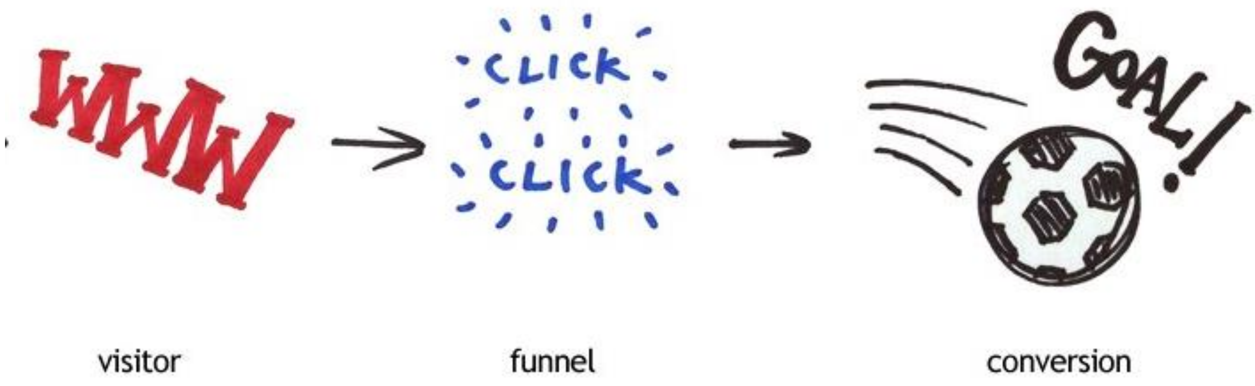


Predicting Click Conversion Rate

By:

- Abhijeet Sharma
- Keerthi Bai Reddy
- Saranya Chintalapati
- Shuvrangshu Mukhopadhyay

Contents

Executive Summary:

The objective of this SDM project is to develop a predictive model for click conversion rate using product-level data. Click conversion rate measures the percentage of users who click on a product ad and complete a desired action, such as making a purchase. By creating a predictive model, we aim to identify which product features have the greatest impact on click conversion rate and tailor our strategies accordingly to improve business revenue.

In today's highly competitive ecommerce landscape, it is essential for businesses to maximise their revenue generation by optimising their product targeting strategies. By leveraging the power of advanced data analytics and machine learning techniques, businesses can gain valuable insights into the factors that drive click conversion rates and use this information to develop targeted and effective product strategies. This SDM project aims to contribute to the development of such techniques and highlight the importance of using data-driven approaches to optimize revenue generation.

To achieve this, we will collect a large dataset of user interactions with products on an Indonesian ecommerce website. The data will include information on product features such as category, brand, price, and availability, and user demographics, interests, and behavior. We will use advanced statistical and machine learning techniques to analyse the data and identify patterns that can help us predict click conversion rates for various products.

Our analysis will focus on understanding which product features are most influential in driving click conversion rate. By identifying these features, we can develop a predictive model that can accurately forecast click conversion rates for different products and user segments. This model can help businesses tailor their product targeting strategies and optimise their revenue.

Overall, this SDM project aims to demonstrate the value of using advanced data analytics and machine learning techniques to optimize revenue generation using product-level data. By understanding the factors that drive click conversion rates and using predictive models to forecast outcomes, businesses can develop effective product targeting strategies and optimize their revenue.

Problem Definition & Significance:

The target client for this SDM project is any ecommerce business that aims to optimize revenue generation by improving their product targeting strategies. Specifically, we are addressing the business problem of low click conversion rates, which can show that the product ad is not reaching the right target audience or that the product features are not meeting user needs. Low click conversion rates can lead to wasted marketing efforts and decreased revenue, making it an essential problem for ecommerce businesses to address.

This is an interesting and important problem because click conversion rate is a crucial metric that measures the percentage of users who click on a product and complete an action, such as making a purchase. According to a recent study by Smart Insights, the average click-through rate (CTR) for ecommerce ads across all industries is around 2%, while the average conversion rate is approximately 3%. This highlights the importance of optimizing click conversion rates to improve revenue generation in the ecommerce industry.

Furthermore, the ecommerce industry is highly competitive, with new businesses entering the market every day. According to Statista, the global ecommerce market is expected to reach $6.5 trillion by 2023. As such, ecommerce businesses need to continually optimize their product targeting strategies to remain competitive and maximize revenue generation.

Overall, the problem of low click conversion rates is an interesting and important one for ecommerce businesses to address. By improving click conversion rates, businesses can optimize revenue generation and remain competitive in a rapidly growing industry.

Prior Literature:

Study Title	Predictors	Findings	Citation
Success Factors of E-Commerce – Drivers of the conversion rate and basket value.	• Conversion Rate (DV) • Basket value (DV) • Website design (including usability, appearance, and navigation) • Product presentation (including quality of product images, product descriptions, and user reviews) • Pricing strategy (including level of prices, promotions, and discounts) • Customer service (including response times, availability, and helpfulness)	The website design has the strongest impact on conversion rates, followed by pricing strategy and customer service, while product presentation has a smaller but still significant impact. Additionally, using multiple sales channels is associated with higher conversion rates and basket values compared to using a single sales channel. Therefore, e-commerce businesses should focus on improving these factors to increase their conversions and basket values.	Darius Zumstein and Wolfgang Kotowski
Post-Click Conversion Rate – Predictive model on E-Commerce Recommender System.	• Click-through rate (DV) • User demographics • Past purchase history • Features of the recommended item.	The study utilized three statistical tests for CTR prediction, including linear regression, XG Boost algorithm, and random forest. Linear regression and random forest showed similar performance with a mean squared error (MSE) of around 2%, while the XG Boost algorithm had a slightly higher MSE of around 2.59%. The most important features for CTR prediction across all three models were CTR during the past 7 days, followed by average position during the past 7 days and current average position, and they all had similar levels of prediction variability.	Yuhe Ding University of North Carolina, Chapel Hill North Carolina, December - 2018
Predictive Analytics of E-Commerce	• Conversion Decision (Yes - 1 or No - 0) (DV)	Random Forest was found to be the best	Niu, X., Li, C., & Yu, X. (2019)

search behavior for conversion.	• QueryLength • CurrentQueryPosition • NumQuery • ClickPosition • AvgClickPosition • NumSearchResults • NumClickQuery • NumClickSession • ClickEntropyQuery AvgClickEntropySession • PageDwellTime • SessionDwellTime • UserType • Device • HourOfDay	performing model, achieving an accuracy of 76% after tuning the 'mtry' parameter. Logistic Regression was used as the base model and achieved an accuracy of 61%, but after addressing multicollinearity among predictors and selecting only four variables, it was outperformed by Random Forest.	
The determinants of conversion rates in SME e-commerce websites.	• Conversion Rate (DV) • Free Shipping (0 No - 1 Yes) • Free Returns (0 No - 1 Yes) • Discounts • Season (0 regular - 1 Sales) • Speed of load (0-100) • Luxury websites (0 non-brand products; 1 branded products) Week (0 = Saturday & Sunday - 1 = weekdays) • Free Shipping (0 No - 1 Yes) • Free Returns (0 No - 1 Yes) • Discounts • Season (0 regular - 1 Sales) • Luxury websites (0 non-brand products; 1 branded products) • Week (0 = Saturday & Sunday - 1 = weekdays, Monday to Friday.		Di Fatta, Davide Patton, Dean Viglia, Giampaolo 2018/03/01
Analyzing conversion rates in online hotel booking. The role of customer reviews, recommendations and rank order in search listings.	• Conversion Rate (DV) • Price • Rank • Recommendation Number • Location rating • Service Rating • City	This study analyzed the factors that influence hotel booking decisions and click conversion rates. The results suggest that customers prioritize location rating over star rating and service rating when choosing a hotel. Additionally, high numbers of recommendations can offset the negative impact of a low rank in search listings on conversion rates. Logistic regression was	Asunur Cezar and Hulisi Ögˇüt Department of Business Administration, TOBB University of Economics and Technology, Ankara, Turkey

		used, and beta distribution was chosen for modeling fractional data. Factor analysis was also performed to reduce the substantial number of variables into a smaller set of factors.	
Impact of different platform promotions on online sales and conversion rate: The role of business model and product line length.	• Conversion Rate (DV) • Sales (DV) • Direct promotions • Gift promotions • Price promotions • Product type ('reseller'==1,'marketplace'==0) • SKU • Quantity promotion • Bundle promotion, • Coupon promotion • Length (product line length) • Weekend, • Holiday, • Brand	This study examines the impact of different promotions on sales and conversion rates in a three-level hierarchical promotion structure using data from JD.com, one of the largest online retailers in China. Monetary promotions, including direct and quantity promotions, were found to have a stronger impact on sales and conversions than other types of promotions. The study also highlights three key characteristics of e-commerce platforms, including the use of digital information and technologies, collection and use of transaction data, and network effects among users in the online community. OLS regression was used to analyze the impact of independent variables on conversion rates, and feature engineering was performed on different types of promotions. Interactions were also performed on resellers with types of promotion and brand.	Tingting Tong, Jianjun Xu Xun Xu Nina Yan
Analyzing Factors of Users Click Behavior on Ads	• Click Through Rate (DV) • Gender (Men==1, women =2) • Age • Consumption Level (Low, medium and high) • Shopping Depth (new, general and regular) • Occupation – student (Undergrad==1, non-graduate == 0) • Brand Tendency	This study analyzed data about advertisement and users' profiles on Taobao using logistic regression. The model found that women tend to click on online ads more than men, older age people are more likely to click on ads, and customers who buy low and medium-priced products tend to click on ads	**Sitong Zhou** College of Urban Transportation and Logistics, Shenzhen Technology University, Shenzhen 518118, China

		online. Additionally, customers are more likely to click on non-branded products than branded ones.	
An examination of antecedents of conversion rates of e-commerce retailers	• Conversion Rate (DV) • Purchase Intention score • Website satisfaction score • Average monthly visits • average unique monthly visits • average ticket price	This study explores the factors that affect website satisfaction and purchase intention score, which ultimately impact conversions. Indicators such as information quality, system quality, and service quality, as well as website ease-of-use, response time, language customization, website layout, and transaction capabilities were found to affect website satisfaction. Purchase intention score is influenced by trust, perceived ease of use, perceived usefulness, and presentation quality. The analysis shows that both website satisfaction and purchase intention score have significant weightage on conversions.	Naveen Gudigantala Pelin Bicen Eom Robert B. Pamplin

Data Source:

The data we used in this analysis is of an E-commerce giant in South-East Asia owned by Shopee, Indonesia. This data is proprietary and was readily available on AWS marketplace for research purposes. The data consists of 4 levels that include product level, category level, department level and store level. The store level and brand level are masked data. We will be considering category level in the analysis as it is more granular.

Data Dictionary:

PRODUCT LEVEL
ITEM_ID
ORIGINAL_PRICE
SALE_PRICE
REVIEW_RATING
REVIEW_COUNT
SOLD
VIEW_COUNT
HISTORICAL SOLD
STOCK
LIKED_COUNT
SHOW_DISCOUNT
DISCOUNT_PERCENT
CREATE_TIME
PRODUCT_NAME
MONTH_YEAR

STORE LEVEL
SHOP_ID
SHOP_LOCATION

BRAND LEVEL
BRAND

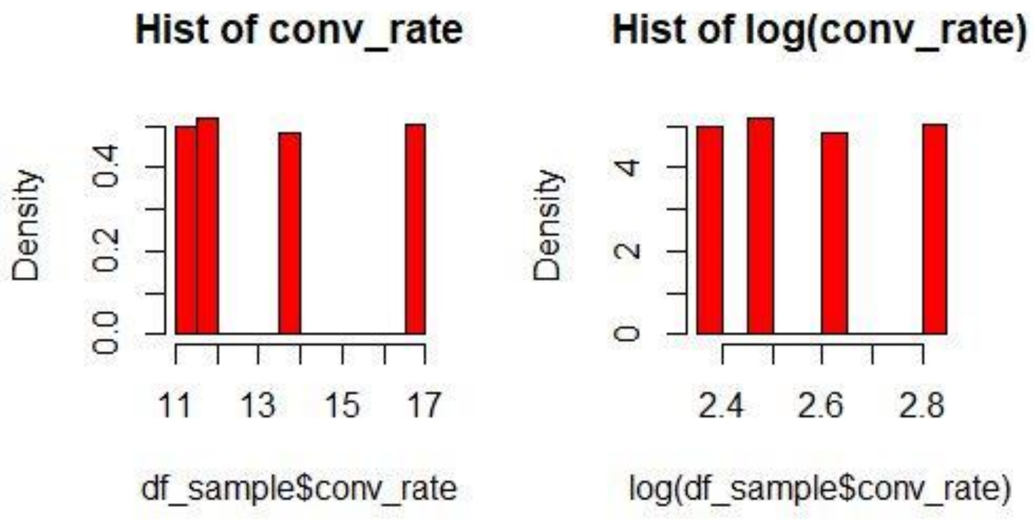
CATEGORY LEVEL
CATEGORY_ID
CATEGORY
DEPARTMENT

Variable Choice / Predictor Table:

Level of data	Variable	Sign of Effect	Rationale
Product Level	ITEM_ID	Just an identifier	NA
	ORIGINAL_PRICE	Will include sale_price instead as that is post discount	NA
	SALE_PRICE	Increase in sale_price reduces conversion rates, as customers will opt for more affordable products	-
	REVIEW_RATING	Larger review_rating influences the customer to purchase products, hence, increased rates	+
	REVIEW_COUNT	More there are review_count, the customers will likely purchase products, hence, increased rates	+
	SOLD	This is used to calculate the click conversion rates	NA
	VIEW_COUNT	This is used to calculate the click conversion rates	NA
	HISTORICAL_SOLD	The more the product was sold in the past, higher the conversion rate	+
	STOCK	With bigger value of stocks, companies expects to sell more products, hence increase click conversion rates	+
	LIKED_COUNT	Products with more likes will likely sell more, hence, increase click conversion rates	+
	SHOW_DISCOUNT	Products with discount will sell more, hence, increase click conversion rates, but including discount_percent as it is numeric	NA
	DISCOUNT_PERCENT	Products with discount will sell more, hence, increase click conversion rates	+
	CREATE_TIME	Year is extracted and we will control for time	+/-
	PRODUCT_NAME	Name is an identifier, hence excluded from analysis	NA
	MONTH_YEAR	Same as create_time, hence excluded	NA
Category Level	CATE_ID	Its an category identifier, hence excluded	NA
	CATEGORY	We are interested in looking at fixed effects of category, conversion rates will vary based on category	+/-
Department Level	DEPARTMENT	Since we are considering category, excluded department	NA
Store Level	SHOP_ID	Shop identifier, excluded	NA
	SHOP_LOCATION	We have masked data, hence, we are excluding it	NA
Brand Level	BRAND	We have masked data, hence, we are excluding it	NA
Currency Level	CURRENCY	Currency does not have anything to do with click conversion rate	NA

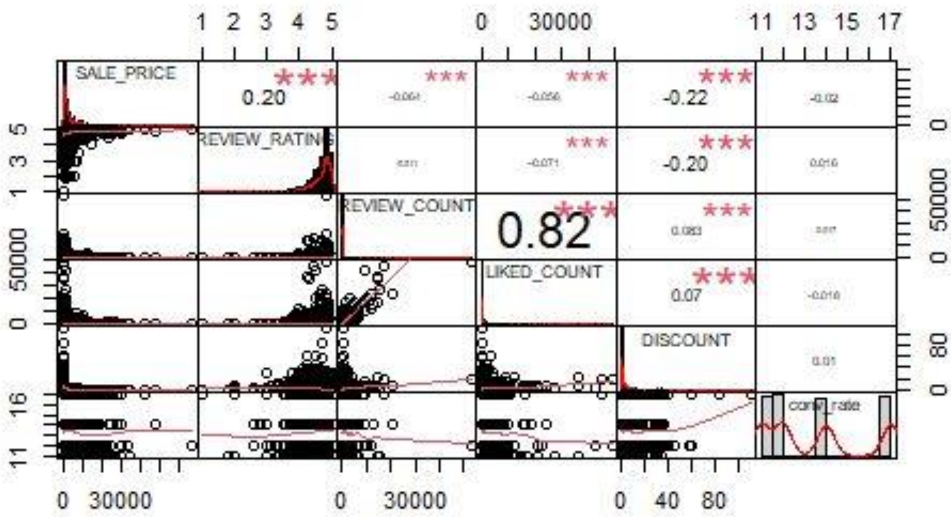
Descriptive Analysis/ Data Visualization:

Distribution of Target Variable:



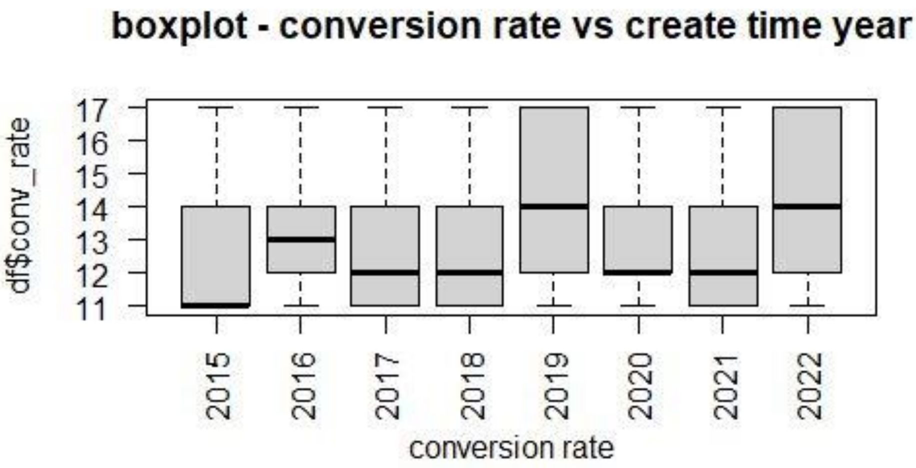
We do not see the normal and transformed plots following any known distributions, hence, we cannot use GLM or LM models to predict click conversion rates.

Correlogram Plot:



As per the above correlogram plot, we can see there is no evidence of correlation existing among the independent variables, except for liked_count and review_count. Since both are important predictors, and we must include them in our analysis.

Conversion Rate for all the years:



Based on the annual breakdown, we can conclude that the conversion rate is not showing a clear trend of growth or decline. Nonetheless, certain spikes in the conversion rate can be attributed to specific years. For instance, in 2016, the company may have been actively promoting its newly launched products. In 2020, during the Covid pandemic, the conversion rate dropped as the country was uncertain about how to deal with the new circumstances.

Models:

Justification: Our target variable is clicking conversion rate, which is a ratio of Count of Purchase and Count of Impressions. Since, the distribution does not follow a certain kind of distribution, and the conversion rate is bounded between 0 and 100, we will use tobit models. We also want to investigate if there is any sort of interactions present and how it is affecting the overall conversion rate.

Furthermore, we saw that none of the models we had trained on the complete dataset were converging, despite adjusting their hyperparameters and using various optimization algorithms. Therefore, we decided to take a different approach and randomly sample a smaller subset of the data based on a particular department (Men's Clothing) to fit the models. We found that this method helped reduce overfitting and improve the generalization performance of the models, as they converged and achieved higher accuracy on the test set. As a result, we drew more reliable and robust conclusions from the experiments and supplied better insights for our research question.

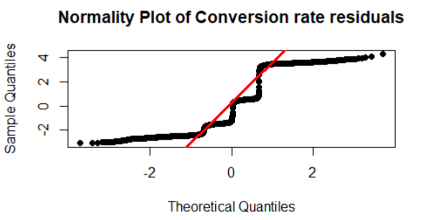
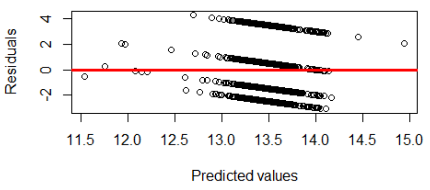
```
lm1 = lm(conv_rate ~ SALE_PRICE + REVIEW_RATING + REVIEW_COUNT + LIKED_COUNT +
HISTORICAL_SOLD + STOCK + DISCOUNT + CATEGORY + CREATE_TIME_YEAR, data =
df_sample)

tobit1 = tobit(conv_rate ~ SALE_PRICE + REVIEW_RATING + REVIEW_COUNT +
LIKED_COUNT + HISTORICAL_SOLD + DISCOUNT + STOCK + CREATE_TIME_YEAR + CATEGORY,
left = 0, right = 100, data = df_sample)
```

```
tobit2 = tobit(conv_rate ~ SALE_PRICE+ STOCK + DISCOUNT*REVIEW_RATING +
REVIEW_COUNT + CATEGORY*LIKED_COUNT + HISTORICAL_SOLD + CATEGORY*REVIEW_RATING
+ CREATE_TIME_YEAR, left = 0, right = 100, data = df_sample)
```

```
=========================================================================================
                                                   Dependent variable:
                                ---------------------------------------------------------
                                                        conv_rate
                                           OLS                        Tobit
                                           (1)                  (2)                (3)
-----------------------------------------------------------------------------------------
SALE_PRICE                       -0.00001 (0.00001)   -0.00002 (0.00001) -0.00001 (0.00001)
REVIEW_RATING                       0.026 (0.082)        0.150 (0.118)      -0.061 (0.173)
REVIEW_COUNT                      0.00001 (0.0001)     -0.0001 (0.0001)   -0.0001 (0.0001)
LIKED_COUNT                      -0.00000 (0.00002)   -0.00000 (0.00003) -0.00002 (0.00003)
HISTORICAL_SOLD                  -0.00001 (0.00001)   0.00001 (0.00002)  0.00001 (0.00002)
STOCK                            -0.00000 (0.00000)   -0.00000 (0.00000) -0.00000 (0.00000)
DISCOUNT                            0.001 (0.004)        0.006 (0.006)      -0.122 (0.099)
CATEGORYT-Shirts                   -0.013 (0.053)        0.008 (0.077)      -1.193 (1.083)
CREATE_TIME_YEAR2016                0.340 (1.385)        0.018 (1.447)       0.045 (1.447)
CREATE_TIME_YEAR2017                0.338 (1.342)        0.929 (1.360)       0.967 (1.359)
CREATE_TIME_YEAR2018                0.340 (1.333)        0.632 (1.339)       0.674 (1.338)
CREATE_TIME_YEAR2019                0.555 (1.330)        0.554 (1.332)       0.590 (1.332)
CREATE_TIME_YEAR2020                0.429 (1.328)        0.349 (1.329)       0.387 (1.328)
CREATE_TIME_YEAR2021                0.413 (1.327)        0.360 (1.327)       0.401 (1.327)
CREATE_TIME_YEAR2022                0.591 (1.329)        0.455 (1.330)       0.496 (1.329)
DISCOUNT:REVIEW_RATING                                                       0.028 (0.022)
CATEGORYT-Shirts:LIKED_COUNT                                              0.0001 (0.00004)
REVIEW_RATING:CATEGORYT-Shirts                                               0.252 (0.231)
Constant                          12.958*** (1.382)   12.423*** (1.438)  13.369*** (1.540)
-----------------------------------------------------------------------------------------
Observations                            9,362                4,681              4,681
R2                                      0.001
Adjusted R2                            -0.0003
Log Likelihood                                           -10,525.050        -10,522.550
Residual Std. Error               2.296 (df = 9346)
F Statistic                     0.792 (df = 15; 9346)
Wald Test                                              15.541 (df = 15)   20.563 (df = 18)
=========================================================================================
Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```

Quality Check / Assumptions Testing:

Multicollinearity (Fail)	**SALE_PRICE** 1.23 **REVIEW_RATING** 1.15 **REVIEW_COUNT** 15.59 **LIKED_COUNT** 3.31 **HISTORICAL_SOLD** 10.68 **DISCOUNT** 1.11 **STOCK** 1.03 **CREATE_TIME_YEAR** 1.18 **CATEGORY** 1.22	During our analysis, we found a strong collinearity between the variables REVIEW_COUNT and HISTORICAL_SOLD. However, we also recognized that both variables were crucial predictors for our analysis, and dropping either of them could significantly affect the results. Therefore, we decided not to remove any of these variables, but instead explored alternative strategies to address the collinearity issue. It is worth noting that although multicollinearity can lead to unstable and unreliable

		estimates of the coefficients, it does not necessarily invalidate the predictive power of the variables. Therefore, it is important to assess the collinearity and its impact on the analysis, but not necessarily remove the variables if they are considered relevant and informative for the research question.
Independence (Pass)	**durbinWatsonTest(residuals(tobit2))** **1.946362**	This value is close to 2 which means there is no evidence of autocorrelation that exists.
Normality (Fail)		To diagnose the normality assumption, we used various statistical tests such as the Shapiro-Wilk test and Q-Q plots, which showed significant deviations from the expected normal distribution. This could be due to various factors such as outliers, skewness, or heavy-tailed distributions in the data. As a part of future work, we will address the issue of non-normality, we explored alternative methods such as non-parametric tests, data transformations, or bootstrapping. These methods can help improve the robustness and validity of the statistical inferences, even with non-normality.
Equality of Variance (Fail)		In addition to the issues of collinearity and non-normality, we also found evidence of heteroskedasticity in our regression model. Specifically, the variance of the residuals appeared to be increasing or decreasing across the range of predictor variables, indicating that the error terms were not constant. This could be due to various factors such as outliers, measurement error, or the presence of unobserved variables that affect the variance of the residuals.

		As a part of future work, we will utilize WGLS and FGLS to address the issue of violation of this assumption

Model Interpretations:

Dependent – Conversion Rate				
Models ------>	OLS	Tobit	Tobit (Interactions)	Marginal Effect Interpretations
SALE_PRICE	-0.00002	-0.00002	-0.00001	Not Significant
REVIEW_RATING	0.15	0.15	-0.061	Conversion Drops by 6.1 Unit if review rating increase by 100 unit
REVIEW_COUNT	-0.0001	-0.0001	-0.0001	Not Significant
LIKED_COUNT	0	0	-0.00002	Not Significant
HISTORICAL_SOLD	0.00001	0.00001	0.00001	Not Significant
STOCK	0	0	0	Not Significant
DISCOUNT	0.006	0.006	-0.122	Conversion drops 1.2 Units if Discount is increased by 1 Unit in Category Men's clothing
CATEGORYT-Shirts	0.008	0.008	-1.193	Conversion drops 11.93 Units if item be9ing sold is from Category t-shirts in comparison to Category Outerwear
CREATE_TIME_YEAR2016	0.018	0.018	0.045	If 100 products are being sold in 2016, conversion rate increase by 4.5 units in comparison to base year which 2015
CREATE_TIME_YEAR2017	0.929	0.929	0.967	If 10 products are being sold in 2017, conversion rate increase by 9.6 units in comparison to base year which 2015
CREATE_TIME_YEAR2018	0.632	0.632	0.674	If 10 products are being sold in 2018, conversion rate increase by 6.7 units in comparison to base year which 2015
CREATE_TIME_YEAR2019	0.554	0.554	0.59	If 10 products are being sold in 2019, conversion rate increase by 5.9 units in comparison to base year which 2015
CREATE_TIME_YEAR2020	0.349	0.349	0.387	If 10 products are being sold in 2020, conversion rate increase by 3.8 units in comparison to base year which 2015
CREATE_TIME_YEAR2021	0.36	0.36	0.401	If 10 products are being sold in 2021, conversion rate increase by 4 units in comparison to base year which 2015
CREATE_TIME_YEAR2022	0.455	0.455	0.496	If 10 products are being sold in 2016, conversion rate increase by 4.9 units in comparison to base year which 2015
DISCOUNT:REVIEW_RATING	0.028			D(Conv)/D(DISCOUNT:REVIEW_RATING)
CATEGORYT-Shirts:LIKED_COUNT	0.0001			D(Conv)/D(CATEGORYT-Shirts:LIKED_COUNT)
REVIEW_RATING:CATEGORYT-Shirts	0.252			D(Conv)/D(REVIEW_RATING:CATEGORYT-Shirts)
Constant	12.423**	12.423*	13.369**	

D(Conv)/D(DISCOUNT:REVIEW_RATING)= -0.006(Discount)-0.118(Review_Rating)+0.028(Discount:Review_Rating)

D(Conv)/D(CATEGORYT-Shirts:LIKED_COUNT)= -0.077(CATEGORYT-Shirts)-0.00003(Liked_Count) + 0.0001(Category-Shirts:Liked_Count)

D(Conv)/D(REVIEW_RATING:CATEGORYT-Shirts)= -0.077(CATEGORYT-Shirts) -0.118(Review_Rating)+0.252(REVIEW_RATING:CATEGORYT-Shirts)

Recommendations:

1. Think about expanding your product offering: T-shirts convert at a rate that is 12 units lower than outerwear. To evaluate if you can improve overall conversion rates, it would be worthwhile to investigate other product categories.
2. Concentrate on boosting your sales volume because they are strongly associated with the conversion rate. In succeeding years, both sales and the conversion rate rose. Consider spending money on marketing and advertising initiatives to raise brand awareness and boost sales.
3. Examine the effects of extraneous variables: Although these figures offer insightful information about conversion rates, it's necessary to consider extraneous variables that might have influenced the outcomes. For instance, modifications in consumer behavior or the state of the economy could have an effect. You can better understand how to maximize your marketing efforts by examining these elements.
4. Keep track of conversion rates over time because they can change a lot from year to year. You can modify your marketing plan by keeping a careful check on these figures and spotting trends. For instance, you might need to invest in innovative marketing strategies if you see that conversion rates are declining over time.

Future Work:

- We can include other categories in the analysis by not just limiting it to only men's clothing. For instance, we can have women's clothing, children's clothing, electronics, home goods, or other categories of products. By analyzing click conversion rates across multiple product categories, we can get better insights into the factors that influence customer behavior and the effectiveness of their marketing strategies.
- Shoppe has multiple locations across the world, and we can analyze conversion rates across different shoppe locations like Indonesia, Thailand, Brazil, Mexico, and other countries to understand how conversion rates vary by location.
- The data we used for this analysis was not randomly sampled, so the models failed to meet the assumptions, and the insights gained from the study may need to be more right. So, for the future, we would like to directly reach out to the e-commerce websites and obtain a more representative sample that includes information on customer demographics, purchase history, and transactions that gives better insights and helps understand customer behavior on conversion.

References:

Paper reference links:

[1] https://www.emerald.com/insight/content/doi/10.1108/IJCHM-05-2014-0249/full/html

[2] https://www.emerald.com/insight/content/doi/10.1108/MRR-05-2014-0112/full/html

[3] https://www.researchgate.net/publication/340417892_Success_Factors_of_E-Commerce_-_Drivers_of_the_Conversion_Rate_and_Basket_Value

[4] https://core.ac.uk/reader/210609030

[5] https://core.ac.uk/download/pdf/301371722.pdf

[6] https://www.sciencedirect.com/science/article/pii/S0969698917306525

[7] https://www.sciencedirect.com/science/article/pii/S0167923622000173

[8] https://www.atlantis-press.com/proceedings/isemss-22/125982064

Dataset link:

https://aws.amazon.com/marketplace/pp/prodview-y2bqrnxiikw6c?sr=0-3&ref_=beagle&applicationId=AWSMPContessa#offers

Appendix:

R code for the analysis:

```
rm(list = ls())
library(rio)
df = import("Final SDM Project File.xlsx")
View(df)
str(df)
table(df$CURRENCY)
table(df$BRAND)
table(df$SHOP_LOCATION)
table(df$CATEGORY)


#Check for NULLs
colSums(is.na(df))
#review rating, review count, sold, historical sold, Liked count has NULLs

#show_count has lot of NULLS so dropping them from analysis
df$SHOW_DISCOUNT = NULL

#removing NULLs
df <- df[complete.cases(df),]
colSums(is.na(df))

#subset the data by removing sold = 0 and view_count = 0 from dataset
df = subset(df, df$SOLD!=0 | df$VIEW_COUNT!=0)
df = subset(df, df$DEPARTMENT == "Men Clothes")
table(df$DEPARTMENT)
View(df)

#factors
df$CATEGORY = factor(df$CATEGORY)
df$SHOP_LOCATION = factor(df$SHOP_LOCATION)
df$DEPARTMENT = factor(df$DEPARTMENT)
df$CURRENCY = factor(df$CURRENCY)

#controlling for time, we need to extract year and month from create time
variable
df$CREATE_TIME_YEAR = format(df$CREATE_TIME, "%Y")
df$CREATE_TIME_YEAR = as.factor(df$CREATE_TIME_YEAR)
df$CREATE_TIME_MONTH = format(df$CREATE_TIME, "%b")
df$CREATE_TIME_MONTH = as.factor(df$CREATE_TIME_MONTH)


#Predictors
```

```
#Sale price
#review rating
#review count
#historical sold
#stock
#discount %
#liked_count
#Create_year
#category

#response variable - conversion
df$conversion = df$SOLD/df$VIEW_COUNT
#conversion rate in %
df$conv_rate = df$conversion*100
df$conv_rate = round(df$conv_rate,0)

summary(df$conv_rate)

#sampling data
set.seed(12)
index = sample(1:nrow(df), size = 0.5*nrow(df), replace = F)
df_sample = df[index,]
dim(df_sample)


#exploratory data analysis

#histogram of conversion rate
hist(df_sample$conv_rate, col = "red", probability = T ,
     main = "Hist of conv_rate")
hist(log(df_sample$conv_rate), col = "red", probability = T,
     main = "Hist of log(conv_rate)")

table(df$CATEGORY)
table(df$DEPARTMENT)
table(df$CREATE_TIME_YEAR)

#data visualization
library(lattice)
bwplot(~df_sample$conversion | df_sample$CATEGORY)
boxplot(df_sample$conv_rate ~ df_sample$CATEGORY,NAMES=NULL,
        xlab = "categories", ylab = "Conversion rate")
bwplot(~df_sample$conv_rate | df_sample$CREATE_TIME_YEAR, xlab = "conversion
rate")

#check for correlations
df_num = c("SALE_PRICE","REVIEW_RATING","REVIEW_COUNT","LIKED_COUNT",
           "DISCOUNT","conv_rate")
library(PerformanceAnalytics)
chart.Correlation(df_sample[,df_num])

## Models
summary(df$conv_rate)
```

```
#Conversion rate is a censored data
#Using tobit model

#Since DV: Conversion is sold/impression which is count/count
# a percentage, therefore, we can use OLS regression

#base model - linear model withh all predictors

lm1 = lm(conv_rate ~ SALE_PRICE + REVIEW_RATING + REVIEW_COUNT + LIKED_COUNT +
           HISTORICAL_SOLD + STOCK + DISCOUNT + CATEGORY + CREATE_TIME_YEAR,
         data = df_sample)

summary(lm1)

#tobit model

library(AER)
tobit1 = tobit(conv_rate ~ SALE_PRICE + REVIEW_RATING + REVIEW_COUNT +
              LIKED_COUNT + HISTORICAL_SOLD + DISCOUNT + STOCK +
CREATE_TIME_YEAR + CATEGORY,
              left = 0, right = 100, data = df_sample)
summary(tobit1)
AIC(tobit1)

tobit2 = tobit(conv_rate ~ SALE_PRICE+ STOCK + DISCOUNT*REVIEW_RATING +
REVIEW_COUNT +
                 CATEGORY*LIKED_COUNT + HISTORICAL_SOLD +
CATEGORY*REVIEW_RATING
               + CREATE_TIME_YEAR, left = 0, right = 100,
               data = df_sample)
summary(tobit2)
AIC(tobit2)

library(stargazer)
stargazer(lm1, tobit1, tobit2, type = "text", out = "out.txt", single.row =
TRUE)


#interaction between stock and price to check when price is low
#and stock is high - does conversion improve?
#how distinct categories affect conversion with discount and higher review
rating respectively

#assumptions

#multicollinearity
round(vif(tobit1),2)
#high multi collinearity between historical sold and revieq_count

#Independence
library(car)
durbinWatsonTest(residuals(tobit2))
```

```
#no sign of auto correlation - PASS

#Heteroscedasticity
residuals <- residuals(tobit2)
fitted_values <- fitted(tobit2)
plot(fitted_values, residuals, pch = 1,xlab = "Predicted values", ylab =
"Residuals")
abline(0,0,col = "red",lwd = 3)

#Normality
qqnorm(residuals, pch = 19, main = "Normality Plot of Conversion rate
residuals")
qqline(residuals, lwd =3, col = "red")
```