Proceeding Book of ***The 6th ICMEM 2021****, 11-13 August 2021, Bandung, Indonesia*
ISBN*: 978-623-92201-2-9*

Customer Churn Rate Analysis in Fashion E-Commerce Using Logistic Regression and Random Forest

Muhammad Hadi* and Sri Rahayu Hijrah Hati

Faculty of Economics and Business, Universitas Indonesia, Indonesia.

E-mail address: mhadialathas10@gmail.com

Abstract - **E-commerce companies think about long-term customer relationships in terms of conversion rates and repeat purchase rates. It is more cost-effective to retain existing customers than to acquire new ones, which is why it is crucial to track customers at high risk of turnover and target them with retention strategies. This research aims to identify what factors play a significant role in e-commerce to the customer churn rate in the next month. This study is based on 77,841 transactions data collected from Indonesia's fashion company through their e-commerce sales channel. In processing the data, descriptive statistics and predictive analytics with logistic regression and random forest models are used to achieve the research's objective, as both models have a good level of accuracy in making predictions and classifications. This study shows several factors such as gender, customer length of stay, order amount, and shipping cost significantly influencing the churn rate. This study recommends that the company and the fashion e-commerce industry manage customer churn by made appropriate strategic and business steps after knew the factors that cause it.**

Keywords - **Customer Churn Rate; E-Commerce; Big Data; Logistic Regression; Random Forest**

I. INTRODUCTION

According to [1], Indonesia's eCommerce market value stood at US$ 32 billion in gross market value in 2020, growing from US$ 21 billion in 2019. Google, Bain, and Temasek estimate it could grow to US$ 83 billion in 2025. In addition, based on the report, Indonesia's eCommerce market value makes Indonesia the biggest eCommerce market in Southeast Asia.

As websites have developed into distribution platforms, minimal research has explored the variables that have positively affected the intention to buy. In recent years, the competitive e-commerce environment has been sharpening up. Each e-commerce company makes some promotional scheme to draw customers to keep visiting and purchasing on its websites.

Many studies have shown [2] that the "abandonment" of online shopping carts is one of the most critical issues facing e-commerce. Abandonment applies to clients who add items to the shopping cart, but they usually leave the website when it comes time to pay and checkout. Several factors can influence consumers to leave their shopping carts.

Customer churn in E-commerce can be challenging to track because it is not usually something that happens overnight. A shipping fee is one of a crucial component that is likely to influence consumers' online purchasing behavior and decision-making [3]; it will help to increase the conversion rate and also the customer's willingness to re-visit and purchase again in e-commerce.

In the new digital age, many large and small e-commerce firms offer free shipping strategies to attract and retain customers. Under this scheme, for orders equal to or greater than a certain amount or a certain number of goods, e-commerce bears the expense of delivery but charges a fixed fee otherwise [4].

Besides shipping fee, there are still many factors that can influence customer churn. Understanding churn within a business can reinforce the retention efforts by pinpointing where and how customers are falling out. For years, the primary focus in e-commerce was customer acquisition. The acquisition is essential, particularly for newer brands or retailers looking to grow their shopper base. But it is not a sustainable way to increase the revenue in the long-term without effective customer retention.

PT. XYZ Company Profile

PT. XYZ is a fashion company from Indonesia and was founded in Jakarta in 1995. This company distributes bags, clothing wallets, fashion accessories, cosmetics with direct sales methods through members as partners to market its products. This company has several brands where each brand represents the characteristics of the intended target market. Each month the company release a product catalog which contains majority of new product and some product in the previous period that most selling and favorite by the customers.

Based on the business model and marketing distribution using a direct selling method that is unique comparing other fashion companies in the industry, two types of customers and causes trigger customer purchase intention in PT XYZ. First, a casual customer who purchases the product because they like and need it. Second, reseller customers who buy the product and sell the product again to get incentives and bonuses if they achieve the monthly target.

The growing use of online media is sufficient for various industries to start an online business, especially the fashion industry, where the company can sell and distribute the products with a broader reach. In 2012, PT.

ISBN*: 978-623-92201-2-9*

XYZ started a digital transformation where they started to build an e-commerce website as their sales channel. Apart from being a point of sale, this website functions as a digital tool to help members monitor their sales performance, recruitment, and incentives. The company is also taking advantage of its offline stores and applying the online-to-offline concept by integrating offline stores on e-commerce websites to order the products they want online and pick them up at the nearest offline store or also in convenience store.

Using one sample of big data set in a fashion company, this study is being conducted with the two intentions. First, identify what factors play a significant role in customer churn rate using logistic regression and random forest. Second, to provide insight into the term of predictive modeling for the related industries in alleviating retention rate.

II. METHODOLOGY

Data Collection

Data collection is the initial stage before big data analysis can be carried out to predict customer churn based on test variables that cause churn. The data used for this study were obtained from the real dataset of a company whose name was disguised, namely XYZ Inc. This company is engaged in the fashion industry and sells items such as bags, wallets, and shoes, where customers can buy these products through the website.

The data source is based on the problems that cause current churn and research [5], which uses churn indicator variables such as customer profile information, promotions, and payments and [6], that uses data related to shipping fees. This study combines some of these variables. Second, the data source was dependent on the data available that was recorded in the company. The variables used involve customer data (customer id, age, and gender), length of stay, number of orders, order amount, order discount amount, shipping fee, discounted shipping fee, and churn indicators.

This study using a customer churn period in 1 month. According to [7], the difficulty in defining the customer churn period is that the prediction period duration should be set in a way. First, it captures the activity/inactivity of customers with a fairly long inter-purchase time, and second, it captures the defection of those with a short average inter-purchase time as soon as possible.

To avoid the problems mentioned above, the duration of the prediction period in this study chooses because of several factors. First, as the company implement the direct selling marketing strategy and applies a monthly bonus to reseller customer who achieved the monthly sales target from their re-selling program, this monthly period it is easier to identify the type of reseller customer. Second, according to [8], if the window period is too short, it will show an artificially high churn rate. If the window period is too long, it is challenging to know when the customers lose until long after they are gone. As the company releases a new product and a new promotional campaign every month, one common way of identifying when the company has lost a transactional customer is using the period in 1 month. Finally, to recheck and ensure the time of window period is appropriate, the prediction period is set to be approximately equal to the average inter purchase time of the last customer in the mass of the sorted customer base [7].

Data are taken from Indonesia eCommerce is attractive as Indonesia has a huge market in eCommerce and is still growing. According to [1], Indonesia's internet economy grew from USD 40 billion in 2019 to USD 44 billion in 2020. Out of the USD 44 billion, USD 32 billion came from Indonesia's eCommerce sector.

The data taken is demographic and transactional data from web visitors' activities recorded using the database system, which shows the pattern of each customer's service use and transactions. This research will use historical data between January – February 2020 before the COVID-19 pandemic happened in Indonesia.

TABLE I
CHARACTERISTICS AND TYPE OF DATA

Variable Name	**Data Type**	**Description**
Customer ID	*Ratio*	*Unique id for each customer*
Age	*Ratio*	*Age of Customer*
Gender	*Nominal*	*Gender of Customer*
Length_of_Stay	*Ratio*	*Customer length of stay since first register*
Num_of_Order	*Ratio*	*Total number of order in 1 month*
Total_Order_Amount	*Ratio*	*Total number of order in 1 month*
Total_Discount_Amount	*Ratio*	*Total order amount in 1 month*
Total_Shipping_Amount	*Ratio*	*Total Shipping amount in 1 month*
Total_Shipping_Discount_Amount	*Ratio*	*Total Shipping Discount Amount in 1 month*
Churn	*Ratio*	*information about whether the customer make an order on February 2020 (1 = YES, 0 = NO)*

The data that has been provided is transaction data that is still raw, so it requires further data checking. Furthermore, the data is analyzed, and after going through cleaning, filtering, and processing data, a dataset is produced, which is predicted to generate understanding and information for customer churn analysis.

Data Analysis

This research aims to predict customer churn and its variables that influence it with the help of big data analysis modeling results that can be used in the following strategic step for the company to retain their customer retention rate will not be significantly reduced. There are several steps required to develop a model which can be used to predict customer churn at XYZ Inc. Stages these

include, dataset collection, data cleansing / preprocessing, identify variable and customer churn model selection, model creation and finally model finalization.

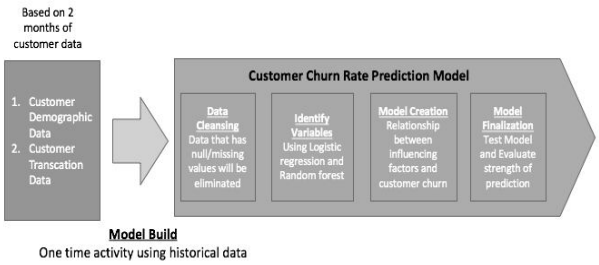

Fig. 2. Data Analysis Flow

After going through the data collection and data grouping stages, the next step is extracting data to obtain a data set. By applying the concepts of business and statistics, the data sets that have been obtained need to be analyzed to understand that can answer business questions. The results of this analysis are ultimately expected to help the decision-making process in business. Data analysis can be divided into two categories, namely descriptive analysis, and predictive analysis.

The primary algorithm used in this study is the logistic regression algorithm. Using the random forest algorithm as a companion in this study is because the random forest algorithm has a good level of accuracy in making predictions and classifications so that it can be used as a comparison of the results of forming predictive models using logistic regression algorithms. However, the random forest algorithm has a weakness in explaining the effect of independent variables on the observed dependent variable. Another reason is that the two methods were chosen because they are quite popular, relatively easy to use, and have a low error rate.

Logistic Regression

Logistic Regression is a Machine Learning algorithm used for classification algorithms and used mainly as data analysis and inference tools to assign observations to a discrete class set. The goal is to understand the role of the input variables in explaining the outcome. The logistic regression model was developed by David Cox in 1958 and has become one of the popular methods for predicting the value of the dependent variable in categorical form.

The Logistic Regression uses a more complex cost function defined as a Sigmoid function to map predicted values to probabilities [9]. The function maps any real value into another value between 0 and 1. In addition, The cost function represents the optimization objective so that we can develop an accurate model with minimum error.

Random Forest

Random forests [10] is a substantial modification of bagging that builds an extensive collection of de-correlated trees and averages them. A random forest consists of multiple random decision trees. The trees incorporate two distinct types of randomization. First, each tree is constructed using a random sampling of the original data. Second, a subset of features is randomly selected at each tree node to obtain the optimal split.

Random forests perform similarly to boosting on many problems, and they are easier to train and tune. As a result, random forests are widely used and are implemented in a variety of packages. There are some that we need to understand when we want to interpret the random forest algorithm in machine learning, such as feature importance, partial dependency plot, and inTrees.

III. LITERATURE REVIEW

Customer Relationship Management

Fostering good relationships with customers is one of the priorities of various companies at this point. The approach taken is through Customer Relationship Management (CRM), which leads to the management and analysis of customer interactions through multiple practices, strategies, and technologies needed.

The purpose of CRM is to improve good business relationships with customers, increase company value and trademarks, direct sales growth, and help manage customer retention [11]. In the end, the impact of good CRM management is to increase the value of customer satisfaction. When customers are satisfied with the company's services or products, it is expected that customer loyalty will get better.

Customer Retention and Customer Churn

One of the topics that concern many companies today is Customer retention. In business, especially in industries where customers have relatively low switching costs, retaining customers is something that companies need to pay attention to in doing business because there are so many factors that can encourage customer churn. Moreover, in some mature companies, the costs involved in acquiring new customers can be more expensive to retain existing customers [12].

Churn rate, also known as the rate of attrition, is the percentage of customers who stop using a business or have left your service over within a given period. Customer churn is a harsh fact that all companies must cope with. Even the biggest and most profitable businesses suffer from consumer churn, and permanent, sustainable business growth needs to recognize what causes once-loyal customers to abandon their business.

Some various approaches and formulas can be used to measure it. Quoting from the article How to Calculate Customer Churn Rate and Revenue Churn Rate [13], one straightforward approach to calculating it is to use the

ISBN*: 978-623-92201-2-9*

calculation of the number of customers, namely the number of customers at the beginning of the service period and the number of customers remaining at the end of the period.

To calculate the percentage, the number of customers at the beginning of the period minus the number of customers at the end of the period, then the difference between the number of customers is divided by the number of customers multiplied by 100.

Understanding customer churn's nature is essential in making an overall evaluation of marketing efforts that have already been spent and customers' satisfaction levels. It is more cost-effective and easier to maintain loyal customers than acquiring new ones. Seeing the importance of retaining customers, customers have turned into a factor that needs special attention. One of them is to identify customers who fall into the category of churn and analyzing the cause. For this purpose, we need a model that can predict customers who do not want service and find the factors that influence it.

Factor Affecting Customer Churn

Various factors can influence customers to remain loyal to use the service or leave it. These factors, if drawn from the source, can come from external or internal factors. External factors influence the customer to stop service, for example, because the customer moves to use competitors' services. Other external factors beyond economics, such as social and political factors, also affect the market. On the other hand, internal factors include the company's promotions, sales, services, quality, and values.

Length of Stay

There's are several factors that taken for this study. First, Length of stay which has definition the time since customer register and make first transaction in the platform. The length of stay also represent the relationship quality, as the customer already usage and the service and has intention to using the service again which explain the customer already has a good relationship quality in the services.

According to previous research [14], there was created a theoretical model of customer retention. The newly created model combines the previously mentioned authors thought and shows that customer retention is influenced by three factors customer satisfaction, relationship quality and switching costs.

Shipping Fee and Shipping Fee Discount

Prior study analysis reveals that promotional free shipping increases order incidence but reduces order values, whereas threshold-based free shipping encourages larger orders but has little impact on order incidences [6]. Thus A shipping fee is an essential component that is likely to influence consumers' online purchasing behavior and decision-making; it will help to increase the conversion rate and also the customer's willingness to re-visit and purchase again in e-commerce.

As an impact, Many e-commerce offers threshold free shipping policies to attract and retain consumers. The threshold of free shipping here means the amount that deducts into the customer's shipping fee. Free shipping policies can be divided into two types: (1) unconditional free shipping, where the e-commerce absorbs all shipping costs for all orders [15] and (2) using threshold free shipping (TFS), where the e-commerce bears the cost of shipping for orders equal to or larger than a predetermined amount (i.e., a threshold value), but charges a fixed fee otherwise [3].

Gender and Age

Demography is usually one of the most important factors for market segmentation by marketers to deliver the correct information to invite the customer to purchase. Besides that, it is found that online purchase intention can be the primary determinant based on demographic variables [16].

A prior study from [17] found that gender significantly influences online purchase intention. Males have a stronger online purchase intention than females. The study results reflect that males have a stronger online purchase intention which may be attributed to males having a higher comfort level using technology. Meanwhile, in study from [16] also found a correlation between gender and purchase intention in social media with a different outcome: females tend to use social media than males for online shopping.

Other demographic variables found in a study from [16] are the age that stated millennials are more intent to use social media for online shopping compared to elders. Another study from Radka [18] found that younger generations shop online more often and revealed that online shopping was statistically significantly higher after taking the 16-34 years and 35-54 years categories than the 55+ category.

IV. RESULTS

A. Descriptive Statistics

ISBN*: 978-623-92201-2-9*

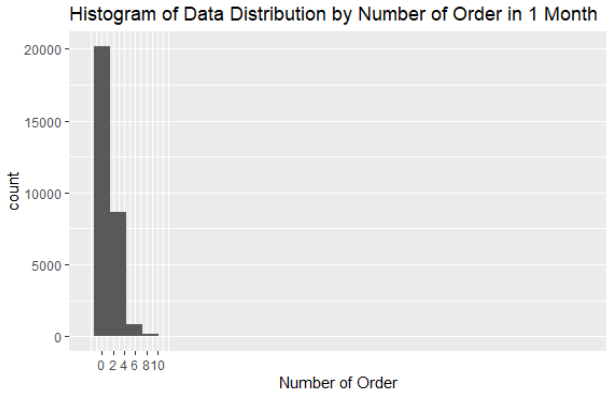

Fig. 3. Histogram of Data Distribution by Number of Order in 1 Month

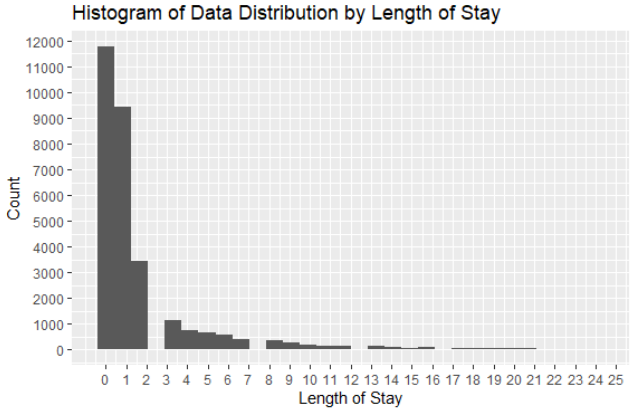

Fig. 4. Histogram of Data Distribution by Lenght of Stay

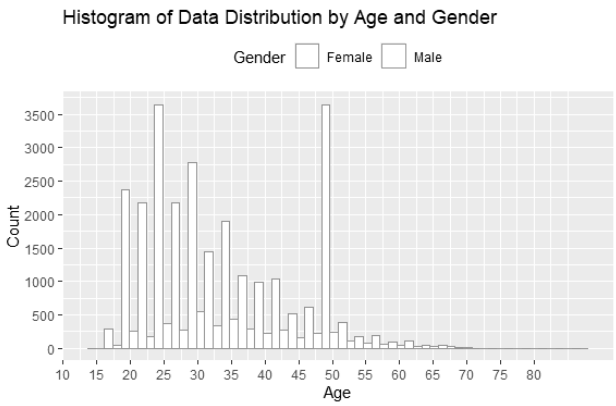

Fig. 5. Histogram of Data Distribution by Age and Gender

This study also explores and visualizes the data set by distributing independent variables to understand the data patterns better and potentially form hypotheses. A few observations can be made based on the histograms for numerical variables. First, most of the customers in the dataset are around 20-30 years old. In contrary, there is a lot customer in age 50, most of the reason why the age is 50 is in huge number, it occurs because the data in old ERP system is set the default as 50 years old. Those data is carried over when customers activate their old accounts when they activate their online accounts on the website, and they did not change it.

Another finding is that most of the customers make less than two orders in 1 month. In addition, there are many new customers in the organization (less than 2 Years old), it seems the company makes a massive acquisition in the last two years. Gender distribution shows that the dataset features a relatively has a considerable proportion of female customers. Female customers also dominated in each group of age and length of stay in the company.

The original premise of the business case is when a customer who has used the service for more than two years is a customer who allegedly understands how to use XYZ services, knows the quality of the product and gets value from the use, and has a lower tendency to leave the service.

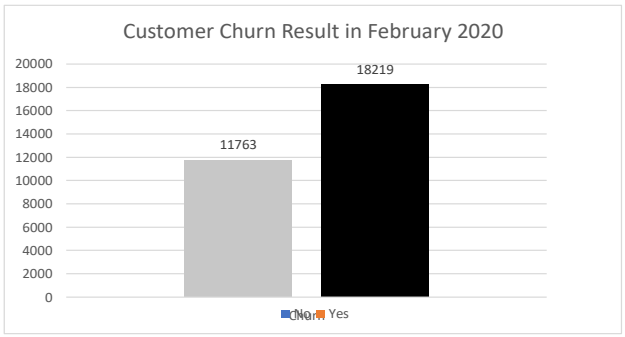

Fig. 6. Visualization for Identifying Number of Customer Churn

The number of customers who leave service or continue to use XYZ service can be mapped in the form of tables and plots using R with a value of 1 for churning customers and a value of 0 for those who continue to use the service. Data shows that from 29982 customers, 18219 customers left XYZ Inc. services. This means that from customers who were observed during that period, 60.76% of customers left the service.

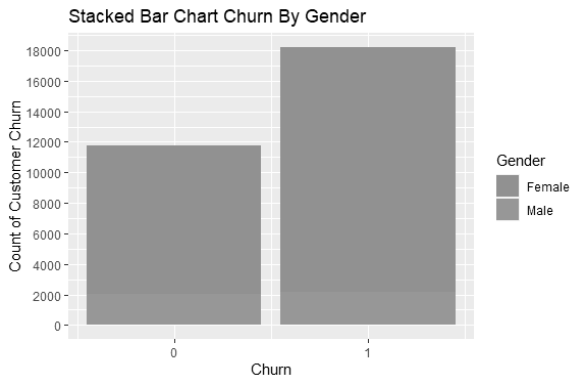

Fig. 7. Stacked Bar Chart Churn by Gender

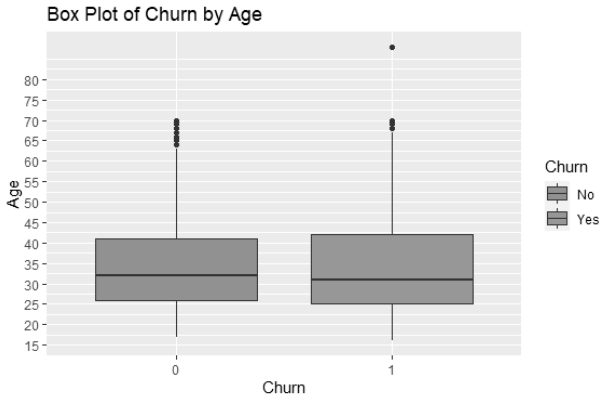

Fig. 8. Box Plot Churn by Age

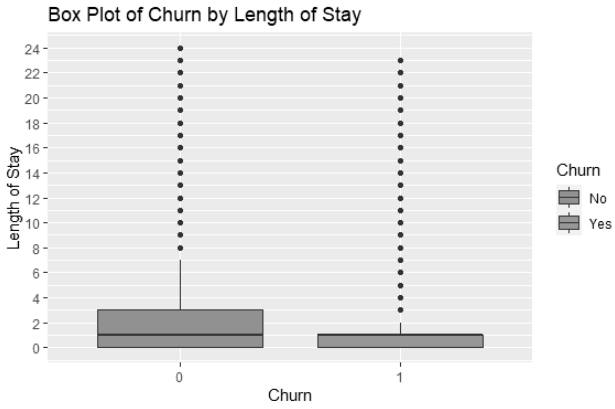

Fig. 9. Box Plot Churn by Length of Stay

From the stacked bar and box plots above, this study learns several things. First, in alignment with the gender distribution, female customers dominated the number of customer churn. Second, the distribution of customers that churn based on age is wider than the one that does not churn, but the difference is not huge. Also, a customer who has a length of stay in the company for more than one year tends not to churn. In addition, among the customers who churn, most seem only to make one or fewer orders in 1 month.

B. Logistic Regression Model

The method used to build a customer churn prediction model in this paper uses two methods: logistic regression and random forest analysis. They are quite popular and have a good average level of accuracy. Various other methods can be used to form a customer churn prediction model, such as decision trees, AdaBoost, neural networks, support vector machines (SVM), and others. Various other studies compare and evaluate the results of these various models.

The logistic regression model was built by defining the dependent variable and the independent variable. The dependent variable used is Churn, in the form of a churn value containing 0 or 1. The variable ID, the customer id, is not included as a variable used in the calculation. The independent variable used is the length of the customer subscribe to XYZ.Inc service, customer age when in January 2020, customer gender, customer order amount in January, customer discounts during January, shipping customer amounts during January, and shipping discounts during January.

TABLE 2
LOGISTIC REGRESSION RESULT IN FIRST PHASE

Variables	**Coefficients**	**Std. Error**	**Z-Value**	**Sig.**
Intercept	1.52	4.7	32.3	***
Gender	-3.34	3.58	-9.32	***
Age	2.32	1.20	1.93	.
Num_of_Order	-5.61	1.75	-31.97	***
Length_of_Stay	-1.07	4.74	-22.55	***
Total_Order Amount	-1.13	2.12	-5.36	***
Total_ Discount_Amount	4.73	4.46	-4.51	***
Total_Shipping _Amount	6.75	9.87	6.84	***
Total_Shipping_Discount_Amount	N/A	N/A	N/A	

From the regression results, it is found that the variables that have a significant influence on the prediction results of customer churn by looking at the results of the calculation of the P-value, which has a value below 5% which are gender, length_of_stay, num_of_order, total_order_amount, total_discount_amount, and total_shipping_amount. Because not all variables have a significant value to be used in the logistic regression calculation formula, this logistic regression model was rebuilt by eliminating the age and total_shipping_discount_amount variables.

From the logistic regression results above, the results are obtained $\sum \alpha i Xi$= 1.596 – 3.307 (Gender) – 1.056 (Length_of_stay) – 5.621 (Num_of_order) – 1.135 (Total_order_amount) – 2.027 (Total_discount_amount) + 6.638 (Total_shipping_amount).

Newdata is a sample of 1 customer that is included in a variable data frame type. To test more than one customer, import a file containing the required data into the R data frame. In this example, we can get the probability churn value with the predict function. Based on that function, we got the probability value of a customer churn from the data entered through the new variable is 54%

ISBN*: 978-623-92201-2-9*

C. Evaluation and Testing Logistic Regression Model

Testing the customer churn prediction model built with logistic regression analysis is carried out to determine the level of accuracy in predicting customers who will leave the service. Testing is done by dividing the dataset obtained into two partitions, namely train data and test data. Train data is used to build logistic regression model equations, while test data is used to test whether the model that has been built has a good level of accuracy in predicting customer churn. In this modeling, 70% of the data will be used as train data, and a test data of 30% is used to test each data with the proportion of customer comparisons with the same ratio of conditions of churn = 1 and churn = 0.

From the description above, it can be seen that the comparison composition of the test data and train data for the conditions of churn = 1 and churn = 0 has the same proportion. The train data consisted of 20987 customers, while the test data consisted of 8995 customers. Train data obtained from the data sharing is used to form a logistic regression model. Furthermore, the prediction of customer churn is carried out with the model that has been formed, namely XYZModel with input test data.

The prediction results of the XYZModel model with input test data have an error rate of 31.4% and an accuracy rate of 68.6%. The calculation of predictions using the script above is done by calculating the probability value for a P-value greater than 0.5 to be included in the churn category and other values as the no-churn category. Comparison of the predicted results and the actual results, whether the customer leaves the service or not, can be presented by combining the predicted value and the actual value of the churn variable.

D. Random Forest Model

The formation of the customer churn prediction model with the random forest algorithm uses the classification approach as the decision tree. The main difference is that this algorithm results from observations of many decision trees made and then use the average results of the various decision trees as the final model. Estimated error values are made for cases that are not used in forming the tree. This value is referred to as the OOB (out of the bag) error expressed.

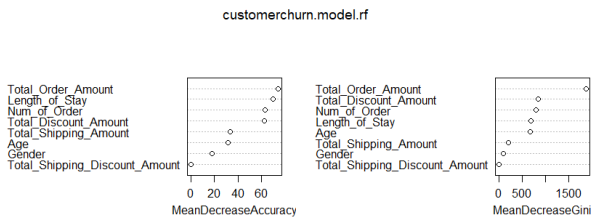

Fig. 10. Random Forest Mean Decarease Accuracy

The predicted value in the random forest algorithm is the churn variable. The independent variables used for classification are age, gender, length_of_stay, num_of_order, total_order_amount, total_discount_amount, total_shipping_amount and total_shipping_discount_amount.

From the function formed by tree modeling, then the customer churn prediction model is obtained. The importance level of the variables used in the random forest is seen from the mean decrease accuracy value. When compared with the results of the significance level variable of the logistic regression model, results are not contradictory. Variables such as total order amount, length of stay, number of orders, and total order discount amount have the highest score than other variables.

To use the model in a prediction function, the prediction in the R function can be used to calculate the approximate results. Similar to the steps in the logistic regression section, various customer data variables to be tested are included in a data frame.

E. Evaluation and Testing Random Forest Model

Like the logistic regression, testing the customer churn prediction model built with random forests was carried out to determine the accuracy in predicting customers who will leave the service. Testing is done by dividing the dataset obtained into two partitions, namely train data and test data. After dividing the dataset based on that, the next step is to check the accurate level of the random forest model using a confusion matrix. Based on the confusion matrix, the accuracy level is 77.5%

F. Comparison Logistic Regression and Random Forest Model

After get the result of each model and the level of accuracy, the result of Logistics Regression has an accuracy level of 68.6%. Meanwhile, in the Random Forest model, the accuracy level is 77.5%. Based on the accuracy level from those predictive models created, the Random Forest model would be a better choice for predicting the customer churn in this study as that model has a better level of accuracy.

With the existing consumer insights through data, companies can predict customers' possible needs and issues, define proper strategies and solutions against them, meet their expectations, and retain their business. Based on the predictive analysis and modeling, businesses can focus their attention on the targeted approach by segmenting and offering them customized solutions and analyzing how and when the churn is happening in the customer's lifecycle with the services will allow the company to come up with more preemptive measures

V. CONCLUSION

According to the research results, there are several conclusions as follow. Based on logistric regression and random forest, variables that influence customer churn in this research data are the length of stay, gender, shipping fee, the total number of orders, order amount, and order discount amountThe length of stay has a significant influence on customer churn. As customers already have trust and loyalty, thus they will stay longer to use the services, and the longer the customer uses the service, the less change the customer will churn.

The results suggest that gender has a significant influence on customer churn. Meanwhile, It is inferred that age does not have a significant influence on online customer churn. It also comes into focus that Shipping amount has a significant influence on customer churn. The result supports various previous studies; thus, A shipping fee is an essential component that will likely influence consumers' online purchasing behavior and decision-making

The study results reflect that total number order and order amount have a significant influence on customer churn. In other words, when customers already trust and feel safe with the e-commerce platform, they have tended to repurchase intention and also increase the number of transactions.

One of the goals of this study is how predictive analytics can predict the likelihood of customer churn according to the variables observed by increase understanding of a condition and the need to predict future behavior through the analysis performed. One of the effective ways to avoid customer churn is to prevent it before it occurs. The approach that can be taken is to find customers who might leave the service and then analyze the variables significantly based on usage patterns and service patterns that affect the decision. This study shows that the company can get insight into descriptive statistics and make actionable actions. In addition, his study presents two models to analyze and predict the future behavior of their customers, namely logistic regression and random forest. In summary, the modeling carried out by both methods gives good results, with the highest accuracy value is the random forest method. Companies can use this method for future churn modeling.

With the development of technology, companies in the fashion industry can leverage their sales channel conversion and expand their revenue using the e-commerce platform. However, it also increases the competition in the fashion industry, as customers have more accessible access to look for and compare the multiple brands before making their purchase.

Thus, customer retention program implementation is one of the business world's priorities and more important to keep customers using the product/service. Therefore companies need to foster good relationships with customers, with customers who have good customer lifetime value to keep using the service. This good relationship can be managed by understanding the needs and problems faced by customers to maintain customer satisfaction.

By paying attention to customer satisfaction, companies can avoid potential loss of revenue due to customer churn. The following are the research result implication for the management is need to consider some variables that will affect customer churn. Thus management also needs to take action that will minimize the customer churn. There are some recommendations based on the study that can help management to minimize customer churn.

First, create an exciting scheme of a loyalty program for the existing member as the data that we found the length of stay is one of the significant factors that influence the churn rate. Second, Offer an incentive that combines a scheme to increase basket size and Free shipping discount policy. Finally, Give personalized recommendations for the products or promotions based on the user profile.

This research has several limitation which need to be improved for the future researches which have similar topics and discussion. The limitation of this research are as follow; First, the data source was dependent on the data available that was already recorded in the company. This limitation resulted in an incomplete variable that can be explored to get other factors that can influence the customer churn. Second, the company controls the free shipping policy, and many companies will have different shipping policies that will influence the result. Third, the study periods are before the pandemic COVID-19, As the pandemic tends to affect the fashion industry, making customers change the prioritization they should purchase.

Based on the limitation explained, there are some recommendations for future research as the following; first, It will be beneficial in future research to be done by trying to comparing another models and adding another variables that can influence the customer churn. Second, It will be beneficial in future research to use data from longer period and various economic situations (such as pandemic, crisis, governmental changing, etc.) Third, Future research may be carried out to further explore how consumers would react to different Threshold Free Shipping (TFS) policies in a more complicated business situation where a wide assortment of product categories and different price ranges are offered

REFERENCES

[1] Google and Temasek. (2020). e-Conomy SEA 2020 Report. https://economysea.withgoogle.com/ (accessed 02 August 2021).
[2] Boone, T. & Ganeshan, R. (2013). Int. J. Production Economics: Exploratory analysis of free shipping policies of online retailers.
[3] Koukova, N.T., Srivastava, J., Steul-Fischer, M., (2012). Journal of the Academy of Marketing Science: The effect of shipping fee structure on consumers' online evaluations and choice.
[4] Huang, W.-H., Cheng, Y.-C., (2015). Transportation Research Part A: Threshold free shipping policies for internet shoppers.
[5] Baragoin, C. (2001). Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data. IBM Redbooks.
[6] Lewis, M., (2006). Journal of Retailing: The effect of shipping fees on purchase quantity, customer acquisition, and store traffic.
[7] Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing Churn Prediction Techniques and Assessing Their Performance: A

Contingent Perspective. *Journal of Service Research*, *19*(2), 123–141. https://doi.org/10.1177/1094670515616376

[8] Outlier AI. (2017). How to calculate churn rates across industries. https://towardsdatascience.com/how-to-calculate-churn-rates-across-industries-68c692b64605

[9] Pant, A. (2019). *Introduction to Logistic Regression*. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[10] Breiman, L. (2001) Random forests. Machine Learning, 45, 5-32. doi:10.1023/A:1010933404324

[11] Chai, W. (2020). CRM (customer relationship management). https://searchcustomerexperience.techtarget.com/definition/CRM-customer-relationship-management

[12] Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: quality comes to services. Harvard Business Review, 68(5), 105–111.

[13] Salesforce. (2020). How to Calculate Customer Churn Rate and Revenue Churn Rate. https://www.salesforce.com/resources/articles/how-calculate-customer-churn-and-revenue-churn/ (accessed 03 May 2021).

[14] Hennig-Thurau, T., Gwinner, K. P., & Gremler, D. D. (2002). Understanding Relationship Marketing Outcomes: An Integration of Relational Benefits and Relationship Quality. Journal of Service Research, 4(3), 230–247. https://doi.org/10.1177/1094670502004003006

[15] Becerril-Arreola, R., Leng, M., & Parlar, M. (2013). Online retailers' promotional pricing, free-shipping threshold, and inventory decisions: A simulation-based analysis. European Journal of Operational Research, 230(2), 272–283. https://doi.org/10.1016/j.ejor.2013.04.006

[16] Sharma, B. K., & Parmar, S. (2018). Impact of Demographic Factors on Online Purchase Intention Through Social Media- With Reference To Pune , Maharashtra. 05(01), 45–50.

[17] Uj, S., & Rs, S. (2013). Does Demography Influence Online Purchase Intention ? Evidence from North-West India. International Journal of Advance in Management and Economics, 6(2), 01–08.

[18] Bauerová, R. (2018). Are Online Purchases Affected By Demographic Factors in the Czech Republic? Acta Academica Karviniensia, 18(1), 5–16. https://doi.org/10.25142/aak.2018.001