Gender of customer	Character
9	HourSpendOnApp	Number of hours spend on mobile application or website	Numeric
10	NumberOfDeviceRegistered	Total number of devices registered per customer	Numeric
11	PreferedOrderCat	Preferred order category of customer in last month	Character
12	SatisfactionScore	Satisfactory score of customers on service	Numeric
13	MaritalStatus	Marital status of customer	Character
14	NumberOfAddress	Total number of added addresses per each customer	Numeric
15	Complain	Any complaint has been raised in last month	Numeric
16	OrderAmountHikeFromlastYear	Percentage increases in order from last year	Numeric
17	CouponUsed	Total number of coupons has been used in last month	Numeric
18	OrderCount	Total number of orders placed in last month	Numeric
19	DaySinceLastOrder	Day Since last order by customer	Numeric
20	CashbackAmount	Average cashback in last month	Numeric

Table 1: Data dictionary

8. PROJECT ANALYSIS

8.1. EXPLORATORY DATA ANALYSIS

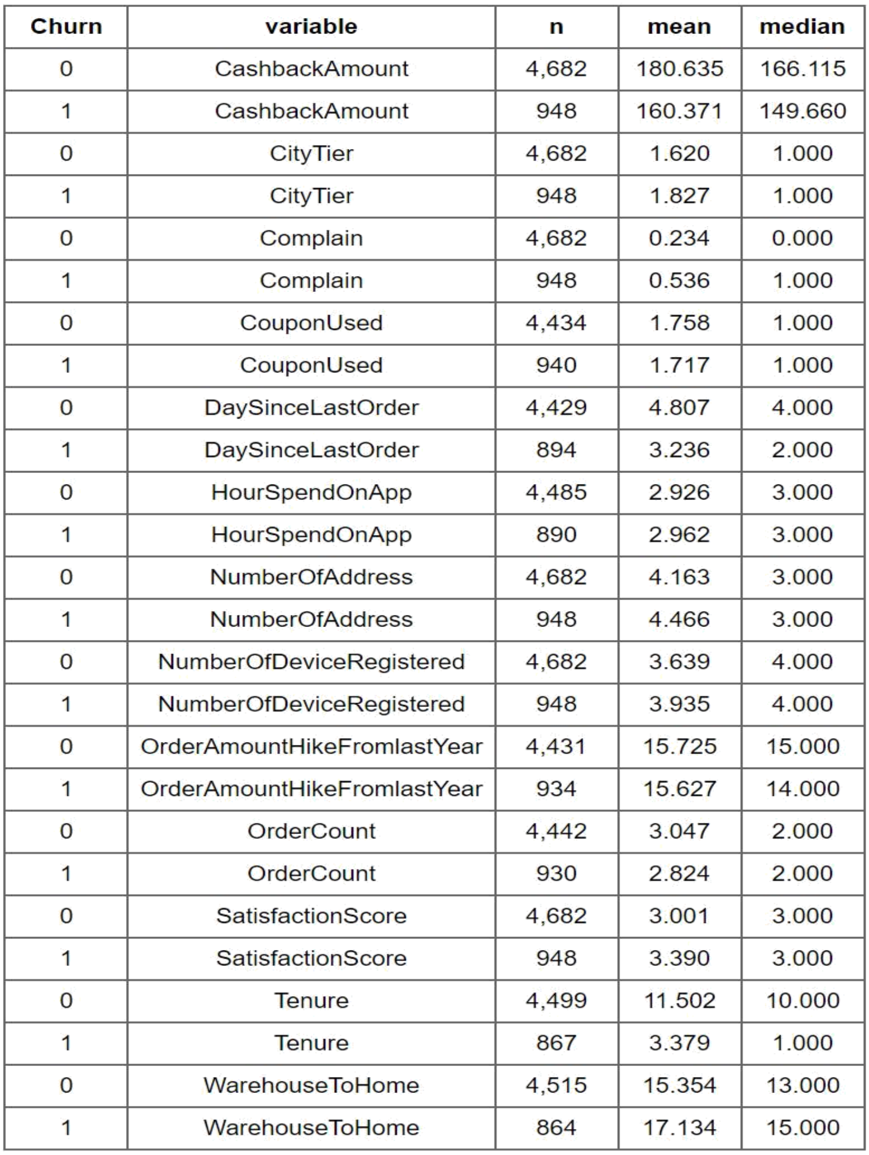
The first step in data exploration is to import several libraries in R to explore and visualize the data. then the numerical and categorical columns will be explored in addition to identification of missing data.The outcome of our data set is Churn, and there are no missing values in "churn" column. However, the outcomes variables are imbalanced due to the high number of retained customers in comparison to churned customers as shown in the table below.



churn	frequency
0	4,682
1	948

Table 2 : Churned versus retained customers

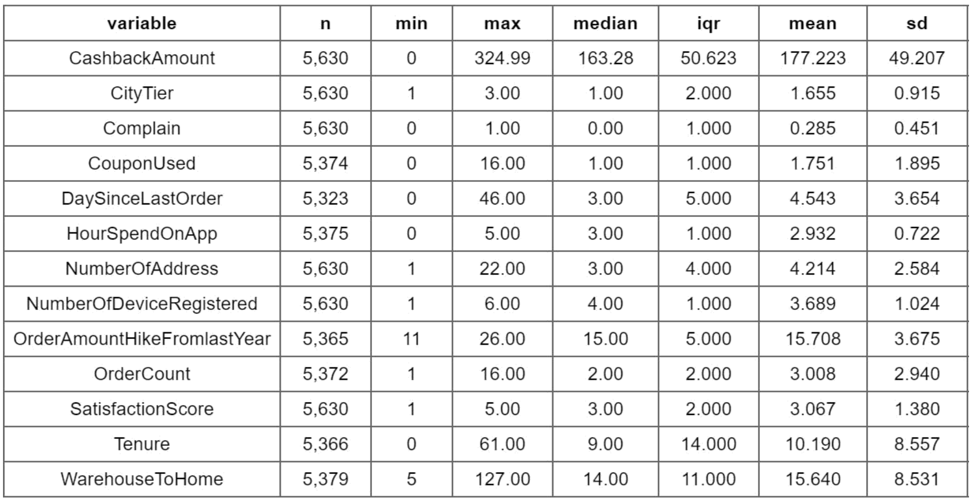
The following table contains the summary statistics of all numeric columns in our data. any column that has n=5630 shows that there are no missing values as we have 5630 unique customer IDs.



Churn	variable	n	mean	median
0	CashbackAmount	4,682	180.635	166.115
1	CashbackAmount	948	160.371	149.660
0	CityTier	4,682	1.620	1.000
1	CityTier	948	1.827	1.000
0	Complain	4,682	0.234	0.000
1	Complain	948	0.536	1.000
0	CouponUsed	4,434	1.758	1.000
1	CouponUsed	940	1.717	1.000
0	DaySinceLastOrder	4,429	4.807	4.000
1	DaySinceLastOrder	894	3.236	2.000
0	HourSpendOnApp	4,485	2.926	3.000
1	HourSpendOnApp	890	2.962	3.000
0	NumberOfAddress	4,682	4.163	3.000
1	NumberOfAddress	948	4.466	3.000
0	NumberOfDeviceRegistered	4,682	3.639	4.000
1	NumberOfDeviceRegistered	948	3.935	4.000
0	OrderAmountHikeFromlastYear	4,431	15.725	15.000
1	OrderAmountHikeFromlastYear	934	15.627	14.000
0	OrderCount	4,442	3.047	2.000
1	OrderCount	930	2.824	2.000
0	SatisfactionScore	4,682	3.001	3.000
1	SatisfactionScore	948	3.390	3.000
0	Tenure	4,499	11.502	10.000
1	Tenure	867	3.379	1.000
0	WarehouseToHome	4,515	15.354	13.000
1	WarehouseToHome	864	17.134	15.000

Table 3: Summary statistics for numerical columns

The following table studies the relationship between each the frequency of each attribute with respective of the outcome which is either churn or no churn. In addition, it reflects the mean, median of each attribute, Further details are listed below:



variable	n	min	max	median	iqr	mean	sd
CashbackAmount	5,630	0	324.99	163.28	50.623	177.223	49.207
CityTier	5,630	1	3.00	1.00	2.000	1.655	0.915
Complain	5,630	0	1.00	0.00	1.000	0.285	0.451
CouponUsed	5,374	0	16.00	1.00	1.000	1.751	1.895
DaySinceLastOrder	5,323	0	46.00	3.00	5.000	4.543	3.654
HourSpendOnApp	5,375	0	5.00	3.00	1.000	2.932	0.722
NumberOfAddress	5,630	1	22.00	3.00	4.000	4.214	2.584
NumberOfDeviceRegistered	5,630	1	6.00	4.00	1.000	3.689	1.024
OrderAmountHikeFromlastYear	5,365	11	26.00	15.00	5.000	15.708	3.675
OrderCount	5,372	1	16.00	2.00	2.000	3.008	2.940
SatisfactionScore	5,630	1	5.00	3.00	2.000	3.067	1.380
Tenure	5,366	0	61.00	9.00	14.000	10.190	8.557
WarehouseToHome	5,379	5	127.00	14.00	11.000	15.640	8.531

Table 4 : Relationship between numerical columns and churn

8.2. DATA CLEANING

Some way were performed after data disquisition which will help in creating more accurate machine literacy models and exclude bias. and after data cleaning, our data is composed of 3774 rows and 20 columns. Originally, in Preferred Order Cat column, there were two variables indicating the same meaning “Mobile” and “Mobile Phone” which were grouped in one position labeled as “Mobile”. also, In Preferred Payment Mode column “ Cash on Delivery ” and “ COD ” situations that have been grouped in one position labeled as “ Cash on Delivery ” in addition to “ Credit Card ” and “ CC ” situations have been grouped in group called “ Credit Card ”. Eventually, all missing values have been neglected to exclude crimes while erecting the models.

8.3. DATA VISUALIZATION

In the bar charts below, we illustrated all the numerical attributes in our dataset to study their distribution, further description is shown below.

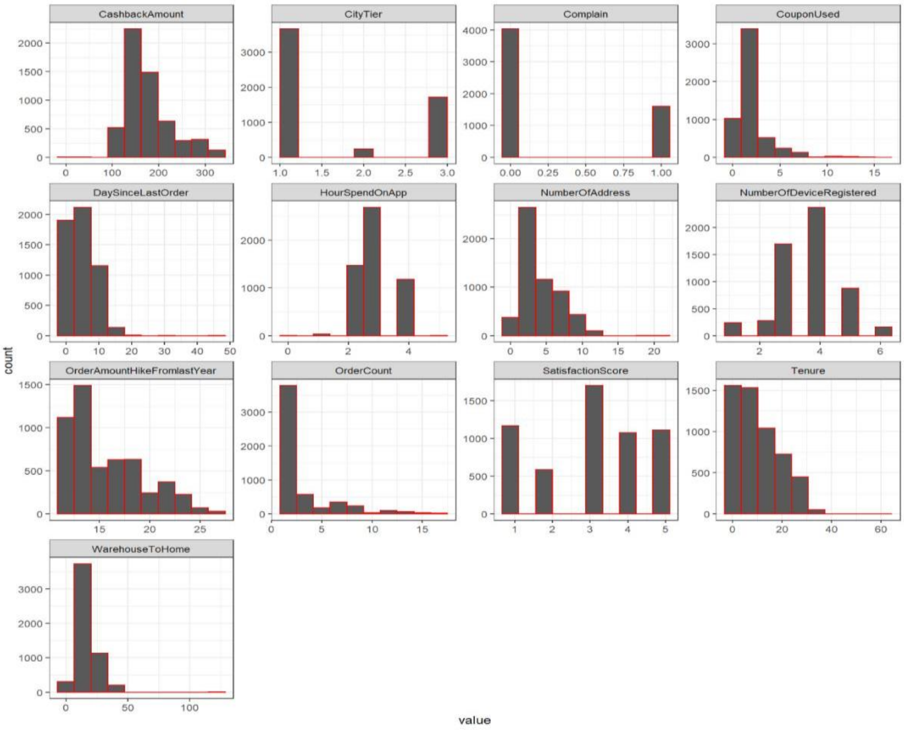


Figure 2: Numerical variables bar plot

Some observations on the visualizations listed above:

1) Most of the columns are right-skewed such as CouponUsed, DaySinceLastOrder and Tenure.

- Some variables have limited values like CityTier (3 values) and Complain (2 values). Box plots give an indication of the distribution of the values while taking in consideration mean, median and outlier. in the following box plot, numerical attributes were plotted in box plot and description is mentioned as below:

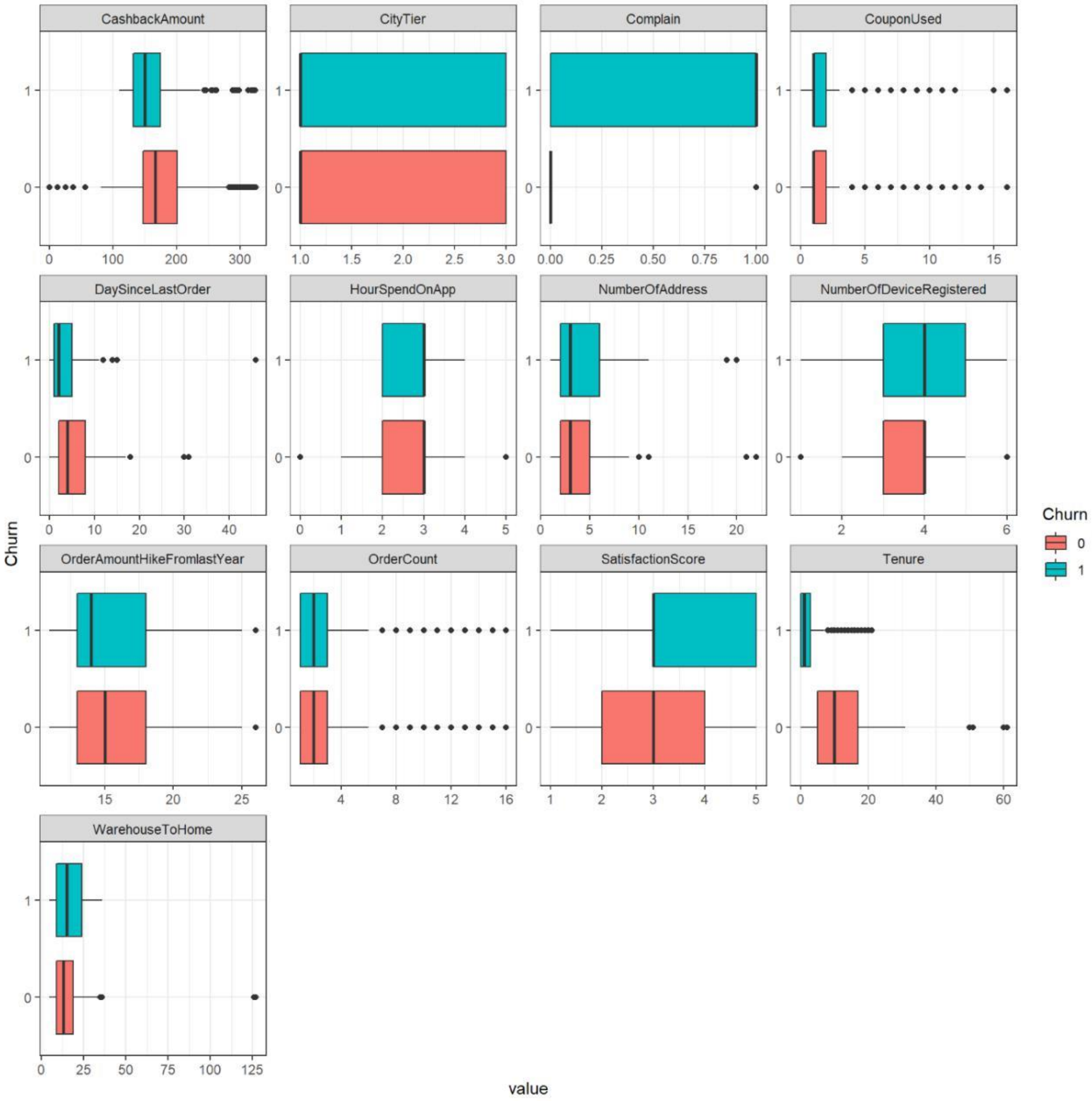


Figure 3: Numerical variables box plot

9. FUTURE WORK

In unborn systems, I would like to make real time analysis for a original e- commerce platform and link it to post marketing software. This integration will enable associations to automate their offers with churners which are clearly going to minimize client waste. Also, I would like to go in depth in retained guest's geste and utmost favored goods. Hence, studying retained client geste will reflect greatly on the company's income.

10. REFERENCES:

[1]. Zhang, D. (2015). Establishment and application of customer churn prediction model. Beijing Institute of Technology.

[2]. Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019, January). Churn prediction using neural network-based individual and ensemble models. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 634 -639). IEEE.

[3]. Wu, X. J., & Meng, S. S. (2017). Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost. Industrial Engineering, 20(02), 99-107.

[4]. Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp, 79, 403-408.

[5]. Shao, D. (2016). Analysis and prediction of insurance company's customer loss based on BP neural network. Lanzhou University

[6]. Lu, N., Liu, X. W., & Lee, L. (2018). Research on customer value segmentation of online shop based on RFM. Computer Knowledge and Technology, 14(18), 275-276, 284.

[7]. Huang, J. (2018). A Comparative Study of Social E-Commerce and Traditional E-commerce. Economic and Trade Practice, (23), 188-189.

[8]. Feng, X., Wang, C., Liu, Y., Yang, Y., & An, H. G. (2018). Research on customer churn prediction based on comment emotional tendency and neural network. Journal of China Academy of Electronics Science, 13(03), 340 -345

[9]. Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion, 54, 128-144.

[10]. Dhote, S., Vichoray, C., Pais, R., Baskar, S., & Shakeel, P. M. (2020). Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce. Electronic Commerce Research, 20(2), 259-274.

[11]. Agrawal, S., Das, A., Gaikwad, A., & Dhage, S. (2018, July). Customer churn prediction modelling based on behavioural patterns analysis using deep learning. In 2018 International conference on smart computing and electronic enterprise (ICSCEE) (pp. 16). IEEE.

[12]. Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. IEEE Access.

[13]. Ho, T. K. (1995). Random decisions forest. Proceedings of 3rd International Conference on Document Analysis and Recognition (pp. 278-282). New Jersey: IEEE.

[14]. Geetha, V., Punitha, A., Nandhini, A., Nandhini, T., Shakila, S. and Sushmitha, R., 2020, July. Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-5). IEEE.

[15]. Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W., 2019. A churn prediction model using random forest: analysis of machine learning te chniques for churn prediction and factor identification in telecom sector. *IEEE Access*, *7*, pp.60134-60149.

[16]. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory. Heidelberg: Springer; 1995. p. 23–37.

[17]. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310.

[18]. Decision Trees. (2022, 04 22). Retrieved from Sickit learn: https://scikit-learn.org/stable/modules/tree.html#:~:text=Decision%20Trees%20(DTs)%20are%20a,as%20a%20piecewise%20constant%20approximation.

[19]. Thomas W. Edgarm, D. O. (2017). Research Methods of Cyber Security.

[20]. Random Forest. (2020, December 7). Retrieved from IBM: https://www.ibm.com/cloud/learn/random-forest#:~:text=%20What%20is%20random%20forest%3F%20%201%20Decision,bagging%20method%20as%20it%20utilizes%20both...%20More%20